基于 DNA 自动机的串行二进制进位加法的实现*)

李汪根[1,3]　丁永生[1,2]

(东华大学信息科学与技术学院　上海 200051)[1]　(数字化纺织服装技术教育部工程研究中心　上海 200051)[2]

(安徽师范大学数学计算机学院　芜湖 241000)[3]

摘　要　提出了一种基于 DNA 自动机的串行二进制进位加法的实现方法。对于一位二进制的进位加法，通过预先设计的 DNA 自动机模型在一个试管中以自动机的方式完成。对于 n 位二进制的进位加法，通过将 n 个类似的试管按照从低位到高位的顺序组成串行网络；将低位加法操作产生的进位转移到高位试管，组成高位自动机的输入符号串，完成高位的加法操作。这种运算方式类似于电子计算机中加法运算系统，为 DNA 计算机实现算术运算提供了一种新颖的方法。

关键词　DNA 自动机，串行，进位加法，DNA 编码

Implementation of Serial Binary Carry-Save Adder Based on DNA Automaton

LI Wang-Gen[1,3]　DING Yong-Sheng[1,2]

(College of Information Sciences and Technology, Donghua University, Shanghai 200051))[1]

(Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education Donghua University, Shanghai 200051)[2]

(College of Mathematics and Computer Sciences, Anhui Normal University, Wuhu 241000)[3]

Abstract　The implementation of a kind of serial binary carry-save adders based on DNA automaton is proposed. For one bit binary, the addition will be automatically completed in one test tube according to DNA automaton designed in advanced. For n bits binary, it will be automatically completed according to the following strategy: constructing a serial network of n test tubes from the lower position$(m-1)$ to the higher position(m), transferring the carry-save bit produced when the addition at the lower position is completed from the$(m-1)$-*th* test tube to the m-th test tube, and forming input string of the DNA automaton, in which the addition at the higher position will be completed according to the DNA automaton. This process has an analogy with the addition system implemented in electronic computer. It provides a novel method for performing arithmetic operations in DNA computer.

Keywords　DNA automaton, Serial, Carry-save addition, DNA encoding

1　引言

使 DNA 计算机具备常规的算术运算能力是 DNA 计算机研究的主要内容之一。加法运算是算术运算中最基本的运算，是其它算术运算的基础。1996 年，Guarnieri 等人在“Science”上发表文章[1]，首次给出了 DNA 计算机运算系统中正有理数的进位加法运算。他们首先给出了加数和被加数的 DNA 编码方法，然后在此基础上完成了进位信息的传递以及二进制加法的 DNA 分子生物运算过程的构造和控制，建立了 DNA 计算机系统中的加法运算模型。后来，Guarnieri 等人又将以上结果扩展到一般的 3 进制乃至 k 进制的情况[2]。1999 年，Yurke 等人提出了一种新的 DNA 计算机系统中的加法运算模型[3]，这种模型与 Guarnieri 等的工作相比，其特点是对输入串与运算串的 DNA 分子编码是分离的。

2001 年，Shapiro 等人提出了一种可编程、可自治 DNA 有限状态自动机模型[4]，它的所有部件，包括硬件、软件、输入、输出都是生物分子。硬件由限制性内切酶和连接酶组成，而软件(转移规则)和输入分子都是双链 DNA 分子。通过对输入分子实施一系列的酶切、绑接和连接循环等操作，以 DNA 分子的形式产生代表最终状态的输出分子。

以前的 DNA 计算机系统中的加法运算模型其操作数的 DNA 编码规则复杂，实现加法运算的生物操作体系复杂。本文提出一种基于 DNA 自动机的串行二进制进位加法的实现方法。模型中操作数的编码非常简单，只要分别对二进制数 0 和 1 进行编码，就可以完成操作数的编码；实现运算的生物操作简单，多位二进制数的加法仅仅是一位二进制数加法的循环。模型的运算方式类似于电子计算机中加法运算系统，很容易实现一位或 n 位二进制的进位加法运算。

