DNA 计算机的研究和展望*

陈惟昌[1)**] 陈志华[2)] 邱红霞[1)] 王自强[1)]

([1)]中日友好临床医学研究所生物物理研究室，北京 100029;
[2)]中日友好临床医学研究所生物化学及分子生物学研究室，北京 100029)

摘要 DNA 计算机是计算机科学和分子生物学互相结合、互相渗透而产生的新兴交叉研究领域. 目前已取得较大进展. DNA 计算机是以编码的 DNA 序列为运算对象，通过分子生物学的运算操作以解决复杂的数学难题. DNA 计算机的重要特点是信息容量的巨量性和密集性，和处理操作的高度并行性，通过强力搜索策略迅速得出正确的答案，从而使其运算速度大大超过常规计算机的计算速度. 介绍了 DNA 计算机的近期进展和工作原理及其分子生物学的运算操作过程. 并对 DNA 计算机的未来发展前景及在生物信息学中的意义，进行了分析和讨论.

关键词 DNA 计算机，NP 完全问题，并行处理，强力搜索策略，互补 DNA 链

学科分类号 Q617

计算机科学家将数学问题分为三大类：a. 容易计算问题 (easy computational problems)，b. 难解计算问题 (hard computational problems)，c. 不可计算问题 (incomputable problems). 多项式函数问题 (polynomial function problems，P 问题)，其计算所需时间随着多项式变量数目的增加呈比例增加 (线性关系)，属于易解问题. 例如，证明数 d_1 和 d_2 是否是数 N 的因子，即是易解问题. 非多项式问题 (non-polynomial problems) 又称 NP 问题，则属于难解问题. NP 问题的计算时间，则随着变量数目的增加呈指数增加. 推销员的哈密尔顿路问题 (Hamiltonian path problem，HPP 问题)，即属于此类难解问题. 所谓 HPP 问题是指推销员在一个具有 n 个城市和 m 条城市间航线的地图 (有向图) 中，从某一城市出发，寻找一条通路，到达另一城市，而且经过其他所有城市仅仅一次 (图 1). HPP 问题属于搜索问题 (search problems). 搜索问题是 NP 问题的一种，属于 NP 完全问题 (non-polynomial complete problems). 其他如寻找图论中的最大连结顶点集问题 (maximal clique problem)，以及布尔代数式 (Boolian formula) 中变量取值的“满足”问题 (satisfaction problem，SAT 问题) 等都属于 NP 完全问题. 计算机科学已经证明，对所有 NP 完全问题，并不存在统一的通用的有效求解算法. 至于不可计算问题是指那些不能通过常规计算机计算求解的数学问题，例如要证明哥德巴赫猜想，即是不可计算问题.

1994 年，Adleman 用 DNA 计算方法以多项式的计算时间解决了推销员的 HPP 问题[1]. 他的工作开辟了生物分子计算机的新纪元，使计算机学界深受鼓舞. 其后 Lipton[2] 提出用 DNA 计算方法求解布尔代数式的 SAT 问题和 Ouyang 等[3] 用 DNA 计算方法求解图论中的最大连结顶点集问题. 应用芯片固相固定技术进行 DNA 计算的方法，亦有报道[4~6]. Cox 等[7]，Karli 等[8] 以及 Ogihara 等[9] 对 DNA 计算机的特点及存在问题，均进行了系统的评述. 一个研究 DNA 计算机的高潮，正在到来.

1 DNA 计算机的原理

DNA 计算机的主要原理是大规模的并行运算操作. 据 Adleman 估计，生物计算机一步可完成 10^{20} 次运算，使其运算速度大大超过电子计算机的运算速度. Adleman 还计算出，生物计算机每消耗 1 焦耳的能量，可以完成 10^{19} 次运算，其能量损耗及能量效率亦远比电子计算机为优. Adleman 在计算具有 7 个城市和 13 条航线的哈密尔顿路问题的具体方法是 (图 1)，第一步，先用长度为 20 个核苷酸不同的 DNA 序列，编码 7 个顶点 (城市) 设为 O_i (i = 0、1、2、…6). 例如，O_2 的编码为 TATCGGATCGGTATATCCGA，O_3 的编码为 GCTATTCGAGCTTAAAGCTA，O_4 的编码为 GGCTAGGTACCAGCATGCTT. 然后对 13 条 $i \rightarrow j$ 边进行编码 $O_{i\rightarrow j}$，$O_{i\rightarrow j}$ 的前 10 个核苷酸是 O_i 的

