甲骨文检索的粘贴 DNA 算法

栗青生,杨玉星
LI Qing-sheng,YANG Yu-xing

安阳师范学院 计算机与信息工程学院,河南 安阳 455002
School of Computer and Information Engineering,Anyang Normal University,Anyang,Henan 455002,China
E-mail:aylqs@163.com

LI Qing-sheng,YANG Yu-xing.Sticker DNA algorithm of Oracle-Bone Inscriptions retrieving.Computer Engineering and Applications,2008,44(28):140-142.

Abstract: In order to research and protect Oracle-bone Inscriptions better,a DNA coding method suited to DNA computer was designed;and a sticker DNA algorithm of Oracle-bone Inscriptions retrieving was proposed.The coding method of to-be-retrieving Oracle-bone Inscriptions and the standard Oracle-bone Inscriptions library was complementary,so the bio-chemical operations could execute better.The simulation results indicated that the algorithm was feasible and validated.
Key words: Oracle-Bone Inscriptions;DNA computing;sticker model;DNA coding

摘 要: 为了能更好地研究和保护甲骨文,设计了一种适合 DNA 计算机的甲骨文编码方式,并据此提出了进行甲骨文检索的粘贴 DNA 算法。根据 DNA 双链分子具有双螺旋结构的特性,甲骨文标准字库的编码和待检索文字的编码采用了互补的方式,以利于生化操作的执行。仿真结果表明该算法具有可行性和有效性。
关键词: 甲骨文;DNA 计算;粘贴模型;DNA 编码
DOI: 10.3778/j.issn.1002-8331.2008.28.047 **文章编号:** 1002-8331(2008)28-0140-03 **文献标识码:** A **中图分类号:** TP301.6

1 引言

甲骨文是我国历史文化长河中的一支奇葩,是目前所发现的世界上最古老的文字。为了保护历史文化遗产以及探究中华文字的起源问题,对甲骨文的研究方兴未艾,特别是计算机的介入,使得对甲骨文的识别及编码技术的研究成为一个热点。文献[1]提出对甲骨文进行线性编码;文献[2]使用 26 个字母和 10 个数字对甲骨文进行编码,探讨了象形码输入法的原理;文献[3]使用可视化的方法检索甲骨文,设计了一种可视化的甲骨文输入法。然而,晶体管计算机的发展始终受电子元件物理极限的制约,其发展将在不远的将来达到这个极限值,届时,晶体管计算的发展将不再遵循摩尔定律。因此,科学家正探索一种新的计算机——DNA 计算机,这种新型的计算机具有运算的并行性高、信息存储量大、能耗低等优点,将来有望得到推广和使用。但目前的 DNA 计算机还属于专用型计算机,仅限于处理 NP 问题、数据加密问题等,如果将来能够使用 DNA 计算机对甲骨文进行研究,不仅对甲骨文的研究与保护起到非同寻常的作用,也是对 DNA 计算机应用的一个推广。

本文对甲骨文标准库进行 DNA 编码,对已经能够识别并释义的文字作为一个编码库(编码字库 A)。另外,本文基于粘贴 DNA 模型提出了一种对新发现甲骨文字的识别算法,若不是标准库中已有的文字,则将其加入到一个新的字库(编码字库 B)中,供相关专家进行研究,若是字库中已有的文字,则在字库中检索出对应的文字。

为了便于理解甲骨文标准库的 DNA 编码方案及甲骨文的 DNA 检索算法,下面首先对粘贴 DNA 模型的基本知识进行简单介绍。

2 粘贴 DNA 模型

在粘贴 DNA 模型中用单、双链 DNA 分子进行编码,分别对应于传统计算机的 0 和 1。在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。例如,子链数为 3,子链长度为 6 个碱基的位串:

CCAGCA AACGTT GTCGAT
GGTCGT CAGCTA

对应的二进制串是 101。

粘贴模型在位串上定义了四种基本操作[4]:

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

基金项目: 河南省社科规划项目(No.2006FLS007)。
作者简介: 栗青生(1966-),男,副教授,硕士生导师,研究方向:智能计算、多媒体数据智能处理;杨玉星(1981-),男,助教,研究方向:DNA 计算、生物计算。
收稿日期: 2007-11-21 **修回日期:** 2008-02-27

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”的位串(即，DNA 链)的集合。广义的分离操作是指根据若干连续的位元的状态将存储合成物的集合 T 分解为两个集合：$+(T,i,j)$ 和 $-(T,i,j)$，其中 $+(T,i,j)$ 为从第 i 个位元起的 j 个位元均为指定状态的位串的集合为“1”的位串的集合，$-(T,i,j)$ 为至少有一位不是指定状态的位串的集合。

