基于粘贴模型的两类全排问题的 DNA 算法

栗青生[1],杨玉星[1],马季兰[2]
LI Qing-sheng[1],YANG Yu-xing[1],MA Ji-lan[2]

1.安阳师范学院 计算机与信息工程学院,河南 安阳 455000
2.太原理工大学 计算机与软件学院,太原 030024
1.School of Computer and Information Engineering,Anyang Normal University,Anyang,Henan 455000,China
2.College of Computer and Software,Taiyuan University of Technology,Taiyuan 030024,China
E-mail:jade_star@163.com

LI Qing-sheng,YANG Yu-xing,MA Ji-lan.DNA algorithm of two kinds of full permutation problem based on sticker model.Computer Engineering and Applications,2010,46(4):46-48.

Abstract: Sticker DNA algorithms of linear full permutation and circle full permutation are proposed based on the vast parallelism of sticker model,and the differences between the algorithms are illustrated.The operation steps are given through an instance,and simulation experiments are carried out to illustrate the biochemical processes.The final correct results are gotten.Consequently,the feasibilities of the algorithms are proved.At last,the complexities are analyzed.
Key words: full permutation;circle permutation;DNA computing;sticker model

摘 要:基于粘贴模型的巨大并行性,分别给出了线性全排列和圆周全排列问题的粘贴 DNA 算法;分析了两类问题的 DNA 算法的不同之处;通过一个实例给出了实验操作步骤,并对生化实验进行了模拟,得出了正确的结果,从而证明了算法的可行性。最后,对算法的操作复杂度进行了分析。
关键词:全排列;圆排列;DNA 计算;粘贴模型
DOI:10.3778/j.issn.1002-8331.2010.04.014 **文章编号**:1002-8331(2010)04-0046-03 **文献标识码**:A **中图分类号**:TP301.6

1994 年,美国加利福尼亚大学的 Adleman 教授用现代分子生物技术,通过 7 天的生化试验,解决了 7 个节点的有向哈密尔顿路径问题[1],开创了 DNA 计算的新纪元。由于 DNA 计算具有并行性高、存储量大、能耗低、DNA 资源丰富等优势,随后,有关 DNA 计算的研究在诸多国家展开,使用 DNA 方法解决图论、数学问题是其中的一个热点。

文献[2]提出 DNA 标号图的概念;文献[3]解决了 0-1 整数规划问题;文献[4]提出了一种基于分治的背包问题的 DNA 算法;由于在 Adleman-Lipton 模型(实质上是一种粘贴模型)中没有明确提出 0 和 1 的概念,不便使用原有的分离操作,文献[5]提出广义的分离技术;该文基于粘贴模型及广义的分离与广义的多级分离技术,提出了线性全排列问题及圆周全排列问题的 DNA 算法。

1 粘贴 DNA 模型

1.1 粘贴 DNA 模型介绍

由于粘贴模型具有“不需要延伸 DNA 分子链”、“不需要酶的参与”以及“材料可以重复利用”等优点,成为 DNA 计算中备受关注的计算模型之一。

粘贴模型的存储区中放置着由存储链和粘贴链组成的存储合成物。存储链是一个由 n 个不重叠的子链组成的单链 DNA 分子,而每个子链包含 m 个碱基。每个粘贴链也是由 m 个碱基构成,而且每个粘贴链均与存储链中的某一个子链满足 Watson-Crick 互补关系。当一个存储合成物中的某一个位元为单链时表示 0,为双链时表示 1。

有时,也可以使用指定的 DNA 单链分子表示某一位元的状态(即,真和假)。例如:若有子链数为 3,子链长度为 6 的碱基位串,指定其第一个位元(子链)用 5′-GAGACT-3′表示真,用 5′-TTACGA-3′表示假;第二个位元用 5′-CCCTAG-3′表示真,用 5′-GAGACT-3′表示假;第三个位元用 5′-ACCAGT-3′表示真,用 5′-GTTGGT-3′表示假;则碱基位串 5′-GAGACTGAGACTACCAGT-3′对应的二进制串为 101。

粘贴模型在位串上定义了合并(Merge)、分离(Separate)、设置(Set)、清除(Clear)等四种基本操作[6],下面仅对将用到的

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60973051);河南省教育厅自然科学研究项目(the Natural Science Research Project of Education Department of Henan Province,China under Grant No.2008B520001)。
作者简介:栗青生(1966-),男,副教授,硕士生导师,研究方向:智能信息处理;杨玉星(1981-),男,硕士,研究方向:DNA 计算;马季兰(1948-),女,教授,研究方向:操作系统,智能计算。
收稿日期:2008-09-02 **修回日期**:2008-10-31

