

# DNA 并行计算中的 B-树数据结构的设计与实现<sup>\*</sup>

邹超君<sup>1,2</sup>, 朱雅莉<sup>2</sup>

(1. 湖南大学 信息科学与工程学院, 长沙 410082; 2. 衡阳师范学院 计算机科学系, 湖南 衡阳 421002)

**摘要:** DNA 分子特性使得 DNA 计算具有极大的存储密度和高度的计算并行性。不管何种计算模型, DNA 分子的选择和 DNA 编码都十分重要。提出了 DNA 计算中的 B-树的数据结构设计方法。首先给出了 B-树定义及其操作的形式化描述, 接着介绍了本计算模型采用的 3D 结构 DNA 分子—— $k$ -arms 分子结构, 详细给出了一棵  $m$  阶 B-树的构造步骤, 最后实现了其查找、插入和删除等操作。提出了 DNA 分子计算的 3D 结构和分治策略, 具有一定的可扩展性和并行性, 对 DNA 计算的其他模型有参考价值。

**关键词:** DNA 计算; 可扩展性; 并行计算; 分治策略; DNA 编码

**中图分类号:** TP311      **文献标志码:** A      **文章编号:** 1001-3695(2012)05-1775-03

doi:10.3969/j.issn.1001-3695.2012.05.046

## Design and implementation of B-tree data structure in DNA parallel computing

ZOU Chao-jun<sup>1,2</sup>, ZHU Ya-li<sup>2</sup>

(1. College of Information Science & Engineering, Hunan University, Changsha 410082, China; 2. Dept. of Computer Science, Hengyang Normal University, Hengyang Hunan 421002, China)

**Abstract:** The best virtues of DNA computing are the maximum density of storage and the huge parallel of computing because of the features of DNA molecules. It is important for any DNA computing model to choose some DNA molecules and code them. This paper proposed the method of designing B-tree in DNA computing. Firstly, formally described the definition of B-tree and operations. Then, described the 3D DNA molecules,  $k$ -arms DNA molecules in detail. Furthermore, proposed the approach of creating a B-tree with  $m$  sub-trees. Finally, proved the operations, such as searching, inserting and deleting. DNA molecules with 3D structure and divide-and-conquer strategy were scalable and parallel, and could be used in other DNA computing model.

**Key words:** DNA computing; scalability; parallel computing; divide-and-conquer strategy; DNA encoding

## 0 引言

自 1994 年 Adleman<sup>[1]</sup> 博士首次用 DNA 分子解决了一个有向图哈密尔顿问题以来, DNA 计算得到了迅速发展<sup>[2,3]</sup>。DNA 计算是以 DNA 分子作为数据存储载体, 以生物酶分子和生物化学反应作为信息处理工具的一种新的计算模式。DNA 分子由两条脱氧核糖核酸链组成, 携带大量生物遗传信息。正是由于 DNA 的巨大并行性、自我复制性及 Watson-Crick 的互补结构, 使得这种新的计算模式具有很大优越性, 即极大的存储密度和高度的计算并行性。众所周知, 电子计算机是依靠电流接通或断开产生“二态”, 其中一种状态用“0”来表示, 另一种状态用“1”来表示, 由此产生 0-1 序列来进行信息处理; 而 DNA 计算则是对 DNA 序列通过可控的生化反应来进行信息处理。电子计算机中所有的信息都是用二进制来存储的, 数据与数据之间呈现出各种关系, 这个关系及其表示称为数据结构。数据结构在计算机中占有重要的地位; 在 DNA 计算中同样需要合理的数据结构来组织数据——DNA 分子。许进等人<sup>[4]</sup> 详细描述了分子生物计算中的数据结构与特性; 李汪根等人<sup>[5]</sup> 研究了 DNA 计算机中队列的数据结构实现, 给出了

DNA 计算机中的广义表设计与实现<sup>[6]</sup>; 朱雅莉等人<sup>[7]</sup> 描述了 DNA 计算机中堆栈及二叉树的数据结构设计。本文探讨了基于 DNA 并行计算的数据结构设计, 构建用于文件系统及查找的 B-树数据。