2　具有栈式存储结构的 DNA 自动机

自动机是一种描述和执行算法的工具，本节首先介绍具有栈式存储结构 DNA 自动机的概念，然后详细描述其上运用到的生物操作。

2.1　具有栈式存储结构的 DNA 自动机

一个典型的自动机包含一个有穷控制器、一条输入带和一个读头。根据读头当前在输入带中所处的位置以及自动机的当前状态，按照一定的规则，自动机会自动完成一次操作，这个操作包括改变自动机的状态以及移动读头。

*)国家自然科学基金(60474037，60004006)、教育部新世纪优秀人才支持计划、教育部高等学校博士点专项基金(20030255009)。**李汪根**　博士研究生；**丁永生**　博士，教授，博士生导师。

一个具有栈式存储结构的自动机就是在自动机的基础上增加了栈的操作，又称为下推自动机。它由3部分组成：有限状态控制器、输入符号串和栈。同有限状态自动机一样，下推自动机也是从初始状态开始。在有限状态控制器的控制下，根据下推自动机的当前状态、读头读入输入符号串的符号和栈顶符号选择相应的动作。根据动作规则，下推自动机执行一次动作将会完成以下操作：1)改变下推自动机的状态；2)通过入栈或出栈操作修改栈顶元素；3)决定读头是否移动。如果输入符号串处理结束时，下推自动机处于可接受状态集 F 的某一个状态，且栈的内容为空，则表示自动机接受该字串；否则自动机不接受该字串。

具有栈式存储结构的DNA自动机[5]，其代表程序的输入字符串、表示动作规则的转移分子、自动机的状态以及栈都用DNA分子表示。限制性内切酶和连接酶来完成对DNA分子的生物操作，它们共同组成自动机的硬件。通过对表示输入串的DNA片段实施一系列的酶切、绑接和连接等操作以DNA片段的形式产生代表自动机最终状态的输出分子。同时通过对表示栈的DNA片段实施一系列的酶切、绑接和连接等生物操作实现入栈或出栈操作，以修改栈顶元素。

2.2 DNA自动机的生物操作

下面我们介绍在DNA自动机中完成自动机的运行所需要的一些生物操作及其实现方法，这些生物操作包含了对DNA链进行处理的一系列生化反应[6,7]。在下面的描述中，如果 x 表示一条 DNA 单链，则 $\bar{x}$ 表示 x 的互补链。

(1)合成。在一定长度范围内，合成可以将A、T、C、G 4个碱基按照预定的顺序排列在DNA单链上。对于超过目前技术许可长度的更长DNA单链，在不需要知道碱基排列顺序的前提下，可以从其他质粒上直接提取。为了描述的方便，用操作 $Syn(a)$ 表示合成DNA单链 a。

(2)退火。实验室合成的具有互补碱基对的两条DNA单链，经温浴后自然冷却，其互补碱基对会绑接而形成有双链结构的DNA。用操作 $Ann(a,b,c)$ 表示具有互补碱基对的单链 a,b 经退火绑接合成双链结构的 c 片段。其中满足 $a=x\bar{b}y$ 或 $b=x\bar{a}y$。

(3)混合。将含有DNA片段的两个试管的缓冲液倒入另一个试管中，形成包含两个试管所有DNA片段的混合液。用操作 $Mel(T1,T2,T)$ 表示将试管 $T1$ 和 $T2$ 的溶液倒入试管 T 中。

(4)复制/放大。通过聚合酶链式反应(PCR)可以在实验室对一特定DNA片段进行快速、大量复制。用操作 $Pcr(a)$ 表示对模板链为 a、经PCR反应后得到其足量的DNA片段。

(5)分离。DNA片段在凝胶电泳缓冲液中带负电，在电场的作用下会产生迁移，其迁移的速率与片段的碱基个数相关：碱基个数大的迁移速率小，反之则大。经凝胶电泳后，溶液中的DNA片段会按照碱基个数的大小排列在凝胶带上。用操作 $Sep(a)$ 表示分离出DNA片段 a。

(6)抽提。通过亲和力，将以包含给定序列的DNA片段作为子串的DNA链提取出来。用操作 $Abs(a)$ 表示抽提出含有片段 a 的DNA链。

(7)切割。对含有某种限制酶识别位点的DNA片段在这种酶的作用下，在切割位点处被切割成两个DNA片段。用操作 $Cut(a,m,a_r,a_l)$ 表示DNA片段 a 在酶 m 的作用下切割形成两个片段 a_r,a_l。

