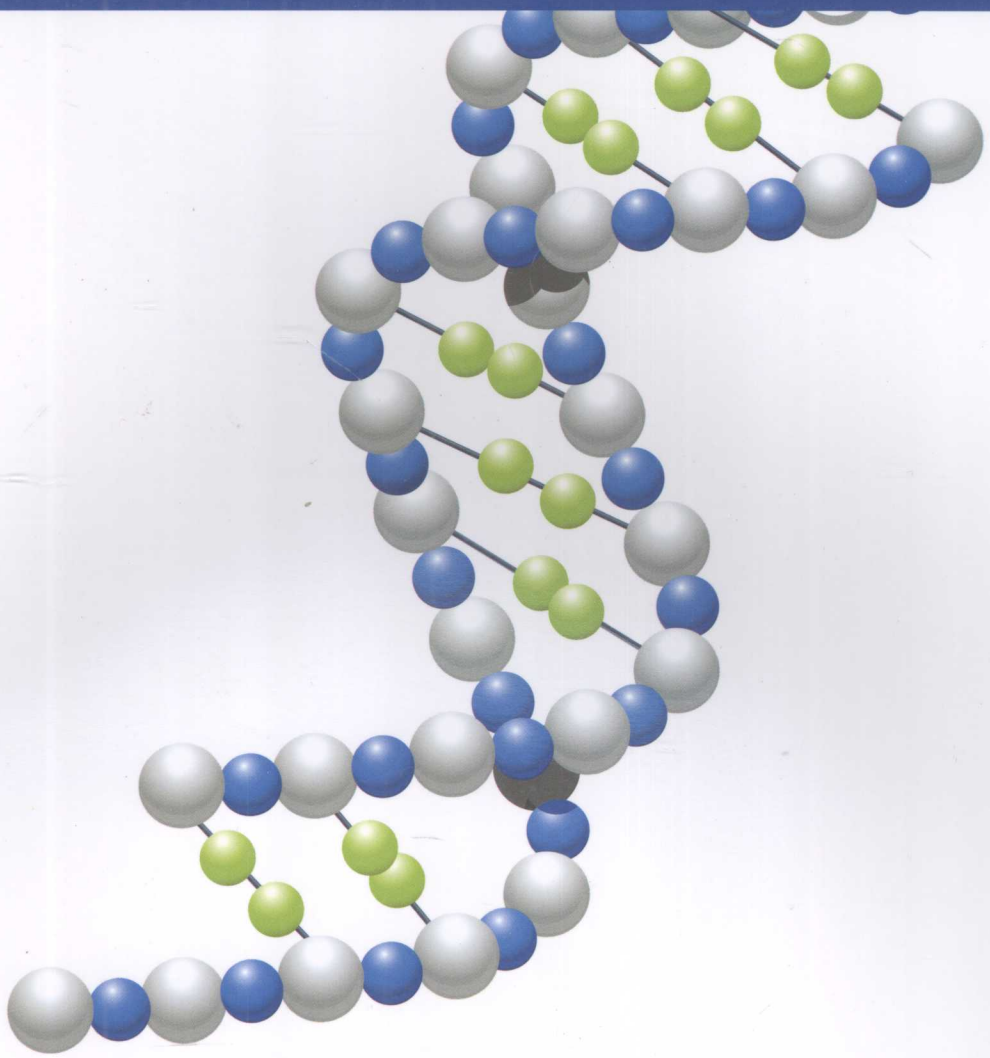


# DNA编码序列的设计与优化

王延峰 崔光照 著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

郑州轻工业学院
河南省信息化电器重点实验室

# DNA 编码序列的设计与优化

王延峰　崔光照　著




電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

## 内 容 简 介

本书全面系统地介绍了 DNA 编码理论的基本内容及 DNA 编码序列的各种设计方法，集中涵盖了作者近年来在该领域内的研究成果。全书共 14 章，在较系统地介绍了 DNA 计算产生的背景、意义、基本思想、与 DNA 计算相关的分子生物学基础、DNA 编码问题的定义和 DNA 编码的分子生物学约束及其相关研究进展的基础上，提出了一种基于统计学原理的、无须实验就可确定各评价指标的权重系数的方法——组合权重法，并据此建立了一套 DNA 编码系统评价模型。同时，详细探讨了启发式算法（如 Hopfield 神经网络算法、模拟退火遗传算法、文化遗传算法、非支配排序遗传算法、蚁群算法、粒子群优化算法、人工鱼群算法、野草算法）和搜索算法（如剪枝算法、随机产生实时过滤算法）等在 DNA 编码序列设计中的应用研究。

本书既可作为信息与计算科学、分子生物学、应用数学、计算机、系统科学等专业的研究生教材或高年级本科生的选修课教材，也可供相关专业科研人员参考。



**图书在版编目（CIP）数据**

DNA 编码序列的设计与优化 / 王延峰，崔光照著. —北京：电子工业出版社，2013.9
ISBN 978-7-121-21324-3

Ⅰ. ①D… Ⅱ. ①王… ②崔… Ⅲ. ①脱氧核糖核酸—计算方法—编码—设计 Ⅳ. ①Q523-39

中国版本图书馆 CIP 数据核字（2013）第 198758 号

责任编辑：董亚峰
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×1 092　1/16　印张：13.00　字数：333 千字
印　　次：2013 年 9 月第 1 次印刷
定　　价：38.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