## 1 基于 DNA 计算的 B-树的设计

### 1.1 B-树的定义

B-树是一种广泛用于文件系统平衡的多路查找树。一棵  $m$  阶的 B-树, 或为空树, 或为每个节点至多有  $m$  棵子树的  $m$  叉树<sup>[8]</sup>, 并且除根之外的所有非终端节点至少有  $\lceil m/2 \rceil$  棵子树。其非终端节点中包含下列信息数据  $(n, A_0, K_1, A_1, K_2, A_2, \dots, K_N, A_N)$ , 如表 1 所示。

表 1 B-树非终端节点的结构

|     |       |       |       |       |       |         |       |       |         |       |       |
|-----|-------|-------|-------|-------|-------|---------|-------|-------|---------|-------|-------|
| $n$ | $A_0$ | $K_1$ | $A_1$ | $K_2$ | $A_2$ | $\dots$ | $K_i$ | $A_i$ | $\dots$ | $K_N$ | $A_N$ |
|-----|-------|-------|-------|-------|-------|---------|-------|-------|---------|-------|-------|

### 1.2 B-树的节点的 DNA 编码

如前定义所描述, B-树的非终端节点包含一个以上的关键字及指示其左右的指针。因此, 本文使用  $k$ -arms ( $k=3$ ) DNA 分子和双链 DNA 分子来表示 B-树的节点和关键字及其连接。

收稿日期: 2011-09-15; 修回日期: 2011-11-07      基金项目: 国家自然科学基金资助项目(90715029)

作者简介: 邹超君(1981-), 女, 湖南衡阳人, 讲师, 硕士研究生, 主要研究方向为 DNA 计算(zouchaojun@126.com); 朱雅莉(1979-), 女, 湖南宁乡人, 讲师, 硕士, 主要研究方向为 DNA 计算、计算机应用技术。

Jonoska 等人<sup>[9]</sup>于1999年提出使用顶点连接的  $k$ -arms 分支 DNA 分子来形成图并描述和解决图的可满足性与图顶点着色问题。 $K$ -arms DNA 分子是双链 DNA 分子的一种变异性结合。自然界中已经存在类似于  $k$ -arms DNA 分子结构的 DNA 分子,如 Holliday 分子。 $K$ -arms 具有一定的稳定性, $k$  取值可为 2、3 或 4,其结构模拟如图 1 所示<sup>[9]</sup>。

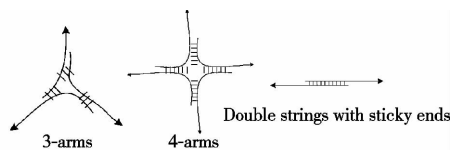


图1  $K$ -arms DNA分子和双链DNA分子的结构

本文使用 3-arms DNA 分子和双链 DNA 分子两种结构类型来表示 B-树的节点<sup>[6]</sup>,如图 2 所示。

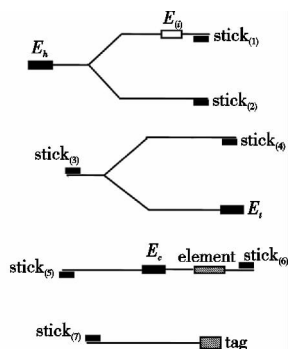


图2 构建B-树节点的DNA片段

1) tag 节点关键字个数的标志,即节点的大小,tag 等于  $n$ 。由长度为含有 12 的  $n$  倍碱基对的双链 DNA 分子表示。

2) element 节点的关键字元素。由长度为 12 的不同的碱基序列的双链 DNA 分子表示。

3)  $E_{(i)}$  一种限制性内切酶的识别点。作为节点的识别点,在遍历和查找时起到标志性作用。限制性内切酶是生物体内能识别并切割特异的双链 DNA 序列的一种内切核酸酶。它是可以将外来的 DNA 切断的一种蛋白质,即能够限制异源 DNA 的侵入并使之失去活力,但对自己的 DNA 却无损害作用,这样可以保护细胞原有的遗传信息。由于这种切割作用是在 DNA 分子内部进行的,故名限制性内切酶(简称限制酶)。大多数酶是非常专一的,只能催化一种化学反应,并且效率极高。自然界已经创造出多种多样的、在处理 DNA 过程中非常有用的酶<sup>[10]</sup>。

