图的最大权团的 DNA 计算

马润年[1], 张 强[2,3], 高 琳[4], 许 进[3]
(1. 空军工程大学电讯工程学院, 陕西西安 710077; 2. 大连理工大学机械工程学院, 辽宁大连 116024;
3. 大连大学先进设计技术中心, 辽宁大连 116622; 4. 西安电子科技大学计算机学院, 陕西西安 710071)

摘 要: 给定顶点赋权的无向图, 图的最大权团问题是寻找每个顶点都相邻的顶点子集(团)具有最大权. 这个问题是寻找无权图的最大团问题的推广. 图的最大团和最大权团都是著名的 NP 完全问题, 没有非常有效的算法. 1994 年 Adleman 博士首先提出用 DNA 计算解决 NP 完全问题, 使得 NP 完全问题的求解可能得到解决. 本文给出了基于质粒技术的无向图的最大权团问题的 DNA 算法, 依据 Head T 等的实验手段, 本文提出的算法是有效并且可行的.

关键词: DNA 计算; NP 完全问题; 最大权团
中图分类号: TN4 文献标识码: A 文章编号: 0372-2112 (2004) 01-0013-04

Using DNA to Solve the Maximum Weight Clique of Graphs

MA Run-nian[1], ZHANG Qiang[2,3], GAO Lin[4], XU Jin[3]
(1. *The Telecommunication Engineering Institute, Air Force Engineering University, Xi' an, Shaanxi* 710077, *China*;
2. *School of Mechanical Engineering, Dalian University of Technology, Dalian, Liaoning* 116024, *China*;
3. *Advanced Design Technology Center, Dalian University, Dalian, Liaoning* 116622, *China*;
4. *School of Computer, Xidian University, Xi' an, Shaanxi* 710071, *China*)

Abstract: Given an undirected graph with weights on the vertices, the maximum weight clique problem is to find a subset of mutually adjacent vertices (i.e., a clique) having the largest total weight. This problem is a generalization of the problem of finding the maximum cardinality clique of an unweighted graph. Owing to the maximum cardinality clique problem and the maximum weight clique problem of graphs to be NP-complete, there are no effective methods to solve these two problems. Doctor Adleman introduced firstly the DNA computing in 1994, with which the NP-complete problems are likely to be solved. This paper introduces the DNA solution to the Maximum Weight Clique Problem of an undirected graph based on the plasmoid. On the basis of Head T et al, the algorithm is an effective and feasible method.

Key words: DNA computing; NP-complete problem; maximum weight clique

1 引言

最大团问题[1] (Maximum Clique Problem, 简记为 MCP) 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, 而且因为它在理论和实践上都有着重要的意义, 如计算机视觉、信息恢复、容错等领域有着广泛的应用. 另外, 有许多 NP 完全问题都可转化为 MCP, 如可满足性问题、独立集问题、顶点覆盖等问题. MCP 的一个重要推广是图的顶点赋权的最大权团问题[1] (Maximum Weight Clique Problem, 简记为 MWCP). MWCP 是寻找权最大的团 (注意 MWCP 不要求团的顶点最多), MWCP 在计算机视觉、模式识辨和机器人技术等方面有着重要的应用. 很明显, 若每个顶点的权相等, 则 MWCP 就是顶点没有赋权的 MCP. 由于 MCP 和 MWCP 都是著名的 NP 完全问题, 这些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.

然而, Adleman[2] (1994) 第一次利用现代分子生物技术, 在试管中进行了 DNA 实验, 解决了有向图的哈密尔顿路问题 (Hamiltonian Path Problem, 简记为 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 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] (1995) 修正了 Adleman 的实验方法, 解决了著名的"可满足性"问题 (Boolean Satisfaction Problem, 简记为 SAT). Ouyang[4] 等 (1997) 给出了 MCP 的 DNA 解; Head[5] 等 (2000) 用基于质离的 DNA 计算求解了最大独立集问题; Liu Qinghua[6] 等一直致力于表面上的 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

收稿日期: 2002-04-11; 修回日期: 2003-08-30
基金项目: 国家自然科学基金 (No. 69971018); 陕西省自然科学基金 (No. 2001X05)

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.