# 前　言

1994 年，美国南加州大学的 Adleman 教授针对图论中的一个 NP 完全问题——有向哈密尔顿路问题，首次利用 DNA 分子，通过 DNA 编码，并借助连接、变性、复性、PCR 扩增、电泳等一系列生物实验操作，完成了对该问题的求解。该研究成果随即引起了计算机科学、数学、分子生物学等领域科学家们的极大兴趣。该成果的重要意义在于其采用了一种全新的计算介质——DNA 分子，以分子生物学技术实现了目前传统计算机无法解决的困难问题的求解，并开发了该计算模式本身所固有的潜在的巨大并行性。

DNA 计算的核心是将编码后的 DNA 序列作为输入信息，在试管内或其他载体上经过一定时间的可控生物化学反应，然后从反应产物中得到全部最优解。DNA 计算的最大优点在于 DNA 分子中的遗传密码及其生化反应的巨大并行性。DNA 计算的诸多优点及其应用前景极大地吸引了不同学科、不同领域的众多科学家，特别是计算机科学、分子生物学、数学、物理学、化学以及信息学领域内的科学家们。DNA 计算有望成为人类科学发展史上的又一个里程碑，因为 DNA 计算有可能解决目前在传统计算机上无法解决的许多问题，如密码破译、NP 困难问题以及工程领域中最大的难题——局部极小值问题等。目前，国际上 DNA 计算领域以及由此衍生出来的 DNA 自组装、DNA 纳米技术等领域的研究成果层出不穷，极大地促进了数学、计算机科学和生物学等学科领域的相互交叉与渗透。

编码问题是 DNA 计算及其相关研究领域中的核心问题。编码质量不仅决定着能否按预期目标进行特异性杂交（DNA 计算）能否顺利进行，而且还决定着杂交反应后所生成的解空间的大小。从分子生物学角度来看，DNA 序列自身的物理化学属性决定了 DNA 编码序列的存在形式；而 DNA 编码序列的热力学属性则是杂交反应的动力源泉。基于此，本书主要围绕 DNA 序列的物理化学属性以及热动力学特性进行 DNA 编码理论的研究与设计。本书的主要内容包括：第 1 章介绍了 DNA 计算产生的背景、意义、研究进展、基本思想以及 DNA 计算研究所面临的问题。第 2 章介绍了与 DNA 计算及 DNA 编码相关的分子生物学基础，包括 DNA 的理化性质及 DNA 杂交反应的热力学基础。第 3 章详细阐述了 DNA 计算中的编码问题、DNA 编码的分子生物学约束及其相关研究进展。第 4 章提出了一种随机产生实时过滤算法的 DNA 编码序列设计方法。该方法在综合考虑 DNA 计算中编码序列的约束条件的基础上，根据约束条件之间的相互制约关系，采用整体优化的思想，首先将各约束条件进行归类；然后依据各约束条件的计算时间复杂度和约束强弱程度进行优化组合排序，最后采用随机产生实时过滤算法产生所需数目的 DNA 编码序列。在综合考虑计算编码序列生物约束条件之间的制约关系的基础上，第 5 章提出了一种基于统计学原理的、无须实验就可确定各评价指标的权重系数的方法——组合权重法，并据此建立了一套 DNA 编码序列系统评价模型。利用该评价模型可以对编码序列集合进行合理、客观的评价。另外，该系统评价模型对采用演化策略进行 DNA 编码序列的设计研究时构造适应度函数也具有重要的指导意义。第 6 章至第 14 章详细讨论了 Hopfield 神经网络算法、模拟退火遗传算法、

文化遗传算法、非支配排序遗传算法、蚁群算法、剪枝优化算法、粒子群优化算法、人工鱼群算法、野草算法等在 DNA 编码序列设计与优化中的应用。

本书第 1～10 章由王延峰执笔撰写，第 11～14 章由崔光照执笔撰写。在本书的写作过程中，得到了作者单位、同事及家属们的有力支持与帮助。感谢作者的同事王子成博士和张勋才博士，其中，王子成博士参与了本书第 11 章的方案及实验设计工作，张勋才博士参与了本书第 14 章的方案及仿真工作，不仅如此，作为本书的最初读者，他们还对本书的撰写风格、部分章节的编排等提出了许多宝贵意见。另外，牛莹、申永鹏、牛云云、周君和、孙军伟、郑艳、卢伟丽、魏东辉、白学文、侯贺伟、叶盟盟、田桂花、韩琴琴、赵涛涛、王燕等硕士研究生也参与了部分方案和文字工作。在此，作者对上述帮助者一并致谢。

本书是作者在整理攻读博士学位期间以及工作后研究成果的基础上撰写而成的。首先感谢我们的博士生导师许进教授。无论是在攻读博士学位期间，还是在之后的工作过程中，许老师睿智的思想、严谨的治学态度以及平易近人的高尚品格时刻都在感染着我们，从他身上，我们不仅学到了专业方面的知识，同时还学会了如何做人，如何治学。感谢博士生导师潘林强教授，潘老师高度的责任心、敬业的精神、乐于助人的品格以及严谨求实的治学态度都给我们留下了深刻的印象。衷心感谢潘教授给予我们的莫大帮助。感谢我们的工作单位郑州轻工业学院，是她为我们开展科研工作提供了一个宽松的平台，一个舒适的载体，一个自由发挥的舞台。

