基于质粒 DNA 匹配问题的分子算法*

高 琳[1)**] 马润年[1)] 许 进[2)]

([1)]西安电子科技大学电子工程研究所，西安 710071; [2)]华中科技大学系统科学研究所，武汉 430074)

摘要 给定无向图，图的最小极大匹配问题是寻找每条边都不相邻的最大集中的最小者，这个问题是著名的 NP-完全问题. 1994 年 Adleman 博士首次提出用 DNA 计算解决 NP-完全问题，以编码的 DNA 序列为运算对象，通过分子生物学的运算操作解决复杂的数学难题，使得 NP-完全问题的求解可能得到解决. 提出了基于质粒 DNA 的无向图的最大匹配问题的 DNA 分子生物算法，通过限制性内切酶的酶切和凝胶电泳完成解的产生和最终接的分离，依据分子生物学的实验手段，算法是有效并且可行的.

关键词 质粒，DNA 计算，NP-完全问题，最大匹配

学科分类号 TP301.6，Q78

最小极大匹配问题 (maximum maximal matching problem, MMMP) 是一个著名的组合优化问题，这不仅仅因为它最早被证明是 NP-完全问题之一[1]，而且因为它在理论和实践上有着重要的意义 (如在最优分配问题中的应用). NP-完全问题的求解一直困扰着人们，近些年，人们用神经计算，进化计算等方法来求解 NP-完全问题也取得了一些进展. 1994 年，Adleman[2] 第一次利用分子生物学技术，在试管中进行了 DNA 的实验，解决了有向图的哈密尔顿路问题 (Hamiltonian path problem, HPP). 虽然在实验室进行了 7 天的实验，才使一个只有 7 个顶点的有向图的哈密尔顿路问题得到解决. 但是由于他首先提出用 DNA 计算的方法来解决 NP-完全问题，开辟了求解 NP-完全问题计算的新领域，因而在国际上引起了巨大的轰动. Lipton[3] 修正了 Adleman 的实验方法，解决了著名的“可满足性”问题 (Boolean satisfaction problem, SAT). Ouyang 等[4] 给出了最大团的 DNA 解; Head 等[5] (2000 年) 用基于质粒的 DNA 计算求解了最大独立集问题; Liu 等[6] 一直致力于表面上 (类似于生物芯片) 的 DNA 计算，成功地解决了 SAT 问题，在生物实验的手段和方法上更加完善，减少了早期试管实验的差错率. 相信，随着生物芯片 (biochip) 技术的不断发展，DNA 计算将会更加简单和方便.

本文提出基于质粒的 DNA 计算求解 MMP，将图的边编码为双链 DNA 片段，将其作为外源 DNA 片段连接在合适的质粒载体上，以形成新的质粒，然后采用质粒的重组、提取、纯化等技术[7]，通过基本的生物操作如质粒的连接、扩增、凝胶电泳及生物酶等完成解的生成及最终的解分离[8]. 根据文献 [5] 的实验手段和步骤，本文提出的算法是完全有效和可行的.

1 质粒计算的概念

质粒是游离于细菌染色体之外的具有自行复制子的双链闭环 DNA 分子，用于重组 DNA 技术的质粒是经过改造的，具有复制子、选择标志、克隆位点等[7,8].

设 P 是一个质粒，k 是一个正整数，s_1，s_2，…，s_3 是 P 的 k 个相互不重叠的子段. 对于每个 i，核苷酸序列 s_i 不能出现在质粒 P 的其余位置上，并称 s_i 是质粒 P 的“位置”. 通过切割和粘贴，质粒在“位置”处不断地修改，相应的核苷酸序列 s_i 要么在质粒上，要么不在，分别用 1 和 0 表示，这就相当于电子计算机的 k 比特的寄存器. 本文正是利用质粒所具有的特征提出图的最大匹配的 DNA 算法，其基本的生物操作是[7]:

a. 连接 (ligating): 在连接酶的作用下，将目的基因的 DNA 片段连接在开口的质粒上以形成闭环状的质粒，或者将酶切后的质粒重新环化; b. 放大 (amplifying/copying): 必须将重组的 DNA 分子导入宿主菌中，通过细菌培养来扩增所需的 DNA. 通常采用的宿主菌为大肠杆菌，根据不同

* 国家自然科学基金和陕西省自然科学基金资助项目 (69971018, 2001X05).

