计算机. DNA . 碱基

(3)

303-309

计算机科学发现新大陆 —— DNA

Cipra, B

Q.523

Barry Cipra

邹声元

去年春天在麻省理工学院演讲时，伦纳德·阿德勒曼 (L. Adleman) 从他的口袋里掏出了一个超级计算机，他把这台机器称为 TT-100. 这台机器虽然只做过一些简单的计算，但阿德勒曼和其他人相信它蕴涵的理念可使计算的方式产生“进化”.

TT-100, 这个名字代表 100 微升试管，是一个粗短的、铅笔大小的塑料小瓶，含有微小的微处理器溶液，这些溶液就是通常所说的 DNA. 在《科学》杂志的研究报告中，南加州大学的计算机科学家阿德勒曼论证了利用 DNA 解决数学问题的可行性.

这个思想使阿德勒曼的许多同事满怀激情. “我想，我们将找到碱基的认知能力，人人都将对它感到兴奋，” 普林斯顿大学的计算机科学家理查德·利普顿 (R. Lipton) 说，“我们将更多地了解计算，我们也将更多地了解 DNA. 那肯定棒.”

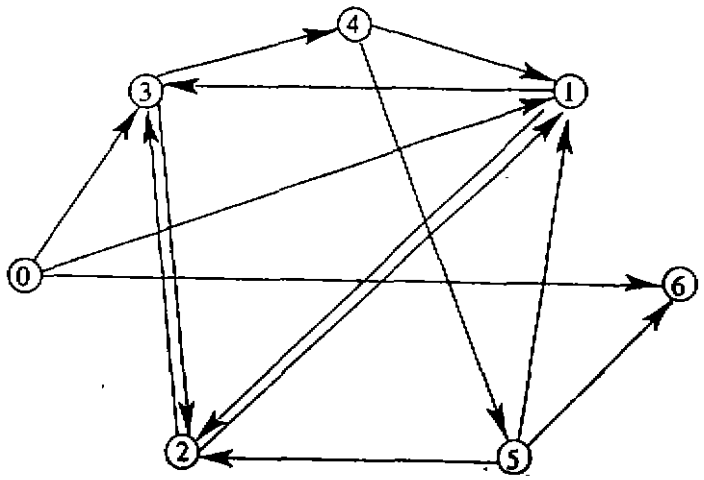

图1.阿德勒曼的开创性的图. 沿这个图的单向路径，找到从城市“0”到城市“6”的唯一哈密顿通路，DNA 花了将近一个星期的时间 (摘自《科学》，266 卷，1994 年 11 月 11 日，1022 页).

如果它在实用层次上获得成功 —— 这是非常可能的假设，DNA 计算将彻底改变计算机“硬件”的性质. 过去五十年来，计算几乎成了电子学的同义词. 但是，阿德勒曼

原题：Computer Science Discovers DNA. 译自：What's Happening in the Mathematical Science, Vol. 3, AMS, 1996, pp. 27–37.

的 DNA 计算机依赖生物化学而不是半导体物理学来执行作为计算基础的逻辑运算. 实际上，它用每个 DNA 片段作为一个独立的处理器，依照描述 DNA 分子特征的碱基的顺序对信息编码，利用自从 1953 年沃森 (Watson) 和克里克 (Crick) 发现双螺旋结构以来生物化学家发展起来的精巧过程进行计算.

在《科学》所述论证中，阿德勒曼用 DNA 解决了一个简单的组合问题 —— 哈密顿通路问题. 这个问题也能表述为连接一个城市集合的单向路径系统. 从一个特定的城市开始，到另一个城市终止，沿这条路依次访问每个城市刚好一次是可能的吗？当城市数目很大时，哈密顿通路问题变得异常困难. 它是许多所谓的“ NP- 完全”问题中的一个，对这些问题计算机科学家假设不存在有效的算法 (参看《数学科学中在发生什么》第一卷的“复杂性基础”). 然而阿德勒曼的例子只有七个城市 (见图 1). 人解决这个例题不是难事，只要一会儿功夫就可以“看出”结果. 但是，DNA 没有视觉系统，让它去寻找答案要求完全不同的看待问题的方法.