* 国家自然科学基金资助项目 (39770210).
** 通讯联系人.
Tel: 010-64221122-4434，E-mail: chenwch@mail.east.net.cn
收稿日期：2000-04-06，接受日期：2000-06-07

3′端的 10 个核苷酸，$O_{i\to j}$ 的后 10 个核苷酸是 O_j 的 5′端的 10 个核苷酸，例如 $O_{2\to 3}$ 边的编码为：GTATATCCGAGCTATTCGAG. 同理可编码 $O_{3\to 4}$. 注意在编码时应使 $O_{3\to 2}$ 与 $O_{2\to 3}$ 的编码不同. 然后以 O_3 的互补 DNA 序列 $\overline{O_3}$，即 CGATAAGCTCGAATTTCGAT 将 $O_{2\to 3}$ 边和 $O_{3\to 4}$ 边联结起来. 在分子生物学反应中，用 50 pmol 的 $\overline{O_i}$ 和 50 pmol 的 $O_{i\to j}$ 进行混合，以 $\overline{O_i}$ 作为夹板，使对应的边进行连结反应，由此产生大量随机的各种联结通路. 第二步，应用 O_0 及 $\overline{O_6}$ 作为引物，对上述随机产生的 DNA 序列进行扩增，只有那些从顶点 0 点起始而终于顶点 6 点的通路 DNA 序列能被扩增，其余的通路因不满足条件而不被扩增. 第三步，将 PCR 扩增产物进行琼脂糖电泳，选取分子质量为 140 bp 的双链 DNA (dsDNA)，纯化后再用 PCR 扩增，所得的 DNA 序列即是代表通过 7 个城市的通路的 DNA 序列. 第四步，将 dsDNA 加热变性成单链 DNA (ssDNA)，再用含有 $\overline{O_1}$ 的磁珠进行结合，选取那些通过 1 点城市的通路，然后再热变性，相继重复用 $\overline{O_2}$、$\overline{O_3}$、$\overline{O_4}$、$\overline{O_5}$ 的磁珠进行分离，提取同时通过 1、2、3、4、5 各点城市的通路，此即满足从 0 点起始，经过 1、2、3、4、5 点到达 6 点的哈密尔顿路. 第五步，将第四步的答案用 PCR 扩增后，并进行电泳分离. 所得结果，代表所求的正确答案. (0→1→2→3→4→5→6). Adleman 的计算表明，DNA 计算机计算 HPP 时的操作运算步骤，与城市节点的数目呈线性关系，但用电子计算机计算时，其运算步数与节点的数目呈指数增加.

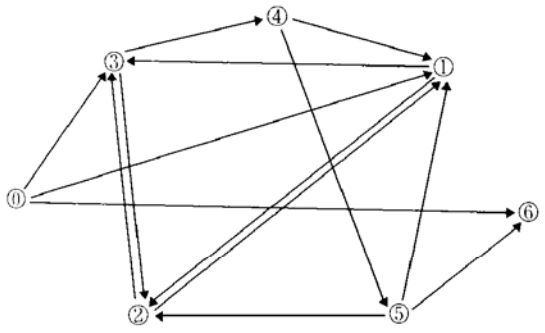

Fig. 1　Hamiltonian path problem[1]

图 1　哈密尔顿路图[1]

Liu 等[10]用芯片法，在固相表面进行 SAT 问题的 DNA 计算，取得令人瞩目的结果. Liu 等解决的 SAT 问题是：

$$F = (w \vee x \vee y) \wedge (w \vee \overline{y} \vee z) \wedge (\overline{x} \vee y) \wedge (\overline{w} \vee \overline{y})$$

式中，w，x，y，z 是布尔变量，只能取值为 0 (假) 或 1 (真). $\vee$ 为逻辑或 (Logical OR) 运算，$x \vee y = 0$ 仅当 $x = y = 0$. $\wedge$ 为逻辑与 (Logical AND) 运算，$x \wedge y = 1$ 仅当 $x = y = 1$. $\overline{x}$ 是 x 的逆运算. $x = 0$ 时 $\overline{x} = 1$，$x = 1$ 时 $\overline{x} = 0$. 公式 F 共有 4 个子句，子句内部的变量以 $\vee$ 连接，例如 $(\overline{x} \vee y)$，只要 $x = 0$ 或 $y = 1$ 或二者成立，则该子句为真. 子句之间以 $\wedge$ 连结，这表明若要 F 为真，则 4 个子句之值都必需同时为真. 由于每个变量可有 0 或 1 两种取值，则 n 个变量共有 2^n 个可能取值的组合. $n = 4$ 时，共有 $2^4 = 16$ 种变量取值组合，以 $wxyz$ 4 个变量而言，可有 0000，0001，0010，…1111 等 16 种取值组合. 现在要求解的是，这 16 种组合中，那些能满足 $F = 1$. Liu 等采用的第一步称为建立阶段 (make) (图 2). 即建立用 8 个编码寡核苷酸的 DNA 序列，对这 16 种取值进行编码，如 0000 为 S_0= CAACCCAA，0001 为 S_1= TCTCAGAG，…1111 为 S_{15}= ACTGGTCA 等. 然后，合成一系列的用于表面固定的寡核苷酸序列成为 DNA 单词，每个单词的编码如下：

