DNA计算及其在医学领域中的应用

李汪根 著

冶金工业出版社
Metallurgical Industry Press

DNA计算及其在医学领域中的应用

李汪根 著

北京
冶金工业出版社
2012

内 容 简 介

本书面向生物计算的学科前沿，在现有 DNA 计算成果的基础上，从生命现象的本质特征来研究 DNA 计算的容错性、可复制性以及随机性，探讨了 DNA 计算在基因网络、自复制以及疾病检测中的应用，为 DNA 计算在生物医学领域上的应用提供研究思路。

本书可作为高等院校相关专业高年级本科生或研究生的教材及参考书，也可供从事生物计算、计算机科学、自动控制、智能科学、系统科学、应用数学等领域研究的教师和科技工作者参考。

图书在版编目（CIP）数据

DNA 计算及其在医学领域中的应用/李汪根著. —北京：冶金工业出版社，2012.6

ISBN 978-7-5024-6020-4

Ⅰ.①D… Ⅱ.①李… Ⅲ.①脱氧核糖核酸—计算方法 Ⅳ.①Q523

中国版本图书馆 CIP 数据核字(2012)第 130143 号

出 版 人 曹胜利
地　　址 北京北河沿大街嵩祝院北巷 39 号，邮编 100009
电　　话 (010)64027926 电子信箱 yjcbs@cnmip.com.cn
策划编辑 姜晓辉 责任编辑 姜晓辉 美术编辑 李 新
版式设计 孙跃红 责任校对 卿文春 责任印制 牛晓波
ISBN 978-7-5024-6020-4
三河市双峰印刷装订有限公司印刷；冶金工业出版社出版发行；各地新华书店经销
2012 年 6 月第 1 版，2012 年 6 月第 1 次印刷
148mm×210mm；3.75 印张；100 千字；109 页
15.00 元
冶金工业出版社投稿电话：(010)64027932 投稿信箱：tougao@cnmip.com.cn
冶金工业出版社发行部 电话：(010)64044283 传真：(010)64027893
冶金书店 地址：北京东四西大街 46 号(100010) 电话：(010)65289081(兼传真)

前　言

DNA 计算是以 DNA 分子及生化酶为物质基础，施以适当的生化操作来解决计算问题的一种新型计算模式。DNA 计算有望实现现有电子计算机所无法实现的大规模并行处理和组合运算功能，是解决包括 NP 等复杂问题的突破口之一。尤为重要的是，DNA 计算的研究已经超越了纯粹计算机研究的范畴，在研究生命的本质过程、疾病诊断、新材料开发等领域也有重要应用。

本书是作者在东华大学信息科学与技术学院博士后流动站工作期间和在国家自然科学基金（61070060）资助下取得的研究成果。全书共分 7 章对 DNA 计算及其在生物医学领域的应用进行了系统地阐述和讨论。

本书研究内容涉及多学科交叉领域，且理论密切结合实际。本书结构安排合理，既照顾到面，又照顾到点，有深度和广度。读者既可以了解到这一领域的前沿研究进展，又可以深入到某一较深的研究方向。

感谢合作导师东华大学丁永生教授，感谢东华大学信息科学与技术学院网络智能研究室，感谢冶金工业出版社为本书的出版所付出的辛勤工作，没有他们的指导和帮助，本书的出版不可能如此顺利。