阿德勒曼“编写”他的 DNA 计算机程序是通过数组寡聚核苷酸 (这个生物技术术语指 DNA 单链的短片段). 第一组，含有七个不同的片断，每个片断 20 个碱基长，表示城市. 第二组，表示单向路径，这些片断也是 20 个碱基长. 每个路径片断的前 10 个碱基与对应此路径的起始城市片断的后 10 个碱基互补，而路径片断的后 10 个碱基与此路径到达城市的前 10 个碱基互补 (见图 2).

用生物技术的优点是 DNA 十分丰富而且 (至少，每个单位) 生产便宜：现在花 40 美元你可以买大约十亿亿 (10^{17}) 份寡聚核苷酸. 阿德勒曼每种只“拈”一点，但即使如此，片断的数量也十分巨大，这就意味着他要有把握期待从 DNA 的大海中捞出针来.

计算的第一步很容易：阿德勒曼只是把所有单链 DNA 片断堆在一起. 城市和路径寡聚核苷酸的互补部分开始迅速结合在一起，A 和 T 匹配，C 和 G 匹配，从而生成更长的双链 DNA 片断序列. 每个这样的序列表示城市到城市的一条可能的路径. 问题是它们中是否有一条是从城市“ 0 ”到城市“ 6 ”的哈密顿通路.

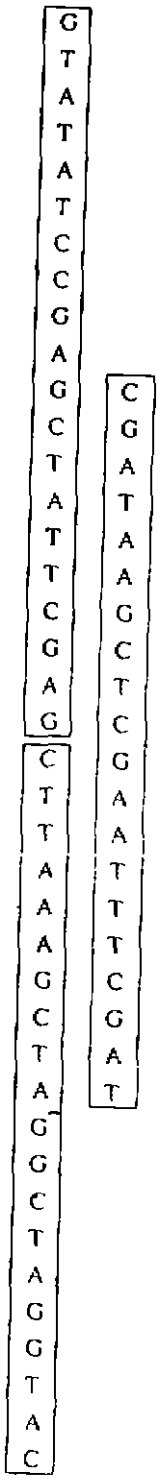

图2.左边排列的是两条寡聚核苷酸“路径”、右边是一个与之互补的 DNA 片段“城市”

为了找到答案，阿德勒曼利用作为 DNA 指纹基础而众所周知的聚合酶链式反应 (PCR), 首先“放大”所有在要求的段上开始和终止的序列. 然后，由于琼脂糖凝胶能按分子尺寸区分分子，他利用琼脂糖作用于所得到的生成物，而且只保留对应 140 个碱基对的部分，因为只有这些 DNA 链对应连接七个城市的通路. 他要再放大和用胶洗涤这种 DNA 数次. 其后，阿德勒曼把双链 DNA 片断分解成单链片断并且用每个城市的寡聚核苷酸“孵化”它. 他每次对一个城市这样做一遍，同时洗掉任何不包括城市 DNA 片断的任何 DNA.

最后，分别做了六次放大反应，阿德勒曼标示出了哈密顿通路本身 (通路唯一). 比如第四次反应，合成了 140 个碱基对通路的 DNA 放大链，它从城市“ 0 ”开始，到城市“ 4 ”结束. 每个子链 (用一琼脂糖凝胶测量到) 的长度告诉阿德勒曼依什么次序访问城市.

整个计算花了一个星期 —— 可算是创记录. 但阿德勒曼达到了他的目的：证实了生物技术能用于解决一个与生物化学无关的“抽象”问题. (亚历山大·格雷厄姆·贝尔 (A. G. Bell) 也没有真隔着城镇与他的助手通过电话.)

DNA 超越普通电子计算机的可喜之处纯粹在数量上：硬件工程师认为十亿是个相当大的数字，但生物化学家惯常处理大于其百万亿倍的量，例如阿伏伽德罗数，稍大于 6×10^{23}. 一次讨论中，阿德勒曼争辩道，甚至一微摩尔的 DNA 也有在各个层次上的并行处理能力.