设置(Set)：对存储合成物的集合 T 中的所有位串的某一特定位置 i 的位元，$Set(T,i)$ 的含义是：若其状态为“0”，将其状态设置为“1”。

清除(Clear)：对存储合成物的集合 T 中的所有位串的某一特定位置 i 的位元，$Clear(T,i)$ 的含义是：若其状态为“1”，将其状态设置为“0”。

各种操作的物理实现方法可参见文献[5]，模拟实现方法参见文献[6]。

基于粘贴模型的计算模式就是将问题的所有可能解用位串来编码，得到一个“数据池”，对该数据池中的位串通过上述操作的某一种或者几种操作的排列组合，筛选出结果串，如果结果串为空，则表明问题无解。

3　编码与检索

3.1　甲骨文标准库的 DNA 编码

要使用粘贴 DNA 模型进行甲骨文的检索，首先要对标准库进行编码，编码的原则是重码少、长度短。

步骤 1　对甲骨文进行标准化处理。即，求取该甲骨文的最小外接圆(关于求解最小外接圆的方法，文献[7]做了较为详细的介绍，本文不再赘述)。

步骤 2　将该圆形区域作为图片处理，网格化该图片后即可采集编码点。

网格化采点的方法有：正方形网格化采点，同心圆网格化采点，极坐标网格化采点等。

由甲骨文的结构可以得知，采用正方形的采点方式进行编码易产生重码，而极坐标网格化采点的方法又过于烦琐，故本文使用同心圆网格化采点的方法进行编码，下面将介绍这种采点编码的方法。

设标准化后的甲骨文图片的半径为 r，以该图片的几何中心为原点作平面直角坐标系，并以 $i*r/n$ 为半径画 n 个同心圆(这里，$i=1,2,\cdots,n$)，坐标轴与同心圆的交点即可作为采集点(坐标原点也作为一个采集点)。

步骤 3　对采集点编码。若采集点与文字的线条相交，则以 P_j^{T}(其中，$j=1,2,\cdots,4n+1$)编码，否则，以 P_j^{F} 编码。由于目前已经识别并能准确释义的文字只有一千多个，因此，当 n 取 3 时，足可对这些文字进行编码(此时，可编码的甲骨文的理论数量为 2^{13}，即 8 192 个)。

每个采集点采用 10 个碱基进行编码，各采集点的编码如下：

P_1^{T}=TTACACCAAT，P_1^{F}=CCTACAATTC；

P_2^{T}=CCAACATACT，P_2^{F}=ACCTAATACT；

P_3^{T}=TCCTCTAACA，P_3^{F}=CCCTACAATC；

P_4^{T}=ACTCCACTTA，P_4^{F}=TTATAACTTC；

P_5^{T}=AACCACAAAC，P_5^{F}=CAATACCACC；

P_6^{T}=CCAATAACCT，P_6^{F}=ATACACTTAC；

P_7^{T}=ACCCCAATAA，P_7^{F}=CTCATACTAC；

P_8^{T}=CTATTTCACC，P_8^{F}=TATTCTCACC；

P_9^{T}=ACACCTAACT，P_9^{F}=ATCAACATCA；

P_{10}^{T}=CTATTCTACT，P_{10}^{F}=CCTTTACCTC；

P_{11}^{T}=ATCTTTCCCC，P_{11}^{F}=AAATAACATT；

P_{12}^{T}=TCCATTTCTC，P_{12}^{F}=CACCCCTATA；

P_{13}^{T}=TTCCATGGGA，P_{13}^{F}=GGAGACAGAT。

为了标识是否具有重码，在这 130 个碱基组成的单链之前再链接上一个位段，若有重码则用一个特殊的 DNA 单链 P_0^{T}=CCCCCCCCC，若无重码，则用 P_0^{F}=TTTTTTTTTT 标识。故，一个甲骨文将最终由一个长度为 140 个碱基的 DNA 单链来编码。例如，甲骨文的“人”字经过标准化处理及编码点采集后如图 1 所示。

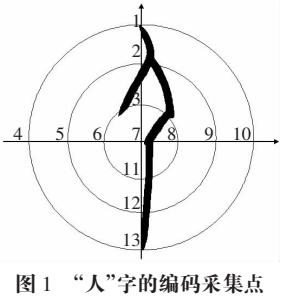

图 1　“人”字的编码采集点

由图 1 可知，采集点 1、2、12、13 与文字相交，故这些点的编码取 P_j^{T}($j=1,2,12,13$)，其余各采集点的编码取 P_j^{F}($j=3,4,\cdots,11$)，没有和该字重码的甲骨文字，故重码识别位取 P_0^{F}=TTTTTTTTTT，“人”字的 DNA 编码为：TTTTTTTTTTTTACACC-AATCCAACATACTCCCTACAATCTTATAACTTCCAATACCAC-CATACACTTACCTCATACTACTATTCTCACCATCAACATCAC-CTTTACCTCAAATAACATTTCCATTTCTCTTCCATGGGA。