4)  $E_h$  一种限制性内切酶的识别点。在一定的活性条件下,从  $k$ -arms 分支点切断,并留下一粘性末端,用于连结前一个节点。

5)  $E_e$  一种限制性内切酶的识别点,用于识别关键字元素的值。在 B-树的查找、删除及插入操作中起到关键作用。

6)  $E_i$  一种限制性内切酶的识别点。在一定的活性条件下,从  $k$ -arms 分支点切断,并留下一粘性末端,用于连结下一个节点。

7) stick 粘性末端。一种特殊的单链 DNA 分子。通过连接酶的作用,在连接反应下与其他 DNA 片段连接。

为了构造一棵 B-树,  $E_h$  和  $E_i$  产生的粘性末端必须为互补的单链。两条单链多核苷酸通过互补碱基之间的氢键形成双链分子,从而完成连接。通过 PCR 技术,可以复制多份相同 DNA 分子的溶液,放在不同试管备用。连接反应和识别节点及元素应该在不同试管进行<sup>[10]</sup>。

### 1.3 基于可扩展并行 DNA 计算的 B-树的设计

基于以上编码,可以构造一棵  $m$  阶 B-树,过程如下:

a) 初始化一棵  $m$  阶 B-树。首先初始化一棵  $m$  阶 B-树。用一个表示连接节点的 3-arms DNA 分子作为 B-树的空节点。在一个空节点的 3-arms DNA 分子的带有  $E_h$  识别点的分支做荧光标记 sign,表示根节点  $E_{(0)}$ 。

b) 连接反应 1。连接反应的方法有几种,如平头连接反应和粘端连接反应<sup>[10]</sup>。本文使用端连接。所有的空节点在一定活性酶下,可以并行连接各自的 tag。如图 2 所示,tag 是带一个粘性末端的双链 DNA 分子。其粘性末端是与空节点的  $stick_{(1)}$  互补,但每个节点的粘性末端的编码都不相同。这可以通过碱基对的不同排序来表示。

c) 连接反应 2。如图 2 所示,同样用带有两个粘性末端的 3-arms DNA 分子连接关键字元素。Stick<sub>(4)</sub> 与 stick<sub>(5)</sub> 为互补的单链,每个元素的粘性末端的编码都不一样,并行连接反应。B-树的每个节点的关键字都是排序好的,每个节点可以并行地连接排序好的多个关键字元素。

连接反应 1、2 可以同时在不同试管内并行处理。当以上三个阶段完成,就可通过节点连接设计出一棵  $m$  阶 B-树。其步骤是:在限制性内切酶  $E_h$  和  $E_i$  的活性条件下,分别产生粘性末端。前一个节点的  $E_i$  粘性末端和后一个节点的  $E_h$  粘性末端互补,并且每个节点的编码都不同。由于每种酶的活性条件不同,因此可以将分治策略应用到节点的连接中。首先复制溶液,在不同的试管中并行处理,两两连接,再两两混合,直至最终得到一个  $m$  阶 B-树。图 3 所示为一棵 DNA 编码的 4 阶 B-树。要注意的是,此处叶子节点均省略。叶子节点可以用一个带粘性末端的双链 DNA 分子表示。其连接反应类似。

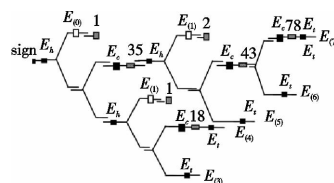


图3 一棵DNA编码的4阶B-树

其中:■表示关键字个数; $E_{(3)}$ 表示节点4,包含一个关键字 11; $E_{(4)}$ 表示节点5,包含一个关键字 27; $E_{(5)}$ 表示节点6,包含一个关键字 39; $E_{(6)}$ 表示节点7,包含一个关键字,分别是 47、53、64; $E_{(7)}$ 表示节点8,包含一个关键字 99。