本书得到国家自然科学基金（60573190，60773122，61070238，60970084，61272022）、河南省基础与前沿技术研究计划项目（082300413203，092300410166，12230041321）、河南省创新型科技人才队伍建设工程支持项目（094100510022，124200510017）、中国博士后科学基金以及郑州轻工业学院博士基金（2010BSJJ002）等的资助。

由于作者水平有限，加之时间仓促，书中难免有不妥之处，欢迎广大读者批评指正。

作者 E-mail：yanfengwang@yeah.net。

作　者

郑州轻工业学院

河南省信息化电器重点实验室

2013 年 5 月

# 目　录

# 第1章 DNA计算

量子理论、生物基因工程和传统计算机技术（本书把以微处理器为基础的电子计算机统称为传统计算机）被认为20世纪的三大科学技术革命。自1942年世界上第一台“ABC计算机”（Atanasoff-Berry Computer）诞生至今，尽管只有短短70年的历程，但对人类的生产活动和社会进步产生了极其重要的影响。

理论计算机科学家们将计算问题划分为3类：容易类、困难类和不可计算类[1]。对于容易类问题，传统计算机足以胜任。但在处理困难类问题时，由于受传统计算机基本工作原理（按串行运算、线性存储方式进行符号处理）的限制，其运算所需要的时间会随着待求解问题规模的增大而呈指数级增长。

20世纪下半叶，传统计算机蓬勃发展的主要基础是基于硅基材料的微电子器件的不断更新换代。目前，建立在以硅基材料为基础、互补金属氧化物半导体（Complementary Metal Oxide Semiconductor, CMOS）器件为主流的半导体集成电路技术中，其主流产品的特征尺寸已缩小至0.18～0.1μm。硅基技术已高度成熟，硅基CMOS芯片的应用也日益扩大，以硅平面的加工工艺技术作为技术基础的高新加工技术也将持续发展下去。据国际权威机构预测，到2016年，微电子器件最小特征尺寸将缩小到13nm。然而，硅基CMOS的发展同任何事物一样，都有其产生、发展、成熟和衰亡的过程，不可能按照摩尔定律所揭示的规律长期发展下去。随着特征尺寸的缩小，最终将达到器件结构的诸多物理限制。量子物理学理论已经成功地预测出芯片微处理能力的增长不可能长期地保持下去[2]。与此相对应，由于受运算速度和存储容量的限制，传统计算机目前尚无法对困难类问题进行有效求解。与此同时，随着社会和科学技术的进步与发展，工程领域内也在不断地涌现出许多新的复杂或巨复杂系统，在这些复杂系统的研究中，各种棘手的困难类问题又随处可见，如背包问题、大数分解问题、椭圆曲线上的离散对数问题等。这些困难类问题不仅应用广泛，而且在计算机理论科学的研究中也具有十分重要的地位。特别地，对于其中的某些困难类问题，如多数公钥系统中的难解问题，由于要对基于其上的密钥系统的安全性进行分析，所以必须找出问题的精确解。

为了满足日益增长的大规模和超大规模计算需求，发展高性能计算和提高计算性能迫在眉睫。目前发展高性能计算有两条途径。第一条途径是基于现有的半导体集成电路技术，即微处理器技术，该途径可以通过提高并行处理能力来实现。事实上，随着超级计算机体系结构和算法研究的日臻成熟，利用超级计算机所提供的并行计算资源，通过资源或空间重叠的方式来求解困难类问题已经取得了重大进展，主要体现在利用并行处理技术可以将待求解困难类问题的规模成倍增长。尽管超级计算机技术仍在不断进步，但利用并行处理技术来求解困难类问题的能力仍然不够。发展高性能计算第二条途径是突破硅半导并体器件框架，发展非传统的新技术，如超导计算[3]、量子计算[2, 4]、生物计算[5-9]与光计算[10]等。提高计算性能可以从两方面着手，即自顶向下规划和自底向上设计（Top-Down Planning & Bottom-Up Design），前者致力于硬件方面的研制；而后者则致力于计算模型和算法设计，从根本上突破传统计算机与电子技术的局限。比如量子计算和 DNA 计算都为求解复杂困难类问题开辟了崭新的思路。

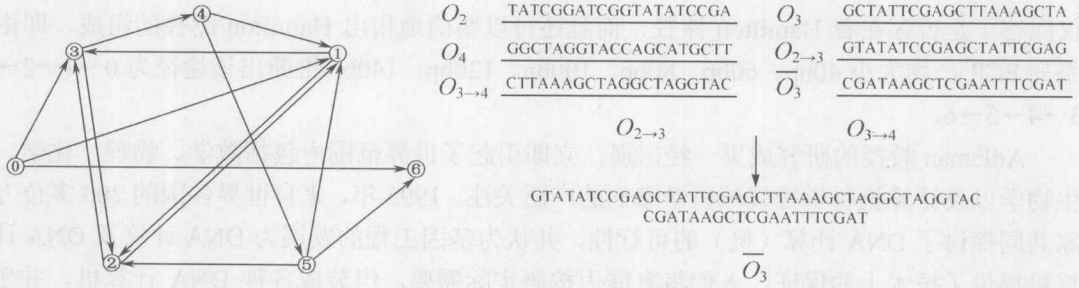
## 1.1 DNA 计算的出现

在分子水平上进行计算的概念最早由 Feynman 于 1959 年提出[11]。与此同时，在生物学研究领域，分子生物学正逐渐出现并日趋成熟，使生物学的研究深入到了分子水平。进入 20 世纪 80 年代，随着人们对分子生物学理论的了解日益加深以及现代生物化学和生物工程技术的日趋完善，进行分子计算的物质基础已基本具备。