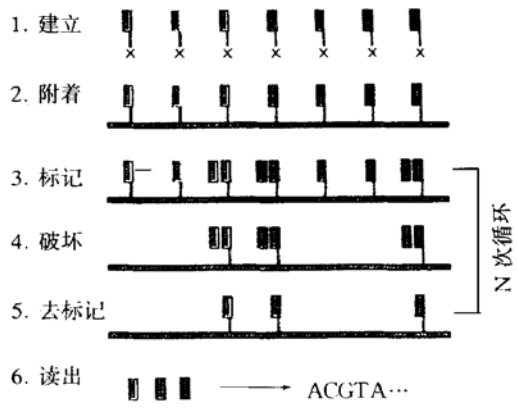

Fig. 2　DNA chip computation[10]

图 2　DNA 芯片计算过程[10]

5′ −HS − C_6 − T_{15} GCTTvvvvvvvvTTCG-3′ (S_i，i = 0，1，…15)，S_i 称为沃森带 (Watson strands). SH 是经硫氢基修饰的功能基团，T_{15} 是由 15 个胸腺嘧啶组成的重复序列，称为分隔序列，将功能编码的寡核苷酸与芯片表面隔离. GCTT 以及 TTCG 是单词的标签 (label). vvvvvvvv 代表 8 个编码寡核苷酸的 DNA 序列 (S_i)，以代表不同的变量取值的组合等. 第二步为附着阶段 (Attach)，即将上述合成的各种寡核苷酸序列 S_i，随机固定于经过马来酰胺活化的镀金玻片表面. 随机固定的优点是在单位芯片面积上可附着更多的 DNA 序

列. 第三步为标记阶段 (Mark), 即合成与 S_i 互补的 DNA 序列 $C_i = \overline{S_i}$, C_i 称为克里克带 (Crick strands). 在第一循环中, 加入满足 $w = 1$, $x = 1$, $y = 1$ 即满足第一子句为真的互补核苷酸序列, 例如, $w = 1$ (C_8, C_9, C_{10}, C_{11}, C_{12}, C_{13}, C_{14}, C_{15}); $x = 1$ (C_4, C_5, C_6, C_7, C_{12}, C_{13}, C_{14}, C_{15}); $y = 1$ (C_2, C_3, C_6, C_7, C_{10}, C_{11}, C_{14}, C_{15}), 即共加入 14 种互补序列 C_i, (C_2 …C_{15}) 与 S_2 …S_{15}14 个 DNA 序列形成 dsDNA, 而 S_0, S_1 则仍为 ssDNA. 第四步为破坏阶段 (Destroy), 即用大肠杆菌外切核酸酶 (*E. coli* exonuclease) 破坏芯片表面的 ssDNA, (即 S_0 及 S_1), 而保留其余 dsDNA. 第五步为去标记阶段 (Unmark), 即将芯片加热, 使余下的 dsDNA (即 S_2, …S_{15}) 解离为 ssDNA, 然后重复标记、破坏和去标记步骤. 但加入的互补 DNA 序列 C_i 是满足公式 F 中第二个子句为真的 DNA 序列, 经过第二次循环, 清除了 S_2 和 S_6, 再经过第三次循环, 清除了 S_4, S_5, S_{12} 和 S_{13}. 经过第四次循环 (即满足第四个子句为真), 清除了 S_{10}, S_{11}, S_{14} 和 S_{15}. 最后剩下的 S_3 (0011); S_7 (0111); S_8 (1000), S_9 (1001) 即为所求之答案. 第六步为读出阶段 (Readout), 即将满足 4 个子句均为真的变量取值读出, 例如, S_3 为: $w = 0$, $x = 0$, $y = 1$, $z = 1$. $y = 1$ 满足第一子句 ($w \vee x \vee y$) 为真, $z = 1$ 满足第二子句 ($w \vee \overline{y} \vee z$) 为真, $x = 0$ 或 $y = 1$ 使第三子句 ($\overline{x} \vee y$) 为真, $w = 0$ 使第四子句 ($\overline{w} \vee \overline{y}$) 为真, 故 F 为真. 同理可证 S_7, S_8, S_9 均满足公式 $F = 1$. 在芯片表面留下 S_3C_3, S_7C_7, S_8C_8, S_9C_9 等 4 种 dsDNA, 通过加热, 使 C_3, C_7, C_8, C_9 从芯片表面解离, 并用 PCR 方法加以扩增并进行荧光标记. 然后将这些 PCR 产物与另一位置排列整齐的 S_0, S_1, …S_{15}芯片杂交, 通过测定各点的荧光强度, 可以检出 S_3, S_7, S_8, $S_9$4 点的荧光强度最强, 从而读出所求的解答.

