A new DNA algorithm to solve graph coloring problem

Jiang Xingpeng[1], Li Yin[2], Meng Ya[3] and Meng Dazhi[2] *

(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China; 2. College of Applied Sciences, Beijing University of Technology, Beijing 100022, China; 3. School of Science, Beijing Institute of Technology, Beijing 100081, China)

Accepted on November 30, 2006

Abstract Using a small quantity of DNA molecules and little experimental time to solve complex problems successfully is a goal of DNA computing. Some NP-hard problems have been solved by DNA computing with lower time complexity than conventional computing. However, this advantage often brings higher space complexity and needs a large number of DNA encoding molecules. One example is graph coloring problem. Current DNA algorithms need exponentially increasing DNA encoding strands with the growing of problem size. Here we propose a new DNA algorithm of graph coloring problem based on the proof of four-color theorem. This algorithm has good properties of needing a relatively small number of operations in polynomial time and needing a small number of DNA encoding molecules (we need only $6R$ DNA encoding molecules if the number of regions in a graph is R).

Keywords: **DNA computing, NP-hard problem, graph coloring problem.**

Adleman[1] described the first successful experiment with standard tools of molecular biology to solve a 7-vertex instance of Hamiltonian Path problem. By using molecular techniques to execute computational operations, Adleman extended the conventional way of performing and looking at computations greatly and made a great step to the "sub-micro" computer. DNA computing appeals to many researchers in the past 12 years[2—8].

The four-coloring problem is closely related to the famous four-color theorem. Coloring problems are generally NP-hard problems. DNA computing has great advantages in the resolution of NP-hard problems[9—12]. All of the current DNA algorithms of graph coloring problems have polynomial time complexity, but there are two difficulties which we cannot overcome. The first one is that the encoding molecules they need are all exponentially increasing with the growing of problem' s size[13—15]. In fact, we know that the amount of molecules in one tube is limited, so it is difficult to solve big-size problems with these algorithms. The other difficulty is that these algorithms often need complicated experimental operations to generate resolution spaces[13—15].

Enlightened by the idea of "rib" which is the key of proof of the four-color theorem, we introduce the concept of "rib group" in this paper, and we prove that any edge 3-coloring of a smooth triangulation can be covered by a rib group. Based on this concept we present a new DNA algorithm for edge 3-coloring which can be converted to vertex 4-coloring. The algorithm has not only polynomial time complexity but also a small number of encoding molecules which is $6n$ if the graph has n regions. Our algorithm needs not complicated operations to generate resolution spaces, it requires about $R+13$ operations comparing with previous $O(n+m)$[13] and $O(\log(n)+n)$[15], here R is the number of regions, n and m are numbers of vertices and edges of a graph. It is more realizable for processing big size problems comparing with other algorithms.

1 Rib group and smooth triangulation

A graph G consists of a finite set $V(G)$ of vertices, a finite set $E(G)$ of edges, and an incidence relation between them. One edge is incident with two vertices, called its ends. A triangulation is one graph whose regions are all triangles. Without generality, a graph can be converted to a triangulation by the addition of edges. And a coloring of triangulation can be easily converted to a coloring of its origin graph[9, 10].

Definition 1. If one triangulation whose every triangle is incident to at least other two triangles, we call it a smooth triangulation.

If one edge of a smooth triangulation is incident

* To whom correspondence should be addressed. E-mail: dzhmeng@bjut.edu.cn

to only one triangle, we say it is on the boundary of the smooth triangulation. For a map, every city (or country) can be regarded as vertex and the incidence of two cities (or countries) can be connected by one edge, consequently, by adding edges it is converted to a triangulation. Furthermore by deleting those single regions and regions which are incident with only one region, a triangulation can become a smooth triangulation. We will focus on the coloring of smooth triangulation in this paper.

Definition 2[9, 10]. A graph G consists of a finite set $V(G)$ of vertices, and κ: $V(G)\rightarrow\{1, 2, 3, 4\}$ is a mapping. For every edge of G, if it is incident to vertex u, v, and $\kappa(u)\neq\kappa(v)$, then we say G is four-colored, and κ is called a four-coloring of G.

Definition 3[9, 10]. A graph G consists of a finite set $E(G)$ of edges, and κ: $E(G)\rightarrow\{1, 0, -1\}$ is a mapping. For every vertex of G, if it is incident to edges e, f, and $\kappa(e)\neq\kappa(f)$, then we say G is three-colored, and κ is called a three-coloring of G.