两种操作进行介绍。

合并(Merge):定义存储合成物的集合 T_1 与 T_2 的合并为 T,则 $T=T_1\cup T_2$。

分离(Separate):根据存储合成物中第 i 个位元的状态(即单、双链)将存储合成物的集合 T 分解为两个集合:$+(T,i)$ 和 $-(T,i)$,其中,$+(T,i)$ 为该位元为"1"的位串的集合,$-(T,i)$ 为该位元为"0"的 DNA 分子的集合。

各种操作的物理实现方法文献[5]已有详细介绍,这里不再赘述。

基于粘贴模型的计算模式就是将问题的所有可能解用 DNA 分子来编码,得到一个数据池,对该数据池中的 DNA 分子通过上述操作的某一种或者几种操作的排列组合,筛选出结果 DNA 分子链,如果结果链为空,则表明问题无解。

1.2　广义的分离

在 Adleman-Lipton 模型[1](其实质是一种粘贴模型)中,没有明确提出 0 和 1 的概念,所以不便使用原有的分离操作,基于这一原因,文献[5]提出了广义的分离概念。

广义的分离(Extended Separate,E-Separate):根据存储混合物中的 DNA 位串是否包含子链 x_i,将存储合成物的集合 T 分解为两个集合:$+(T,x_i)$和$-(T,x_i)$,其中,$+(T,x_i)$为包含子链 x_i 的位串的集合,$-(T,x_i)$表示不含子链 x_i 的位串的集合。若要根据是否包含子链 x_i 与子链 x_j 顺接而成的较长子链,则可表示为 E-Separate $+(T,x_ix_j)$ and $-(T,x_ix_j)$。

下面使用基于粘贴模型的这些操作,求解集合的线性全排列和圆周全排列。

2　线性全排列问题的 DNA 算法

2.1　问题的描述

排列问题是组合数学中一个重要的问题,设 $A=\{a_1,a_2,\cdots,a_n\}$是 n 个不同元素的有限集合,整数 r 满足 $0\leqslant r\leqslant n$,从 A 中任取 r 个数排成有序的一列,称为 A 中取 r 个元素的一个排列。

特别的,当 $r=n$ 时,所有不同的排列称为 A 的线性全排列。

2.2　问题的算法

集合 A 的线性全排列问题的 DNA 算法可描述如下:

步骤 1　生成初始数据池 T_0。

对集合 A 中的每一个元素使用文献[7]的编码方法进行 DNA 编码,这里,将 a_i 的 DNA 编码简记为 X_i,每个元素的 DNA 编码的长度为 m 个碱基。使用文献[1]所述的方法,加入引物,在连接酶的作用下,使各元素的 DNA 编码链接起来。

以 $X_i(i=1,2,\cdots,n)$为引物和其余元素的编码的补码为引物,通过聚合酶连接反应(Polymerase Chain Reaction,PCR)扩充所有 DNA 分子链,生成初始数据池 T_0。

步骤 2　删除长度不为 $m*n$ 个碱基的 DNA 分子。

使用凝胶电泳技术执行按长度分离操作,分离出长度为 $m*n$ 个碱基的 DNA 分子,以其作为新的数据池,仍记为 T_0,这样得到的 DNA 链包含 n 个元素。

步骤 3　删除包含重复元素的 DNA 分子链。

在步骤 2 中得到的 DNA 分子链均包含 n 个元素,只要不包含 $a_1,a_2,\cdots,a_n$ 中的任何一个元素,则该 DNA 分子链必有重复元素。

这一操作可以使用广义的分离操作实现,实现方法可表示如下:

For $i\leftarrow 1$ to n do

E-Separate $+(T,X_i)$ and $-(T,X_i)$

$T\leftarrow +(T,X_i)$

End

经过上述操作步骤后,试管 T_0 中的 DNA 分子链即为集合 A 的线性全排列所对应的 DNA 分子链,使用相关生物技术检测出结果。

3　圆周全排列问题的 DNA 算法

3.1　问题的描述

设 $A=\{a_1,a_2,\cdots,a_n\}$是 n 个不同元素的有限集合,整数 r 满足 $0\leqslant r\leqslant n$,从 A 中任取 r 个数在圆周上排成有序的一列,称为 A 中取 r 个元素的一个圆周排列。