(8)连接。由互补序列经退火形成的DNA双链中，在相邻的碱基位置处会形成缺口。这些缺口在连接酶的作用下，会连接在一起而形成完整的DNA片段。用操作 $Lig(m,a)$ 表示在连接酶 m 作用下修复片段 a 的缺口而形成完整片段。

(9)检测/读取。若给定的试管中包含指定的DNA片段，则解释为“是”，反之则为“否”。用操作 $Det(T,a)$ 表示试管 T 中是否含有特定片段 a。

(10)杂交。具有互补粘性末端的两个DNA片段经退火后，互补末端会绑接在一起形成双连结构。用操作 $Hyb(a,b,c)$ 表示具有互补末端的片段 a,b 杂交形成片段 c；或者用 $Hyb(a,b,c,d)$ 表示具有双粘性末端的片段 b 与另两个分别具有与其互补末端的片段 a,c 杂交形成片段 d。

3 基于DNA自动机的二进制串行加法

这里，首先描述在DNA计算机系统中对于加数和被加数都是一位二进制数的进位加法的实现方法，这个运算过程是基于DNA自动机的。然后在此基础上给出操作数是 n 位二进制数的串行进位加法的实现方法。

3.1 基于DNA自动机的一位二进制进位加法

对于两个操作数都是一位二进制数的进位加法，其加法的运算法则可用表1表示。

为了实现基于自动机的二进制进位加法，下面我们描述具有栈式存储结构的DNA自动机 M_{adder} 的构造方法。

表1 一位二进制进位加法运算法则

加数	被加数	低位的进位	结果	
			和	进位
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

用 S_0 表示 M_{adder} 的初始状态，两种可能的终止状态 S_0、S_1 分别对应加法运算结果的和为0和1的情况。两种可能的输入符号为0和1，分别对应于加数、被加数和进位是0和1的情况。栈始终含有一个元素，用来记录加法运算所产生的进位，其可能符号为0或1，分别对应进位数是0和1的情况。栈的初始符号是0。对于一位二进制的加法，一旦运算产生进位，也就是栈顶元素变为1后，栈顶将不会变化。我们用函数 $Tran(S,X,y)=(T,A)$ 表示当前状态是 S、栈顶符号是 X、当前输入符号是 y 时，在控制器的控制下，自动机的状态转变为 T，栈顶符号变为 A。M_{adder} 所有可能的状态转移函数的集合是：

$set_M=\{\ Tran(S_0,0,0)=(S_0,0);\ Tran(S_0,1,0)=(S_0,1);\ Tran(S_0,0,1)=(S_1,0);\ Tran(S_0,1,1)=(S_1,1);\ Tran(S_1,0,0)=(S_1,0);\ Tran(S_1,0,1)=(S_0,1)\}$。

下面我们描述自动机 M_{adder} 的各个组成部件的DNA分子编码方式。

M_{adder} 的输入符号串由两部分构成：一部分是代表加数和被加数所组成的DNA片段。给定进位加法的两个操作数后，这部分是已知的，可以按照符号编码规则预先合成；另一部分是低位上进位加法所产生的代表进位的DNA片段，这部分片段可以从低位进位加法的运行结果中分离出来。

M_{adder}的栈初始化时仅含有一个元素，就是元素 0，可以按照栈元素的符号编码规则预先合成代表元素 0 的 DNA 片段。将这些可以预先合成的 DNA 片段按照自动机的规则合成在一起，栈的元素在合成的 DNA 片段的左边，输入符号串在合成的 DNA 片段的右边，中间是背靠背分别用来实现对栈和输入符号串进行生物操作的 *FokI* 的识别位点。在合成的 DNA 片段的右端包含一个特定的 4bp 粘性末端，用于连接从第 $n-1$ 位加法产生的进位片段。

M_{adder}的转移函数是预先按照 DNA 编码规则设计好的 DNA 片段，这些片段均含有粘性末端。在自动机执行一次操作后，有且仅有唯一的转移函数，其粘性末端正好与经 *FokI* 酶切后产生的片段的末端是互补的。为了便于检测，M_{adder}的最后运行结果由检测分子来完成。检测分子包含两部分：一部分是代表自动机的最终状态的片段，这个片段与二进制进位加法所产生的和对应；另一部分是代表栈顶元素的片段，这部分片段与二进制进位加法所产生的进位对应。为了产生这些片段，需要另一种内切酶来剪切，这里使用 *BsmFI* 来实现。为了分离代表二进制进位加法所产生的和与进位的 DNA 片段，这里采用内切酶 *SmiI* 的酶切操作来实现。