本文提出用基于质离的 DNA 计算求解 MWCP, 将图的顶点按权的大小编码成双链 DNA 片段作为外源 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, 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. 根据 T. Head et al[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. 文中用到的生物计算中的一些概念和用语见文献[2~ 8].

2 质粒计算的概念

质粒[5, 8]是游离于细菌染色体之外的具有自行复制子的双链的 DNA 分子, 其大小范围从 1kB 至 200kB 以上不等. 实验室中用于重组 DNA 技术的质粒是经过改造的, 通常它具有特征: 复制子、选择标志、克隆位点等.

设 P 是一个质粒, k 是一个正整数, $s_1, s_2, \cdots, s_k$ 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". 每次计算都开始于盛水的试管或缓冲器中含有大量的具有相同的 k 个"位置"的质粒. 在计算过程中, 质粒在不断地修改, 直到结果的读取. 需要说明的是质粒的修改只在所谓的"位置"处进行, 主要的方法是切割和粘贴. 在计算过程中, 每个核苷酸序列 s_i 要么在质粒上要么不在质粒上, 我们用 1 表示在质粒上, 而用 0 表示不在质粒上. 在某种程度上, 它相当于电子计算机的 k 比特的存储器. 本文正是利用质粒所具有的特征提出图的最大权团的 DNA 算法, 其基本的生物操作是:

(1) 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.

(2) 放大(Amplifying/Copying): 将重组的 DNA 分子必须导入宿主菌中, 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, 根据不同载体的需求, 选择不同品系的大肠杆菌.

(3) 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.

(4) 分离(Separation): 通过凝胶电泳依据质粒的链长对 DNA 分子进行分离.

(5) 提取(Extracting): 通过亲和纯化法, 提取包含某种特性的 DNA 分子链.

(6) 检测(Dectecting): 从反应产物读取表示解的 DNA 序列.

3 DNA 算法

3.1 问题描述

设 $G=(V, E, W)$ 是一个顶点赋权的无向图, 其中 $V=\{1, 2, \cdots, n\}$ 是顶点集合, $E\subseteq V\times V$ 是边集合, 且 $W\in Z^n$ 是顶点的权向量, 它的第 i 个元素 w_i 对应于顶点 i. 假设所有的 w_i 都是正整数. $\forall S\subseteq V$, 则 S 的权定义为

$$W(S)=\sum_{i\in S} w_i$$

一般地, 规定 $W(\Phi)=0$.

设 $S\subseteq V$, 若 $\forall x, y\in S$, x 和 y 在图 G 中都相邻, 则称 S 是图 G 的一个团. 若没有一个严格包含 S 团的团, 则称 S 是极大团. 最大团, 也就是序数最大的团, 是指顶点数最多的团. 而最大权团是指权最大的团. 由于假设每个 $w_i>0$, 因此最大权团也就是指权值最大的极大团. 很明显, 最大团是极大的, 反之不一定成立. 同样, 由于假设每个 $w_i>0$, 因此最大权团也是极大团, 反之不一定成立. 当每个 $w_i=1$ 时, 图的最大权团就是图的最大团, 因此图的最大团是特殊的最大权团. 本文主要研究图的最大权团的 DNA 计算. 下面的图 1 和图 2 给出赋权无向图 G 及它的补图 $\overline{G}$, 其中圆圈表示顶点, 圆圈中的第一个数字表示该顶点的编号, 而括弧中的数字表示该顶点的权重. 补图 $\overline{G}$ 的顶点集和图 G 的相同, 而补图 $\overline{G}$ 中的顶点间有边的充要条件是在图 G 中相应的顶点间没有边.

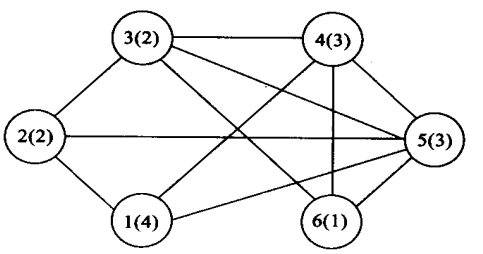


图 1 无向图 G

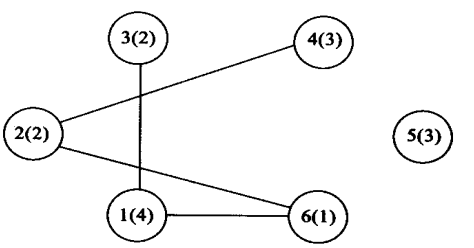


图 2 图 G 的补图

3.2 DNA 算法步骤

DNA 计算在解决问题时主要可分为三个阶段: (1) 对问题进行适当的编码, 也就是将要求解的问题映射到 DNA 链上; (2) 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; (3) 解的提取.

