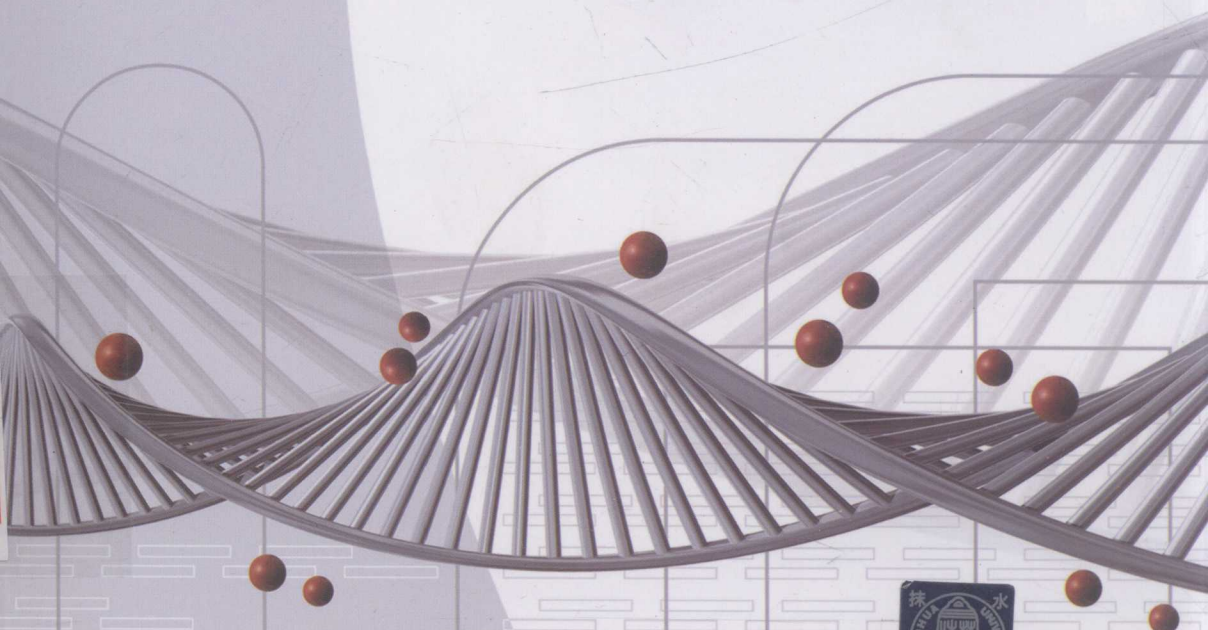
DNA 计算中的编码方法

朱翔鸥 刘文斌 著



清华大学出版社

DNA 计算中的编码方法

朱翔鸥
刘文斌 著

清华大学出版社
北 京

内 容 简 介

编码问题是DNA计算中的基本问题，也是关键问题。在DNA计算模型中，数据通过DNA编码表示，数据计算和处理通过DNA分子间的特异性杂交来完成，DNA编码质量直接影响DNA计算的精确度。本书介绍了DNA计算模型和应用，阐述了DNA计算的编码问题，针对编码方法展开讨论，研究了线性编码方法以及构造和计数问题，研究了模板、模板框和单模板等编码方法，最后建立了DNA解链温度的预测模型。

本书适合从事DNA计算及相关领域的科研人员参考，也可以供高校、科研机构的研究生学习参考。

图书在版编目(CIP)数据

DNA计算中的编码方法 / 朱翔鸥，刘文斌　著. —北京：清华大学出版社，2012.6
ISBN 978-7-302-29169-5

Ⅰ. ①D… Ⅱ. ①朱… ②刘… Ⅲ. ①脱氧核糖核酸—应用—编码 Ⅳ. ①O157.4

中国版本图书馆CIP数据核字(2012)第135262号

责任编辑：李万红
封面设计：华方齐煜传媒
责任校对：邱晓玉
责任印制：杨　艳

出版发行：清华大学出版社
网　址：http://www.tup.com.cn，http://www.wqbook.com
地　址：北京清华大学学研大厦A座　　**邮　编**：100084
社总机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印　装　者：北京国马印刷厂
经　　销：全国新华书店
开　　本：169mm×230mm　　**印　张**：8　　**字　数**：111千字
版　　次：2012年6月第1版　　**印　次**：2012年6月第1次印刷
印　　数：1～1000
定　　价：29.80元