** 通讯联系人.

Tel: 029-8201631, E-mail: lgao@mail.xidian.edu.cn

收稿日期: 2002-02-05, 接受日期: 2002-03-31

载体的需求，选择不同品系的大肠杆菌；c. 酶切(cutting)：在质粒上用特殊的内切酶将表示顶点的某些“位置”切割掉；d. 分离(separation)：通过凝胶电泳依据质粒的链长对DNA分子进行分离；e. 检测(detecting)：从反应产物读取表示解的DNA序列.

2 DNA算法

2.1 问题描述

在无向图 $G=(V, E)$ 中，对边集 E 的任一子集 $M \subseteq E$，如果 M 中任意两条边在 G 中均不邻接，则称 M 是 G 的一个匹配(matching)[1]. M 中的一条边的两个端点叫做在 M 下是配对的. 若匹配 M 的某条边与顶点 v 关联，则称 M 饱和顶点 v，并且称 M 是 v 饱和的，否则称 v 是 M 非饱和的. 极大匹配是不能再加入边的匹配，一个图可以有许多不同的极大匹配，其中边数最多的极大匹配为最大匹配，边数最少的极大匹配为最小极大匹配. 如图1中 $M=\{e1, e3, e5\}$ 是最小极大匹配.

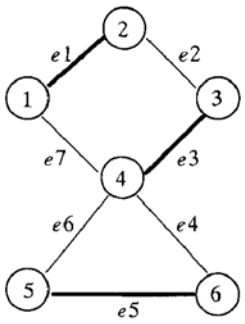

Fig. 1 Undirected graph and its maximum match (bold lines)

2.2 DNA算法步骤

DNA计算在解决问题时分为三个阶段：a. 对问题进行适当的编码，将要求解的问题映射到DNA链上；b. 生物实验，依照算法模型的步骤完成各种实验操作，生成问题的解；c. 解的提取.

步骤1：输入，对图1中的每条边进行编码. 将所有边1，2，…，m 依次编码在一条DNA双链上，每条边用30~50 bp不等或相等的核苷酸片段编码，其原因之一是考虑DNA链的特异性，另一方面要求两条边的碱基之差不能大于其余的边；每条边编码的链两端都有相同的特殊酶切位点的限制性内切酶片段. 而不同的两条边 i 和 $i+1$ 之间都有两种不同的内切酶，而这两种酶之间也可夹一些寡聚核苷酸片段(这个片段可有可无)；

步骤2：把步骤1所产生的DNA片段插入到开口的质粒中，形成闭环状的质粒，然后转入大肠杆菌扩增这样的质粒；

步骤3：检查质粒中任意两条边之间是否有顶点相连，若都没有顶点相连，则转入步骤4；否则将步骤2所产生的结果分成相等的两个试管，在两个试管中分别加入切割有顶点相连的两条边对应的内切酶，再把切割下来的小片段和质粒分离开来，使质粒重新环化后合成一个试管返回步骤3；

步骤4：用凝胶电泳技术找出链最长的质粒，它所含的边集即是最大匹配所对应的匹配；

步骤5：输出结果.

3 算法实现

以图1为例说明算法的实现.