由于本书内容涉及多个学科前沿，知识面广，且时间仓促，再加上作者学识有限，书中的有些观点和提法，难免有不妥之处，恳请广大同行、读者给予批评指正。

作　者

2012 年 6 月

目　录

1　绪论 …… 1

1.1　DNA 计算概述 …… 1

1.2　DNA 计算研究现状 …… 2

1.3　DNA 计算的原理及其特点 …… 5

1.3.1　DNA 计算的原理 …… 5

1.3.2　DNA 计算的特点 …… 6

1.4　DNA 计算的分子操纵手段 …… 7

1.4.1　变性 …… 7

1.4.2　复性 …… 8

1.4.3　剪切 …… 8

1.4.4　绑结和连接 …… 9

1.4.5　分离和测量长度 …… 9

1.4.6　特异性检测 …… 10

1.4.7　扩增（复制） …… 10

1.4.8　测序 …… 11

1.4.9　合成 …… 12

1.5　DNA 计算的实现途径 …… 12

1.5.1　基于试管的 DNA 计算 …… 12

1.5.2　基于表面的 DNA 计算 …… 13

1.5.3　基于芯片的 DNA 计算 …… 13

2　多功能生物芯片反应装置 …… 16

2.1　引言 …… 16

2.2 芯片扫描系统 …… 17
2.2.1 芯片扫描系统概述 …… 18
2.2.2 自动定位芯片扫描系统 …… 22
2.3 芯片电泳系统 …… 25
2.3.1 芯片电泳的工作原理 …… 25
2.3.2 多路芯片电泳系统 …… 27
2.4 芯片 PCR 系统 …… 28
2.4.1 芯片 PCR 概述 …… 28
2.4.2 智能芯片 PCR 系统 …… 29
2.5 智能芯片反应装置的实施 …… 30

3 DNA 计算模型 …… 32

3.1 引言 …… 32
3.2 模型 …… 33
3.2.1 输入模块 …… 33
3.2.2 运算模块 …… 35
3.2.3 存储模块 …… 35
3.2.4 输出模块 …… 36
3.2.5 控制模块 …… 36
3.3 数据结构 …… 38
3.3.1 堆栈 …… 38
3.3.2 队列 …… 45
3.3.3 广义表 …… 48
3.4 存储系统 …… 53
3.4.1 存储载体 …… 54
3.4.2 信息编码 …… 55

3.5 运算系统 …… 56
3.5.1 一位二进制进位加法 …… 56
3.5.2 n 位二进制串行加法 …… 60

4 DNA 遗传算法及其在癌症分类中的应用 …… 61

4.1 引言 …… 61
4.2 DNA 遗传神经网络 …… 62
4.2.1 遗传神经网络算法 …… 62
4.2.2 GA 优化 BP 网络的权值和阈值 …… 63
4.2.3 基于 GA-BPNN 的特征选择 …… 64
4.3 基于 GA-BPNN 的乳腺癌分类算法 …… 65
4.3.1 方法 …… 65
4.3.2 实验分析 …… 68

5 随机 DNA 计算及在基因网络中的应用 …… 71

5.1 引言 …… 71
5.2 确定 DNA 有限状态自动机 …… 71
5.3 不确定 DNA 有限状态自动机 …… 73
5.4 DNA 下推自动机在回文识别中的应用 …… 75
5.4.1 接受回文语言的下推自动机 …… 75
5.4.2 可自治 DNA 下推自动机实现 …… 76
5.5 不确定 DNA 有限状态自动机在基因网络中的应用 …… 81
5.5.1 基因表达的不确定有限状态自动机模型 …… 81
5.5.2 不确定 DNA 有限状态自动机的实现 …… 82

6 容错 DNA 计算及自修复机理 …… 84

6.1 引言 …… 84

6.2 DNA计算的自复制性 …… 86
6.2.1 DNA片段自组装 …… 86
6.2.2 二维DNA分子元胞自动机 …… 87
6.3 DNA计算的可逆性 …… 88
6.3.1 基于DNA计算的布尔电路 …… 89
6.3.2 可逆容错门电路 …… 89
6.3.3 基于DNA计算的可逆容错门电路 …… 91

7 DNA计算在医学检测上的应用 …… 93

7.1 引言 …… 93
7.2 败血症基因芯片检测模型 …… 93
7.2.1 采用的方法 …… 94
7.2.2 实验步骤 …… 95
7.3 基于DNA计算的疾病基因诊疗模型 …… 95