并行处理与传统的顺序处理计算是完全不同的. 即便你一次打开了几个“视窗”，你的家用计算机仍然一次只做一个计算. 相反，在并行计算中，计算机同时做许多计算. 这是通过在许多独立的处理器中分解任务达到的. 其基本思想是一千个便宜的计算机比一个超豪华的机器更快地做一件事情，即使一个便宜的计算机比一个昂贵的慢一百倍.

DNA 计算把计算的概念推到了“一个几乎荒谬的极限”，阿德勒曼说：每个“处理器”可能比它的高技术的电子对手慢万亿倍，为了一个单步“计算”，都要求数分钟到数小时. 但是，当乘以比如 10^{20}(这个数通常被认为接近现实的分子计算机中能够应用的极限),DNA 计算反比其电子对手快一亿倍. 根据阿德勒曼的说法， DNA 计算与传统的计算机比，在能量效率上可能高出万亿倍；与录象带这样的媒体相比，在空间上可能只占万亿分之一.

阿德勒曼确信他开创了一个令人激动的新领域. 但他并未指望分子计算的思想会很快流行.

利普顿是第一个加盟的. 阿德勒曼的论文在《科学》上刊出后不久，他和阿德勒曼在桑蒂斐的一次会议上相遇. 一个星期后，阿德勒曼从电子邮件上得知利普顿关于这个主题已经写了一篇短文. “我十分惊喜，”阿德勒曼说. “利普顿已经简化了整个观点，抽象出了什么是重要的，而且论证了分子计算可以广泛地应用于各类问题，其广泛程度比一个人开始可能想到的大得多. ”

利普顿注意到 DNA 计算归结为四个基本过程. 其一只是检验一个试管到底是否含有任何 DNA. 更复杂的过程是用一个特定的碱基子串分离所有的 DNA 片断，这很像以

邮政交易为目标的程序可能要用一给定的邮编和收入等级筛选数据库以确定住址. 第三个过程，很简单，把两个试管倒一起. 第四个过程是放大 DNA 数量，譬如加倍，也许重复多次.

与用 DNA 寻找一条未知的哈密顿通路不同，利普顿的理论依靠大量的 DNA 在一个特定的、高度结构化的图中产生所有的通路 (见图 3). 这些通路依次用于表示 n- 比特的数. 例如，图 3 中从 a_1 到 a_4 有 8 条通路. 通路 $a_1x_1'a_2x_2a_3x_3a_4$ 能翻译成二进制数 011, $a_1x_1a_2x_2'a_3x_3a_4$ 表示 101, 等等.

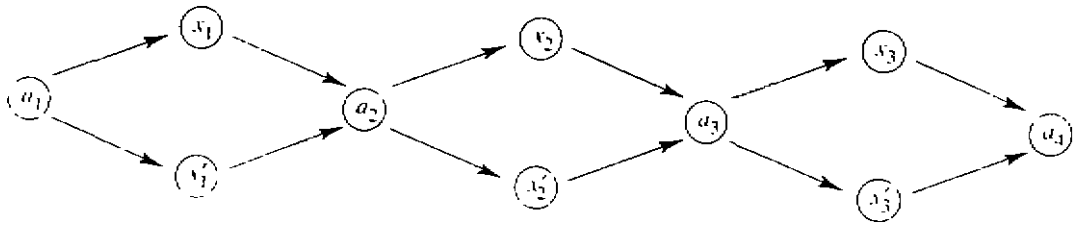

图3. 一个其通路表示二进制数的图

利普顿说、利用 DNA, 考虑编制数量大至 70- 比特的数是可能的. 这使人相信可能攻克如下的著名计算机科学难题：决定一给定的带数十个“布尔”变量的逻辑表达式是否“可满足” (satisfiability), 即是否存在一个对变量的赋值使表达式为真. “布尔”变量只取两个值 (真 / 假，或按计算机的表示， 1/0).

