最大加权独立集问题的 DNA 算法

吴 雪　赵 艺
(华东理工大学电子与通信工程系　上海　200237)

摘　要：该文基于分子生物技术提出了一种求解最大加权独立集(MWIS)问题的 DNA 算法。MWIS 是最大独立集(MIS)的母问题，而 MIS 是著名的 NP 完全问题。该算法的关键技术是基于变长的 DNA 序列来对所给图中的加权顶点进行合理的编码，并在建立初始完备数据链中采用并行重叠放大(POA)技术，然后应用变性、退火、 聚合酶链式反应(PCR)、酶切反应和凝胶电泳等一系列的 DNA 生物操作和计算生成可行解和分离出所要求的最大加权独立集。最后给出了该算法的计算机模拟仿真结果，得到了所给问题的最大加权独立集，对算法的可行性进行了验证和总结。

关键词：DNA 计算；独立集；NP 完全问题；生物技术
中图分类号：TP301.6　　**文献标识码**：A　　**文章编号**：1009-5896(2007)11-2693-05

DNA Solution of the Maximum Weighted Independent Set

Wu Xue　Zhao Yi
(*Dept. of Electronic & Communication Engineering, East China University of Science and Technology, Shanghai* 200237, *China*)

Abstract: The DNA solution of the Maximum Weighted Independent Set (MWIS) problem based on biological technology is primarily presented in this paper. The MWIS problem is a well-known NP complete problem. The crucial point in the algorithm is to use of direct proportional length-based DNA strands to encode the vertices in weighed graphs and POA to build complete data pool, respectively, then the result comes out by a series of biological reaction and computation such as denaturation, anneal, Polymerase Chain Reaction(PCR), gel electrophoresis. And the maximum weighted independent set of the graph is found. Finally, the computer program is given to simulate this algorithm and the MWIS of the graph is also found , and the feasibility of the algorithm is validated and summarized.

Key words: DNA computing; Independent set; NP-complete problem; Biological technology

1　引言

最大加权独立集(Maximum Weighted Independent Set, MWIS)问题作为NP完全类问题中的核心问题，形成时间由来已久。它在工业过程控制，网络设计，自组织网，大规模集成电路设计及经济模型分析等诸多方面都有着广泛的运用，同时它也是组合优化研究的核心问题，有着极高的研究价值。

DNA算法思想是由Adleman于 1994 年提出的，并且利用这个算法解决了Hamilton路问题(HPP)[1]。Adleman证明了利用DNA中的碱基编码并采用生化反应作为计算工具可以并行式地，快速地找到问题的最优解。这使得人们在解决组合优化问题中有了一种全新的思路。而DNA算法也因为其超大规模的并行计算功能，逐渐地成为解决NP完全问题的一个主流算法。随着Adleman采用现代分子生物技术解决HPP之后，世界上掀起了研究DNA算法的高潮[2]，出现了一些基于DNA分子特性的新算法，用以解决典型的NP完全问题。例如旅行推销员问题[3]、最大团问题[4]和最小顶点覆盖问题[5]等。

在本算法提出之前，已经有很多算法来解决非加权的最大独立集问题。如基于图的邻接矩阵来寻找图的最大独立集的算法[6]，一种求解最大独立集问题的混合神经演化算法[7]，基于质粒的DNA算法解决最大独立集问题[8]等。在这些算法中，尤其值得一提的是基于质粒的DNA算法。该算法是Head和他的几位同事在2000年提出的。在生物科学里，作为基因载体的质粒含多种单一酶切位点，在这些含单一酶切位点上可以插入特定的DNA序列片段。Head和他的同事们将代表所要解决的图的各个顶点的DNA序列片段插入到所选的质粒的酶切位点上。通过在质粒上进行一些酶的反应，找出问题的最优解。