为了求解图的最大权团问题, 首先假设所有的 w_i 的最大公约数为 m, 则 $w_i=ml_i$. 下面给出算法步骤:

Step 1: 输入, 对图中的每个顶点及顶点的权值进行编码. 将所有顶点 1, 2, …, n 依次编码在一条 DNA 双链上, 顶点 i 用 $20l_i$bp 的核苷酸片段编码, 并且仍用 i 表示该片段; 每个顶点的两端都有相同的特殊酶切位点的限制性内切酶. 而不同的两个顶点 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;

Step 2: 把 Step1 所产生的 DNA 片段插入到开口的质粒中, 形成闭环状的质粒, 然后转入大肠杆菌进行扩增这样的质粒;

Step 3: 检查质粒中的任意两个顶点之间是否有边相连, 若都有, 则转入 Step 4; 若没有, 比如说在顶点 1 和顶点 2 之间没有边相连, 则把含所有由 Step2 所产生的质粒的实验杯子分成相等的两杯, 在第一杯和第二杯中分别加入切割顶点 1 和顶点 2 的内切酶, 再把切割下来的小片段和质粒分离出来, 并使质粒重新环化后再合成一个杯子返回 Step 3;

Step 4: 用凝胶电泳技术找出链最长的质粒.

Step 5: 确定链最长的质粒所含的顶点集即是最大权团所对应的团;

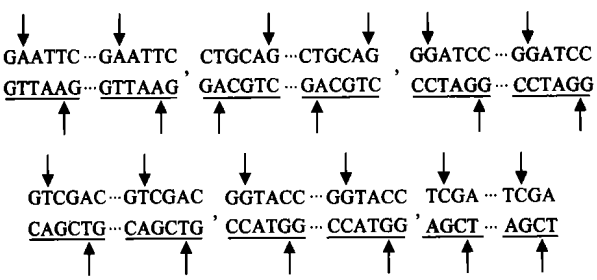
Step 6: 输出结果.

4　算法实现

以图 1 为例说明算法的实现.

Step 1: 输入, 对图中的每个顶点及顶点的权值进行编码;

(1) 对顶点编码: 人工合成 DNA 链, 其大小是$(m_1+\cdots+m_6)\times 20+5\times 6+4=334$bp, 这主要包括每个顶点的编码 DNA 链和每个内切酶的长度之和, 即由下面的 6 部分组成. 而顶点 1 至顶点 6 即所对应的编码分别为:



其中箭头所指的位置分别为内切酶 EcoRⅠ, PstⅠ, BamHⅠ, SalⅠ, KpnⅠ 和 TaqⅠ 的切割位点. 上下链的两个酶切位点之间的长度为 80bp, 40bp, 40bp, 60bp, 60bp 和 20bp. 当然除了规定的以外, 链的其他任何位置上都不能出现给定的内切酶.

(2) 这样将合成的大小为 334bp 的 DNA 链插入到已开口的质粒中(这个开口的质粒要大于 334bp, 如可取质粒 pOK12), 形成闭环状的质粒(如图 3), 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的质粒.

