DNA 分子计算模型

李 燕 王秀峰
(南开大学信息技术学院 天津300071)

The Model of DNA Computing

LI Yan WANG Xiu-Feng
(School of Information Technology, Nankai University, Tianjin 300071)

Abstract The field of practical DNA computing opened in 1994 with Adleman's paper, in which a laboratory experiment involving DNA molecules was used to solve a small instance of the Hamiltonian Path problem. The characteristic of this computation is its powerful ability in parallelism, its huge storage and high energy efficiency. This paper mainly introduces the principles of DNA computing and the sticker computing model.

Keywords Computing science, DNA computing, Sticker computing model

1. 引言

20世纪30年代，由于受到构造性数学学派的影响和数理逻辑的发展，K. Gödel、A. Church、A. M. Turing、E. L. Post 在研究中陆续提出了一批计算模型，如递归函数、λ演算、图灵机、波斯特系统等。进一步的研究发现，这些模型在能力上是等价的。如果是这些计算模型解决不了的问题，任何算法也解决不了；但凡是能用算法方法解决的问题，也一定能用这些计算模型解决。这些模型被称为用算法方法解决问题的极限。在这些计算模型中，以图灵的研究与实际计算机更为接近[6]。半个多世纪以来，计算科学已经取得了重大进展，但还存在许多尚未解决的问题，仍需在理论上和实践中进行深入的研究。

1994年，美国加州大学的 Leonard M. Adleman 在《Science》杂志上发表了首篇关于 DNA 计算的文章[1]，大胆地冲破了传统计算的思想束缚，提出了 DNA 计算的概念。Adleman 应用分子生物学技术，建立了解决哈密尔顿路径问题的 DNA 模型，并成功地进行了实验。这一实验为分子计算的研究打下了坚实的基础。这个奇迹表明了采用 DNA 分子进行特定目的计算的可行性[8~10]。在 Aldeman 实验成功后不久，Lipton 提出了用于求解布尔代数式的满意问题的 DNA 模型。Lipton 发明了一种编码方案，能将 DNA 碱基对翻译成0、1代码，该方案能使 DNA 分子模拟电子逻辑门进行 yes-no 的判断[12]。随后，研究人员相继提出一些不同的 DNA 模型及计算方法。这一将计算技术与生物科学相结合的全新领域，正以其蓬勃发展的趋势为计算科学开辟新的研究方向。

2. 计算科学的新领域：DNA 计算

2.1 DNA 分子的结构

DNA(Deoxyribonucleic)是一种高分子化合物，称为脱氧核糖核酸，是遗传信息存储的媒介。组成 DNA 的基本单位是核苷酸，许多核苷酸连在一起构成了 DNA 聚合体。每一个构成 DNA 的核苷酸是由三部分组成的：①一个脱氧核糖分子(S)；②一个磷酸基团(P)；③一个含氮碱基(A、T、G、C)。组成核苷酸的含氮碱基有4种，它们是腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)、胸腺嘧啶(T)。相应地由不同碱基组成的脱氧核苷酸也就分别称为脱氧腺苷酸(A)、脱氧鸟苷酸(G)、脱氧胞苷酸(C)、脱氧胸苷酸(T)。脱氧核糖有5个碳原子，为了避免糖上的碳与碱基上的碳相混淆，将糖上的碳编码为1′到5′。DNA 分子是由两条头尾方向相反的脱氧核苷酸长链，以右手螺旋的方式盘绕着同一中心轴而形成的。两条链上的碱基通过氢键连接起来，碱基对的连接严格依据 *Watson-Crick* 碱基互补配对原则，即腺嘌呤(A)通过两个氢键与胸腺嘧啶(T)配对，鸟嘌呤(G)通过三个氢键与胞嘧啶(C)配对。反过来也必定是 T 与 A 配对、C 与 G 配对。整个 DNA 分子是有方向性的。如果 DNA 分子起始位是5′位碳原子，那么另一端必定是3′位碳原子。如图1所示。科学实验表明，生物的特性及遗传信息都储存在 DNA 分子的碱基对中，因为 DNA 的分子碱基对的排列组合方式是千变万化的，所以就构成了 DNA 分子的多样性。