简单的逻辑表达式，例如“ (A 或 B) 和 (非 A 或非 B) ” (符号化为， (A ∨ B) ∧ (¬ A ∨¬ B)), 可满足或不可满足，是容易检查的，既然容易尝试所有可能的赋值 (在这个例子中，让 A 为真 B 为假是个可满足的赋值). 但对于带许多变量的复杂表达式，检查是否可满足就不是这么容易. 事实上，它是最早的 NP- 完全问题.

然而，利普顿显示了相当多的可满足问题能用 DNA 计算机解决，且要求的步数大致正比于逻辑表达式的长度. 根本上说，一系列的“合成”和“分离”过程分解所有 DNA 片断，这些 DNA 片断对应满足一给定布尔表达式的赋值，并且最后一个“检测”过程指出是否有任何东西留下.

可满足问题在实践中并不常出现 —— 也许，阿德勒曼说，因为一直没有实用的解决方法. 但是，许多计算机算法含有搜索例行程序，而至少在理论上，搜索例行程序能够通过利用 DNA 的超并行机制加速. 加速因子可能大至万亿.

一个可能的应用是在密码学领域. 利普顿的研究生丹·邦纳 (D. Boneh) 和克里思托弗·邓沃思 (C. Dunworth) 已经计划冲击一个所谓的数据加密标准 (DES) 系统. DES 由国家安全署发明供政府使用. 它的设计类似于标准的组合锁：基本的左右转动过程对每把锁都是相同的，但每个用户有一个各异的秘密组合. 对 DES, 这个组合，或者说“秘钥”，是一个 1 和 0 组成的 56- 比特的单串，它意味着有大约 10^{17} 种不同的可能. 在可以预见的未来，这对防御使用硅片的计算机那是足够了，但它却在一小罐 DNA 能力所

及的范围内失去作用.

邦纳和邓沃思苦心完成了对应于执行 DES 算法的生化过程序列. 他们发现用小于一千步来破译密码是可能的 —— 用目前的实验室技术大约是四个月的工作.

利普顿提到, DES 的分析更多的是一个理想实验而不对密码学构成任何真正威胁. 通过加倍 DES 秘钥的比特数, 邦纳和邓沃思的这种靠蛮力的 DNA 冲击很容易被击退: 甚至化学家也认为 10^{34} 是一个很大的数.

提高 DNA 计算技术来与目前的电子计算机竞争的难易程度也还不清楚. 可靠性是着重关心的: 生物化学家习以为常的错误率足以使一个现代计算机制造商破产. 不过, 利普顿指出, 回忆靠真空管运行机器的日子, 电子计算机在可靠性上也曾有过它自己的麻烦. 生物化学也可能经历一场像晶体管发明一样的戏剧性变革.

的确, 利普顿说, 即使 DNA 计算机本身永远不成功, DNA 计算也将可能刺激生物技术的进步. 一方面, 大规模可靠系统的计算需求将推动分子生物学家发展越来越好的方法; "我们将以他们 (生物学家) 未曾经历过的方式向他们施加压力," 利普顿说. 另一方面, 当计算机科学家里里外外更多地了解分子生物学, 他们应该能够把从计算机科学中取得的洞察力转换成对生物学家有用的知识. "如果足够透彻地理解操纵 DNA 时分子生物学在做什么, 那么它就是在做一种奇妙的计算," 利普顿解释道. "假如你在排列 DNA、你正在做实验, 你也正在推理, 在某种通常的感觉上这就是计算. 也许, 作为计算机科学家, 我们可以用一种不同的、更有效和更容错的方式帮助组织这件事." 以同样的方式, 计算机科学家已经在集成电路方面极大地帮助了固体物理学家. "我们正尝试在相互交往的道路上做更多的了解," 利普顿说.

最后, 阿德勒曼和利普顿说, DNA 计算面临的决定性问题是计算机科学家是否能找到这样的科学上 (或经济上) 有重要意义的课题, 与不厌其烦地一个接一个飞快抓动开关相比, 利用 DNA "疯狂" 的并行机制, 这些课题能更有效地解决. 他们一旦找到这种课题后, 还会有什么问题将出现呢?