本文针对加权的最大独立集问题提出了一种有效的DNA算法。算法的关键技术是基于变长的DNA串来对所给图中的加权顶点进行合理的编码，并采用并行重叠放大技术(POA)[9]来建立初始数据池，再通过聚合酶链式反应(PCR)、

2006-05-08 收到，2006-11-16 改回

酶切反应[10]和凝胶电泳等一系列分子生物操作对数据池进行运算，生成了可行解和分离出所要求的最大加权独立集，从而有效地解决了加权的最大独立集问题。

本文首先介绍算法中应用的DNA分子生物操作，然后分别给出最大加权独立集问题的DNA算法和算法的生物实现，最后对算法进行计算机模拟仿真和总结。

2 DNA 分子生物操作

DNA(Deoxyribo Nucleic Acid)是一种由许多单体牢固地连接在一起的被称为脱氧核糖核酸的聚合物，是在活性细胞中起决定作用的分子，具有由两条单链相互盘绕而成的双螺旋结构。DNA的基本结构单位是核苷酸，随其碱基的不同而不同。碱基有两种类型：嘌呤和嘧啶。嘌呤分为腺嘌呤(Adenine，缩写为A)和鸟嘌呤(Guanine，缩写为G)；嘧啶分为胞嘧啶(Cytosine，缩写为C)和胸腺嘧啶(Thymine，缩写为T)。在双螺旋的DNA中，分子链是由互补的碱基配对组成，DNA的碱基排列配对方式只能是A与T配对，C与G配对，这种配对原则称为Watson-Crick碱基互补原则。通过磷酸二酯键可以组成一条DNA单链，而利用Watson-Crick碱基互补原则，单链很容易形成双链。在DNA计算中，存在许多可行的DNA分子生物操作。单链和双链可以根据不同的生物操作发生变化。

2.1 变性(Denaturation)

由于两个互补碱基之间的氢键比同一链内相邻核苷酸之间的磷酸二酯键要弱得多，则可在不破坏单链的前提下，通过对 DNA 溶液加温到 95℃左右使 DNA 分子双链分离，此过程称为变性。

2.2 复性(Renaturation)

与变性过程相反，将加热的 DNA 溶液从 95℃到 55℃慢慢进行冷却，分开的两条单链又会通过氢键结合在一起，这一过程称为复性，也称为退火(annealing)过程。

2.3 切割(Cut)

限制性内切酶可在特定子序列处切割 DNA 链。外切酶通过每次从 DNA 分子的末端去掉一个核苷酸来缩短 DNA 分子；内切酶通过破坏 DNA 分子内部的磷酸二酯键来剪切 DNA 链。各种类型的外切酶和内切酶将对应不同的切割对象(单链或双链)、切割识别位和切割方式。

2.4 聚合酶链式反应(Polymerase Chain Reaction, PCR)

PCR主要是解决DNA分子的复制问题。PCR具有令人难以置信的灵敏度与有效性：即使开始时只有该分子的一条链，PCR也能在很短的时间里产生出数百万个所需的DNA分子的拷贝。假设有DNA分子a，通过PCR来扩增a将通过重复 3 个步骤构成的基本循环来完成变性，加载引物和延伸。重复基本循环n次后，至少在理论上将产生 2^n条a的拷贝。

2.5 凝胶电泳法(Gel electrophoresis)

凝胶电泳的主要目的是测量 DNA 分子的长度。电泳技术基于 DNA 分子带有负电荷这个事实，如果将 DNA 分子置放在电场中，则 DNA 分子将向正电极方向移动。当 DNA 分子穿过凝胶向正电极方向移动时，由于凝胶网孔起着分子筛选的作用，因此小分子在通过凝胶时比大分子容易移动(快)，显然，同长度大小的分子移动时具有相同的速度，而知道了一个分子迁移的距离，就能计算出它的长度。