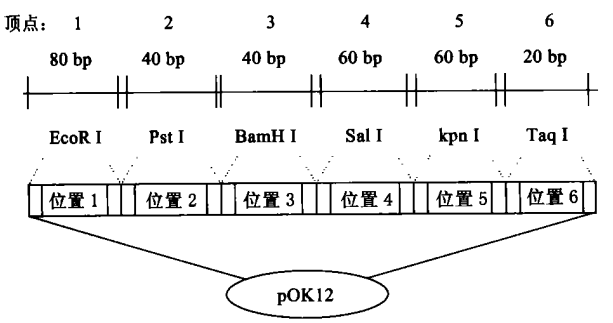


图 3　质粒的结构图

(3) 检查新的质粒中的任意两个顶点之间是否有边相连(或者检查它的补图中的任意两个顶点之间是否无边相连). 由于顶点 1 和顶点 6 之间没有边相连, 因此顶点 1 和顶点 6 不可能同时出现在同一个团中, 因此, 把含(2) 所产生的质粒的实验杯 T_0(可取 80ng) 分成两个等量的杯子 T_1 和 T_2. 在 T_1 中加入内切酶 EcoRⅠ 切掉顶点 1 所代表的带有粘性末端的 DNA 片段(链长为 80bp), 利用质粒分离技术把切下来的 DNA 片段和质粒分离, 并使 T_1 中的质粒重新环化. 这样重新环化的质粒的链长为原来质粒的链长减去 80bp. 由于(2) 中开口的质粒在整个实验过程中是不发生变化的, 因此, 我们无妨记重新环化的质粒的链长为 334- 80= 254bp. 同样, 在 T_2 中加入内切酶 TaqⅠ 切掉顶点 6 所代表的带有粘性末端的 DNA 片段(链长为 20bp), 把切下来的 DNA 片段和质粒分离, 并使 T_2 杯子中的质粒重新环化. 这样重新环化的质粒的链长为 334- 20 = 314bp. 这时含有顶点 6 的团全部在 T_1 中, 而含有顶点 1 的团全部在 T_2 中. 然后把 T_1 和 T_2 混合在一起得新的 T_0. 由于已经检查过图 G 中顶点 1 和顶点 6 之间没有边, 这时我们认为已经检查过的顶点之间加了一条边, 直到任意的两个顶点之间都有边连. 在新的图中, 由于顶点 1 和顶点 3 之间没有边相连, 可将 T_0 分成等量的两个 T_1 和 T_2. 在 T_1 中加入内切酶 EcoRⅠ 切掉顶点 1 所代表的带有粘性末端的 DNA 片段(链长为 80bp), 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. 这样重新环化的质粒的链长为 314- 80= 234bp 和 254bp. 同样, 在 T_2 中加入内切酶 BamHⅠ 切掉顶点 3 所代表的带有粘性末端的 DNA 片段(链长为 40bp), 把切下来的 DNA 片段和质粒分离, 并使 T_2 杯子中的质粒重新环化. 这样重新环化后的质粒的大小为 314- 40= 274bp 和 254- 40= 214bp. 然后再把 T_1 和 T_2 混合在一起得新的 T_0, 这时 T_0 中的质粒有四种: 含有顶点 2、3、4、5、6; 含有顶点 2、3、4、5; 含有顶点 2、4、5、6; 含有顶点 1、2、4、5 的团. 而这时顶点 2 和顶点 4 之间也没有边连, 可将 T_0 分成等量的两个 T_1 和 T_2. 在 T_1 中加入内切酶 PstⅠ 切掉顶点 2 所代表的带有粘性末端的 DNA 片段(链长为 40bp), 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. 环化后的质粒的链长分别为 214- 40= 174bp, 234- 40= 194bp, 254- 40= 214bp, 274- 40= 234bp. 同样, 在 T_2 中加入内切酶 SalⅠ 切掉顶点 4 所代表的带有粘性末端的 DNA 片段(链长为 60bp), 把切下来的 DNA 片段和质粒分离, 并使 T_2 杯子中的质粒重新环化. 这样重新环化的质粒的链长分别为 214- 60= 154bp, 234- 60= 174bp, 254- 60= 194bp, 274- 60= 214bp. 这样再将 T_1 和 T_2 中的混和在一起得新的 T_0. 由于顶点 2 和顶点 6 之间还没有边, 可将 T_0 分成等量的两个 T_1 和 T_2. 在 T_1 中加入内切酶 PstⅠ 切掉顶点 2 所代表的带有粘性末端的 DNA 片段(链长为 40bp), 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. 环化后的质粒的链长分别为 154- 40= 114bp, 174- 40= 134bp, 194- 40 = 154bp, 214- 40= 174bp, 174bp, 194bp, 214bp, 234bp. 同样, 在 T_2 中加入内切酶 TaqⅠ 切掉顶点 6 所代表的带有粘性末端的 DNA 片段(链长为 20bp), 把切下来的 DNA 片段和质粒分离, 并使 T_2 杯子中的质粒重新环化. 这样重新环化的质粒的链长分别为 154- 20= 134bp, 174- 20= 154bp, 174bp, 194- 20= 174bp, 194bp, 214- 20= 194bp, 214bp, 234bp. 经过这样实验后图 G 的任意两个顶点之间都有边相连, 这时 T_1 和 T_2 中的质粒都代表图 G 的团, 最大权团一定在里面, 现在的问题是如

何把它分离出来.