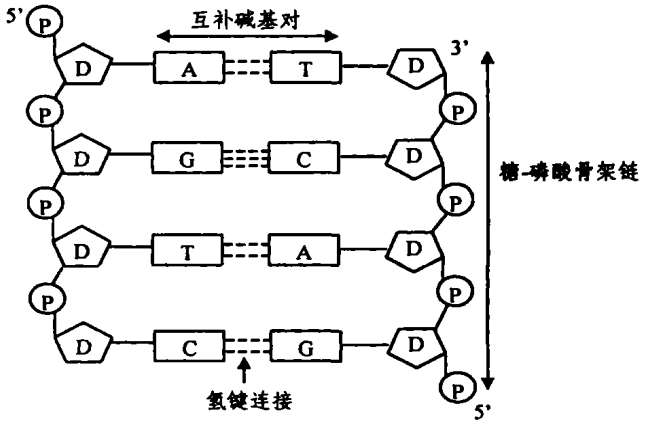

图1 DNA 分子结构图

2.2 DNA 计算机理

DNA 计算是一种关于计算的新的思维方式，本质上就是利用大量不同的核酸分子杂交，运用分子生物技术产生类似某种数学过程的运算结果。一个 DNA 单链可以看作是由4个不同的符号 A、T、C、G 连成的一个串。所以可以使用4个字母

李 燕 博士研究生，主要研究方向：智能预测与控制、进化计算。王秀峰 博士生导师，主要研究方向：复杂系统建模、智能控制、进化计算。

的集合$\sum=\{A,T,C,G\}$，对问题进行编码；或者把 T 看作0，C 看作1，G 看作2，A 看作3，于是一个核苷酸链就成了一个四进制数。电子计算机中的编码"0"和"1"可代表的信息，当然更能用四进制数来表示。

在计算过程中，可以利用酶对 DNA 序列进行操作，不同的酶相当于 DNA 串上的不同算子。如限制内切酶(restriction endonucleases)可作为分离算子，它识别链中特定的短序列，并在这个位上进行"切割"；连接酶(ligase)可作为连接算子，它把刚切过的 DNA 的粘端与其它链连接在一起；聚合酶(polymerases)可作为复制算子，进行 DNA 链的复制；外切核酸酶(exonuclease)可作为删除算子；还有转移酶(terminal transferase)、修饰酶(modifying enzymes)等，根据不同的过程可以采用不同的酶。在计算过程中需要用到的分子生物技术有：聚合酶链式反应 PCR(Polymerize chain reaction)、凝胶电泳(使 DNA 链按长度分离)、亲和层析(过滤析取)、加热/退火(合成 DNA 双链及其逆过程解链)、克隆、磁珠分离、标记、合并、连接、切割、检测等。

2.3 DNA 计算的基本特点

(1)*运算速度快*　目前最快的巨型机每秒能执行10^{12}次操作，DNA 计算由于其强大的潜在并行性，已达到每秒10^{20}个次操作。

(2)*超低能耗*　生化反应所需要的能量消耗很小，完成同样的运算 DNA 计算所消耗的能量是大型机的十亿分之一。

(3)*存储容量高*　DNA 存储信息的密度是1bit/立方纳米。而现在计算机存储一位信息需要10^{12}立方纳米。

3. 粘接计算模型

自 Adleman 的试验成功之后，研究者们又提出了多种 DNA 计算模型[10]：有的是对 Adleman 的方案进行改进，使之操作更简便、通用、计算速度更快、能容错等[2,5]；有的则不同于 Adleman 的模型如基于表面化学的 DNA 计算模型，或者基于 RNA 翻译和转录的生物化学反应的分子计算模型[3,7]。本文介绍一种粘接计算模型[4]，这个模型的设计方法不同于 Adleman 的设计模型[1]。

粘接模型是应用 DNA 链作为信息表示的物理基础，它的计算是基于 Watson-Crick 的补码变化规律。粘接模型具有随机存储功能，理论上认为 DNA 材料可以重复使用。

粘接模型中用于表示信息的 DNA 单链分为两种，分别称为存储链和粘接短链。假设存储链的长度为 n(即由 n 个碱基组成)，它包含 t 个互不重叠的子串，各个子串的碱基排列顺序要求不同，子串之间互相相连，中间不含有任何碱基。每个子串有 s 个碱基。每一个长度是 s 的粘接短链要求与存储链中的某个子串是互补的。如图2所示。

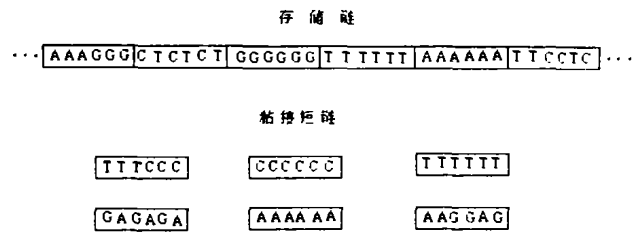

图2　粘接模型的存储链及粘接短链示意图

在计算过程中，粘接短链与存储链中的某些子串会发生退火反应，形成存储联合体。如图3所示。定义存储联合体中的每一个子串对应一个布尔值(0或1)。对于与粘接短链发生退火反应的子串，用一位二进制代码"1"表示，规定这时子串是处于打开状态；相反的，如果某一子串没有粘接短链与它发生退火反应，则用"0"表示，此时这一子串被称为是处于关闭状态。这样一个存储联合体就代表一个二进制串。图3 表示的存储联合体是 n=36，t=6，s=6。第①条存储联合体中的所有子串都处于关闭状态，第②、③、④条存储联合体中的子串有的处于打开状态，有的处于关闭状态。它们所表示的二进制串分别为：000000、001000、110000、011001。需要说明的是，在退火反应前，存储链中的子串既不是处于打开状态，也不是处于关闭状态。