特别的,当 $r=n$ 时的排列称为 A 的圆周全排列,简称 A 的圆周全排。

3.2　问题的算法

圆排列与线性排列的不同之处在于圆排列首尾相接,这也是圆周全排问题与全排列问题的不同之处。例如,3 个元素的集合 $A_1=\{a_1,a_2,a_3\}$的线性全排列为:$a_1a_2a_3,a_2a_3a_1,a_3a_1a_2,a_1a_3a_2,a_2a_1a_3,a_3a_2a_1$。然而,因为 $a_1a_2a_3,a_2a_3a_1,a_3a_1a_2$ 属于同一种圆周排列,$a_1a_3a_2,a_2a_1a_3,a_3a_2a_1$ 属于同一种圆周排列,所以集合 A_1 的圆周全排列只有两种,即 $a_1a_2a_3$ 和 $a_1a_3a_2$,如图 1 所示。

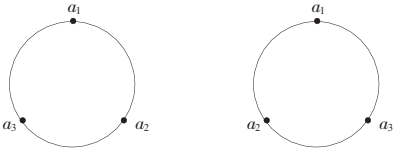

图 1　集合 A_1 的圆周全排列

圆周全排问题的 DNA 算法与全排列问题的 DNA 算法的不同之处在于二者的初始数据池的构建有所不同,其余步骤均相同。圆周全排列问题的 DNA 算法的初始数据池应为那些第一个元素为某一特定元素的 DNA 单链,不妨设第一个元素为 a_1,圆周全排列问题的步骤 1 可描述如下:

步骤 1　生成初始数据池 T_0。

对集合 A 中的每一个元素使用文献[7]的编码方法进行 DNA 编码,这里,将 a_i 的 DNA 编码简记为 X_i,每个元素的 DNA 编码的长度为 m 个碱基。使用文献[1]所述的方法,加入引物,在连接酶的作用下,使各元素的 DNA 编码链接起来。

以 X_1 为引物和其余元素的编码的补码为引物,通过 PCR 扩充第一个位元为 a_1 的编码的 DNA 分子链,生成初始数据池 T_0。

其余步骤和全排列问题的 DNA 算法相同。

下面,将结合实例,给出这两类排列问题的 DNA 算法的具体实验操作过程。

4　实例

4.1　实例描述

已知集合 $A_2=\{a_1,a_2,a_3\}$,求 A_2 的线性全排列和圆周全排列。

4.2 求解线性全排列

求解集合 A_2 的线性全排列的 DNA 算法及相应的生化实验操作如下：

步骤 1 生成初始数据池 T_0。

首先，对 a_1、a_2、a_3 进行 DNA 编码。每个元素采用长度为 10 个碱基的 DNA 单链编码，记 a_1、a_2、a_3 的编码分别为 X_1、X_2、X_3，采用的编码如下：

X_1=5′–TGATAGGACC–3′

X_2=5′–CCGTGTATAG–3′

X_3=5′–ACAGTTGTGC–3′

其次，对上述编码进行 PCR 扩充，使其在引物和连接酶的作用下充分反应，生成各种可能的排列。再对表示所有可能的排列的 DNA 链进行 PCR 扩充，生成初始数据池 T_0。

步骤 2 删除长度不为 30 个碱基的 DNA 分子。

将试管 T_0 中的溶液倒入加有电场的多级分离装置中，使 DNA 分子通过凝胶。由于 DNA 分子带负电，DNA 分子将向正极方向移动，在凝胶的摩擦力的作用下，按长度形成一条条 DNA 分子带。据此分离出长度为 30 个碱基的 DNA 分子，以其作为新的数据池，仍记为 T_0。

步骤 3 删除包含重复元素的 DNA 编码的 DNA 分子。

在步骤 2 中得到的 DNA 分子链均包含 3 个元素，只要不包含 a_1、a_2、a_3 中的任何一个元素，则该 DNA 分子链必有重复元素，将这些包含重复元素的 DNA 分子链删除，即：

For i←1 to 3 do

E–Separate +(T_0,X_i) and –(T_0,X_i)

T_0←+(T_0,X_i)

