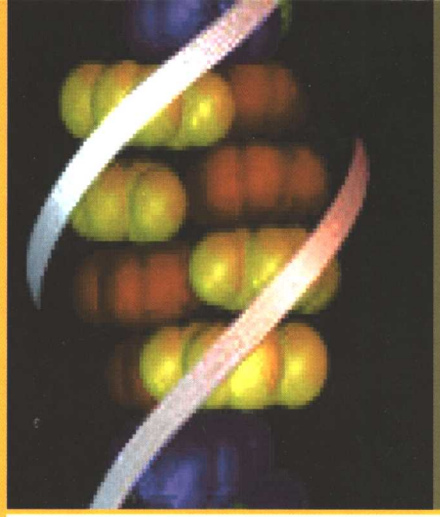
DNA Computing

New Computing Paradigms



# DNA计算

## 一种新的计算模式


G. Păun

（德） G. Rozenberg 著

A. Salomaa

许 进 王淑栋 潘林强 译



清华大学出版社

Springer

DNA Computing
New Computing Paradigms

# DNA计算
## 一种新的计算模式



清华大学出版社
北京
Springer

## 内容简介

目前在大规模并行计算模式方面主要有两种新模式：量子计算模式和生物计算模式. 本书即是对生物计算模式(DNA 计算模式)的详尽介绍，内容涉及粘贴系统、Watson-Crick 自动机、插入-删除系统、剪接系统、有穷 H 系统的通用性、剪接循环串、分布式 H 系统等. 本书内容组织合理，介绍由浅入深，并给出了所需的语言学和生物学方面的基础知识.

本书可作为生物信息学等专业的教材，也是一本该领域研究人员的极好的参考书.



**图书在版编目(CIP)数据**

DNA 计算：一种新的计算模式/(德)珀温(Păun, G.)，(德)罗森贝格(Rozenberg, G.)，(德)萨洛马(Salomaa, A.)著；许进，王淑栋，潘林强译. —北京：清华大学出版社，2004.9
书名原文：DNA Computing: New Computing Paradigms
ISBN 7-302-08658-3

Ⅰ.D…　Ⅱ.①珀…②罗…③萨…④许…⑤王…⑥潘…　Ⅲ.并行算法　Ⅳ.TP301.5

中国版本图书馆 CIP 数据核字(2004)第 046955 号

出 版 者：清华大学出版社　　地　　址：北京清华大学学研大厦
　　http://www.tup.com.cn　　邮　　编：100084
　　社 总 机：010-62770175　　客户服务：010-62776969
责任编辑：薛　慧
印 刷 者：清华大学印刷厂
装 订 者：三河市新茂装订有限公司
发 行 者：新华书店总店北京发行所
开　　本：175×245　印张：22.75　字数：437 千字
版　　次：2004 年 9 月第 1 版　2004 年 9 月第1 次印刷
书　　号：ISBN 7-302-08658-3/TP・6209
印　　数：1～2000
定　　价：39.00 元

# Preface to the Chinese Translation

It is a great honor and pleasure for us to see our book translated into Chinese, and we take this opportunity to welcome our new readership. We also wish to thank Professor Xu Jin and his associates (Wang Shudong, Pan Linqiang) for the work of translating the book. In view of the continuously growing importance of Chinese both generally and as a language of science, we are very happy to see our work published in Chinese. We have had many Chinese scientific collaborators in the past. For instance, one of us (Salomaa) has published a book about the Chinese Remainder Theorem with Cunsheng Ding and Dingyi Pei.

In the few years since the first appearance of our book, the interest in and activity around DNA computing has been growing continuously. Membrane systems due to one of us (Păun) have gained in importance. Apart from theoretical work, laboratory techniques have been developed towards the specific needs of DNA computing. Many new research groups work in the area in Europe, North America and Asia. Particularly close to us has been the European Molecular Computing Consortium, http://www.tucs.abo.fi/EMCC, which consists of groups in various countries and has overall meetings at regular intervals. Our book has appeared also in Japanese and Russian translations.