参考文献 …… 100

1 绪 论

1.1 DNA 计算概述

DNA 计算是一种关于计算的新的思维方式，它是以 DNA 分子及生化酶为物质基础，施以适当的生化操作来解决计算问题的一种新型的计算模式[1]。自 1994 年 Adleman 博士提出运用 DNA 分子进行计算并成功地运用分子生物实验解决了著名的 Hamilton 路径问题[2]以来，利用 DNA 分子作为计算手段的研究取得了很多成果[3~6]。DNA 计算的研究目前已涉及 DNA 计算的能力、计算模型和算法等许多方面，如基于 DNA 生物实验方法的求解 NP 完备问题的 DNA 算法[2,7~11]、基于 DNA 计算的密码破解[12~15]、DNA “通用计算机”的构造等[16~21]。也有学者将 DNA 计算与遗传算法、神经网络、模糊系统和混沌系统等智能计算方法相结合[22~30]。

近年来，生物芯片技术得以迅速发展[31~34]，使临床快速鉴定细菌成为可能。生物芯片主要是运用分子生物学、基因信息与分析化学等原理进行设计制造，采用微机电技术制备微小化装置，来进行生物反应或分析，其最大优点：高通量分析、高灵敏度检测、特异性强。目前，国外几乎所有的主要制药公司都不同程度地采用了生物芯片技术，应用生物芯片来寻找药物靶标，查检药物的毒性或副作用。用芯片作大规模的筛选研究可以省略大量的动物试验，缩短药物筛选所用时间，从而带动创新药物的研究和开发。

败血症（septicemia）是致病菌或条件致病菌侵入血液循环而

引起的急性或亚急性全身感染。败血症可引起感染性休克、心肾功能不全等并发症，使许多组织器官受到损害，可形成脑膜炎、骨髓炎、肺脓肿、肝脓肿等，是一种十分危急的病症。如果患者不能及时得到正确的诊断和治疗，将会危及生命，病死率可高达30%～50%。

及时、正确检测血液中的病原微生物对治疗败血症、挽救患者的生命至关重要。目前，医学上检验血液样品中有无病原微生物存在的普遍方法是血培养检查（Blood Culture Examination）。但该方法耗时长，检查病菌的种类需要3～5天时间，且容易出现假阳性或假阴性。也有用基因芯片通过杂交荧光PCR定量的方法进行检测的报道，但一方面仅限于实验室研究；另一方面也不是从DNA计算和计算机算法的角度进行判定。

虽然，DNA计算取得了很多研究成果，但DNA计算还处在初级阶段，DNA计算的实用性还远远达不到要求。一个可能的原因就是，现有的DNA计算模型没有考虑到参与计算的DNA分子本身的生物学特性。本书就是在研究现有DNA计算成果的基础上，从生命现象的本质特征来研究DNA计算的容错性、可复制性以及随机性，以提高DNA计算的实用性，更好地揭示生命现象的本质，并为DNA计算在生物信息学、医学等生命领域上的应用提供研究思路。

1.2 DNA计算研究现状

20世纪科学史上最伟大的发明之一就是电子计算机的出现。从第一台电子计算机的诞生起，计算机经历了从电子管、晶体管、集成电路到超大规模集成电路的发展历程。随着科学技术的发展，大约每18个月计算机芯片的速度就会提高一倍，而尺寸会

减少50%。今天的计算机无论是性能价格比，还是性能体积比都是第一台电子计算机的成千上万倍。随着因特网的建立和计算机的普及，电子计算机影响着人类生活的各个方面。然而，进入21世纪，随着社会和科学技术的发展，许多复杂系统不断出现，比如：蛋白质结构预测、药物筛选等非线性问题和NP问题等。电子计算机在解决这类复杂系统时常常显得无能为力，主要体现在以下两个方面：（1）与要解决的实际问题相比，电子计算机的运算速度太慢，无法在有效时间内解决这些实际问题；（2）目前的电子计算机存储容量太小，在现有计算机体系结构和算法下，其内存远远不能满足解决这些复杂问题的实际需要。虽然，电子计算机也正向高速度、大容量、小体积方向飞速发展，但由于集成电路的复杂性，硅芯片的存储极限，以及传统计算机本身计算方法的局限性，使得计算机在实现超微结构，超大存储容量和运算速度的提高等方面存在很大困难。基于这些原因，科学家们一直在寻求新的计算技术，以满足这些新的需要。