1994 年，美国南加州大学的 Adleman 教授以脱氧核糖核酸（Deoxyribonucleic Acid，DNA）分子为计算介质，以现代分子生物技术为手段，在生物实验室内成功地解决了 7 个顶点的有向 Hamilton 路径问题（Hamiltonian Path Problem，HPP）[6]，开创了分子计算、特别是 DNA 计算的新纪元。

在介绍 Adleman 教授开拓性工作之前，首先了解一下 Hamilton 路径问题。Hamilton 路径问题是指在一个有多个城市的地图网络中，寻找一条从既定起点到既定终点，并且沿途恰好经过所有其他城市仅一次的路径。即对于一个给定的网络，在确定起点和终点后，如果存在着一条路径穿过这个网络，就说这个网络存在着 Hamilton 路径。虽然有很多算法尝试解决 Hamilton 路径问题，但都面临着“指数爆炸”（又称“组合灾难”）的计算复杂性问题。Hamilton 路径问题在 20 世纪 70 年代初被证明是“NP 完备”的，即针对该问题，目前还没有有效的，即所谓的“多项式时间”的求解算法。

针对有向 Hamilton 路径问题，Adleman 给出了一个如图 1-1 所示的由 7 个顶点和 14 条有向线段组成的简单有向 Hamilton 路径问题的 DNA 算法实证。在图 1-1 中，设起点 $V_{in}$=0、终点 $V_{out}$=6。可以很直观地看出，图中存在着一条 Hamilton 路径：0→1→2→3→4→5→6。若有向线段 2→3 不存在，则没有 Hamilton 路径；若 $V_{in}$=3、$V_{out}$=5，则也找不到 Hamilton 路径。




（a）有向图　　（b）利用DNA分子对图进行编码

图 1-1　Adelman 实验

当 $V_{in}$=0 和 $V_{out}$=6 时，存在着唯一的 Hamilton 路径：0→1→2→3→4→5→6，如图 1-1（a）所示。对于图中的每个顶点 $i$，对应于一个随机产生的 20bp 的单链 DNA 片断 $O_i$，如图 1-1（b）所示。（图中，$O_2$、$O_3$ 和 $O_4$ 分别对应于顶点 2、3 和 4），对于图中的边 $i\to j$，一个寡聚核苷酸片段 $O_{i\to j}$ 由顶点 $i$3′ 端的 10 个碱基和顶点 $j$5′ 端的 10 个碱基组成（图中边 2→3 对应于 $O_{2\to3}$，边 3→4 对应于 $O_{3\to4}$）。对于图中的每个顶点 $i$，$\overline{O_i}$ 是 $O_i$ 的 Watson-Crick 补（图中 $\overline{O_3}$ 与 $O_3$ 相互补，$\overline{O_3}$ 作为一个夹板用于链接 $O_{2\to3}$ 和 $O_{3\to4}$），除 $\overline{O_3}$ 以外，所有的寡聚核苷酸均书写为 5′ → 3′ 。

针对上述实例，Adleman 设计了以下算法：

Step1：组建图中所有可能存在的路径集合；

Step2：筛选出以 $V_{in}$ 开头、以 $V_{out}$ 结尾的途径；

Step3：筛选出共经历 $n$ 个顶点的路径；

Step4：筛选出经历了每个顶点的路径；

Step5：经过上述 3 次筛选，考察有否路径存在，即回答了 Hamilton 路径问题。

以 DNA 分子为介质，以现代分子生物学方法为手段，Adleman 通过生化实验成功地完成了上述算法步骤。具体方法如下。

首先，设计并合成若干个 20bp 长的寡聚核苷酸以对应于每个顶点和有向线段。代表顶点 $i$ 的寡聚核苷酸记为 $O_i$；代表有向线段 $i\to j$ 的记为 $O_{i\to j}$，其前 10bp 为 $O_i$ 的后 10bp，其后 10bp 为 $O_j$ 的前 10bp。另外，顶点 $O_i$ 的互补链记为 $O_i$。然后将 $\overline{O_i}$ 、$O_{i\to j}$ 共 21 种寡聚核苷酸混合、连接。容易想象，在 $\overline{O_i}$ 的指导下，各条 $O_{i\to j}$ 相互连接，构成了所有可能路径的总集。即完成了算法的第一步。

接着，应用聚合酶链式反应（Polymerase Chain Reaction，PCR）技术，以第一步连接产物为模板，以 $O_0$、$\overline{O_6}$ 为引物进行扩增，所得产物即为算法第二步的结果。

随后，将 PCR 产物经过 Agarose 凝胶电泳，可以将相差 20bp 的各个条带分离开，其中，140bp 条带显然就是经历了 7 个顶点的路径集，切胶回收即完成了算法第三步。