3 最大加权独立集问题的 DNA 算法

3.1 问题的描述

若加权图 $G=(V, E)$的顶点集 V的子集 S中的任何顶点之间都不相邻，则称 S为图 G的独立集，顶点个数最多的独立集称为最大独立集。各顶点权重之和最大的独立集称为图 G的最大加权独立集。如图 1 所示，顶点集合(4,3,1)是该图的一个独立集，且是最大独立集，但不是最大加权独立集。顶点集合(5,2)是独立集，且是最大加权独立集。最大加权独立集问题已被证明为是个 NP 完全问题。

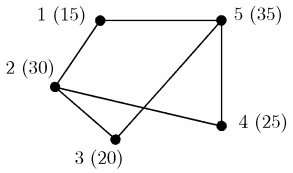

图 1 5 顶点加权图(括号内数字是权值)

3.2 最大加权独立集问题的 DNA 算法

DNA算法主要是具有并行式计算的能力，能有效解决图灵机很难解决的问题，例如NP完全问题就是在多项式时间里不能用图灵机解决而可以用DNA算法解决的问题。这主要是因为随着输入变量的增加，计算步骤呈指数级增加，而采用DNA计算却能大大提高执行速度。我们首先设计DNA序列来对问题的变量进行编码，然后通过DNA的变性、退火、切割、PCR、凝胶电泳[11]等一系列的生物操作和计算来得到问题的最优解。

本文提出用如下 DNA 算法来找出图的最大加权独立集。

步骤 1 用 N位二进制数对具有 N个顶点的图的独立集进行编码。在独立集内的顶点编成 1，在独立集外的顶点编成 0。对于图 1 中的顶点，点集(5,2)是所给图的一个最大加权独立集，它的二进制编码是 10010。点集(4,3,1)是个独立集，它的二进制编码是 01101。 依此可用 N 位二进制数对具有 N个顶点的图的所有可能独立集进行编码，本文称其为初始完全数据池。

步骤 2 删除相邻顶点都为 1 的编码。由独立集的定义知道，在独立集内不可能存在两个节点是相邻的，所以要在编好的码中去掉相邻顶点的码都为 1 的编码。例如，编码 xxx11(x 为 1 或 0)，由于图 1 中顶点 1 和顶点 2 是相邻的，

所以它们不可能存在于同一独立集中，因此两个点的码不可能同时为 1。像这样的编码还有： 11xxx，1x1xx，1xxx1，x1x1x，xx11x，因而要删除类似这样的编码。删除了这些编码之后，剩下的所有的编码就是独立集的编码。

步骤 3　从剩下的独立集的编码中，找出编码中各个 1 所对应的顶点的权重之和最大的编码，也就是最大加权独立集所对应的编码。

4　算法的生物实现

步骤 1　用 DNA 序列对图 1 中的 N 个顶点进行编码。

首先对图 1 中的 N 个顶点用 DNA 序列进行编码。每个二进制位用两个 DNA 部分来对其编码，一部分是表示此二进制位的状态(V_i)，另一部分表示此二进制位的位置(P_i)。对于代表如图 1 所示的 5 顶点的 DNA 分子，有 5 个状态部分($V_1 \sim V_5$)和 6 个位置部分($P_1 \sim P_6$)。每个顶点的码为 $P_iV_iP_{i+1}$ (i 是奇数时)和 $\overline{P_{i+1}V_iP_i}$ (i 是偶数时)。V_i 代表第 i 个顶点的状态，有两个值：0 和 1。以图 1 为例，下面用 DNA 序列对各加权顶点进行编码。各加权顶点 DNA 序列的长度按如下规则产生：