The field of DNA computing in its present form started with Adleman's celebrated experiment in 1994. The details are found below in the book. On March 14, 2002 (on the 60th birthday of one of us, Rozenberg!), Adleman and his associates published another experiment: a DNA-based computer was used to settle the satisfiability problem for propositional formulas in 3-conjunctive normal form, with 24 clauses in 20 variables. This involves going through $2^{20}$, roughly one million, possibilities.

People have for centuries tried to enhance their computational abilities by manufactoring various devices. Adleman's second experiment seems to provide the first molecular device for such an enhancement.

In spite of the rapid development in the field, we believe that our book is still quite up-to-date. This is due to the fact that the material consists mainly of the foundations, upon which further research can be built.

We hope that our new readers will find the book interesting and useful!

Tarragona, Leiden and Turku in June 2002,

Cheorghe Păun Grzegorz Rozenberg Arto Salomaa

# 序　言

看到本书被译成中文，我们感到非常高兴，而更令我们欣慰和荣幸的是该书又增添了许多新的读者. 我们还十分感谢许进教授和他的学生王淑栋、潘林强为本书翻译所做的大量工作. 中文作为一种科学交流的语言起着越来越重要的作用，因此，我们十分高兴地看到自己的研究成果有了中文版. 过去，我们与中国的许多科技合作伙伴保持着良好的联系，例如， Salomaa 就出版过 Cunsheng Ding 和 Dingyi Pei 先生所著的《中国剩余定理》一书.

从本书的首版到现在的几年间， DNA 计算的研究兴趣和研究成果取得巨大的发展. 课题组成员 Păun 先生建立的膜系统就是其中的一个重大突破. 除了理论研究的工作外，先进实验技术的发展为 DNA 计算的特殊需要提供了研究保障. 新的研究团体也在欧洲、北美洲和亚洲积极地开展工作. 特别是欧洲的分子计算课题组与我们保持密切的联系. 在网址 http://www.tucs.abo.fi/EMCC 上可以查到各个国家的研究团体和所有定期举行的会议. 另外，本专著还被译成日文和俄文.

目前的 DNA 计算领域始于 1994 年 Adleman 先生的著名实验. 具体细节可参见本书中的内容. 2002 年 3 月 14 日 (Rozenberg 先生的 60 岁生日), Adleman 和他的同事完成了另一个试验： DNA 计算机解决了 20 个变量 24 个子句的 3- 可满足性问题. 这要涉及到 $2^{20}$—— 即大约 100 万种可能性.

几世纪来，人们试图制造各种设备来提高计算能力. Adleman 的第二次试验首次为这样的努力提供了分子设备.

尽管 DNA 计算领域发展很快，但我们仍旧相信该专著的内容是最新的，这是因为该书主要包含基础理论研究和进一步的研究方向.

我们希望读者能从本书中找到乐趣和帮助！

珀温 (Păun,G.)　罗森贝格 (Rozenberg,G.)　萨洛马 (Salomaa,Arto.)

2002 年 6 月于塔拉戈纳省，莱顿和土尔库

# 译 者 序

1984 年，硕士导师王自果教授把我引进运筹学的大门. 从此我与优化计算，特别是图与组合优化方面的优化计算结下了不解之缘. 自那时起，每当在杂志上看到有关图与组合优化方面的学术论文，我都会带着浓厚的兴趣去拜读. 从 1990 年开始，我逐渐掌握了用人工神经网络求解图与优化问题的方法，后来又学会了用遗传算法和模拟退火等算法来求解图与组合优化问题. 通过这些积累，使我在解决优化问题上有了一定的方法、技巧与经验，但同时我也认识到，不管是一些常规的图论算法还是人工神经网络方法、遗传算法、模拟退火算法等，它们在电子计算机这个平台上，面对一般性困难的 NP- 完全问题，实际上都是 "无能为力" 的！这些算法只能是一点点地改进，一点点地推进. 虽然有大量的报道宣称（包括本人在内），用某方法对某 NP- 完全问题在一些特殊条件下，或者在规模较小的条件下解决得很好，但实际上，对于一般情况下大规模的 NP- 完全问题，已有的方法在 **电子计算机** 平台上都是很难解决的！

1996 年年末的某一天，阅读 "Science" 时，无意中发现 Adleman 发表的题为 "Molecular Computation of Solutions to Combinatorial Problems" 的文章. 看了题目和摘要之后，兴趣大发：还有人用 DNA 分子来求解 "我们图论中的有向 Hamilton 路问题" ！我想好好读读此文，但读不懂！原因是我不懂 DNA 分子的性质，更不了解什么是连接酶. 能读懂的只是他的算法步骤，而他的算法好像是一种最笨的枚举法，因此我也就没有在意此文.