然后，应用亲和纯化法，用磁珠标记的 $\overline{O_1}$ 与回收产物（经变性为单链）结合，纯化得到含 $O_1$ 的子集，再依次用 $\overline{O_2}$ 、$\overline{O_3}$ 、$\overline{O_4}$ 、$\overline{O_5}$ 亲和纯化，完成算法第四步的筛选。

再次 PCR 检测即为算法的第五步。

事实上，Adleman 用 $O_0$、$\overline{O_i}$（$i$=1, 2, …, 6）引物对进行了一系列 6 个 PCR 反应，不仅回答了是否存在着 Hamilton 路径，而且还可以明确地指出 Hamilton 路径的组成，即由系列 PCR 产物大小 40bp、60bp、80bp、100bp、120bp、140bp 推断出该途径为 0→1→2→3→4→5→6。

Adleman 教授的研究成果一经出现，立即引起了世界范围内包括数学、物理、化学、生物学以及计算机科学等领域科学家们的广泛关注。1995 年，来自世界各国的 200 多位专家共同探讨了 DNA 计算（机）的可行性，并认为基因工程的发展为 DNA 计算及 DNA 计算机提供了技术上的保证，人们将有能力按照实际需要，组装成各种 DNA 计算机，并乐观地认为，一旦 DNA 计算机研制成功，其运算能力是传统计算机望尘莫及的，是一个极具开发价值的研究领域。

和传统计算机相比，DNA 计算（机）具有以下突出优点[12, 13]：

（1）高度并行。DNA 计算（机）一周的运算量相当于所有传统计算机问世以来的总运算量。

（2）海量存储能力。作为信息的载体其储存容量非常大，1 立方米的 DNA 溶液可存储 $10^{23}$ 个二进制数据，远远超过目前所有传统计算机的总储存量。

（3）低能耗。DNA 计算（机）所消耗的能量只有一台传统计算机完成同样计算所消耗能量的 $10^{\{-10\}}$。

（4）资源丰富。现在从植物和动物中提取 DNA 的技术已经非常成熟，而在自然界中，植物和动物处处可见。

DNA 计算（机）的上述优点及其应用背景极大地吸引了不同学科、不同领域的研究者，尤其是计算机科学、分子生物学、数学、物理学、化学以及信息学领域内的科学家们。由于为解决传统计算机无法解决的许多问题（如密码破译、NP 困难问题以及工程领域中的局部极小值问题等）提供一条可行的潜在途径，所以，DNA 计算（机）有望成为人类科学史上的一个新的里程碑。同时，通过 DNA 计算领域的研究，也可以极大地促进数学、计算机科学、生物学和材料学等学科领域的相互交叉与渗透。

## 1.2 DNA 计算的研究进展

1995 年，美国普林斯顿大学的 Lipton 教授在 Adleman 教授思想的启发下，给出了以下一个简单的可满足性（Satisfiability，SAT）问题的 DNA 计算模型[12]：

$$(x_1 \vee x_2) \wedge (\overline{x_1} \vee \overline{x_2})$$

式中，$\overline{x_1}$ 和 $\overline{x_2}$ 分别为布尔变量 $x_1$ 和 $x_2$ 的补，即 $x_i = 1 \Leftrightarrow \overline{x_i} = 0$。

Lipton 方法与 Adleman 的方法基本相同。通过构造一个接触网络图 $G$，将 SAT 问题的解空间映射为通过接触网络图 $G$ 的始点 $a_1$ 和终点 $a_{n+1}$ 的所有 Hamilton 路径，然后对有向图中的所有顶点和边进行编码。即接触网络图 $G$ 中的顶点和边均用长度为 20bp 的寡聚核苷酸片断表示，任一顶点 $i$ 用 $p_iq_i$ 表示，$p_i$ 和 $q_i$ 分别为顶点 $i$ 的前 10bp 碱基和后 10bp 碱基构成的寡聚核苷酸片断；对任一个有向边 $i \rightarrow j$ 用 $\overline{q_i p_j}$ 表示。将编码顶点和边的核苷酸片断

放入初始试管 $t_0$ 中，经过充分反应后就会形成代表接触网络图 $G$ 中的各种有向路的寡聚核苷酸片断。然后以 $\overline{p_{a_1}}$ 和 $\overline{p_{a_3}}$ 为引物搜索出试管 $t_0$ 中以 $a_1$ 开始以 $a_3$ 结尾的有向路。再从试管 $t_0$ 中搜索出第一位为 1（$x_1$=1）的寡聚核苷酸片断并放入试管 $t_1$ 中，剩下的放入试管 $t_1'$ 中；然后再从试管 $t_1'$ 中搜索出第二位为 1（$x_2$=1）的 DNA 分子放入试管 $t_2$ 中；将试管 $t_1$ 和 $t_2$ 合并为 $t_3$，得到满足第一个子句的寡聚核苷酸片断。从试管 $t_3$ 中搜索出第一位为 0（$\overline{x_1}$ =0）的寡聚核苷酸片断，并放入试管 $t_4$ 中，剩下的放入试管 $t_4'$ 中；然后再从试管 $t_4'$ 中搜索出第二位为 0（$\overline{y}$ =0）的 DNA 分子放入试管 $t_5$ 中；将试管 $t_4$ 和 $t_5$ 合并为 $t_6$，得到满足第二个子句的寡聚核苷酸片断。最后检查试管 $t_6$，如果有寡聚核苷酸片断，说明该问题是可满足的；否则，该问题是不可满足的。