将 M_{adder}的所有组件放在试管 T_a 中，便可以自动完成高位的进位加法操作。这些组件包括：预先合成的含有 4bp 粘性末端的片段 a，它表示了加数和被加数以及栈顶元素为 0 的栈；从低位进位加法所产生的代表进位的片段 b，它含有与 a 互补的粘性末端；所有设计好的含有 4bp 粘性末端的转移分子；*FokI* 酶、*T*4 连接酶和 *Taq* 热聚合酶等生化酶和反应缓冲液，它们是 M_{adder}的计算工具。M_{adder}运行结束后需要添加 *BsmFI* 以及 *SmiI* 两种酶来分离运行结果。以上所有生物操作过程如表 2 所示。DNA 自动机实现的一位二进制加法的试管体系如图 1 所示。

表 2　试管 T_a 中 M_{adder}完成的生物操作流程

步骤	操作	注释
Step1	$Hyb(a,b,c)$	低位进位加法所产生的进位片段 b 与第 n 位的加数、被加数以及栈片段 a 退火连接生成代表输入串和栈的片段 c
Step2	$Lig(T4,c)$	对片段 c 进行连接操作
Step3	$Cut(c,FokI,c_l,c_{r1})$	T_a 中片段 c 中具有正向识别位点的部分在 *FokI* 的作用下切割成 c_l 和 c_{r1} 两个片段
	$Cut(c,FokI,c_{l2},c_l)$	T_a 中片段 c 中具有反向识别位点的部分在 *FokI* 的作用下切割成 c_{l2} 和 c_l 两个片段
Step4	$Hyb(c_{l2},t_i,c_{r1},id_i)$	T_a 中经酶切形成的片段 c_{l2} 和 c_{r1} 分别与合适的转移片段 t_i 杂交形成片段 id_i
Step5	$Lig(T4,id_i)$	杂交后的片段 id_i 在 T_4 连接酶的作用下进行连接操作
		在 *FokI* 的作用下，自动反复执行步骤 step3～step5，直到没有合适的转移分子可供选择
Step6	$Lig(T4,S_i)$	检测分子与最后的酶切片段连接后生成运行结果的报告分子 S_i
Step7	$Cut(S_i,BsmFI,S_{il},S_{ir1})$	报告分子 S_i 经 *BsmFI* 酶切后形成 S_{il}，S_{ir1} 两个片段
	$Cut(S_i,BsmFI,S_{il},S_{ir2})$	报告分子 S_i 经 *BsmFI* 酶切后形成 S_{il}，S_{ir2} 两个片段
Step8	$Cut(S_{il},SmiI,stack,sum)$	片段 S_{il} 在 *SmiI* 酶切下形成代表进位的片段 *stack* 和代表和的片段 *sum*
Step9	$Abs(stack)$	从 T_a 中分离出代表第 n 位加法进位的 *DNA* 片段 *stack*，放入试管 T_c
Step10	$Abs(sum)$	从 T_a 中分离出代表第 n 位加法和的 *DNA* 片段 *sum*，放入试管 T_c

3.2　基于 DNA 自动机的 n 位二进制串行加法

对于两个 n 位二进制数(A_i)和$(B_i)$$(i=0,\cdots,n-1,0$ 表示最低位)，类似两个多位数相加的笔算步骤，串行加法按照从 $i=0$ 至 $i=n-1$ 的顺序逐位进行；每次进行全加的数包括加数、被加数和低位产生的进位数。例如，对第 i 位进行运算，得到第 i 位全加后的和 S_i 及进位 C_i 后才进行下一位 $i+1$ 位的运算，即 A_{i+1}、B_{i+1} 和 C_i 相加。