Definition 4[9, 10] (Fig. 1 (a)). Let H be a smooth triangulation, and κ a three-coloring of it, then we call a series g_0, r_1, g_1, $\cdots$, r_t, g_t a $\{1, -1\}$ linear rib if it satisfies the following constrains:

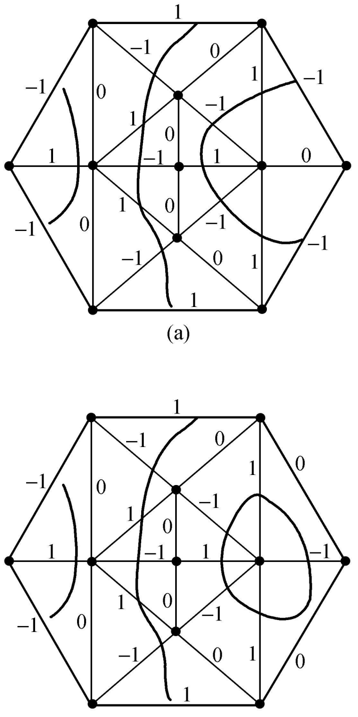

Fig. 1. Three-coloring of graph (a) is composed of three linear ribs, they construct one rib group. Three-coloring of graph (b) is composed of two linear ribs and one circular rib, they construct another rib group.

1) g_0, g_1, $\cdots$, g_t are different edges of H;

2) r_1, r_2, $\cdots$, r_t are different regions of H;

3) g_o, g_t are on the boundary of H;

4) $0< i\leqslant t$, r_i is incident to g_{i-1} and g_i;

5) $0\leqslant i\leqslant t$, $\kappa(g_i)\neq 0$, and $1\leqslant i\leqslant t$, $\kappa(g_i)\neq\kappa(g_{i-1})$.

Definition 5 (Fig. 1(b)). Let H be a smooth triangulation, and κ a three-coloring of it then we call a series g_0, r_1, g_1, $\cdots$, r_t, g_t a $\{1, -1\}$ circular rib if it satisfies the following constrains:

1) g_0, g_1, $\cdots$, g_t are different edges of H;

2) r_1, r_2, $\cdots$, r_t are different regions of H;

3) g_o, g_t are not on the boundary of H;

4) $0< i\leqslant t-1$, r_i is incident to g_{i-1} and g_i, and r_t is incident to g_0, g_{t-1};

5) $0\leqslant i\leqslant t$, $\kappa(g_i)\neq 0$, and $0< i\leqslant t$, $\kappa(g_i)\neq\kappa(g_{i-1})$.

Using the same method, we can define $\{1, 0\}$, $\{-1, 0\}$ linear rib or $\{1, 0\}$, $\{-1, 0\}$ circular rib. They have the same properties because they are at symmetrical position, we take $\{1, -1\}$ rib as our study target for convenience.

Property 1. For any coloring of a smooth triangulation, two different ribs have not a common region[11, 12].

Definition 6 (Fig. 1). A group of ribs which cover all regions of H is called a rib group. In this paper, we discuss only $\{1, -1\}$ rib groups.

Theory 1. For any coloring of smooth triangulation, there is at least one rib group.

Proof. It is enough to prove any region of H belonging to one rib. Let κ be a three-coloring of G. And one region r whose edges e, f, g are colored by 1, -1, 0 respectively. H is a smooth triangulation, so r is incident to at least two other regions, supposing they are r_1, r_2. Without generality, suppose r_1 is incident to edge f of r, then colors of the other two edges of r are -1 and 0 respectively. If f_1 is one of the edges of r_1 whose color is -1, f_1 is not on the boundary of H. Because H is a smooth triangulation,

f_1 must be incident to another region of H, say r_2. So r$_2$ and r_1 have one common edge f_1. The other two edges of r_2 are 1 and 0 respectively. Note that the edge colored by 1 is f_2. Continuing this procedure, we can find the edge of one region colored by 1 or -1 is on the boundary or is just e. If it returns to e, a circle is formed, we call it a $\{1, -1\}$ circular rib.

If the last region has one edge which is on the boundary, we define this edge f_m, and note that the band beginning from e and ending at f_m is a semi-rib, $erfr_1f_1r_2\cdots r_mf_m$, if e is on the boundary of H, then $erfr_1f_1r_2\cdots r_mf_m$ is a rib. If not, e is incident to another region of H, we can find another semi-rib beginning from e and ending at another edge which is on the boundary of H. Combining these two semi-ribs a complete rib is formed and obviously r belongs to this rib. Because one region belongs to only one rib, ribs are not intercrossing[11,12], we can say that H is covered by a group of ribs which are not intercrossing. That is to say, there is a rib group.