步骤1：输入，对图1中的每条边进行编码，每条边的位置及寡聚核苷酸片段的长度如图2所示，

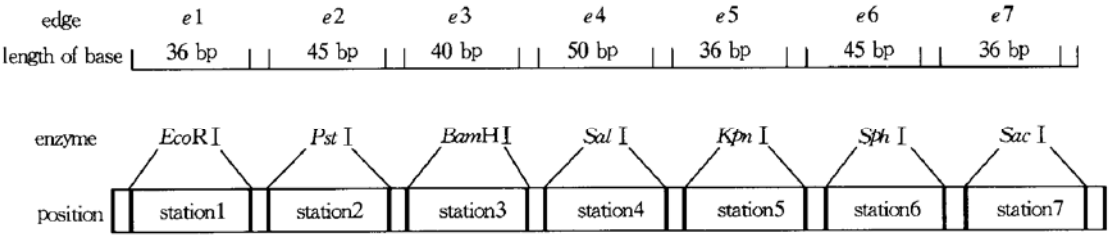

Fig. 2 Position of the edge and its DNA chain

每段的两端都有特殊酶切位点的序列，例如station1表示 $e1$ 所在的位置，其两端含有 *Eco*R I 的识别序列GAATTC，其分割点在G与A之间，这样通过 *Eco*R I 的作用，就可以将 $e1$ 从链上切除. 在下面的算法中，设两条边之间没有寡聚核苷酸片段，人工合成DNA链，其长度为288 bp，边 $e1$~$e7$ 所对应的编码如图3所示.

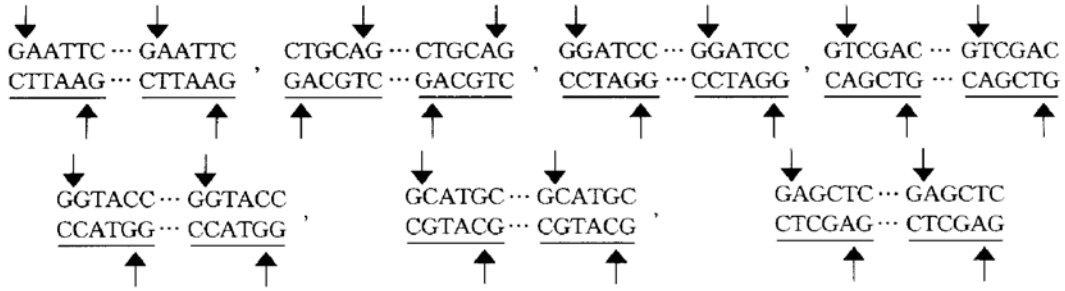

Fig. 3 Encoding the edge

图 3 中箭头所指的位置分别为内切酶 *Eco*R Ⅰ, *Pst* Ⅰ, *Bam*H Ⅰ, *Sal* Ⅰ, *Kpn* Ⅰ, *Sph* Ⅰ 和 *Sac* Ⅰ 的切割位点, 带下划线的序列为互补序列, 省略号表示可用四种碱基 A, T, C, G 随意编码. 需要说明的是, 除了规定的位置外, 其他位置不能出现限制性内切酶的识别序列.

步骤 2: 将合成的长度为288 bp的 DNA 链插入到已开口的质粒中, 形成闭环状的质粒, 然后转入大肠杆菌进行扩增, 以期达到数量足够多的所需的新质粒, 用试管 T_0 表示.