产品编号：047550-01

序　言

遗传算法、人工神经网络和蚁群算法等智能计算方法试图从生物系统及其进化过程中获得灵感，探索解决复杂问题有效算法。1994 年，Adleman 在 Science 上发表的文章“Molecular Computation of Solutions to Combinatorial Problems”标志着一个新的研究领域——DNA 计算的诞生。这篇文章首次利用生化反应的并行性，为解决 NP-完全问题开辟了新的思路。DNA 计算的基本原理是：以编码生命信息的遗传物质——DNA 序列作为信息编码的载体，利用 DNA 分子的双螺旋结构和碱基互补配对的性质，将所要处理的问题映射为特定的 DNA 分子；然后在生物酶的作用下，通过可控的生化反应生成问题的解空间；最后利用各种现代分子生物技术如聚合酶链反应 PCR、聚合重叠放大技术 POA、超声波降解、亲和层析、分子纯化、电泳、磁珠分离等手段获取运算结果。

在这种基于生化反应的新型计算方式中，信息的识别主要是通过 DNA 分子间的特异性杂交来实现的。由于 DNA 分子间的杂交在不完全互补的情况下也有可能发生并进而形成各种不希望的二级结构，从而导致错误的计算结果。因此，如何通过有效的编码来提高 DNA 计算过程中的“信噪比”，一直是 DNA 计算研究中的一个重点和难点问题。国内外许多学者在这方面进行了许多有益的探索。在 Frutos 提出的模板编码的基础上，我们对模板编码方法及编码数的理论进行了深入的研究。本书的内容是我们对这些研究的一个总结，希望能对 DNA 计算及近年来热门的自组装模

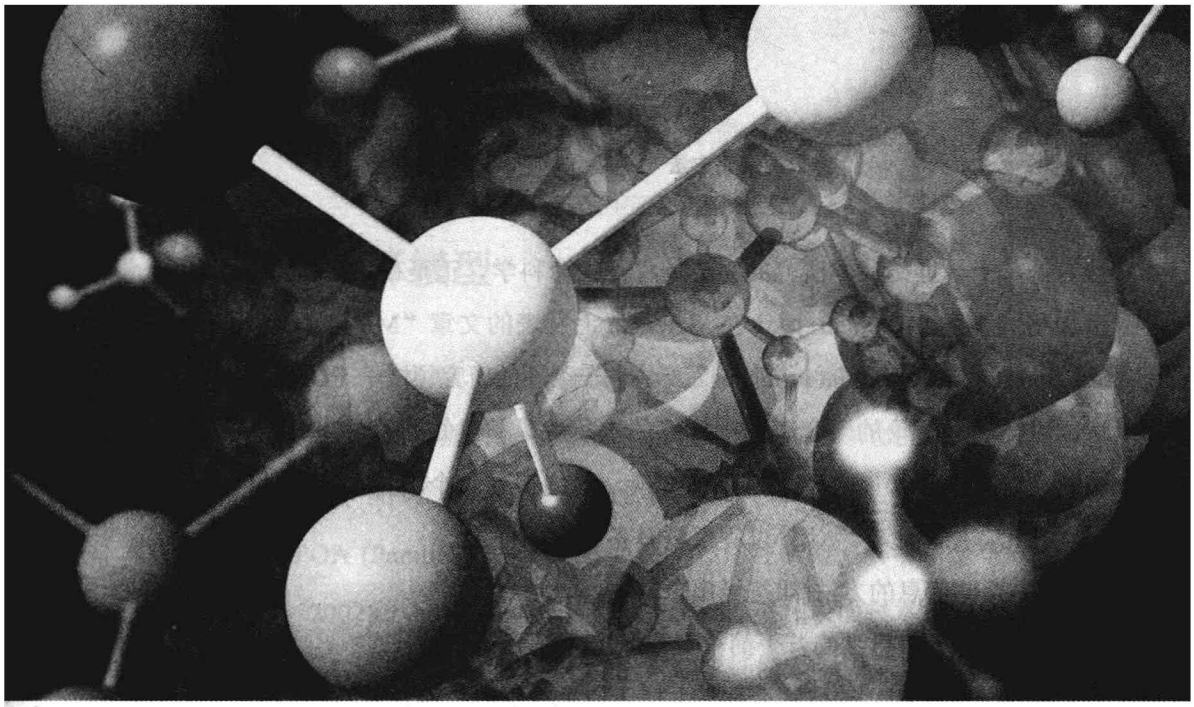
型的研究等有一定借鉴及促进作用。由于我们水平有限，另外时间仓促，书中难免有错误和不准确之处，恳请广大读者批评指正。