生命的发展是一个漫长的过程，生命体在遗传、变异和选择的作用下不断向前发展和进化[35,36]。生命体是一个非常复杂的系统。生命为了维持远离平衡的耗散结构，必须能够进行自组织（Self-organization），构造相应的信息处理系统，而进行自组织所需的指令以DNA的形式存储。从这个意义上说，生命信息系统可看成一个高级信息处理系统，生命自开始就进行着各种复杂的计算。计算机科学中的很多技术都是受到生命信息系统的启发而发展的，这些技术包括元胞自动机、神经网络、进化计算、免疫计算和生物分子计算等。近年来，随着计算机科学与生物科学的发展和相互交叉，一种新型的高度并行的生物分子计算模式——DNA计算备受人们的关注。

自 Watson-Crick 揭开 DNA 的奇妙结构以来，生物、化学和基因工程技术正处在日新月异的巨大进步和发展中，它将提供更多的、新的供计算使用的人工合成酶。Adleman 的实验[2]就是这个正在蓬勃兴起的科学新领域的一次展示。DNA 计算机有望实现现有计算机所无法真正实现的大规模并行处理和组合运算功能，是解决包括 NP 等复杂问题的突破口之一[1,3,4,11]。

极为重要的是，DNA 计算机中的算法研究已经超越了纯粹计算机研究的范畴，实际上是在研究生命的本质过程，即基因组是如何实现通过算法来实现生命过程（发育、衰老、疾病、死亡等）。已有学者将其应用到医疗领域。奥林巴斯公司与东京大学联合开发出了全球第一台能够真正投入商业应用的 DNA 计算机，用于基因的诊断[37]。该计算机由分子计算组件和电子计算部件两部分组成，前者用来计算分子的 DNA 组合，以实现化学反应，搜索并筛选出正确的 DNA 结果；而后者则对这些结果进行分析，并且能将原来人工分析 DNA 需要的 3 天时间缩短为 6 个小时。以色列科学研究院的 Shapiro 教授等人也开发出了应用于癌症治疗和诊断的 DNA 计算机模型[38]。除了在医疗领域外，如新材料开发领域也在探讨 DNA 计算机的应用，力图通过有效的分子自装配达到生产出新材料的目的[39~44]。这些足以说明，DNA 计算机正试图走出只能解决数学问题的有限用途。

但是，目前的 DNA 计算在一定程度上还不十分完善。DNA 计算可能发生误差[45~47]。现代分子生物学提供了诸如 PCR 扩增、DNA 分子纯化、高效电泳及亲和层析、磁珠分离等技术，但它们所消耗的时间和空间复杂性远远比 DNA 计算过程繁复得多。而且随着问题复杂性的增加，算法所需要的核苷酸分子数成指数增加[7,9~11]。如何设计优化的核苷酸编码是一个直接影响计算结果

的关键因素[22-24]，例如，设计不合理的单链核苷酸编码自身可能发生退火反应。同时，由于热力学和动力学的原因，大量的 DNA 聚在一起，偶尔可能发生退火反应甚至可能发生 DNA 链的动力分解，从而导致“伪解”的产生。另外，在 PCR 扩增、退火反应等都可能发生碱基错匹配的错误[47]。

总之，DNA 计算虽然目前还只能解决一些极其简单的问题实例，并且存在许多不足和障碍（如可靠性、灵活性、运输和逻辑等方面），但它在特定的复杂问题领域，已显示出极大的潜力，这一新领域的巨大潜力值得重视和培育。