## 2 基于可扩展并行 DNA 计算的 B-树的基本操作与实现

### 2.1 基于 DNA 计算的 B-树的查找及其分析

由 B-树的定义可知,在 B-树上进行查找的过程与二叉排序树的查找类似。根据给定的键值  $K$ ,先在根节点键值的集合中采用顺序(当  $m$  较小时)或二分(当  $m$  较大时)查找方法进行查找。若有  $K = K_i$ ,则查找成功,根据相应的指针,即可取记录;否则,若  $K$  在  $K_i$  和  $K_{i+1}$  之间,取指针  $A_i$  所指的节点,重复这个查找过程,直到在某节点中查找成功,或在某节点处出现  $A_i$  为空,查找失败。

限制性内切酶的生物学特性使得 DNA 编码的 B-树查找变得非常简单,查找性能大大提高。同样地,包括在 B-树中找节点和在节点中找关键字<sup>[8]</sup>。如图 3 所示,每个节点都有不

同的限制性内切酶的识别点  $E_{(i)}$ , 每个关键字元素都有不同的限制性内切酶的识别点  $E_c$ 。结合荧光技术, 利用特定酶与DNA分子的特性, 通过检测酶切后荧光量进行计算<sup>[11]</sup>。由于限制性内切酶的活性条件不同, 因此, 同时查找两个或者以上节点或者关键字元素时, 可以复制溶液在不同的试管中并行进行。

## 2.2 基于DNA计算的B-树的插入和删除

1) B-树插入的生物操作 查找不成功之后, 需进行插入。显然, 关键字插入的位置必定在最下层的非叶节点, 有下列几种情况:

a) 插入后, 该节点的关键字数  $n < m$ , 不修改指针;

b) 插入后, 该节点的关键字数  $n = m$ , 则需进行“节点分裂”, 令  $s = (m/2)$ , 在原节点中保留  $(A_0, K_1, A_1, K_2, A_2, \dots, K_{s-1}, A_{s-1})$ , 建新节点  $(A_s, K_{s+1}, \dots, K_n, A_n)$ , 并将  $(K_s, p)$  插入双亲节点;

c) 若双亲为空, 则建新的根节点。

插入操作是查找插入点后使用限制性外切酶进行切割, 形成粘性末端<sup>[10]</sup>, 设计新的节点的粘性末端与之互补, 通过连接反应将节点或者关键字顺利插入。以图3的DNA编码的4阶B-树为例, 假设需依次插入关键字30、60。首先从根节点查找起, 如果小于根节点, 则插入左子树节点; 如果大于根节点, 则插入右子树节点。显然30是插入  $E_{(4)}$  节点。直接酶切该节点的右叶子节点, 在其中插入关键字元素30。同样地, 该元素由一个3-arms DNA分子连接一段双链DNA分子组成, 其中双链DNA分子代表键值, 并修改tag值, 如图4所示。最后生成两个叶子节点。通过查找, 关键字60插入节点  $E_{(6)}$ , 插入后, tag大于4, 产生节点分裂。原节点生成两个新的子节点(47)、(60, 64), 关键字53及其指针插入双亲节点  $E_{(2)}$ , 分别修改tag。如图4所示, 若双亲节点的tag大于4, 则建新根节点。

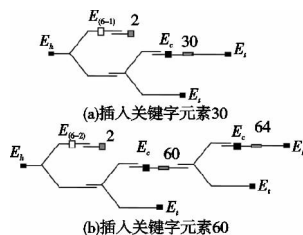


图4 插入关键字元素30和60

2) B-树删除的生物操作 在深度为  $(h+1)$  的  $m$  阶B-树中删除一个键值  $K$ , 首先要查到键值  $K$  所在的节点及在节点中的位置。若  $K$  在非终端节点中, 则把该节点的右边(或左边)指针所指子树中的最小(或最大)键值与  $K$  对调, 使  $K$  移到终端节点。因此, 非终端节点中关键字元素的删除只需要通过切割和连接实现相应的代表关键字元素的DNA双链分子的对调, 然后再对终端节点的关键字进行处理。