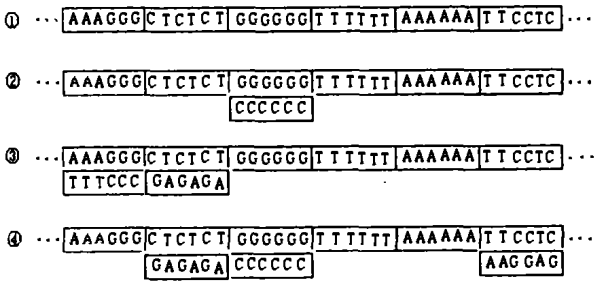

图3　存储联合体示意图

粘接模型是在多重集合上的操作，基本元素是存储联合体。粘接模型的基本操作有：合并(*merge*)、析取(*extracte*)、调整(*set*)、消除(*destroy*)、检测(detect)、读取(*read*)。

merge：将两个试管 N_1和 N_2的溶液倒入一个试管中以达到溶合，形成并集(可以理解为多重集合)。

extract：给出一个试管 N 和一个整数 i，$1\leqslant i\leqslant t$，生成两个新的试管$+(N,i)$和$-(N,i)$。试管$+(N,i)$并行地析取出在初始试管 N 中所有存储联合体的第 i 个子串处于打开状态的链。试管$-(N,i)$并行地析取出在初始试管 N 中所有存储联合体的第 i 个子串处于关闭状态的链。

set：给出一个试管 N 和一个整数 i，$1\leqslant i\leqslant t$，并行地将试管中每一个存储联合体的第 i 个子串打开(即发生退火反应)，生成新的试管 $set(N,i)$。如果这个子串本身已经处于打开状态，则不用发生变化。

destroy：给出一个试管 N 和一个整数 i，$1\leqslant i\leqslant t$，并行地将试管中每一个存储联合体的第 i 个子串置于关闭状态，生成新的试管 $destroy(N,i)$。这个操作就是移去第 i 个子串的粘接短链。

detect、*read*：检测结果试管中是否包含 DNA 链，如果有则返回"是"，并将 DNA 的序列读出；否则返回"否"。

粘接模型的初始化试管中的数据池是存储联合体的多重集合。表示为一个(t,l)数据池，其中$1\leqslant l\leqslant t$。存储联合体包含 t 个子串，前 l 个子串的打开或关闭状态是根据具体问题的输入来确定的，后 $t-l$ 个子串处于关闭状态。这样一个(t,l)数据池，可看为一个包含2^l个不同的存储联合体的多重集合，用二进制串 $w0^{t-l}$表示，其中 w 表示长度为 l 的二进制序列。在初始试管中，存储联合体的前 l 个子串表示实际的输入，后(t,l)个子串将用于中间结果的存储和输出。

粘接模型和 Adleman 的模型虽然都是按照碱基互补配对的原则进行运算的，但他们的设计方法是不同的。Adleman 的实验不是从长链开始的，而是采用短的单链利用粘接末端一步步退火的方式，逐步产生可能的解。粘接模型是从一个很

长的存储链与粘接短链退火，产生了具有特殊双链的存储联合体。Adleman 模型产生的是完整的 DNA 分子的双链，双链中间没有单链，而粘接模型的存储联合体是由双链和单链混合组成的。

4. 目前存在的问题

作为新生事物的 DNA 分子计算方法，引起了科学家们的广泛重视，在近几年的研究过程中，取得了许多令人振奋的成果。但仍有许多不足之处，主要表现在以下三个方面：

(1)目前提出的 DNA 计算模型，绝大多数是在理想的情况下进行计算的。因为现有的分子生物学技术，还不能保证实验操作的准确无误。而 DNA 计算对实验的精度要求相当严格，如果出现错误信号，这些信号会在后面的操作中被级联扩大，致使计算失败。所以如何提高实验精度，减少误差，是人们研究的重要课题之一。

(2)目前 DNA 计算能解决的运算种类并不多，由于还没研制出 DNA 串之间的消息传送机制，许多常规并行计算中的技术还不能用于 DNA 计算，现在 DNA 计算的应用领域主要局限在复杂的查找类问题上。

(3)随着问题复杂性的增加，计算所需的核苷酸及酶的分子数会呈指数级增加。实验所需的材料及材料的重复利用问题还需进一步研究。

DNA 计算具有划时代的意义，DNA 计算不再是一种物理性质的符号变换，而是一种通过切割和粘贴、插入和删除操作的化学性质的符号变换。目前对 DNA 计算的研究正处于起步阶段，相信随着科研人员的不懈努力和探索，DNA 计算会不断地发展和完善。