1997 年，我的一位博士生发现了 Ouyang 等人发表在 "Science" 上的题为《最大团问题的 DNA 解 (DNA Solution of the Maximal Clique Problem)》的文章. 阅读之后，觉得与 Adleman 的文章类似，但还是看不懂. 于是，我带着这篇文章，请教了陕西师范大学生物系的一位张老师. 他读了此文之后觉得有道理. 我决定试着学习 "用 DNA 分子求解图论中的 NP- 完全问题" 这一方法，从此，开始学习有关分子生物学的知识. 待有了一点点门道之后，决定好好在这一领域耕耘，并动员我的几位博士生放弃在神经网络和遗传算法等领域的研究，来从事 DNA 计算方面的学习与研究. 到目前，我已经为祖国培养出了从事 DNA 计算研究的博士后 2 名，博士 7 名.

1999 年，看到 Păun 等人撰写的关于 DNA 计算方面的学术专著，立即开始了阅读学习，并萌发了尽快译成中文介绍给我国学者的想法. 本来本书早就应该与读者见面，但是由于种种原因延误至今才出版.

本书是国际上第一部关于 DNA 计算方面的学术专著. 此专著主要总结了 1998

年以前几个主要 DNA 计算模型的数学理论基础方面的研究成果. 全书共 11 章. 第 1 章给出了 DNA 分子的基本结构与性质. 第 2 章主要介绍了 Adleman 和 Lipton 两人在 DNA 计算方面的开拓性工作. 第 3 章给出了后面章节所用的形式语言的有关基本理论. 从第 4 章开始, 作者介绍了 DNA 计算中的几个模型的数学理论, 其中第 4 章讨论粘贴系统 (sticker system), 它是在由 Roweis 等人提出的粘贴模型 (sticker model) 的基础上抽象出来的一种纯数学模型. 第 5 章研究 Watson-Crick 自动机, 它是有关这种粘贴系统的自动机的理论体系. 第 6 章给出了插入 - 删除系统. 第 7 章介绍了所谓的剪接系统, 此系统是目前学者们所关注的一个研究领域. 第 8 章讨论了有穷 H 系统的通用性, 以期对未来的通用 DNA 计算机理论进行探索. 第 9 章给出了剪接循环系统. 第 10 章给出了基于文法系统分布式结构的分布式 H 系统. 第 11 章针对剪接系统从变量的角度对模型进行了进一步的扩展.

Păun 等人所撰写的这本学术专著对 DNA 计算的研究与发展起到了很大的促进作用. 特别是在短短的 3 ~ 4 年的时间内, 从数学的角度对提出的上述 DNA 计算模型进行了很好的研究, 得出了一些出色成果. 虽然 DNA 计算研究领域发展迅速, 但这本专著出版至今, 仍然从理论上对 DNA 计算的研究具有良好的指导作用. 遗憾的是, 本书未能涉及 DNA 计算的核心内容 ——DNA 计算的生化机理以及 DNA 计算的检测方法与技术等方面.

本书的完成与华中科技大学校领导以及系所领导的支持和关心是分不开的. 他们是: 杨叔子院士, 周济院士, 李培根院士, 丁烈云教授, 陈学广教授, 陈国清教授, 张七一教授, 周细刚教授, 齐欢教授, 孙德宝教授以及系党总支周建波书记.

感谢朝夕相处的费奇教授、冯珊教授、廖晓晰教授、岳超元教授、王红卫教授、赵勇教授等.

我还要感谢本书的作者 Păun 教授, 他对翻译工作给予了大力支持, 给出原版中出现的一些错误, 而且使我们双方的研究队伍作了进一步的合作研究.

参加本书翻译的还有刘文斌博士、董亚非博士、刘西奎博士以及张连珍硕士等.