Suppose H have l rib groups, and the size (number of regions) of ith rib group is m_i, $i=1, 2, \cdots, l$, then every rib group can represent 3×2^{m_i} kinds of coloring by considering three kinds of ribs $\{1, 0\}$, $\{1, -1\}$, $\{-1, 0\}$, then the number of all coloring is $\sum_{i=1}^{l} 3\times 2^{m_i}$.

2 DNA algorithm of three-coloring of graph

The four-coloring theorem has proved[9-12] that every graph has at least one four-coloring and one three-coloring. Our DNA algorithm is under the existence of three-coloring.

2.1 DNA coding of a graph

According to the definition of smooth triangulation, one region has at most one edge on the boundary. For example (Fig. 2), if the ith region has one edge e_i, and it is incident to the jth and the kth regions through common edges e_{ij}, e_{ik} respectively. The code of the ith region has two classes: (i) If one end of its code is e_i, then the other end is e_{ik} or e_{ij}. Every encoding molecule encodes information of one region and two edges. The middle segment of it encodes region i, and two ends of it encode e_i and e_{ik} or e_{ij}(see Fig. 2(c)). (ii) If any end of its encoding molecule is not e_i, then its two ends are e_{ik} and e_{ij}, as the dot curve indicates in Fig. 2(a). Similarly, the middle segment of the molecule encodes region i, and ends of it encode e_{ik} and e_{ij}(Fig. 2(b)). Simultaneously, color of edges must be encoded in the molecule, but color of two edges is different in the same molecule, for example, they can be 1, -1 or -1, 1 ordinarily. In order to form a rib, we should encode three directions of one region (Fig. 2(a)), every direction needs two encoding molecules, then one region must be encoded with six molecules considering the different colors in the same molecule.

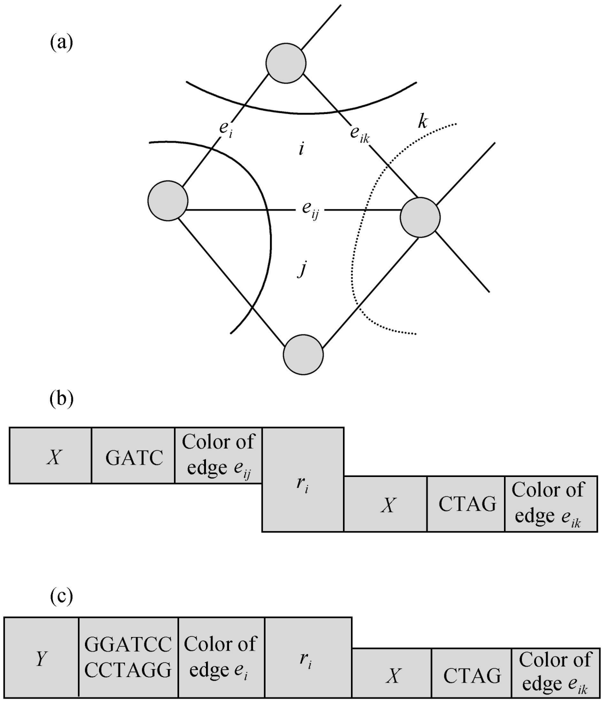

Fig. 2. (a) The coding of the ith region has two classes. Supposing the region has one edge e_i on the boundary, we can get two classes of encoding styles: If one end of its code is e_i, then the other end is e_{ik} or e_{ij}, as the solid curve indicates; if any end of its code is not e_i, then its two ends are e_{ik} and e_{ij}, as the dot curve indicates. (b) The second class of encoding molecules. (c) The first class of encoding molecules. Here X and Y are two segments which are designed to regulate molecular length.

For the purpose of operation on encoding molecules, we insert two restriction endonuclease site *Bam*HI and *Mbo* I (Fig. 3). A linear rib and circular rib can be incised to two molecules with two 4-base sticks respectively by the corresponding enzyme at two sites. These sticks are ready for next Watson-Crick reaction.