参 考 文 献

1 Adleman L M. Molecular computation of solutions to combinatorial problems[J]. Science,1994,226:1021～1024

2 Bach E,Condon A,et al. DNA model and algorithms for NP-complete problems[J]. Journal of Computer and System Sciences, 1998,57(2):172～186

3 Cukras A R,et al. Chess games:a model for RNA based computation[J]. BioSystems,1999,52(1):35～45

4 Roweis S,et al . A sticker based model for DNA computation [J]. Journal of Computational Biology,1998,5(4):615～629

5 Gifford D K. on the path to computation with DNA[J]. Science, 1994,266:993～994

6 赵致琢. 计算科学导论[M]. 北京:科学出版社,1998

7 Allison L,et al. Sequence complexity for biological sequence analysis[J]. Computer and Chemistry,2000,24(1):43～45

8 李人厚,余文. 关于 DNA 计算的基本原理与探讨[J]. 计算机学报,2001,24(9):972～978

9 陈惟昌,等. DNA 计算机的研究和展望[J]. 生物化学与生物物理进展,2001,28(2):156～159

10 高琳,等. DNA 计算的研究进展与展望[J]. 电子学报,2001,29(7):973～977

11 李振刚. 分子遗传学概论[M]. 合肥:中国科学技术大学出版社,1990

12 Lipton R J. DNA solution of hard computational problems [J]. Science,1995,268(5210):542～545

(上接第12页)

中相应词的权重会成倍增加，反之，成倍减少。一种更常用的反馈方式是采用 Rocchio 反馈模型[13]。更有效的用户兴趣文件可由以下公式迭代产生：

$$P_{k+1}=P_k+\beta\sum_{k=1}^{n_1}\frac{R_k}{n_1}-\gamma\sum_{k=2}^{n_2}\frac{S_k}{n_2}$$

其中，P_{k+1}是新的用户兴趣文件，P_k 是旧的用户兴趣文件，R_k 是用户反馈中认为感兴趣的(相关的)文档 k 的内容表示，S_k 是用户认为不感兴趣的(不相关)文档 k 的内容表示，n_1是相关文档数，n_2是不相关文档数，β,γ 值决定了正负反馈的相对作用。

结束语 随着信息资源的迅速增长，过滤技术正得到越来越广泛的应用。各式各样的过滤系统，推荐系统层出不穷，内容涉及 Web 页面，Usenet 新闻，音乐，电影甚至笑话。

本文简要介绍了基于内容和基于协作的两种过滤方式，列举了这两种方式各自的优势与不足，并针对用户兴趣文件的建立、维护等关键环节具体描述了两种过滤方式用到的不同实现技术。

参 考 文 献

1 Lawrence S,Giles C L. Searching the World Wide Web. Science, April. ,1998,280:98～100

2 Balabanovic M,Shoham Y. Fab:Content-Based,Collaborative Recommendation. Communications of the ACM,1997,40(3)

3 Krulwich B, Burkey C. Learning user information interests through extraction of semantically significant phrases. In:Proc. of the AAAI Spring Symposium on Machine Learning in Information Access Stanford,Calif. ,March1996

4 Lang K. Newsweeder: Learning to filter netnews. In:Proc. of the 12th Intl. Conf. on Machine Learning (Tahoe City,Calif.) 1995

5 Harman D. Overview of the 3rd Text Retrieval Conference (TREC-3). In:Proc. of the 3rd Text Retrieval Conf. Gaithersburg,Md,Nov. 1994

6 Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J. GroupLens: An open architecture for collaborative filtering of netnews. In:Proc. of the ACM Conf. on Computer-Supported Cooperative Work,Chapel Hill,NC,1994

7 Hill W,Stead L,Rosenstein M,Furnas G. Recommending and evaluating choices in a virtual community of use. In:Conf. on Human Factors in Computing Systems-CHI'95. Denver,May 1995

8 Shardanand U,Maes P. Social information filtering: Algorithms foe automating word of mouth. In:Conf. on Human Factors in Computing Systems-CHI'95. Denver,May 1995

9 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现

10 王继成，潘金贵，张福炎. Web 文本挖掘技术研究

11 Pazzani M J. A Framework for Collaborative,Content-Based and Demographic Filtering. Department of Information and Computer Science University of California,Irvine Irvine,CA 92697 pazzani @ics. uci. edu

12 刘绍翰，武港山，张福炎. 相关反馈技术在互联网文本信息检索中的应用

13 Rocchio J J. Relevance Feedback in information Retrieval. In:G. Salton,ed. SMART Retrieval System,Prentice Hall,1971. 313～323