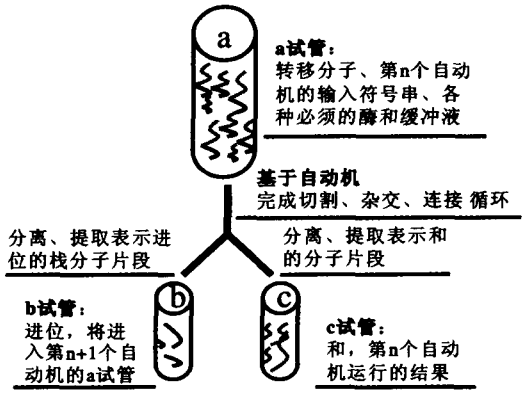

图 1　一位二进制加法的试管体系

对于 n 位二进制的串行加法只要将一位二进制试管体系按照从低位到高位的顺序，即从第 0 位到第 $n-1$ 位顺序排列便可实现。图 2 表示 n 位二进制串行加法的试管体系。图 2 中，S_i 表示第 i 位加法运算所产生的和，C_i 表示第 i 位加法运算所产生的进位。这样，n 位串行加法的运行结果看成字符串就是片段 $C_{n-1}S_{n-1}S_{n-2}\cdots S_0$，通过检测分子分别检测这些 DNA 片段表示的二进制数字，就可以得出二进制运算的结果。

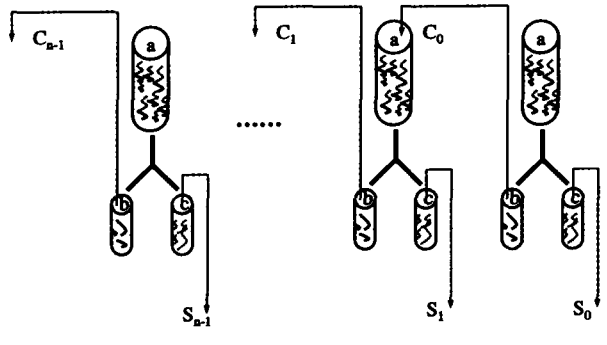

图 2　n 位二进制串行加法的试管体系

4 M_{adder}编码的示例

在基于DNA自动机的二进制串行加法中，用到了3种限制性核酸内切酶，其中*FokI*的识别位点是5'-*GGATG*-3'，剪切位点是(9,13)；*BsmFI*的识别位点是5'-*GGGAC*-3'，剪切位点是(10,14)；*SmiI*的识别位点是5'-*ATTTAAAT*-3'，剪切位点是识别位点内部1/2位置。为了验证基于DNA自动机的二进制串行加法的可行性，我们在表3(a)给出了M_{adder}的所有符号的DNA编码，表3(b)给出了转移函数的DNA编码，表3(c)给出了检测分子的DNA编码。基于已有研究成果[4]，我们不难得到现有的生物技术可以实现基于M_{adder}DNA编码的二进制串行加法运算的结论。

表3(a) M_{adder}符号的DNA编码

(注：T是结束处理的终止符)

0	1	T
ACGAAG	ACGACA	GTATGA
TGCTTC	TGCTGT	CATACT

表3(b) M_{adder}转移函数的DNA编码

(注：编码中方框内数字代表满足条件的碱基序列，这些碱基与相邻序列连接后不含*FokI*识别位点)

转移函数	DNA编码
$Tran(S_0,0,0)=(S_0,0)$	ACGAAC [7] CATCCGGATG TGA TC GTAGGCCTAC ACT GCTT
$Tran(S_0,0,1)=(S_1,0)$	ACGAAG [7] CATCCGGATGTG TC GTAGGCCTACAC GCTG
$Tran(S_1,0,0)=(S_0,1)$	ACGAAG [7] CATCCGGATGTGA TC GTAGGCCTACACTCTTC
$Tran(S_1,0,1)=(S_1,1)$	GACA [9] CATCCGGATGTG GTAGGCCTACACGCTG
$Tran(S_0,1,0)=(S_0,1)$	GAAG [9] CATCCGGATGTGA GTAGGCCTACACTGCTT

表3(c) M_{adder}结果检测分子的DNA编码

(注：编码中方框内数字代表满足条件的碱基序列，这些碱基与相邻序列连接后不含*BsmFI*和*SmiI*识别位点)