We take region i as an example to understand our encoding strategy. The smooth triangulation has

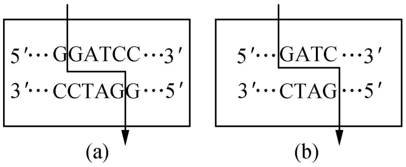

Fig. 3. (a) *Bam*H I restriction endonuclease site; (b) *Mbo*I restriction endonuclease site.

regions i, j, k, and region i is incident to j and k. The length of every encoding segment in one molecule is determined by some rules which we will give at the end of this section.

Encoding molecules of region i are represented by:

$$[\,Y|\,GGATCC|\,\kappa(e_i)|\,r_i]\,|\,\kappa(e_{ij})|\,CTAG|\,X\rangle \tag{1}$$

$$[\,Y|\,GGATCC|\,\kappa(e_i)|\,r_i]\,|\,\kappa(e_{ik})|\,CTAG|\,X\rangle \tag{2}$$

$$\langle X|\,GATC|\,\kappa(e_{ik})|[\,r_i]\,|\,\kappa(e_{ij})|\,CTAG|\,X\rangle \tag{3}$$

Here [ ] and $|NNN\rangle$ denote a double stranded and single stranded DNA molecules respectively; $\rangle$ and $\langle$ denote $3'$ sticky end on the upper strand and lower strand respectively. We use | to separate two neighboring functional segments, X and Y are two segments which are designed to regulate molecular length.

We must indicate that if a region has no edge on the boundary, it cannot be encoded by representation (1) and (2). Its six encoding molecules are all like representation (3). Watson-Crick reaction can occur between two sticks which satisfy the following rules: (i) they encode the same edge; (ii) they encode the same color; (iii) they have complementary endonuclease site.

Now we analyze the appropriate length of encoding molecules. Suppose $|X|$ is the length of segment X. If functional segments of encoding molecules satisfy the following formulas, Watson-Crick reaction will happen and all correct rib groups will have the same length (n regions) which is helpful to extract correct rib groups by biochemical operation: (i) $|X|-|Y|=1$; (ii) $|\kappa(e_{ij})|-|\kappa(e_i)|=1$; (iii) $|\kappa(e_{ij})|=|X|$, $|\kappa(e_i)|=|Y|$.

We can regulate the length of functional segments to satisfy natural biological reaction conditions.

2.2 Experiment of DNA coloring algorithm

First, we put all encoding molecules of one graph into a tube and self-assembly reactions among molecules will occur in the tube. Sticks will hybridize with other sticks if they are complementary. At last reactions will end when these big molecules have no sticks, in other words, molecules become possible ribs. That is to say, all encoding molecules assemble into long dsDNA strands whose two ends are segments like the left of representations (1) or (2). Then we put DNA ligase into this tube and the complete double strands are created. In the strands, the resulting rib regions and edges occur alternatively, and colors of edges incident with the same region are different. In order to avoid these resulting molecules mixed with miss-matched DNA strands, we add DNA exonuclease to degenerate miss-matched DNA strands.

Next, we process the linear rib and circular rib with different restriction endonucleases, their two ends will become 4-base sticks which randomly self assembly each other. The outcomes are all potential rib groups. After this step, molecules come into being a resolution space in tubes. Then we continue:

(i) Separating molecules which encode linear ribs from those which encode circular ribs by low melting agarose gel electrophoresis.

(ii) Collecting all molecules which cannot pass through gels and putting them into tube t_1. Collecting all molecules which enter into gels by DNA Gel Extraction Kit and putting them into tube t_2.

(iii) For t_1: firstly degenerating these circular molecules into circular ssDNA and extracting them by affinity purification systems using primers $\langle X|\,CTAG|\,\kappa(e_{ij})|$. Pooling all molecules extracted to a chip whose codes are $\langle X|\,CTAG|\,\kappa(e_{ij})|$ (prepared previously). Reactions happen between probes and molecules, after reactions are complete, taking out the chip and putting it into another tube which is added with endonuclease *Bam*H I. *Bam*H I cut circular ribs to linear ribs. Resulting strands have two ends with 4-base sticks in the tube. For t_2: putting endonuclease *Mbo* I into the tube, two sticks of linear ribs were formed by this enzyme. (iv) Mixing solutions of t_1 and t_2, and adding DNA strands $[X]\langle GATC|$ and $[X]\,|GATC\rangle$. By lowering down temperature and renaturing, all sticks of ribs and those short DNA strands will complement with others when they satisfy Watson-Crick rules. After that a ligase is added, and the complete strands are generated.