本书由朱翔鸥和刘文斌共同完成，其中朱翔鸥负责第2～5章，刘文斌负责第1、6～9章。感谢清华大学出版社对本书出版给予的大力支持和帮助。本书的研究内容及出版特别得到了国家自然科学基金项目(60403002，60970065)，浙江省自然科学基金(Y1080227，Y105654，R1110261)的资助，在此深表谢意。

朱翔鸥　刘文斌

2012年5月

目　录



第1章 DNA计算的概述

1.1 引 言

计算机科学和生物科学的联姻已经成为信息科学研究中的一个非常热门的前沿领域。遗传算法、人工神经网络和蚁群算法等都是试图从生物系统及其进化过程中获得灵感，探索解决复杂问题的有效算法的经典范例。另一方面，随着人类基因组计划的顺利实施和后基因组计划的开始，涌现出海量的生物分子数据。如何挖掘和发现这些数据的内涵，揭示生命的奥秘已经成为科学家面临的一个新的挑战。有人

预言，二十一世纪科学研究的主战场将是生命科学。正是在这种背景下，1994 年，南加州大学的 Adleman 博士在 Science 上发表的文章“Molecular Computation of Solutions to Combinatorial Problems”标志着一个新的研究领域——DNA 计算的诞生[1]。

可计算性理论的研究表明：计算的本质就是从已知的符号串开始，依据一定的法则，对这些符号串进行一系列的变换，最终得到一个满足预定条件的符号串的过程。因此，原理上实现计算过程仅需具备二个条件：

(1) 存储信息的方法(即符号串)；

(2) 作用于信息上的若干简单操作(即法则)。

现代分子生物学的研究表明，生物体异常复杂的结构正是对由 DNA 序列表示的遗传信息执行简单操作的结果。因此，从本质上二者极为相似。正如计算机中是用“0”和“1”表示信息一样，DNA 单链可以看作是在字母表 Σ={A、G、C、T}上表示和译码信息的一种方法，生物酶及其它一些生化操作则是作用在 DNA 序列上的算子。因此，DNA 计算的出现表明了计算不仅是一种物理性质的符号变换，而且可以是一种化学性质的符号变换。应用 DNA 分子的切割和粘贴、插入和删除等来完成计算的这种变革是前所未有的，具有划时代的意义。

在 DNA 计算诞生的这十多年的时间，DNA 计算的研究取得了很大的进展。随着研究的深入，一方面对于 DNA 计算的研究并不像人们起初认为的那样乐观，其中最大的障碍是如何克服 DNA 计算过程中所产生的“指数爆炸”问题；另一方面，DNA 计算的研究领域在日益扩大。下面主要从 DNA 计算模型、布尔电路的模拟、基于 DNA 的大规模数据库、在生物信息学中的应用及 DNA 计算中的编码问题的研究等几个方面，介绍 DNA 计算近年来的研究和发展状况。最后，我们对 DNA 计算研究的前景和方向进行总结和展望。