"我不认为这样一种并行的体系将解决我们传统上利用超级计算机解决的每个问题 —— 也许本来就不应该这样," 利普顿说. "更可能的是, 其中一些很专门化的问题, 像上述密码学问题, 将按这种方式加以筹划和编写程序. 换言之, 人们将认识到有难以置信的并行方法解决它们. 而且如果对有足够重要性的问题这样做的话, 那么 DNA 计算机就有了某种经济上的促进力."

阿德勒曼的考虑甚至已经开始超出 DNA. 他认为有希望来裁剪计算机以适应全方位的需求. "实际上, 没有通用计算机这样一种东西," 他说. "所有的计算机都是专用的," 对一些问题运作快捷, 对另外一些问题就迟缓. 超快但基本上是串行的电子机器和阿德勒曼的悠闲但超并行的 DNA 计算机处于相反的两个极端. 阿德勒曼谈到, 也许有人会问, 对于某些实际问题, 是否存在某个处于二者之间的东西, 它能最好地处理这些问题, 比如说带百亿个处理器的计算机, 它的每个处理器每秒钟运行小于一百个操作, 却与相连接的比如说一千个毗邻处理器进行 "通讯".

在某种意义上答案是显然的; 这样的计算机已经存在. 当你思考这个问题时, 你就

正在用你自己这台计算机，不是吗！

比特，拜特和碱基对

特玛·施里克 (T. Schlick) 和威尔玛·奥尔森 (W. Olson) 正合作从事一种不同类型的 DNA 计算. 施里克是纽约大学化学系和纽约大学数学科学库朗 (Courant) 研究所联合聘任的应用数学家，目前，她还加入了霍华德·休斯医学院. 奥尔森是鲁特格斯大学的化学家. 他们一直在利用计算机仿真研究在化学力作用下 DNA 和其他高分子 (如蛋白质) 盘绕扭折的反应方式.

经典的 DNA 沃森 - 克里克模型是一个简单的双螺旋楼梯结构，楼梯的每级由互补的碱基对组成，碱基对由嘌呤、嘧啶腺嘌呤、胸腺嘧啶、鸟嘌呤和胞核嘧啶构成，它们联接着两条糖磷酸带 (见图 4). 蛋白质也常常简单地看作氨基酸链. 但真正的情况是，所有的这些分子都有复杂的三维形状，对其无数的生物功能这是很基本的. 例如，假如蛋白质不折叠，它们就不能做太多的事情. 而且，假如 DNA 真分布得像教科书上显示的那么直，它甚至填不进一个细胞中 (人的染色体平均会有四分之三英寸长)，更不用说与蛋白质相互作用了.

施里克和奥尔森对 DNA 像一团乱麻包缠自己的超螺旋现象特别感兴趣. 超螺旋总的起因足够简单：由于寻求能量最低状态，任何长的有弹性的物体都会通过局部扭曲转向整体缠绕. 然而，准确的 DNA 超螺旋动力学还无人知晓. 利用一个简化的 DNA 弹性模型 (其能量函数基于实验测定的弯折扭曲参数)，施里克和奥尔森已经仿真了一个 DNA 闭环的形变，从初始的圆环转变到了第八图 (见图 5).

这个仿真基于由施里克和查尔斯·佩斯金 (C. Peskin)(也来自库朗研究所) 发展的新算法，以得到盘绕高分子运动方程的近似解. 这些新算法，称为兰格芬 / 隐 - 欧拉 (LI) 模式，与传统方法相比，它使获得时间增量更大的近似动态“快照”成为可能.