Finally we generate all rib groups from ribs:

(i) We extract DNA strands whose lengths are equal to n encoding molecules of the region by sodium dodecyl sulfate-polyacrylamide gel electrophoresis.

(ii) Rising temperature to degenerate all molecules and extract all DNA molecules which pass through every region at least once by affinity purification using $\langle r_i|$ as the primer.

(iii) The final results are those molecules which pass every region once and only once. Collect all these molecules and read the color they encode.

2.3 Reading of color on DNA chips

We can use DNA chips whose probes are complement of DNA molecules which encode all edges and their colors to read color of the graph. Corresponding to the probes, we call rib group target molecules. We use two different kinds of fluorescein which can activate each other to tag target molecules and probes respectively. One kind of fluorescein is tagged at $3'$ end of target molecules and the other tagged at $5'$ end of probes. Here the length of probes is variable as we have showed previously that the length of our encoding molecules depends on the needs. When target molecules and probes complement with each other, fluorescein can be activated and fluorescence occurs. We use PCR to extend rib groups, let them react on the chip and then read chips using a fluorescence detector.

The above-mentioned are only $\{1, -1\}$ rib groups, in fact $\{1, 0\}$, $\{-1, 0\}$ rib groups can be discussed similarly. Because of symmetry, it is enough to substitute $\{1, -1\}$ with $\{1, 0\}$ or $\{-1, 0\}$ to get other two kinds of rib groups and all coloring of a graph.

3 Transfer between three-coloring and four-coloring

In the proof of four-color theorem, the vertex four-coloring problem has been transferred to the edge 3-coloring problem. This substitution is supported by the following theorem:

Theorem 1[9-12]. A triangulation H is four-coloring if and only if it is three-coloring. This theorem has two meanings: giving a three-coloring of a triangulation, we can find a four-coloring of it, and *vice versa*. For instance, if we know a three-coloring of H, for any edge e, if it has two vertices u and v, we can construct the following corresponding four-coloring:

$$\phi(e)=\begin{cases}-1 & \{\kappa(u), \kappa(v)\}=\{1, 2\} \text{ or } \{3, 4\} \\ 0 & \{\kappa(u), \kappa(v)\}=\{1, 3\} \text{ or } \{2, 4\} \\ 1 & \{\kappa(u), \kappa(v)\}=\{1, 4\} \text{ or } \{2, 3\}\end{cases}$$

4 An example of three-coloring and simulation of algorithm

Generally DNA computing must be accomplished by biological experiments, but it is time- and labor-consuming. We can take computer simulation as a substitution. In recent years, computer simulation of DNA encoding and biological experiments has speeded up the development of DNA computing. In this study we simulated the experiment of three-coloring of a triangulation which is generated from Beijing map.

Beijing has 18 districts and for convenience we regard four central districts as a whole, then we get a map with 15 districts (Fig.4(a)). For the reason of transition from coloring regions to coloring vertices of graph, we let one vertex represent a district, and edges represent the contiguity between the districts. We get a graph which can be transformed into a triangulation by adding edges just as we have described previously. Then the problem of coloring Beijing map is transferred to a problem of coloring a triangulation with 20 regions, 15 vertices and 34 edges.

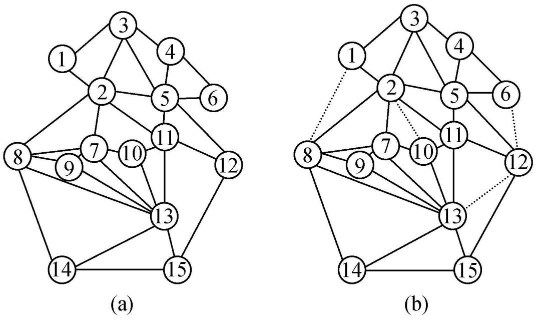

Fig. 4. Transform of Beijing map into triangulation. 1, Yanqing; 2, Changping; 3, Huairou; 4, Miyun; 5, Shunyi; 6, Pinggu; 7, Haidian; 8, Mentougou; 9, Shijingshan; 10, Center; 11, Chaoyang; 12, Tongzhou; 13, Fengtai; 14, Fangshan; 15, Daxing.

We found 1907 ribs which include 1822 linear ribs and 85 circular ribs and these ribs formed rib groups finally in the computer simulation using Matlab7.0.4. Because simulation is time-consuming, we have not found all rib groups but instead we found a part of them (if necessary, you can find all rib groups using our programs). For example, there are 82 rib groups which are composed of two ribs. It means that we can find at least 492 three-coloring for Beijing map simultaneously. And we found that most ribs have middle-size (12—16) regions, and a small number of ribs have fewer regions or many regions.