Lipton 方法分为以下几种计算操作。

（1）Separation（$T$, $O$）：把 DNA 的分子序对分为两部分，一部分包含分子 $O$，另一部分没有包含分子 $O$。

（2）Extraction（$T$, length）：取出长度为第二个自变量值的 DNA 序对。

（3）Cut（$T$，剪切酶）：利用剪切酶，将 DNA 的分子序对剪开。

（4）Anneal（$T$）：将在 $T$ 中的 DNA 单链结合成双链。

（5）Copy（$T$）：将 $T$ 复制一份。

（6）Detect（$T$）：如果 $T$ 中有分子存在，则回答“Yes”，否则回答“No”。

（7）Read（$T$）：读出 $T$ 中的 DNA 分子序对的顺序。

（8）合并（Merge）：这个操作是将两个试管 $P_1$ 和 $P_2$ 注入一个新的试管 $Q$ 中。

（9）附加（Append）：给予一个试管 $P$ 和某种短链 $S$，该操作会把 $S$ 附加到 $P$ 中每个分子的最后面。

Lipton 的主要贡献在于通过对 DNA 序列进行布尔分矢量编码，不仅使 DNA 分子能够模仿数字电子的逻辑门电路做出“是”与“非”的判断，从而具有了逻辑判断能力，更重要的是把基于 Adleman 思想的 DNA 计算模式一般化，使其更具通用性。

1997 年，Ouyang 等利用 DNA 计算解决了另一 NPC 完全问题——图的最大团问题（Maximal Clique Problem，MCP）[1]。数学上，“团”是指每两个顶点之间均有线段相连的一系列顶点所组成的图形，故所谓的 MCP 就是在一个 $N$ 个顶点、$M$ 条线段组成的网络图中，求最大的团具有几个顶点。与 Hamilton 路径问题相类似，MCP 同样是具有“指数爆炸”特征的计算难题。

Ouyang 等利用 DNA 计算解决的是一个如图 1-2 所示的由 6 个顶点和 11 条线段所构成的简单网络。容易看出，顶点（2, 3, 4, 5）组成了最大团，所以这个 MCP 的解为 4。

Ouyang 采用的算法分为以下三步。

Step1：设计一个 6 位的二进制数来描述团，该数第 $n$ 位若为 1 则表示顶点 $n$ 位于此团中，若为 0 则表示顶点不在此团中。如 Clique（4, 1, 0）可描述为二进制数 010011，Clique（5, 4, 3, 2）为 111100。可见，6 位二进制数的集合就是 6 顶点网络中所有可能的团的总集。

Step2：画出“互补图”，即各顶点由原图中所没有的线段相连构成的图。在第一步总集中有这些线段的二进制数显然不是真正的团，应予以排除。由互补图易知这些数为：$xxx1x1$，$1xxxx1$，$1xxx1x$，$xx1x1x$（$x$= 0 或 1）。剩余的数即为原图中真正团的集合。

Step3：排列第二步所得集合，含最多 1 的二进制数就是最大团，其 1 的个数即最大团问题的答案。

Ouyang 通过精心设计，合成了 12 条寡聚核苷酸，分别编码第 1～6 位的“0”和“1”。每一条由三部分组成：两头的 $P_i$, $P_{i+1}$ 部分分别为 20bp 的位置符，中间的 $V_i$ 则为数值符（10bp 表示“0”，0mer 表示“1”）。混合在一起，通过 PCR 循环，互补链部分退火结合、并在 Taq 酶作用下延伸，若干循环后即得到代表二进制数 000000～111111 的所有 DNA 片断。为去除不足 6 位的延伸，用 $P_0$、$P_6$ 为引物进行扩增，便完成了算法第一步。