终止状态	栈顶元素	DNA编码
S_1	0	ACGAGGTATGA TGGTCCC ATTTAAATGGGAC [10] TACT TC CATACT ACCAGGG TAAATTTACCCTG
S_1	1	GACA GTATGATGGTCCC ATTTAAATGGGAC [10] ATAC CTATCT ACCAGGG TAAATTTACCCTG
S_0	1	GACTCTATGATGGTCCC ATTTAAATGGGAC [10] ATAC CATACT ACCAGGGTAAATTTACCCTG
S_0	0	ACGAAGGTATGATG GTCCC ATTTAAAT GGGAC [10] ATAC TCCATACTAC CAGGGTAAATTTACCCTG

结束语 本文提出了一种基于DNA自动机的n位二进制串行进位加法的模型。该模型类似于电子计算机中加法的处理过程，操作数的编码规则简单；实现运算的生物操作在实验室简单可行，多位二进制数的加法仅仅是一位二进制数加法的简单循环。本文为DNA计算机中加法的实现提供了一种可行的模型，通过选择合适的编码，该模型可进一步推广到乘法、除法等其他算术运算，还可以推广到数的补码运算等，对研究DNA计算机的运算系统具有十分重要的意义。

参考文献

1 Guarnieri F, Fliss M, Bancroft C. Making DNA add [J]. Science, 1996, 273: 220～223

2 Guarnieri F, Bancroft C. Use of a horizontal chain reaction for DNA-based addition. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 1999, 44: 105～111

3 Yurke B, et al. DNA implementation of addition in which the input strands are separate from the operator strands [J]. Biosystems, 1999, 52: 165～174

4 Benenson Y, et al. Programmable and autonomous computing machine made of biomolecules [J]. Nature, 2001, 414: 430～434

5 Li W-G, Ding Y-S, Huang Z-D, et al. Stack-type data structure for DNA-based computer. The 11th International Meeting on DNA Computing, London, Ontario, Canada, June, 2005 (Poster)

6 Ding Y-S, Shao S-H, Ren L-H. DNA Computing and Soft Computing. Beijing: Scientific Publishing House, 2002 (in Chinese)

7 许进，张雷. DNA计算机原理、进展及难点(Ⅰ)：生物计算系统及其在图论中的应用. 计算机学报，2003，26(1)：1～11

(上接第166页)

们要求必须是连续而且精确的k-mer子串，影响了算法的应用。

进一步的工作是通过非精确或非连续的k-mer子串以提高算法的灵敏度和研究基于k-mer子串的片段拼接算法及海量片段数据的大规模并行处理技术。

参考文献

1 Myers E W, Sutton G G, Dew LM, et al. A Whole-Genome Assembly of Drosophila [J]. Science, 2000, 287: 2196～2204

2 International Human Genome Sequencing Con- sortium. Initial Sequencing and Analysis of the Human Genome [J]. Nature, 2001, 409: 860～864

3 Kececioglu J D, Meyers E W. Combinatorial Al- gorithms for DNA Sequencing Assembly [J]. Algorithmica, 1995, 13: 7～15

4 Pevzner P A, Tang Haixu, Waterman M S. An Eul -erian Path Approach to DNA Fragment Assembly [J]. Proceedings of National Academy of Sciences, 2001, 98: 9487～9753

5 Batzoglou S, Jaffe D B, Stanley K, et al. ARACHNE: A Whole-Genome Shotgun Asse -mbler In: Proceedings of the Ninth Annual International Conference in Computational Mole- cular Biology (RECOMB), May 2005, Cam bridge, 2005. 177～189

6 Setuball J C, Werneck R F. A Program for Build- ing Contig Scaffolds in Double-barrelled Shotgun Genome Sequencing: [Institute of Computing Technical Report IC-01-05]. Unicamp, 2001

7 张博峰，王正华. DNA片段拼接中基于定长特征子串的重复序列信息屏蔽方法[J]. 国防科技大学学报，2002，24(6)：67～70

8 张春霆. 生物信息学的现状与展望[J]. 世界科技研究与发展，2000，22(6)：17～20

9 涂俐兰，王能超，等. 生物序列拼接及其算法[J]. 生命科学研究，2003，7(2)：79～80

10 Bao, Eddy. 2002. RECON documentation. http://www.genetics.wustl.edu/eddy/recon