由此可见, DNA 计算机进行计算可分为三个主要过程, 第一是编码合成过程, 即对数学问题所有可能的解答, 设计出用不同的 DNA 序列进行编码, 并合成大量的各种 DNA 编码序列. 第二是清除运算过程, 通过 DNA 链的互补结合以及核酸酶的破坏作用, 大量清除那些代表不正确答案的 DNA 序列, 而保留代表正确答案的 DNA 序列. 第三是结果读出过程, 即通过 PCR 方法将正确答案扩增并读出结果.

2 DNA 计算机的评价

常规电子计算机的出现已有半个多世纪, 目前其高速发展的势头, 仍有增无减. DNA 计算机的出现只有短短的 5 年, 其数学计算能力自不能和电子计算机相比. 但由于 DNA 计算机所特有的大规模并行运算的优越性, 使其在解决 NP 问题上引起广泛的重视. DNA 计算机是一项新生事物, 目前还远不成熟, 在理论上和实验上还有许多工作仍待深入研究. 国内尚无这方面的报道, 应急起直追.

3 DNA 计算机的未来展望

DNA 计算机是真正意义的分子计算机, 在理论和使用上都具有很大的发展前途:

a. 关于 DNA 序列的数字编码方面: 二进制数字化编码是信息科学最基本的编码方式. 而目前在 DNA 计算机中使用的代码仍是 4 种碱基的字符编码, 还没有使用真正意义的数字编码. 用 0 (00)、1 (01)、2 (10)、3 (11) 4 个数字对 4 种碱基 (C、T、A、G) 进行二进制编码, 共有 24 种可能的编码组合, 其中 8 种满足碱基互补法则, 它们是拓扑等价的. 我们认为按 4 种碱基分子质量的大小顺序排列的编码格式: 0123/CTAG 是最理想的编码方式[11]. 这一编码格式还可以反映出 4 种碱基的化学性质, 如嘌呤与嘧啶, 氨基与酮基, 强氢键结合与弱氢键结合, 碱基的互补关系等 (图 3). DNA 序列的数字编码, 可以方便地进行各种数学运算, 如用按位加运算可以比较两个 DNA 序列的不同, 求逆运算可求出互补的 DNA 链等. 通过计算 DNA 序列的汉明距离[12], 可以了解在高维空间

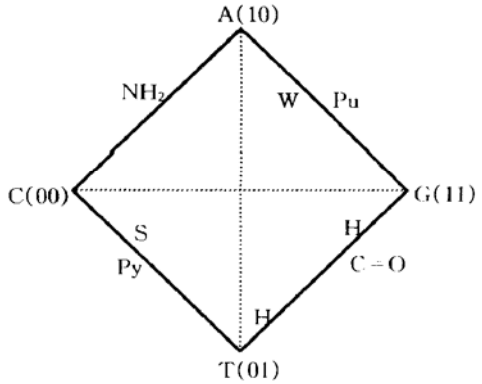

Fig. 3 Binary coding of the nucleotides

图 3 核苷酸碱基的二进制数字编码

Pu: 嘌呤; Py: 嘧啶; NH_2: 氨基; C═O: 酮基; SH: 强氢键; WH: 弱氢键.