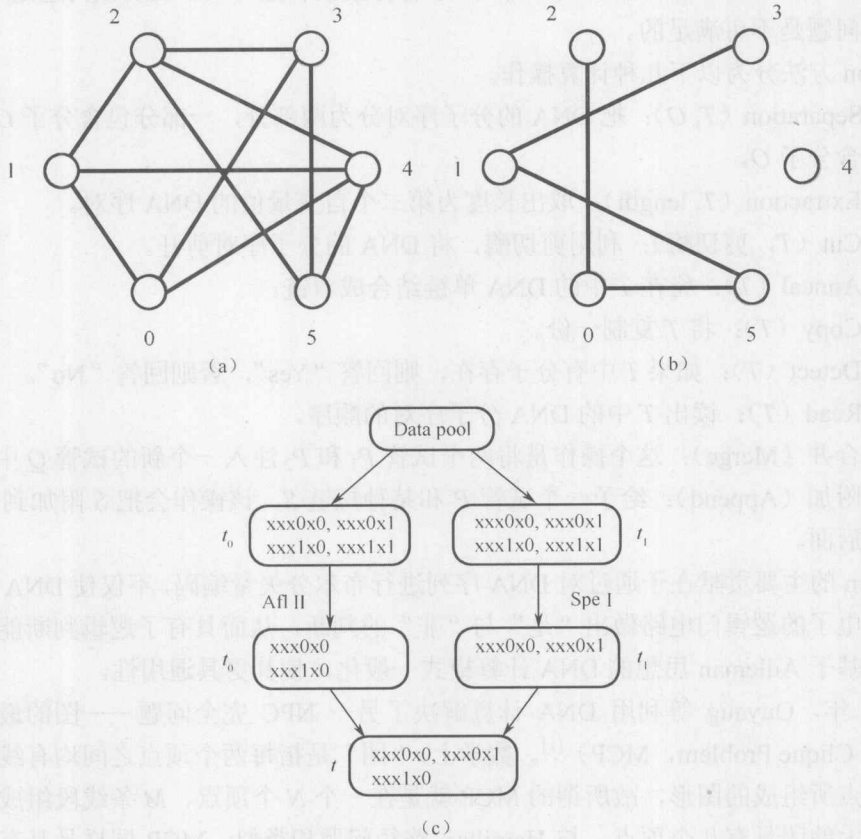



图 1-2　最大团问题

（a）问题实例，其最大团为（5,4,3,2）；（b）（a）的补图，给出的是图（a）中不存在的顶点间的连接；

（c）删除图（b）中与 0～2 边相连的表示团的数字的逻辑过程

根据寡聚核苷酸的设计，若 $i$ 位数值为 1，$V_i$ 是 0bp，$P_i$ 与 $P_{i+1}$ 相邻，正好构成一个内切酶识别位点；若 $i$ 位数值为 0，由于 $V_i$10bp 的插入，破坏了该酶切位点。因此，第二步的去除就可以由内切酶反应来完成了。比如要去除 *xxx*1*x*1，就先用一种内切酶切断 *xxxxx*1，再用另一种内切酶切断 *xxx*1*xx*，合并两种酶切产物，即无 *xxx*1*x*1 的集合了。其他 3 种如法炮制，依次去除。最终再以 $P_0$、$P_6$ 为引物进行 PCR 扩增，所得产物就是第二步的结果。

最后，通过简单的电泳分离即可完成第三步：最短的 DNA 片断所含 1 最多，代表了最大团。实验结果所得到最短的片断是 160bp，扣除 140bp 的位置符长度，数值符总长度为 20bp，表明含有 2 个“0”、4 个“1”。故这个最大团问题的答案为 4。

事实上，进一步基于克隆技术，将该 160bp 条带克隆并测序，就可以回答这个最大团是由哪几个顶点所构成的。测序结果表明此 DNA 片断代表二进制数 111100，正是 Clique (5, 4, 3, 2)。

继上述 3 位学者的工作之后，许多学者给出了不同类型的图与组合优化问题的 DNA 算法。

2000 年，Liu 等对可 SAT 问题进行了基于表面实验的 DNA 计算[14]。2001 年，Wu 对表面计算进行了改进，使得表面计算的操作更具可行性[15]。他们的实验使得 DNA 计算可在固体表面进行，而不仅仅局限于试管溶液中，这样做可以大大降低 DNA 计算的出错率。

利用 DNA 分子的发夹结构，可使 DNA 计算实现自动控制。1997 年，Hagiya 等首次将单链 DNA 所形成的发夹结构用于 Boolean Formula 的学习问题[16]。Sakamoto 等在 1999 年描述了 DNA 分子发夹结构的有限状态变换[17]，并于 2000 年在无任何外加控制的情况下，采用 DNA 分子的发夹构形成功地解决了 6 个变量 10 个子句的 CNF-SAT 问题[18]。他们的算法 减少了计算所需的生物操作步数，不仅提高了计算效率，而且使 DNA 计算朝着自动化方向前进了一步。

2000 年，Head 提出采用质粒 DNA 分子来解决 SAT 问题[19]。2002 年，Gao 给出了图的最大匹配问题（Maximum Matching Problem，MMP）的质粒 DNA 计算模型[20]。2004 年，张连珍等给出了一种基于环状质粒 DNA 计算的新方法[21]，这种计算质粒包含一个特殊的插入 DNA 序列片断，每个片断定位在匹配的限制性内切位点，通过剪切和粘贴实现计算过程，并讨论了 0-1 背包问题（0/1 Knapsack Problem）的质粒 DNA 算法。

除了 DNA 分子外，其他的分子结构或生物材料也可用于计算。Landweber 等首次将核糖核酸（Ribonucleic Acid，RNA）用于分子计算，他们用该方法解决了一个象棋难题[22]。Cukras 等用 RNA 代替 DNA 给出了 SAT 问题的计算模型，并讨论国际象棋问题的 RNA 计算模型[23]。

2001 年，以色列 Weizmann 科学院的 Shapiro 研究小组构造出了基于剪接系统模型的用于 RNA 信息表达分析的可编程有穷自动机[24]，这是截至目前 DNA 计算机研究所取得的最大突破。他们的 DNA 计算机可以对数种信息 RNA 的表达量进行逻辑分析，并依据分析结果释放出能影响基因表达程度的分子。该 DNA 计算机包括 3 个可编程模块（Programmable Modules）：运算模块（Computation Module）、输入模块（Input Module）和输出模块（Output Module）。其中，运算模块扮演随机分子自动机（Stochastic Molecular Automaton）角色；特定信息 RNA 表达量或点突变调控软件分子浓度，亦即自动机转换概率（Automaton Transition Probabilities）通过输入模块完成；输出模块用于控制短的单链 DNA 分子（最终解）的释放。研究小组成功地利用 DNA 计算机鉴定分析与小细胞肺癌和前列腺癌相关基因信息 RNA 的表达量，并根据运算结果释放以抗癌药物模拟出来的单链 DNA 分子。此 DNA 计算机技术可望应用于生化检测、遗传工程、医学诊断与治疗等层面。