(4) 用凝胶电泳技术找出链长最大的质粒, 链长最长的质粒就是所代表的最大权团, 其大小正好是 234bp.

(5) 用分子克隆技术或探针技术或其他生物技术可以确定长度最大的质粒所对应的权最大的团, 其顶点集是{1(4), 4(3), 5(3)}.

(6) 输出结果: 最大权团是顶点集是{1, 4, 5}.

5 结束语

从算法实现的结果中, 我们可以看出最大权团不一定是最大团. 例 1 的最大团是四个顶点 3, 4, 5, 6. 本文提出了基于质粒技术的 MWCP 的 DNA 算法. 首先对问题进行编码, 将图的顶点按权的大小编码双链 DNA 片段作为外源 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, 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. 本文使用 DNA 序列表示顶点权值的大小, 解决了图的 MWCP.

当然对于图 1, 我们也可以构造新的图, 而研究新的图的 MCP 就相当于图 1 的 MWCP. 具体方法是将原来的顶点 i 用 m_i 个顶点来代替, 这 m_i 个顶点之间都有边, 并且原来和顶点 i 相邻的顶点都和这 m_i 个顶点相邻. 可以证明新图的 MCP 相当于图 1 的 MWCP. 但是这样做需要不同的内切酶 15 个. 特别是随着顶点数的增加, 不但内切酶的数量需要增加许多许多, 而且所需的实验步骤也要增多, 当然出错的可能性增大, 所以是不可取的.

为了说明问题方便, 文中的例子权值较小, 对于比较大的权值可以使用同样的编码方法. 但随着求解问题的规模及权的差异性增大, 会出现许多问题: (1) 最优解怎样与其他解分离; (2) 实验过程中可能会导致一些“伪解”或“错解”出现; (3) 对于各种计算问题, 怎样寻找一种 DNA 生物化学反应的运算途径, 使得 DNA 计算适应广阔的问题面, 并具有实用性.

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, 使它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] Bomze I M, Pelillo M, Stix V. Approximating the maximum weight clique using replicator dynamics [J]. IEEE Trans. Neural Networks, 2000, 11(6): 1228- 1241.

[2] Adleman L M. Molecular computation of solution to combinatorial problems [J]. Science, 1994, 266(11): 1021- 1024.

[3] Lipton R J. DNA solution of computational problems [J]. Science, 1995, 268(4): 542- 545.

[4] Ouyang Q, Kaplan P D, Liu S et al. Solution of the maximal clique problem [J]. Science, 1997, 278(17): 446- 449.

[5] Head T, Rozenberg G, Bladergroen R S et al. Computing with DNA by operating on plasmids [J]. Biosystems, 2000, 57: 87- 93.

[6] Liu Q, Wang L, Frutos A G et al. DNA computing on surface [J]. Nature, 2000, 403(13): 175- 179.

[7] 高琳, 许进, 张军英. DNA 计算的研究进展与展望 [J]. 电子学报, 2001, 29(7): 973- 977.

[8] 姜泊, 张亚历, 周殿元. 分子生物学常用实验方法 [M]. 北京: 人民军医出版社, 2000.

作者简介:

马润年 男, 1963 年生于陕西榆林, 1989 年、2002 年分别获山东大学理学硕士学位、西安电子科技大学工学博士学位, 现为空军工程大学电讯工程学院副教授, 主要从事最优化、神经网络、DNA 计算和图论等的研究, 已发表学术论文 40 余篇.

张 强 男, 1971 年生于陕西西安, 1999 年、2002 年分别获西安电子科技大学工学硕士学位、工学博士学位, 现大连理工大学博士后研究, 主要从事神经网络、遗传算法、信号处理和 DNA 计算等的研究, 已发表论文 30 余篇.

高 琳 女, 1964 年生于陕西咸阳, 1990 年、2003 年分别获西北大学理学硕士学位、西安电子科技大学工学博士学位, 现西安电子科技大学计算机学院副教授, 主要从事神经网络、遗传算法和 DNA 计算等的研究, 已发表论文 30 余篇.