步骤 3: 检查质粒中任意两条边之间是否有顶点相连. 由于边 $e1$ 和 $e2$ 之间有顶点相连, 边 $e1$ 和 $e2$ 不可能同时出现在一个匹配中. 因此, 把试管 T_0 中的液体等量分到两个试管 T_1 和 T_2. 在 T_1 中加入内切酶 *Eco*R Ⅰ 切掉边 $e1$ 所代表的带有粘性末端的 DNA 片段 (链长为 36 bp), 把切下来的 DNA 片段和质粒分离, 并使 T_1 中的质粒重新环化. 由于步骤 2 中开口的质粒在整个实验过程中是不发生变化的, 因此, 重新环化的质粒链长为 288 bp− 36 bp= 252 bp. 同样, 在 T_2 中加入内切酶 *Pst* Ⅰ 切掉边 $e2$ 所代表的带有粘性末端的 DNA 片段 (链长为 45 bp), 并使 T_2 杯子中的质粒重新环化. 这样重新环化的质粒链长为 288 bp− 45 bp = 243 bp. 这时含有边 $e2$ 的匹配全部在 T_1 中, 而含有边 $e1$ 的匹配全部在 T_2 中, 然后把 T_1 和 T_2 混合在一起得新的 T_0. 同理, 由于 $e2$ 和 $e3$ 之间有顶点相连, 可将 T_0 分成等量的两个试管 T_1 和 T_2. 在 T_1 中加入 *Pst* Ⅰ 切掉边 $e2$ 所代表的带有粘性末端的 DNA 片段 (链长为 45 bp), 重新环化的质粒链长为 252 bp− 45 bp= 207 bp 和 243 bp. 同样, 在 T_2 中加入 *Bam*H Ⅰ 切掉边 $e3$ 所代表的带有粘性末端的 DNA 片段 (链长为 40 bp), 重新环化的质粒链长为 252 bp− 40 bp= 212 bp 和 243 bp. 把 T_1 和 T_2 混合在一起得新的 T_0, 这时 T_0 中的质粒有三种: 含有边 $\{e1, e3, e4, e5, e6, e7\}$、$\{e2, e4, e5, e6, e7\}$ 和 $\{e3, e4, e5, e6, e7\}$ 的匹配. 而这时边 $e3$ 和 $e4$ 之间也有顶点相连, 用类似的方法加入 *Bam*H Ⅰ 切掉边 $e3$ 所代表的 DNA 片段 (链长为 40 bp), 这时质粒链长分别为 207 bp− 40 bp= 167 bp, 243 bp− 40 bp= 203 bp 和 212 bp. 加入 *Sal* Ⅰ 切掉边 $e4$ 所代表的 DNA 片段 (链长为50 bp), 质粒链长分别为 207 bp− 50 bp= 157 bp, 243 bp− 50 bp= 193 bp 和 212 bp. 经过合并, 得到含有边 $\{e2, e4, e5, e6, e7\}$、$\{e1, e4, e5, e6, e7\}$、$\{e1, e3, e5, e6, e7\}$、$\{e4, e5, e6, e7\}$ 及 $\{e3, e5, e6, e7\}$ 的匹配. 而这时由于 $e4$ 和 $e5$, $e5$ 和 $e6$, $e6$ 和 $e7$, $e7$ 和 $e1$ 等边之间有共同的顶点, 采用同样的方法经过实验后, 所得到的边集中任意两条边之间无顶点相连, 这时 T_1 和 T_2 中的质粒都代表图 1 的匹配, 最大匹配一定在里面. 现在的问题是如何把它分离出来.

步骤 4: 通过凝胶电泳找出链长最长的质粒, 它所代表的编码就是最大匹配, 其链长为 112 bp.

步骤 5: 用分子克隆技术确定长度最大的质粒所对应的最大匹配, 其边集是 $\{e1, e3, e5\}$ 或 $\{e2, e7, e5\}$.

本文提出了基于质粒技术的 MMP 的 DNA 算法. 首先对问题进行编码, 将图 1 的边按一定的编码存储于双链 DNA 中, 然后将双链 DNA 片段作为外源 DNA 连接在合适的质粒载体上, 以形成新的质粒. 然后采用质粒的重组、提取、纯化等技术, 通过基本的生物操作如质粒的连接、扩增、凝胶电泳及生物酶等完成解的生成及最终的解分离. 为了说明问题方便, 文中的例子规模较小, 对于规模大的问题其编码方法和算法的求解过程完全是一致的. 但随着求解问题规模的增大, 会出现下面问题: a. 酶切所需的时间会延长, 因为每个边的切

除对应一种酶，切除不同的边需进行不同的操作；b. 实验过程中酶切有时不能够完全进行，可能会导致一些“伪解”或“错解”出现，即该切除的边由于生物操作的缘故而没有切除；c. 由于限制性内切酶种类有限，不可能解决规模很大的问题，因为每个边必须唯一对应一种内切酶. 尽管在计算中还存在着这样的问题，但随着生物技术的发展，如蛋白质核酸 (protein nucleic acid, PNA) 序列能够压缩限制性内切酶的限制位置，允许同一种酶作用在不同的边上，这样就可以使某些问题得到解决，这就鼓励我们进一步探索质粒计算的技术和方法.