End

以 X_1 的编码的 DNA 补码作为探针，固定在试管 T_1 的试管壁上。将 T_0 中的 DNA 溶液缓缓倒入试管 T_1，经过足够的时间反应后，包含子链 X_1 的 DNA 分子将被吸附在试管 T_1 的试管壁上。倒掉其余的 DNA 溶液，使用缓冲液冲洗试管 T_1 的试管壁；加热，使粘贴链与存储链解链；再使用凝胶电泳技术，即可分离出+(T_0,X_1)，将其倒入试管 T_0 中。仿照上述操作，依次删除不包含子链 X_2、X_3 的 DNA 分子链。

至此，T_0 中的 DNA 分子即为代表 A 的线性全排列的 DNA 分子。

在 Visual C++ 6.0 下对上述实验进行了模拟，检测得到 6 种 DNA 单链分子，最终结果及其对应的排列方案如表 1 所示。

从表 1 得知，该实例的线性全排列方案共有 6 种，得到这一结果需要 3 步广义的分离操作。

4.3 求解圆周全排列

求解集合 A_2 的圆周全排列的 DNA 算法与求解 A_2 的线性全排列的 DNA 算法仅在于初始数据池的构建有所不同。

表 1 线性全排列的结果 DNA 分子及相应方案

序号	5′到 3′的结果 DNA 分子	排列
1	5′–TGATAGGACCCCGTGTATAGACAGTTGTGC–3′	$a_1a_2a_3$
2	5′–TGATAGGACCACAGTTGTGCCCGTGTATAG–3′	$a_1a_3a_2$
3	5′–CCGTGTATAGTGATAGGACCACAGTTGTGC–3′	$a_2a_1a_3$
4	5′–CCGTGTATAGACAGTTGTGCTGATAGGACC–3′	$a_2a_3a_1$
5	5′–ACAGTTGTGCTGATAGGACCCCGTGTATAG–3′	$a_3a_1a_2$
6	5′–ACAGTTGTGCCCGTGTATAGTGATAGGACC–3′	$a_3a_2a_1$

这里仍采用求解线性排列时的编码，对上述编码进行 PCR 扩充，使其在引物和连接酶的作用下充分反应。

接下来，仅对以 X_1 为第一个位元的 DNA 链进行 PCR 扩充，生成初始数据池 T_0。

步骤 2 和步骤 3 与求解线性全排列的操作方法相同。

在 Visual C++ 6.0 下亦对求解圆周全排列的实验进行了模拟，检测得到 2 种 DNA 单链分子，最终结果及其对应的排列方案如表 2 所示。

表 2 圆周全排列的结果 DNA 分子及相应方案

序号	5′到 3′的结果 DNA 分子	排列
1	5′–TGATAGGACCCCGTGTATAGACAGTTGTGC–3′	$a_1a_2a_3$
2	5′–CCGTGTATAGTGATAGGACCACAGTTGTGC–3′	$a_2a_1a_3$

从表 2 得知，该实例的圆周全排列方案共有 2 种，这一实验需要 3 步广义的分离操作。

5 结束语

给出了基于粘贴 DNA 模型的两类全排列问题的生物分子算法。衡量 DNA 算法的复杂度的主要指标为分离操作的步骤数。该文算法所需要的分离操作步骤数均为集合 A 的元素个数 n，即复杂度为 $O(n)$，是一种时间复杂度较低的算法。

参考文献：

[1] Adleman L M.Molecular computation of solutions to combinatorial problems[J].Science,1994,266(5187):1021–1024.

[2] 王世英,原军,林上为.DNA 标号图和 DNA 计算[J].中国科学:A 辑数学,2007,37(9):1059–1072.

[3] 殷志祥,许进.分子信标芯片在 0–1 整数规划问题中的应用[J].生物数学学报,2007,22(3):559–564.

[4] 李肯立,姚凤娟,李仁发,等.基于分治的背包问题 DNA 计算机算法[J].计算机研究与发展,2007,44(6):1063–1070.

[5] Xu Jin,Dong Ya–fei,Wei Xiao–peng.Sticker DNA computer model–Part Ⅰ theory[J].Chinese Science Bulletin,2004,49(8):772–780.

[6] 杨玉星,栗青生,马季兰.一类禁位排列问题的粘贴 DNA 算法[J].计算机工程与应用,2008,44(30):40–42.

[7] Tanaka F,Kameda A,Yamamoto M.Design of nucleic acid sequences for DNA computing based on a thermodynamic approach[J].Nucleic Acids Research,2005,33(3):903–911.