由于译者水平所限, 在翻译过程中难免出现疏漏、不妥乃至错误等, 望各位专家与广大读者不吝赐教.

许 进

2004 年 3 月 26 日于华工园

# 引言　DNA 计算简介

从硅过渡到碳. 从微芯片过渡到 DNA 分子. 在计算机中, 利用有机分子的信息处理能力来代替数字开关部件, 这就是 DNA 计算的基本思想.

以当前的计算机技术要实现微型化存在明显的局限性, 所以要进行大的革新. 很早以前就有人提出现代计算机的基本部件应逐步过渡到分子水平, 这样一来, 它将会比我们利用当前技术制造出的任何东西都要小得多. 量子计算和 DNA 计算是当前这种思想的两种不同表现, 本书主要介绍 DNA 计算.

我们知道, 计算机已有很久的历史, 设计用来便利计算的机械的发明历史就更长了, 已知最早用于进行计算的重要装置是算盘. 当今, 电子计算机 (图 1) 在我们的社会生活中已经取得了支配地位, 没有它们的帮助, 我们的大部分活动都会举步维艰. 然而, 当今的计算机还有许多缺陷, 由于存在大量棘手的问题无法解决, 所以这种计算机还不是漫长发展道路上的终点.

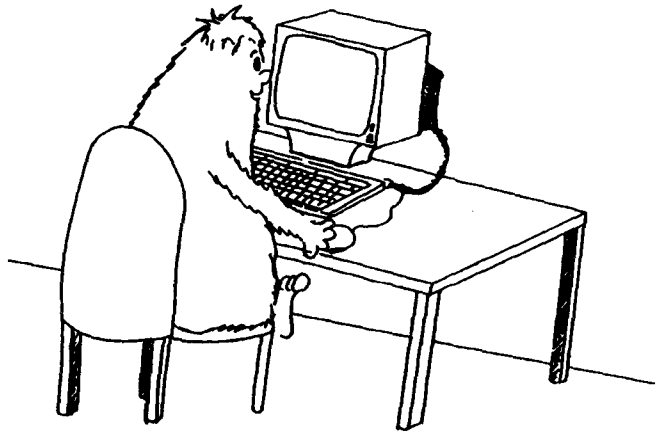


图　1

沿着这一发展道路走下去, 也许就会到达 DNA 计算机阶段, 如图 2 所示, 对试管的所有操作由用户自行完成.

图 3 描述了一种更为先进的模型, 在这里机器人和电子计算与 DNA 计算相结合, 对试管的大部分操作可自动完成, 而无须用户的介入.

当今计算机的著名先驱 Charles Babbage 大约在 1810—1820 年开始着手建立一种自动化的计算机 —— "差分机", 以及功能更强大的机器 —— "通用分析机", 但最终归于失败, 其原因在于缺少足够精确的工具, 缺少只有在 20 世纪才有的机

械和电子设备. 如今，展望 DNA 计算机，也许我们面临着相似的局面：生物技术还没达到足够尖端和精确的水平，特别是，这些技术还没有向着 DNA 计算这种特殊需求的方向适当发展. 不过，这一等待时期很可能要比 Babbage 短得多.