1.2　DNA 计算模型

继 Adleman 的工作之后，1995 年 Lipton 通过构造一个接触网络图，将可满足性(SAT)问题的解空间映射为通过接触网络的始点到终点的所有哈密尔顿路[2]。1997 年，Ouyang 等将另一个 NP-完全问题——图的最大团问题转化为最大独立集问题，利用并行重叠组装 POA (Parallel Overlap Assembly)技术在实验室解决了一个 6 个顶点的最大团问题[3]。2002 年 Adleman 的小组将 Sticker 模型和表面技术相结合，在一个半自动化的装置上解决了一个 20 个变元的 3-可满足性问题[4]。该问题的解空间为 20^{20}，这是到目前为止 DNA 计算通过试验解决的最大规模的 NP-完全问题。有关传统的线性 DNA 分子的计算模型已有很多介绍[5][6][7][8]，在此我们不再详述。下面我们主要从二个方面介绍 DNA 计算模型发展的新趋势。

1.2.1　基于非线性分子结构的计算模型

1. 质粒

质粒(Plasmid)是染色体外能够独立复制、稳定遗传的一种环状双链 DNA 分子，将其导入宿主细胞后，能够随着细胞的繁殖而大量扩增，因此，在基因工程中常用来作为外来信息的载体。图 1.1 给出了一个 pUC19 质粒载体的多克隆位点示意图，其中的 S1，S2，…，S9 等位点可以用来插入所要表示的信息，在这些位点二边的各种限制性酶切位点则使得我们可以对该点表示的信息进行插入和删除操作[9]。

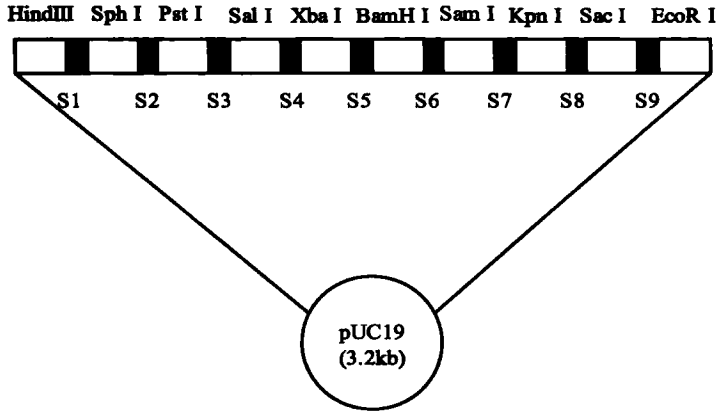


图 1.1 pUC19 质粒载体的多克隆位点示意图

早在 1987 年，Head 就敏锐的洞察到生物操作本身所具有的计算潜能，并在质粒结构的基础上提出了剪接系统(splicing system model)的概念[10]。限于当时的生物技术水平没有引起人们的关注。2000 年，他在“Computing with DNA by operating on plasmids”一文中首次使用质粒 DNA 解决了一个 6 个顶点的最大独立集问题后[11]。此后基于质粒的 DNA 计算才引起研究人员的极大兴趣[12][13]。质粒 DNA 计算模型的主要缺点是所需的限制性内切酶的数量太多，如何克服这方面的缺点将是今后应该研究的一个重点方向。

2. 分子信标

分子信标(molecular beacon)是由一种可用荧光标记的呈茎-环(或“发夹”)结构的寡聚核苷酸序列。环上的寡聚核苷酸序列是分子信标的基因识别部分，它能与靶基因自发地进行特异性杂交。茎部是一段 4-12b 的互补碱基。在分子信标的 5'端和 3'端通过连接臂(linker) 可以连接上荧光基团(如 EDANS)和猝灭基团(如 DABCYL)。正常情况下，茎部的荧光分子与猝灭分子非常接近(7~10nm)，使得荧光分子发出的荧光被猝灭分子吸收并以热的形式散发，因而检测不到荧光信号。当有靶序列存在时，分子信标的环序列即可与靶序列特异性结合，形成比茎-环结构更稳定的双链杂合体。

此时，猝灭分子与荧光分子分开而不能吸收荧光分子发出的荧光，于是可以检测到荧光(图 1.2)。

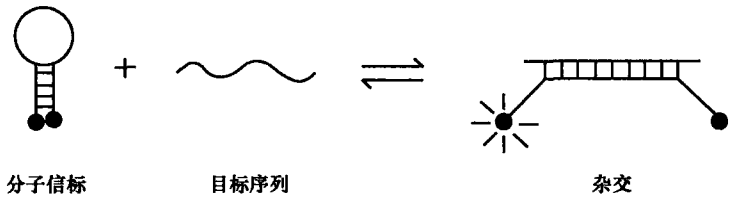


图 1.2　分子信标的作用机制[14]