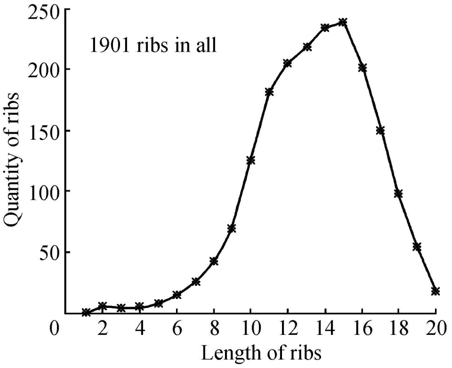

Fig. 5. Distribution of numbers of ribs: most ribs have 10—16 regions, only a small number of ribs have less regions or many regions.

5 Discussion

We propose here a new DNA algorithm of graph coloring problem based on the proof of four-color theorem. Solving NP-hard problems by DNA computing is very important for understanding of the new computer paradigm. However, current algorithms often need a large number of DNA strands to encode the problem. Comparatively, our algorithm needs a small number of DNA strands ($6n$) to encode an n-region graph and needs about $R+13$ operations, which reduced working load when compared with previous $O(n+m)$ and $O(\log(n)+n)$. DNA computing has extreme parallelism when solving NP-hard problem. How to use this parallelism is another problem. Although many algorithms have polynomial time, they often need exponentially increasing molecules with the growing of problems' size and also the resolution spaces are generated by complicated operations. Our algorithm avoids these problems, big-size problems can be resolved by a small quantity of encoding molecules and relatively few experimental operations under our experimental frame. However, the size of solution spaces is difficult to estimate in our method, perhaps it needs more graph theory knowledge. It can be regarded as a general graph problem: how many ribs and how many rib groups are there for a specific graph? Moreover, due to the proof of four-coloring theorem is based on the three-coloring of triangulations, our DNA algorithm of three-coloring may be used as an illumination to find proof of four-color theorem by DNA computing.

References

1 Adleman LM. Molecular computation of solutions to combinatarial problems. Science, 1994, 266: 1021—1024

2 Lipton RJ. DNA solution of hard computational problems. Science, 1995, 268: 542—545

3 Liu Q, Frutos AG, Wang L, et al. DNA computations on surfaces. Nature, 2000, 403: 175—179

4 Winfree E, Liu FR, Wenzler LA, et al. Design and self-assembly of two-dimensional DNA crystals. Nature, 1998, 394: 539—544

5 Braich RS, Chelyapov N, Johnson C, et al. Solution of a 20-variable 3-SAT problem on a DNA computer. Science, 2002, 296: 499—502

6 Benenson Y, Paz-Elizur T, Adar R, et al. Programmable and autonomous computing machine made of biomolecules. Nature, 2001, 414: 430—434

7 Adar R, Benenson Y, Linshiz G, et al. Stochastic computing with biomolecular automata. Proc Natl Acad Sci USA, 2004, 101: 9960—9965

8 Benenson Y, Gil B, Ben-Dor U, et al. An autonomous molecular computer for logical control of gene expression. Nature, 2004, 429: 423—429

9 Appel K and Haken W. Every planar map is four colorable. Part I. Discharging. Illinois J Math, 1977, 429—490

10 Appel K, Haken W and Koch J. Every planar map is four colorable. Part II. Reducibility. Illinois J Math, 1977, 491—567

11 Robertson N, Sanders DP, Seymour P, et al. New proof of the four colour theorem. Electronic Research Announcements of The American Mathematical Society, 1996, 2: 17—25

12 Robertson N, Sanders DP, Seymour P, et al. The four-colour theorem. Journal of Combinatorial Theory (Series B), 1997, 70: 2—44

13 Bach E, Condon A, Glaser E, et al. DNA models and algorithms for NP-complete problems. In: Procedings of 11th Conf. Computational Complexity, IEEE, 1996, 290—300

14 Fu B. Volume bounded molecular computation. Ph. D. Thesis, Department of Computer Science, Yale University, 1997

15 Qiu ZF and Lu M. A new approach to advance the DNA computing. Applied Soft Computing, 2003, 3: 177—189

16 Chen ZJ and Peng XJ. Research progress in DNA fluorescence labeling. Chemical Research and Application, 2005, 17: 154—158