中不同 DNA 编码序列的相异程度，从而对简化和规范 DNA 计算机中的主链 (沃森序列) 和补链 (克里克序列) 的设计与运算，有重要意义.

b. 在生物信息学的理论研究方面：大量的事实表明，大规模并行运算是生物体内普遍存在的基本法则. 例如，大脑神经元网络能对大量的外界输入信息，同时进行大规模的并行运算和加工，从而作出正确的判断[12]，细胞的信号转导系统，能对作用于细胞膜表面受体的多种化学因子 (配基) 的作用进行并行处理，从而调控细胞的整体功能活动. DNA 计算机中多种不同的 DNA 序列及与其有关的生化反应 (如碱基互补结合，酶分解，DNA 链连结延伸，基因扩增等) 在生物体内都是存在的. 因此对生物学中并行运算操作规律的研究，对理解基因组网络的运行规律以及生物复杂系统的调控机理，有重要意义.

c. 目前生物芯片的技术已有很大进展. DNA 计算机和 DNA 芯片技术相结合，有可能应用组合化学 (combinatorial chemistry) 的方法，自动设计与合成大量的编码的 DNA 序列，并使 DNA 的计算操作进一步自动化. 这将使 DNA 计算机向实用化方向大大前进一步. 未来的 DNA 计算机并不仅仅是为了解决数学上的难解问题，而更主要的是，它将以真正的生物大分子计算机芯片的方式，植入生物体内，共同对生命的基本过程进行调节和控制，用以治疗各种机能障碍性疾病.

参 考 文 献

1 Adleman L M. Molecular computation of solutions to combinatorial problems. Science, 1994, **266** (5187): 1021~ 1024

2 Lipton R J. DNA solutions of hard computational problems. Science, 1995, **268** (5210): 542~ 545

3 Ouyang Q, Kaplan P D, Liu S M, *et al*. DNA solution of the maximal clique problem. Science, 1997, **278** (5337): 446~ 448

4 Smith L M, Corn R M, Condon A E, *et al*. A surface-based approach to DNA computation. J Comput Biol, 1998, **5** (2): 255~ 267

5 Frutos A G, Liu Q H, Thiel A J, *et al*. Demonstration of a word design strategy for DNA computing on surface. Nucl Acid Res, 1997, **25** (23): 4748~ 4757

6 Frutos A G, Smith L M, Corn R M. Enzymatic ligation reactions of DNA "words" on surface for DNA computing. J Am Chem Soc, 1998, **120** (40): 10277~ 10282

7 Cox J C, Cohen D S, Ellington A D. The complexities of DNA computation. Trends Biotechnol, 1999, **17** (4): 151~ 154

8 Karli L, Landweber L F. Computing with DNA. Methods-Mol Biol, 2000, **132** (4): 413~ 430

9 Ogihara M, Ray A. DNA computing on a chip. Nature, 2000, **403** (6766): 143~ 144

10 Liu Q H, Wang L M, Frutos A G, *et al*. DNA computing on surface. Nature, 2000, **403** (6766): 175~ 179

11 陈惟昌，陈志华，陈志义，等. 遗传密码的简并及其高维空间的拓扑结构. 自然科学进展，1999，**9** (2)：175~ 178
Chen W C, Chen Z H, Chen Z Y, *et al*. Progress in Natural Science, 1999, **9** (2): 175~ 178

12 Chen W C, Chen Z Y, Wang Z Q, *et al*. Topological structure of the high dimension space and information coding of the biological neural network. Proceedings of 1998 International Conference on Neural Networks and Brain. (Plenary talk on ICNN& B, 1998), Beijing, 1998, PL 19~ 24

Progress in DNA Computer*

CHEN Wei-Chang[1)**], CHEN Zhi-Hua[2)], QIU Hong-Xia[1)], WANG Zi-Qiang[1)]

([1)] *Department of Biophysics, China Japan Friendship Institute of Medical Sciences, Beijing* 100029, China;

[2)] *Department of Biochemistry and Molecular Biology, China Japan Friendship Institute of Medical Sciences, Beijing* 100029, China)

Abstract DNA computer is a new research field which combines both the computer science and molecular biology. DNA computer is proposed to solve a class of hard problems of mathematical complexity by using a set of DNA sequences encoding all candidate solutions to the computational problem of interest and find out the correct answers by serial manipulations of biochemical reactions. DNA computer is exactly a biomolecular computer which stores a vast quantity of information with high density. DNA computer, by means of its huge parallel computation and brute force search strategy, can solve the NP complete problems with polynomial time. The recent advances and principle of DNA computer are introduced. The future development and the bioinformatical significance of DNA computer are also analyzed and discussed.

Key words DNA computer, NP complete problem, parallel computation, brute force search strategy, complementary DNA strands

* This work was supported by a grant from National Natural Sciences Foundation of China (39770210).

** Corresponding author. Tel: 86-10-64221122-4434, E-mail: chenwch@mail.east.net.cn

Received: April 6, 2000 Accepted: June 7, 2000