从 1996 年 Tyagi 和 Krammer 首次建立分子信标探针以来[15]，各种新型的分子信标不断出现。这种荧光探针已在生物技术领域得到了广泛的应用。如何挖掘这种“发夹”(Hairpin)状的二级结构的计算潜力，也就成为 DNA 计算研究的一个方向。2000 年，日本的 Sakamoto 等巧妙的将逻辑运算的约束编码于 DNA 分子，利用这种自发形成的二级结构求解了一个有 6 个变量可满足问题[16]。此外，他们还提出了“鞭子”PCR(Whiplash PCR)技术，通过重复利用“发夹”的形成和打开过程实现了一种有限自动机的状态转换模型[17]。图 1.3 的 GGG 是每次“发夹”茎杆部分延伸的终止点。2003 年，殷志祥等给出了可满足性问题的分子信标计算模型[18]。

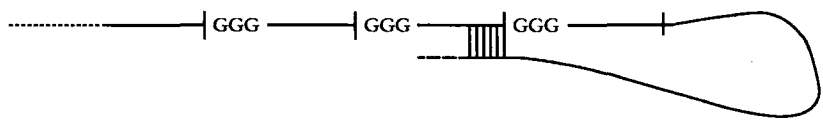


图 1.3　“鞭子”PCR(Whiplash PCR)的示意图

3. DNA 分子的自组装

1982 年，New York 大学的 Seeman 就提出了利用 DNA 分子构造各种简单构件的思想[19]。目前，已经能够在实验室作出一些简单的构件，如三角形、长方形、园等，俗称瓦片(tile)。实际上，Adleman 那篇开创性的文章中就已经包含了分子自组装的思想：代表图中的各种路径是通过线性单链 DNA 分子间的自组装过程产生的！后来，

Winfree 在其博士论文中首次对分子自组装进行了深入的研究，结果表明：基于分子自组装行为的计算模型具有和图灵机一样的计算能力[20][21]。通过对这些简单构件的自组装过程的研究，不仅可以进一步挖掘 DNA 计算解决优化问题的潜力，同时也可能实现对某些生物大分子间相互作用的调控。

基于 DNA 自装配设计的算法相对于其它完成 DNA 计算的算法实验步骤简单，只需要以下四步：

(1) 混合 DNA 输入链形成 DNA 瓦片；

(2) 促使设计好的 DNA 瓦片自装配成分子聚合物；

(3) 将结构中定位的 DNA 链连接起来；

(4) 用一步分离操作提取正确的输出。

本质上，线性自装配是由多层自装配简化而来，与多层自装配相比用线性自装配完成加法有以下优点：(1) 输入和输出链同时装配；(2) 计算对应的真值表每一行用一种 DNA 瓦片表示，输入和输出都编码其上。而对应的多层自装配每一位输入和输出都要分别对应一种 DNA 瓦片；(3) 相邻 DNA 瓦片只需要匹配它们之间的进位：一个 DNA 瓦片向上进位，另一个 DNA 瓦片接收这个进位。用不同的 DNA 黏性末端分别表示进位 0 和 1；(4) 相邻 DNA 瓦片结合必须同时匹配多个黏性末端，不会出现只匹配一些末端仍然结合的情况，降低了出错率。

1.2.2 DNA 进化算法

进化算法是一种在分子水平(即基因或 DNA 分子)模拟生物进化过程来求解复杂问题的一种有效的算法。DNA 计算则是利用实际的生物分子——DNA 的各种生化反应来完成计算过程。因此，二者天生就具有某种必然的联系。1997 年，Deaton 等就应用遗传算法解决 DNA 计算中的编码问题[22]。1999 年，Chen 等首次通过实际的生化试验来完成进化计算过程，从而将 DNA 计算的高度并行性和进化算法中的演化策