(1) 当顶点的状态为 0 时，即顶点不在独立集内时，顶点 DNA 序列的长度为固定值 10；

(2) 当顶点的状态 1 时，即在独立集内时，顶点 DNA 序列的长度为该顶点的权值减去 5。这样，代表各顶点 DNA 串的长度为：$V_1^0 \sim V_5^0$ 是 10；V_1^1 是 10；V_2^1 是 25；V_3^1 是 15；V_4^1 是 20；V_5^1 是 30。P_i 用 20 个碱基对其编码。其中，状态 1 中的 DNA 链含有限制型内切核酸酶(restriction enzyme)的识别位(restriction site)，用来切断不需要的 DNA 链。本文用到的内切酶及其识别位如表 1 所示。这样图 1 中各顶点的 DNA 序列编码如表 2 所示。

表 1　本文用到的限制性内切酶及其识别位(5′ 到 3′ 方向)

Restriction enzyme	Restriction site
BamHI	GGATCC
EcoRI	GAATTC
HaeIII	GGCC
HindIII	AAGCTT
HpaII	CCGG

步骤 2　利用 POA 技术来生成初始完全数据池。

POA技术分若干个阶段反复进行复性、延伸和变性，如图 2 所示。所生成的初始完全数据池如图 3 所示，其中每个顶点或取状态 0 或取状态 1，因而初始完全数据池共有 2^5 种完全数据链。

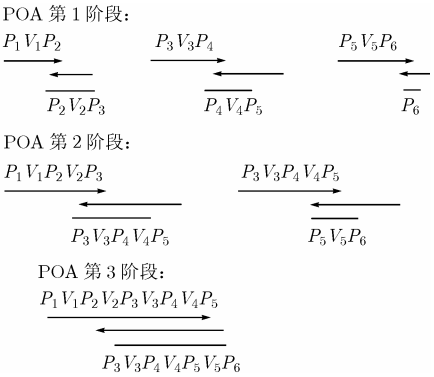

图 2　POA 各循环步骤

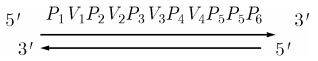

图 3　初始完全数据池(共有 2^5状态)

表 2　各加权顶点编码的 DNA 序列(5′ 到 3′ 方向，粗体的字母是内切酶的识别位)

加权顶点	DNA 序列
$P_1V_1^0P_2$	AATTCGGCTATCGTACGATCacgtcggttcCTA GAGTACGGTAGGTACGT
$P_1V_1^1P_2$	AATTCGGCTATCGTACGATCa**ggatcc**tagCTA GAGTACGGTAGGTACGT
$\overline{P_3V_2^0P_2}$	AACACGTATGATACGATCGTgcatcgatgcACG TACCTACCGTACTCTAG
$\overline{P_3V_2^1P_2}$	AACACGTATGATACGATCGTactca**gaattc**atgc gatcatccagACGTACCTACCGTACTCTAG
$P_3V_3^0P_4$	ACGATCGTATCATACGTGTTatcgacgtacCGC CGTCGGCTGGATATTAG
$P_3V_3^1P_4$	ACGATCGTATCATACGTGTTgtacct**ggcc**tacgg CGCCGTCGGCTGGATATTAG
$\overline{P_5V_4^0P_4}$	GCTCAGCTATTGTATGGCCcatcggatcaCTAA TATCCAGCCGACGGCG
$\overline{P_5V_4^1P_4}$	TGCTCAGCTATTGTATGGCCcataggcata**aagct t**ccatCTAATATCCAGCCGACGGCG
$P_5V_5^0P_6$	GGCCATACAATAGCTGAGCAgatcgatcgaCAG CCAATCGAACGGTCGAA
$P_5V_5^1P_6$	GGCCATACAATAGCTGAGCAtacgatcaactacat **ccgg**tagcgaatcga CAGCCAATCGAACGGTCGAA

步骤 3　以 P_1 和 $\overline{P_6}$ 为引物进行 PCR，使初始数据池中完全数据链大量地被复制，以使数据池中含有大量的完全数据链，减少运算误差的产生。