1.3 DNA 计算的原理及其特点

1.3.1 DNA 计算的原理

DNA 计算的基本思想[1,4]是：以 DNA 链作为信息载体，将原始问题映射成为 DNA 分子链（单链、双链或带有黏性末端的混合链），然后按照一定的规则将原始问题的数据运算高度并行地映射成对 DNA 分子链的可控生化操作。在生化酶的作用下，完成这一系列的生化操作。最后，利用分子生物技术如聚合酶链式反应 PCR、亲和层析、电泳、磁珠分离等，检测所需要的运算结果。因此，DNA 计算可分为 4 个要素：（1）运算载体。从目前的分子生物计算研究情况来看，主要以 DNA 分子进行[2,7,8,10,11]。另外，也有一些学者应用 RNA[48]以及蛋白质[49]等为材料来进行分子计算。在 DNA 分子中，有用诸如质粒 DNA 分子、单链 DNA 分子、双链 DNA 分子、发卡型 DNA 分子等各种类型的 DNA 分子为材料来进行计算；（2）运算工具。DNA 计算是利用各种生化酶来完成的，这些生化酶包括诸如连接酶、核酸内切酶、DNA 聚合

酶、核酸外切酶等；（3）生化反应系统。DNA计算中常用的生化反应包括退火、杂交、连接等。这些生化反应完成一个具体的操作；（4）检测系统。用于检测DNA计算的最后结果，常用的手段包括电泳技术、层析分析技术、分子纯化技术、同位素技术、荧光技术以及激光技术等。

在DNA计算系统中，DNA分子中的密码作为存储的数据，当DNA分子间在某种酶的作用下瞬间完成某种生物化学反应时，可以从一种基因代码变为另一种基因代码。如果将反应前的基因代码作为输入数据，那么反应后的基因代码就可以作为运算结果。这样，通过对DNA双螺旋进行丰富的、精确可控的化学反应，包括标记、扩增或者破坏原有链来完成各种不同的运算过程，就可能研制成一种以DNA作为芯片的DNA计算机。随着人们对DNA计算机研究的不断深入，用于DNA计算所对应的可控生化反应以及PCR扩增技术，特别是关于检测技术的不断提高，生物芯片技术的不断成熟，必将改进诸如Adleman提出的试管实验的方法与步骤，或者改进近期关于表面研究技术等。

1.3.2 DNA计算的特点

DNA计算的核心问题是将经过编码后的DNA链作为输入，经过一定的生物化学反应来完成运算，使得从反应后的产物及溶液中能得到全部的解空间。和传统的电子计算机相比，DNA计算具有如下3个显著特点[1]。

（1）高密度的存储容量。组成DNA的4个碱基的平均长度是0.35nm，平均每个碱基的分子量是33道尔顿，这样大约几十克的DNA分子就可以存储目前全世界所有的信息。

（2）高度的并行性。参加反应的每一个DNA分子都可以看

作为一个独立的处理器。每摩尔的溶液中含有 DNA 分子的数量是 10^{23}，即使在当前的实验条件下生化反应的效率不是 100%，而且生化操作的时间延迟，其计算速度也将是当前超级计算机的 10 万倍以上。在 Adleman 的实验中，通过适当估计，DNA 串的并行操作数目可达 10^{14}。许多研究者认为，用当前技术 $10^{15} \sim 10^{20}$ 个串的并行操作是可以达到的[2,91,92]，而目前最快的巨型机每秒能执行 10^{12} 个操作。虽然，DNA 计算每个操作本身与电子实现相比非常缓慢，但对于当前更强的计算挑战，DNA 反应的巨大并行性足以补偿这一缺点。

（3）DNA 计算机所消耗的能量仅仅是当前电子计算机的十亿分之一[2]。巨型机执行 10^9 次操作需要 1J 能量，而用于实现 DNA 计算操作的酶是在进化中产生的，具有很高的能量效率，1J 能量足以执行 2×10^{19} 次操作。以色列科学家甚至提出了不需要外界提供能量的 DNA 计算模型[50]。

DNA 计算的上述特性，即运算的高度并行性、大容量、低能耗是目前计算机和并行计算机所无法比拟和替代的。从这个意义上说，1994 年由 Adleman 所开创的分子生物计算技术具有划时代的意义。正因为如此，DNA 计算机成为人们所追求的目标。