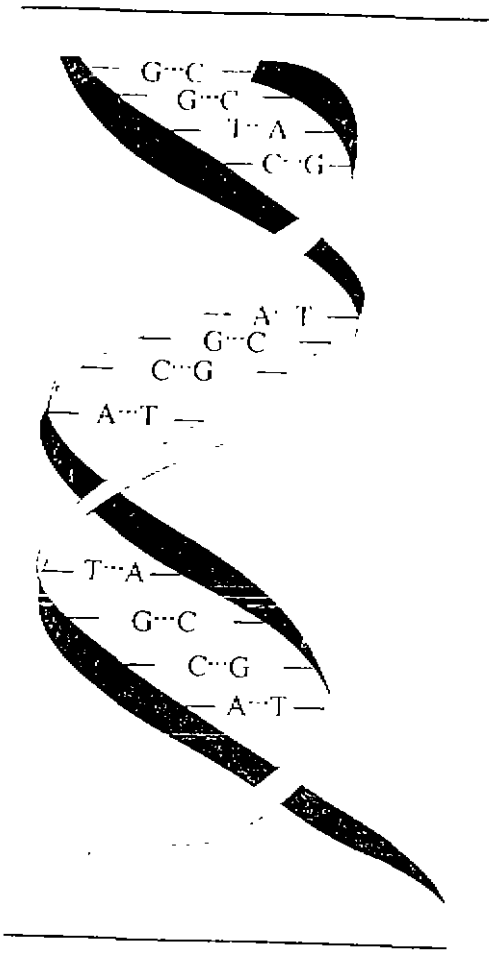

图4. 双螺旋结构由连接糖磷酸“带”的互补碱基对 (A 和 T,C 和 G) 组成

这个结果正促进人们透彻了解 DNA 生物化学——溶剂和盐离子怎样在最短时间内影响超螺旋动力学. 但还有很长的路要走，施里克说. 准备一万幅快照需要花超级计算机数小时，却还只表示一微秒的时间跨度. 然而施里克和纽约大学化学系的张桂华 (音) 开发了一个称为 LIN(兰格芬 / 隐 - 欧拉 / 范式) 的新的计算途径，为更有效的计算提供了希望. 实质上， LIN 的范式部分处理高频组元而 LI 部分维持进行大的时间步骤的能力.

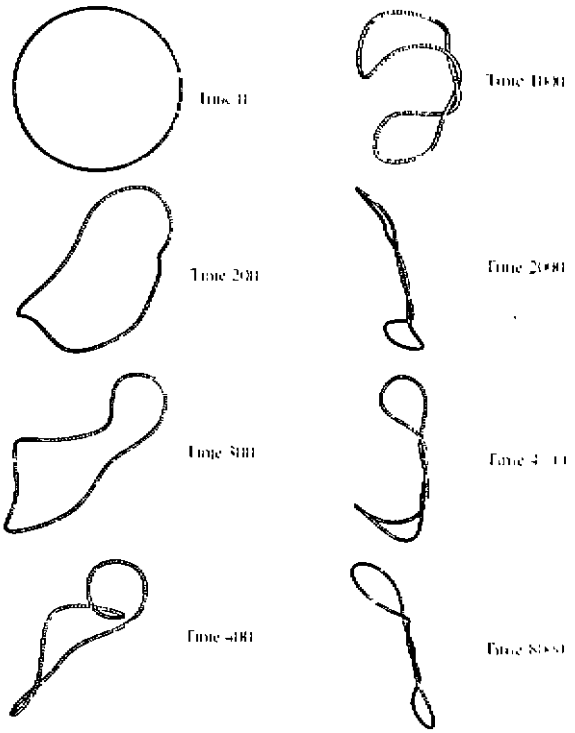

图5. DNA 超螺旋仿真“快照”. (承蒙特玛·施里克和威尔玛·奥尔森提供图片，摘自《分子生物学杂志》， 1992 年， 223 卷， 1089-1119 页，学院出版社)

“最终我们愿意在整个原子细节上研究超螺旋，”施里克说. 他们最后也要研究 DNA 和蛋白质之间的相互作用以及蛋白质本身的缠绕. 然而，这些问题将要求下列几个前沿课题有相当的发展：分子数学建模，计算动力学的算法和自然态计算机能力. 既然电子计算机目前已用于探索固体物理学中的数学模型，那么对于生物化学家的分子动力学模型， DNA 也应该成为一个理想的计算机；这样的幻想倒是很公平的. 另一方面，施里克说，任何时刻研究人员得到一个新的机器，他们的欲望也相应地增长. 按她的经验，对于计算机的能力，有件事情总是真的：“永远不够.”

(邹声元 译　陆维明 校)