步骤 4　对初始完全数据池中的 DNA 串进行筛选。

这一步利用限制性内切酶来进行。在第 1 步的 DNA 序列编码中，将所有状态位 1 的顶点即独立集内的顶点的编码中加入了限制性内切酶的识别位。限制性内切酶在其相应的位点切割。一个限制酶将在特定的识别位与 DNA 粘合，然

后从识别位切割 DNA 链。限制酶通过在一个核苷酸的 3′-端产生 OH 键和在另一个核苷酸的 5′-端产生磷酸键来切割相邻的核苷酸之间的磷酸二脂键。由于各顶点所对应为 1 的状态的 DNA 串中含有不同内切酶的识别位，因此很容易就能够将图中相邻顶点的状态都为 1 的顶点的 DNA 串去掉。

例如，相邻顶点 1 和顶点 2，它们不可能在同一独立集内，因而它们的状态不可能同为 1。首先将初始数据池溶液分在两个试管中，在一个试管中加入内切酶 BamHI，另一个试管中加入内切酶 EcoRI，一段时间反应后，再将两个试管中溶液合并在一个试管中。最后试管中就不可能同时含有状态都为 1 的顶点 1 和顶点 2(如图 4 所示)。依照这种方法，加入不同的内切酶，反复进行，得到的数据池中的各完整 DNA 串所代表的就是所给图的各个独立集。

步骤 5　寻找数据池中最长的 DNA 链。

为了找出代表图中最大加权独立集的DNA链，需要通过凝胶电泳[12]找出数据池中最长的DNA链。根据本文的算法设计，通过凝胶电泳可测定最长链的长度是 205bp，即是代表所给图的最大加权独立集的DNA链长度。

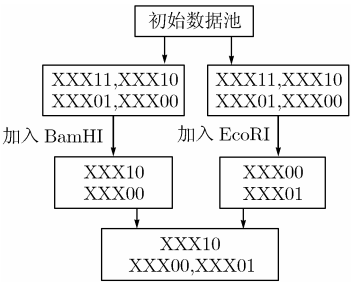

图 4　删除过程(X 为 0 或 1)

步骤 6　输出运算结果。

由上一步骤知道了代表最大加权独立集的DNA链，但还不知道该DNA链具体代表所给图中的哪几个点构成的点集。采用基因工程的方法[13]对代表最大加权独立集的DNA链(如图 5 所示)进行测序，可知最大加权独立集为(5,2) 。

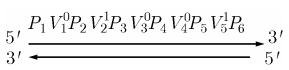

图 5　代表最大加权独立集 DNA 链

5　算法的计算机模拟仿真

由于计算机理的不同，计算机模拟和生化实验反应是有区别的。即计算机是串行机制，因而 DNA 计算的并行性在模拟仿真实现过程中难以体现，但是我们关注的是通过计算机模拟仿真证明本算法设计的正确性。

本文计算机模拟采用 C++语言编程来实现这个算法。限于文章的篇幅，这里只给出仿真的结果。代表所给图的最大加权独立集的 DNA 链如下(图 6)。

```
代表所给问题的最大加权独立集的DNA串是：
P1V10P2V21P3V30P4V40P5V51P6
AATTCGGCTATCGTACGATCacgtcggttcCTAGAGTACGGTAGGTACGTctggatgatcgcatgaattctgagtACGAT
CGTATCATACGTGTTatcgacgtacCGCCGTCGGCTGGATATTAGtgatccgatgGGCCATACAATAGCTGAGCAtacga
tcaactacatccggtagcgaatcgaCAGCCAATCGAACGGTCGAA
Press any key to continue
```

图 6　仿真结果

通过上面的仿真过程知道，代表最大加权独立集的DNA串的长度为 205bp，和预期的生物试验的结果是相同的。据此，找出了代表最大加权独立集的 DNA 链是 $P_1V_1^0P_2\ V_2^1\ P_3V_3^0P_4\ V_4^0\ P_5V_5^1P_6$。由该 DNA 串知，图的最大加权独立集是(5,2)。仿真结果和预期试验结果完全相同，从而验证了该算法的正确性。