1.4 DNA 计算的分子操纵手段

目前，对 DNA 分子的操作既有物理手段，也有化学手段。物理操作实质上是调控生化反应的外部条件，例如温度、酸碱度等。此外是来自各种生化试验手段，尤其是通过各种生物酶的操作。下面我们来介绍 DNA 计算中的一些重要操作[51]。

1.4.1 变性

两个互补碱基之间的氢键结合力比同一链内的磷酸二酯键弱，

这就使得在不破坏单条链的前提下分离两条链成为可能。DNA 分子双螺旋结构的氢键能够在一定的条件下（如加热，极端 pH 值，有机试剂甲醇、乙醇、尿素及甲酰胺等）发生断裂，当所有的氢键都被破坏时，双链 DNA 分子多核苷酸链就完全分开，这一过程就称为 DNA 分子的变性（denaturation）或解链。变性一般采用加热的方法，使 DNA 分子变性的解链温度一般为 85 ~ 95℃。

1.4.2 复性

复性（renaturation）是变性过程的逆过程，即在适当的条件下，两条完全互补的单链在适当的条件下恢复到天然双螺旋结构的过程。热变性的 DNA 一般经过冷却后即可复性。因此，此过程有时也称退火（annealing）。复性温度一般应该比该 DNA 的解链温度低 20 ~ 25℃。

1.4.3 剪切

内切酶能破坏 DNA 分子内部的磷酸二酯键，这种破坏性根据其剪切对象、剪切位置以及剪切方式的不同而不同。例如，SI 内切酶只剪切单链，或在含有单链和双链片断的混合 DNA 分子的单链内剪切，剪切可以在任何地方进行；而内切酶 DNaseI 则既可以在单链分子，也可以在双链分子上的任意位置进行剪切。相对于内切酶而言，限制性内切酶（Restriction endonucleases）却要特殊得多：它们仅剪切双链分子，而且只在一些特定的位置上剪切。通常，限制性酶被绑在 DNA 特定的识别位点（Recognition site）上，然后切开 DNA，大多数是从识别位点内部切开，有时也从外部切开。它切开相邻核苷酸之间的磷酸二酯键。这样，在一个核苷酸的 3′端生成 OH 组，在另一个核苷酸的 5′端生成磷酸酯。剪

切本身可以是齐式的（Blunt）（在两条链上笔直剪下），也可以是交错式的（staggered）。显然，如果一段 DNA 上含有几个识别位点，原则上限制性酶将对它们全部进行剪切。例如，限制性内切酶 ForkI 的识别位点是 5′-GGATG-3′。

1.4.4 绑结和连接

DNA 连接酶（ligases）的功能是封闭双螺旋骨架上的缺口，即当 3′-OH 和 5′-PO_4 彼此相邻，且各自位于与互补链上互补碱基对的二个脱氧核苷酸的末端时，DNA 连接酶会将它们连接成磷酸二酯键。这种连接过程对于 DNA 合成、修复以及遗传重组中 DNA 链的拼接都是十分必要的。此外，T4 连接酶还可以直接将平末端的 DNA 片段连接起来。需要注意的是 DNA 连接酶并不能够连接两条单链 DNA 分子或环化的 DNA 单链分子，且被连接的 DNA 链必须是双螺旋 DNA 分子的一部分。

1.4.5 分离和测量长度

DNA 分子的长度就是其所含核苷酸的数目。为了在 DNA 溶液中分离 DNA 片段或测量 DNA 分子的长度，可采用凝胶电泳（Gel electrophoresis）技术。电泳（electrophoresis）是指带电粒子在电场中向与自身带电相反的电极移动的现象，是分离、鉴定和纯化 DNA 片段的主要方法。DNA 分子是一种强极性分子，因含有大量磷酸，在中性溶液中带负电。由于 DNA 分子所带的负电荷是与其长度成正比例的，分子移动所需的力也是与长度成正比例的。即长度越长，移动时所需的力越大。因此，如果将 DNA 分子置于一个电场中，它们将朝正极方向移动。在理想溶液中，所有分子以相同的速度移动，为了使不同长度的分子能有效分开，就