略结合起来，为解决 DNA 计算面临的最大难题—“指数爆炸”开辟了新的方向[23]。和传统的遗传算法相比，DNA 遗传算法可能存在以下优点[24~26]：

(1) 群体规模巨大，传统的遗传算法中群体规模通常为 10^2~10^5，而在 DNA 遗传算法中群体规模则可达 10^{13}。群体规模的增大将有助于提高群体中个体的多样性，从而使得群体中存在最优解的可能性大大增加。

(2) 由于 DNA 计算的显式并行性，群体规模的增大并不会明显导致计算时间的增加，而在传统的遗传算法中，群体规模的急剧增大将导致计算机在译码过程及评价个体适应度方面的计算量激增。

(3) 在 DNA 计算中，生化反应的不完备性可能会导致各种不可预测的错误结果，而对于 DNA 遗传算法，这些因素则是可以允许的，甚至可以将其视为产生变异的一种因素。

2002 年，Rose 提出了一种基于 Whiplash PCR 的 DNA 的进化算法，用来解决著名哈密尔顿路问题[27]。其主要思想是将图中的所有边编码为一条规则，整个 DNA 链就是一条规则的集合。这样给定图中的起始顶点和终结顶点后，所有的 DNA 链就是不同的自动机，随机的选择一条满足条件的规则进行状态转换。当图中存在此路径的话最终就会有一条 DNA 链会形成图中的哈密尔顿路。

此外，2004 年，Li 等对最大团问题的 DNA 进化算法进行了研究，其中仅用到变异算子。计算机仿真结果表明：基于 DNA 的进化算法能够大大提高传统 DNA 计算模型解决问题的规模和能力[28]。目前，DNA 进化算法的难点是现有的生化技术难以应付进化算法中的各种复杂的适应度函数。

1.3　基于 DNA 的布尔电路模拟

众所周知，电子计算机是由大量简单的逻辑门电路集成而成，从而实现各种计

算及信息处理的功能。为了探索 DNA 计算新的潜在的应用(Killer Application)，1996 年 Ogihara 等首次提出了基于 DNA 的布尔电路的模拟[29]。随后，Amos、Mulawka 等做了进一步的研究。这些模型的最大缺点是在经过一次运算后由 DNA 分子编码的逻辑门也随之被限制性内切酶水解，因而无法重复使用[30][31][32]。分子信标茎-环结构的形成和打开与传统电子计算机中使用的逻辑电路的 0、1 非常相似。因此，研究基于分子信标的逻辑门，并进一步开发基于 DNA 的大规模集成电路将具有得天独厚的优势。

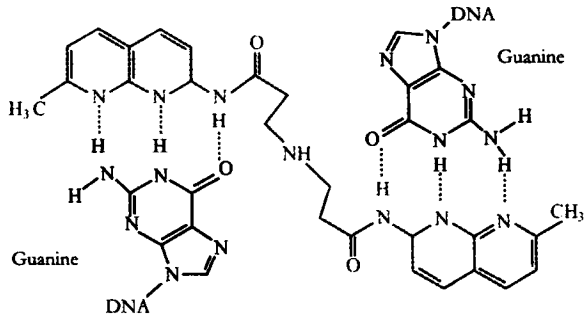


图 1.4　吡啶分子(a naphthyridine dimmer)定向诱导鸟嘌呤 G-G 的配对示意图

2002 年，Corn 的研究小组提出了一种“诱导”型的分子信标结构[33]。其主要思想是：由于吡啶分子(a naphthyridine dimer)的二臂可以分别与二个鸟嘌呤 G 形成稳定的三个氢键(图 1.4)。在分子信标的茎杆部分设计一些鸟嘌呤 G-G 的非匹配对，当加入吡啶分子时可以定向诱导鸟嘌呤非匹配对 G-G 杂交形成分子信标的茎-环结构。这一技术大大提高了分子信标的灵敏度和灵活性。2004 年，刘文斌等提出了一种基于诱导型的分子信标的逻辑与非门(NAND)的 DNA 计算模型[34]，如图 1.5 所示。该模型不仅实现了逻辑门的重复可用性，而且由于分子信标本身有很高的灵敏性，因而其可靠性也大大提高。