6　结束语

本文提出的算法与其它算法相比，具有如下优点：(1)本算法解决了加权的最大独立集问题，使得应用背景更加广泛。(2)算法在对初始数据池处理中运用了并行重叠放大技术(POA)，使得算法的生物实现变得简单，且在产生初始完备解空间方面比Adleman算法中应用的杂交/连接方法更为有效[14]。(3)与一般算法相比，DNA技术是并行式的运算，运算速度快，对于解决顶点数很多的图论问题的效率非常高。(4)本文给出了计算机仿真，使算法的正确性得到了验证。

本文讨论的是最大加权独立集问题，利用 DNA 分子的生物结构及 DNA 分子生物操作使数学上的最大加权独立集问题能够用生物技术的方法很好地加以解决。但具体的生物实现过程亦有不足之处。例如，当权值之间不存在较大倍数关系时，用本文的方法可以很好地解决这个问题，但是当各权值之间存在较大的倍数关系时，本方法似显繁琐。由于 DNA 链的数目随着顶点数的增加而会呈指数级增长，所以当顶点数非常大时，DNA 链的数目将会随之激增，所需要的酶的种类和数量也都会增加，从而使计算的成本增加。本方法需要执行的筛选 DNA 串的操作步骤正比于图中边的数目，且操作中内切酶的切除可能会不完全。这些问题都是需要我们进一步研究解决的。

参 考 文 献

[1] Adleman L. Molecular computation of solution to combinatorial problems. *Science*, 1994, 266(11): 1021–1024.

[2] Maley C C. DNA Computation: theory, practice, and prospects. *Evolutionary Computation*, 1998, 6(3): 201–229.

[3] 董亚非, 谭刚军, 等. 基于粘贴系统求解 TSP 问题. 系统仿真学报, 2005, 17(6): 1299–1306.

[4] Ouyang Q, Kaplan P D, and Liu Shumao, *et al*.. DNA solution of maximal clique problem. *Science*, 1997, 278(17):

446–449.

[5] 高琳, 许进. 最小顶点覆盖问题的 DNA 分子算法. 系统工程与电子技术, 2004, 26(4): 544–548.

[6] 杨铀, 段滋明. 求解图的最大独立集的一种算法. 电脑开发与应用, 2002, 15(6): 13–14.

[7] 李有梅, 徐宗本, 孙建永. 一种求解最大独立集问题的混合神经演化算法. 计算机学报, 2003, 26(11): 1538–1545.

[8] Head T, *et al.*. Computing with DNA by operating on plasmids. *Biosystems*, 2000, 57(2): 87–93.

[9] Kaplan P D, Ouyang Q and Thaler D S, *et al.*. Parallel overlap assembly for the construction of computational DNA libraries. *Journal of Theoretical Biology*, 1997, 188(3): 333–341.

[10] Ibrahim Z. Towards solving weighted graph problems by directproportional length-based DNA computing. Research Report, IEEE Computational Intelligence Society(CIS)Walter J Karplus Summer Research Grant 2004.

[11] G.Paun, G. Rozenberg, et al.. (德)著. 许进. 等. 译. <DNA 计算>. 清华大学出版社, 2004: 1–57.

[12] Amos M. DNA computation[PhD Thesis], The University of Warwick, UK, 1997.

[13] Sambrook J, Fritsch E F and Maniatis T. Molecular Cloning. NY: Cold Spring Harbor Laboratory Press, 1987: 1–56.

[14] Ferretti C, Mauri G, and Zandron C. DNA computing. Heidelberg: Springer Berlin, 2005: 215–223.

吴　雪：　女，1957 年生，副教授，研究领域为 DNA 计算、图论与通信网的性能优化、无线传感器网络系统优化设计、智能信息处理等.

赵　艺：　男，1983 年生，硕士生，研究领域为 DNA 计算.