在终端节点中删除一个键值后, 使得该节点的值个数  $n$  减1, 此时应分以下三种情况进行处理:

a) 若删除后节点中键值数目  $n \geq \lceil m/2 \rceil - 1$ , 在该节点中删去键值  $K$  连同右边的指针。

b) 若删除后节点中键值数目  $n < \lceil m/2 \rceil - 1$ , 且左(或右)兄弟节点的关键字数目大于  $\lceil m/2 \rceil - 1$ , 则把左(或右)兄弟节点中最大(或最小)键值移到父节点中, 再把父节点大于(或小于)上移键值的键值下移到被删关键字所在节点中。

c) 若删除后节点中键值数目  $n < \lceil m/2 \rceil - 1$ , 及其左、右兄

弟节点的键值数目都等于  $\lceil m/2 \rceil - 1$ , 则就必须进行节点的“合并”, 即把应删的键值删去后, 将该节点中的剩余键值和指针连同父节点中指向该节点指针的左边(或右边)一个键值  $K_i$  一起合并到左兄弟(或右兄弟)节点中, 将  $K_i$  从父节点中删去。如果因此使父节点中关键字数目小于  $\lceil m/2 \rceil - 1$ , 则对此父节点作同样处理, 以致于可能直到对根节点作这样的处理而使整个树减少一层。

以图3中所表示的4阶B-树为例。假设需依次删除关键字64、78。首先通过查找得到关键字64所在节点为  $E_{(6)}$ , 删除后节点中键值数目大于1, 因此直接删除。通过查找关键字78在非终端节点, 因此首先通过切割和连接反应将其与左子树最大关键字元素53对调。这样形成新的  $E_{(2)}$  (43, 53) 和  $E_{(6)}$  (47, 78), 再删除78, 因为删除后节点中键值数目等于1, 因此, 同样直接删除。最后结果如图5所示。

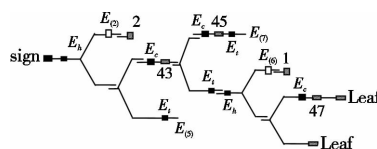


图5 删除关键字元素64

## 3 结束语

本文采用一种3D结构的DNA分子, 构建了用于文件系统及查找的B-树数据库系统。基于DNA计算的B-树的查找和插入等操作具有可扩展性及高度并行性等优势, 计算过程中需要借助的生物实验技术成熟, 具有一定可行性, 但操作过程中存在精确性不高的缺点。此外, DNA分子三维结构稳定性较差, 影响DNA分子生物实验结果的精确性, 这些问题有待解决, 也是DNA计算的未来研究热点之一。本文提出的基于DNA分子3D结构的计算模型和分治策略, 具有一定的可扩展性和并行性, 可用于DNA计算的其他模型。

## 参考文献:

- [1] ADLEMAN L M. Molecular computation of solutions to combinatorial problems[J]. Science, 1994, 266(11): 1021-1024.
- [2] QI Ou-yang, KAPLAN P D, LIU Shu-mao, et al. DNA solution of the maximal clique problem[J]. Science, 1997, 278(17): 446-449.
- [3] 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(4): 1063-1070.
- [4] 许进, 张社民, 范月科, 等. DNA计算机原理、进展及难点(III): 分子生物计算中的数据结构与特性[J]. 计算机学报, 2007, 30(6): 869-880.
- [5] 李汪根, 丁永生. DNA计算机中队列数据结构的设计及实现[J]. 计算机学报, 2007, 30(6): 993-998.
- [6] 李汪根, 丁永生, 任立红. DNA计算机中广义表数据结构的设计与实现[J]. 计算机学报, 2008, 31(12): 2215-2219.
- [7] 朱雅莉, 李肯立. DNA计算机中堆栈数据结构的设计[J]. 计算机工程与科学, 2008, 30(4): 121-123, 127.
- [8] 严蔚敏, 吴伟民. 数据结构[M]. 北京: 清华大学出版社, 2004.
- [9] JONOSKA N, KARL S A, SAITO M. Three dimensional DNA structures in computing[J]. Bio Systems, 1999, 52(1-3): 143-153.
- [10] PAUN G, ROZENBERG G, SALOMAA A. DNA计算: 一种新的计算模式[M]. 许进, 王淑栋, 潘林强, 译. 北京: 清华大学出版社, 2004.
- [11] 张成, 杨静, 王淑栋. DNA计算中荧光技术的应用及其发展[J]. 计算机学报, 2009, 32(12): 2300-2310.