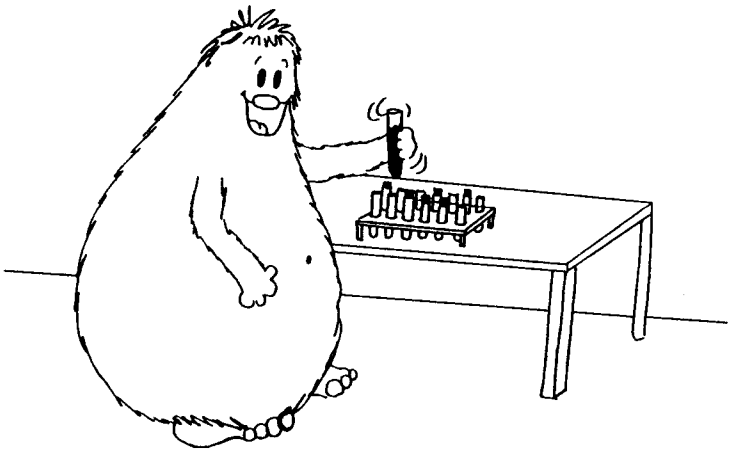


图 2

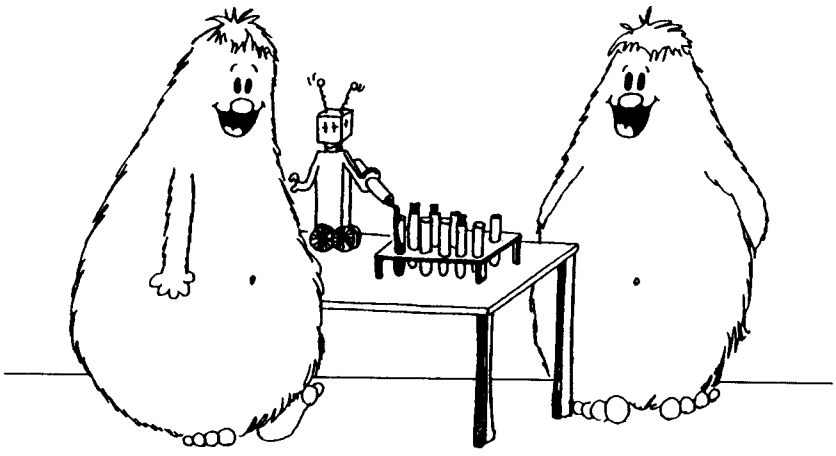


图 3

我们之所以对 DNA 计算的前景抱有如此高的期望，主要基于两点：

(i) DNA 链的巨大并行性;

(ii) Watson-Crick 的互补结构.

现在简要说明这两个特征：

(i) 许多著名的难解计算问题可以通过穷尽所有可能解来解决，然而，这样的搜索规模过于庞大，利用当前的技术是无法完成的. DNA 链却能够高密度地存储信息，并能轻而易举地进行大量拷贝，这也许可以使穷尽搜索成为可能. 一个典型的例子应该是对密文进行密码分析：我们可以同时尝试所有可能的密钥.

(ii) Watson-Crick 的发现是自然界“免费”提供的一个特征 (在理想条件下)，当两条 DNA 单链结合时，相对的两个碱基是彼此互补的. 因此，如果知道了一条链上的一个碱基，也就知道了对应那条链上的另一个碱基，没有必要时时刻刻都去检测. 这一特点提供了一个有力的计算工具，因为互补结构将广泛存在的配位移动语言带到了计算现场，通过将信息以不同形式编码于要结合的 DNA 链中，就能够得出基于结合发生的更进一步的结论.

下面进一步详细阐述这种互补的范例： DNA 由聚合链组成，通常称为 DNA 链. DNA 链由核苷酸构成，核苷酸有 4 个不同的碱基： A(腺嘌呤) 、 G(鸟嘌呤) 、 C(胞嘧啶) 、 T(胸腺嘧啶). 常见的 DNA 双螺旋结构由两条独立的链结合而成，称为 Watson-Crick 互补双链结构. 由于碱基配对相互吸引： A 与 T 连接，G 与 C 连接，形成如图 4 所示的双链结构.

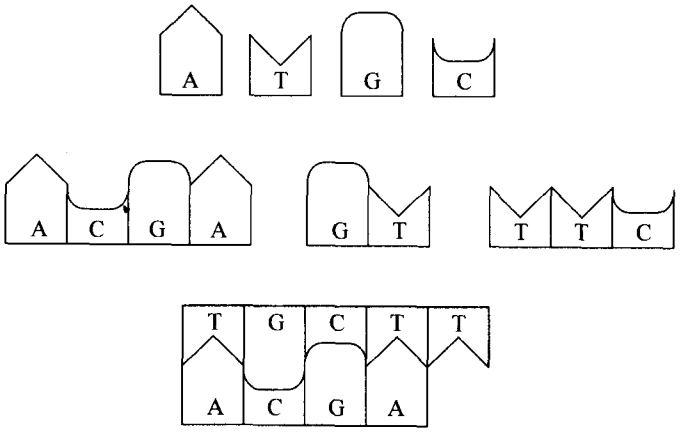



图　4

图 5 和图 6 说明了这种互补结构的重要性. 特别是，如果自然界没有给我们提供这种互补结构，事情将会变得相当复杂. 图 5 中 DNA 计算机用户面临着从大堆单链中寻找匹配这样艰难的任务. 如果图 5 的局面是真实的话， DNA 计算的前景就黯淡无光了，也许本书第二部分提出的理论也就毫无创见性可言了. 但图 5 的情景并非现实，彼此匹配的链互相发现以后， 用户可以很快得到图 6 的结果.