虽然 DNA 计算目前还存在许多问题有待解决，但 DNA 计算观念的提出，向众多领域提出了挑战：对生物学与化学，在于理解细胞和分子机制，使他们成为分子算法的基础；对计算机科学和数学，在于寻找适当的问题和有效分子算法去解决更为复杂的系统模拟与计算问题；对于生理学与工程学，在于构建大规模可信而又易于实现的分子计算机. 正如著名计算机科学家 Lipton[3] 所说，既然人们已开始思考这类问题，就会找到许多方法来适合这个模型，自然科学中最诱人的两个前沿领域——分子生物学与计算机科学联姻，一定会创造出惊人的奇迹!

参 考 文 献

1 Garey M, Johnson D. Computers and Intractability. A Guide to the Theory of NP-completeness. USA: New York, 1979. 192~193

2 Adleman L. Molecular computation of solution to combinatorial problems. Science, 1994, **266** (11): 1021~1024

3 Lipton R J. DNA solution of computation problems. Science, 1995, **268** (4): 542~545

4 Ouyang Q, Kaolan P D, Liu S, *et al*. Solution of the maximal clique problem. Science, 1997, **278** (17): 446~449

5 Head T, Rozenberg G, Bladergroen R R, *et al*. Computing with DNA by operating on plasmids. Biosystem, 2000, **57**: 87~93

6 Liu Q H, Wang L, Anthony G F, *et al*. DNA computing on surface. Nature, 2000, **403** (13): 175~179

7 Turner P C, McLennan A G, Bates A D, *et al*. Molecular Biology. Beijing: Bios Scientific Publishers Limited, 2001. 91~96

8 姜 泊, 张亚历, 周殿元. 分子生物学常用实验方法. 北京: 人民军医出版社, 2000. 7~18
Jiang B, Zhang Y L, Zhou D Y. The Usual Experimental Method of Molecular Biology. Beijing: The Publisher of People's Military Medicine, 2000. 7~18

The Molecular Algorithm of The Matching Problem Based on Plasmid DNA*

GAO Lin[1)]**, MA Run-Nian[1)], XU Jin[2)]

([1)] *Electronic Engineering Research Institute, Xidian University, Xi'an* 710071, China;

[2)] *System Science Research Institute, Huazhong University of Science and Technology, Wuhan* 430074, China)

Abstract Given an undirected graph, the maximum matching problem is to find a subset of mutually non-adjacent edges having the largest number. This problem is a NP-complete and has no effective method. Adleman introduced firstly the DNA computing in 1994, with which the NP-complete problems are likely to be solved. DNA-based algorithm simulates molecular biology structure of DNA by means of molecular biology technological computation. The plasmid DNA contains a specially inserted series of DNA sequence segments, each of which is bordered by a characteristic pair of restriction enzyme sites, the DNA sequence segments of this series were used to represent the vertices of the graph. The solution is reached by applying enzymatic and gel electrophoresis. The DNA solution to the maximum matching problem of an undirected graph based on the plasmid is introduced. On the basis of the experimental method of bio-molecular, the algorithm is an effective and feasible method.

Key words plasmid, DNA computing, NP-complete problem, maximum matching

* This work was supported by grants from The National Natural Sciences Foundation of China (69971018) and Shanxi Natural Science Foundation of China (2001X05).

** Corresponding author. Tel: 86-29-8201631, E-mail: lgao@mail.xidian.edu.cn

Received: Febuary 5, 2002 Accepted: March 31, 2002