2002 年，Braich 等给出了基于粘贴系统模型半自动化自组装 DNA 计算模型，并利用

该模型成功地解决了含有 20 个变元的 3-SAT 问题[25]。他们给出了形如：

$$(\overline{x_3} \vee \overline{x_{16}} \vee x_{18}) \wedge (x_5 \vee x_{12} \vee \overline{x_9}) \wedge (\overline{x_{13}} \vee \overline{x_2} \vee x_{20}) \wedge (\overline{x_9} \vee \overline{x_5} \vee x_{12}) \wedge$$
$$(\overline{x_4} \vee x_6 \vee x_{19}) \wedge (x_5 \vee x_{17} \vee x_9) \wedge (\overline{x_1} \vee x_4 \vee \overline{x_{11}}) \wedge (\overline{x_{19}} \vee \overline{x_2} \vee x_{13}) \wedge$$
$$(x_5 \vee x_{17} \vee x_9) \wedge (x_{15} \vee x_9 \vee \overline{x_{17}}) \wedge (\overline{x_5} \vee \overline{x_9} \vee \overline{x_{12}}) \wedge (x_6 \vee x_{11} \vee x_{14}) \wedge$$
$$(\overline{x_{15}} \vee \overline{x_{17}} \vee x_7) \wedge (\overline{x_6} \vee x_{19} \vee x_{13}) \wedge (\overline{x_{12}} \vee \overline{x_9} \vee x_5) \wedge (x_{12} \vee x_1 \vee x_{14}) \wedge$$
$$(\overline{x_3} \vee x_{20} \vee x_2) \wedge (x_{10} \vee \overline{x_7} \vee \overline{x_8}) \wedge (\overline{x_5} \vee x_9 \vee \overline{x_{12}}) \wedge (x_{18} \vee \overline{x_{20}} \vee x_3) \wedge$$
$$(\overline{x_{10}} \vee x_{16} \vee \overline{x_{18}}) \wedge (x_1 \vee \overline{x_{11}} \vee \overline{x_{14}}) \wedge (x_8 \vee \overline{x_7} \vee \overline{x_{15}}) \wedge (\overline{x_8} \vee x_{16} \vee \overline{x_{10}})$$

的 20 个变元的 SAT 问题的 DNA 算法，他们利用所有的解的组合构造 Sticker 模板，通过杂交反应实现了解的检测与萃取。他们在 Lipton 思想的基础上，将解的萃取和电泳这两步生物操作进行了集成，从而使得他们所采用的生物方法具有自动化的特点。20 个变元的 3-SAT 问题的成功解决为 DNA 计算的高度并行性提供了有力的支撑。

此外，学者们利用 DNA 计算方法还研究了路着色问题[26]、DNA 加法[27, 28]、数据加密问题[29]、计算机代数问题[30]、Petri 网结构[31]、有穷自动机[32]和有界邮政通信问题[33]等。

尽管 DNA 计算（机）的研究已经取得了很大的进展，但其仍处于起步阶段，还有大量的理论挑战和实际问题亟待解决，概括而言，包括：

（1）计算机可以解决哪些问题？确切地说，DNA 计算机是完备的吗？即 DNA 计算机能完成所有的可计算函数吗？

（2）能否设计出可编程的 DNA 计算机？即是否存在着类似于传统计算机通用计算模型那样的通用 DNA 计算系统（模型）？

（3）DNA 计算机的计算能力能否超过图灵机或超级计算机？

为了回答上述问题，科学家们将 DNA 计算模型划分为两类：

（1）生物操作基础上的模型。这些生物操作都可以在实验室成功进行。

（2）形式上的模型。首先将一种或多种生物操作技术抽象为算子，然后在这些算子基础上进行建模。

一般地，形式化模型研究起来比较容易，但建模比较困难，需要与实际生化操作结合起来。目前，许多科学家都在潜心研究这两类 DNA 计算模型。这就类似于在传统计算机诞生之前的 20 世纪 30～40 年代理论计算机的研究阶段。近 20 年来，DNA 计算理论一直是西方发达国家的一个研究热点。科学家也提出了多种 DNA 计算模型，都各有千秋，但是公认的 DNA 计算机的“图灵机”还没有诞生。下面简要地介绍四种 DNA 计算的形式模型及其国内外研究现状。

### 1. 粘贴系统（Sticker System）

1996 年，Roweis 等人构造了一种新的 DNA 计算模型——粘贴模型[34]。此模型有一个可随机访问的存储空间，操作时不需要延伸 DNA 链，也无须酶的参与，并且它的材料在理论上是可以重复使用的。他们提出了用粘贴模型来设计生物计算机的设想，该机器是分子计算中的一种并行机器人工作站，通过一个中心可编程电子计算机控制其中的各种机器人，如液流装置、加热器、制冷器以及一些传统的电子设备等。对于模型中的每一步运算，