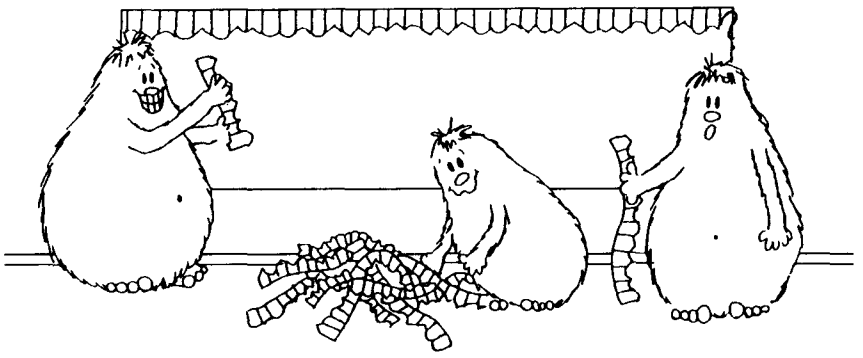


图 5

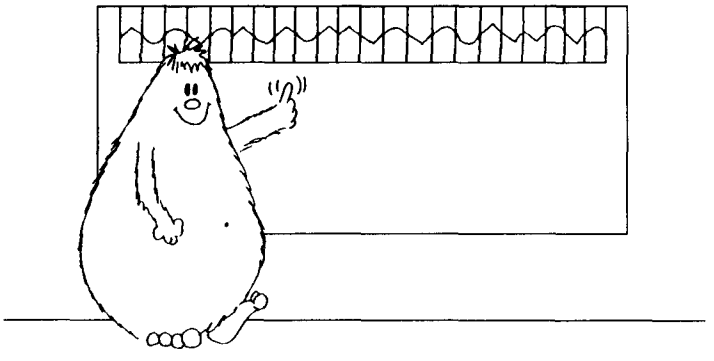


图 6

这种互补范例或一些归纳及修改会在本书的第二部分用一些数学理论来具体阐述. 第一部分是关于 DNA 计算的总体介绍, 包括本书中用到的分子生物学基本概念的介绍 (第 1 章), 同时对实验室研究的前景进行了讨论. 例如, 利用 DNA 链进行操作时, 一些小小的错误可能会使结果大相径庭, 所以, DNA 计算的最终成功很大程度上依赖于相应的实验室技术的发展.

除了利用 DNA 链作为计算支持来解决计算中的困难问题之外, 研究 DNA 计算还有许多原因. 一方面, 理解自然界是如何“计算”的这一点相当重要 (了解生命惊人的精密性和完备性是如何通过 DNA 操作得到的); 另一方面, 在下面的章节中将会发现, “通过 DNA 计算”产生了计算片断: 全新的数据结构, 全新的操作类型, 全新的计算模式, 这一点与当前计算机科学中的习惯有很大不同. 即使证明构造 DNA 计算机不现实 (比如说易发生错误), 而另一种选择可以是在硅框架内执行这一新的计算.

可以进一步理解这些说明: 传统理论上的计算机科学植根于重复写操作, 这

对于大部分自动机械装置和语言理论模式是正确的. 然而, 在计算行为中, 自然界操作 DNA 分子利用的是完全不同的操作类型: 剪切、粘贴、连接、插入、删除等. 我们将会证明, 利用这些操作可以建立计算模型, 并且在功能上等价于图灵机. 于是, 在这种崭新的框架体系中需要重新构建计算理论. 当然, 这一新事物对于计算机科学应用是否具有实际意义是一个为时过早的问题.

# 目　录

# 第一部分

# 数学理论