按照前述的三个步骤对甲骨文的标准字库进行编码，得到甲骨文标准字库(字库 A)的 DNA 编码，将其放入一个大试管 T_0 中。

3.2　待检测甲骨文的 DNA 编码

待检测甲骨文的编码方式与标准字库的编码方式略有不同。

将采集到的甲骨片上的甲骨文拓下来，用扫描仪扫描拓片，将拓片加以标准化处理，利用 3.1 节所述的方法采集编码点。

为了利用 DNA 的双螺旋结构进行甲骨文的检测，待检测文字的编码方式如下：

若采集点 j 与文字交叉，则该点的编码为 P_j^{T} 的 DNA 补码(记作 $\overline{p_j^{\mathrm{T}}}$)，即各碱基通过碱基互补配对原则的补码 $\overline{p_1^{\mathrm{T}}}$；若采集点 j 与文字不相交，则该点的编码为 P_j^{F} 的 DNA 补码(记作 $\overline{p_j^{\mathrm{F}}}$)。例如，若第一个采集点与文字相交，则该待检甲骨文的第一个采集点的编码为 $\overline{p_1^{\mathrm{T}}}$=AATGTGGTTA。这样的编码使得人们能够

通过粘贴 DNA 模型的四种基本操作很方便进行甲骨文的检测,下面介绍检索算法。

3.3 检索算法

将采集到的待检索的甲骨文 k 使用 3.2 节所述的方法进行编码,得到单链 DNA,记作 $code(k)$;并设试管 T_1 中存储的为新发现的甲骨文字的 DNA 编码。下面对编码字库 A(试管 T_0)和编码字库 B(试管 T_1)的 DNA 链进行操作,给出检索字 k 的粘贴 DNA 算法如下:

```
Retrieving(k)
{
    Separate+(T0,2,13) and -(T0,2,13);//其中,从第 2 个位元起的 13 个位元的指定状态与 code(k)的对应位段状态相同。
    if(+(T0,2,13))
    {
     T2=+(T0,2,13);
     Read(T2);//读出 T2 中的 DNA 链,找出对应的文字。
     }
     else
     {Separate +(T1,2,13)and -(T1,2,13);//其中,从第 2 个位元起的 13 个位元的指定状态与 code(k)的对应位段状态相同。
     if(! +(T1,2,13))
     {
       T3=Joint(P0^F, code(k));//拼接两个 DNA 链。
       T1=Merge(T1,T3);//将 k 加入到字库 B 中。
     }
     else
     Read(+(T1,2,13));
   }
}
```

上述算法的物理实现方法如下:

将 $code(k)$作为探针附在多级分离装置[8]的栅层 L 上,通过聚合酶链接反应大量复制标准字库 T_0 的 DNA 编码,将 T_0 中的 DNA 溶液缓缓倒入多级分离装置中,若字库 A 中有相应的甲骨文字,则其编码将被栅层 L 捕获,用缓冲液冲洗栅层 L,加热解链进行检测即可读出结果。

若栅层 L 未捕获任何 DNA 链,则说明在标准字库 A 中检索不到待检文字 k,参照在字库 A 中进行检索的方法,在字库 B 中检索待检文字,若检索到,则读出,否则,以 $code(k)$为模板生成 $code(k)$的补链,并将其链接到 P_0^T(或 P_0^F,根据是否有重码选择)的链后,生成 k 的字库编码,加入到标准编码库 B 中。

4 仿真实例

本文在 Visual C++6.0 的环境下开发了一个仿真软件,对上述检测过程进行了仿真。下面以殷墟出土的一个甲骨文拓片上的文字的检索说明仿真过程。该甲骨拓片如图 2 所示。

图 2　一个待检索的甲骨文字拓片

将该拓片扫描并进行标准化处理后编码如下:AATGAGGTTAGGTTGTATGAGGGATGTTAGAATATTGAAGGTTATGGTGGTATGTGAATGGAGTATGATGATAAGAGTGGTAGTTGTAGTGGAAATGGAGTTTATTGTAAAGGTAAAGTGAAGGTACCCT。

以上述编码作为输入进行检索,检索结果截图如下(图 3)。

图 3　实例运行结果图

由图 3 得知,在标准字库 A 中检索到了和拓片上的文字相符的甲骨文,对应的 DNA 编码为:TTTTTTTTTTTTACACCAATCCAACATACTTACCTCATACTACTATTCTCACCATCAACATCACCTTTACCTCAAATAACATTTCCATTTCTCTTCCATGGGA。即甲骨文的“人”字。

此例是检索到结果并且没有重码的情况,若有重码,则会检索到多个结果供二次选择;若在标准库中检索不到对应的结果,则点击“将甲骨文 k 添加到字库 B”按钮,即可将新发现的文字添加到字库 B 中,供相关专家或学者研究。

5 结束语

本文提出的甲骨文 DNA 编码方案及相应的粘贴 DNA 检索算法可以有效地检索甲骨文字。由于 DNA 计算的高度并行性,该算法最多只需要两步分离操作即可完成,具有较高的执行效率。在为标准字库编码的过程中还存在少量的重码,这个问题可以通过增加采集点的方法得到进一步改善。

参考文献:

[1] 郝文勉.甲骨文编码的线性结构[J].郑州大学学报:哲学社会科学版,2005,38(1):87-92.

[2] 胡金柱,肖明.关于甲骨文象形码输入法的编码原理研究[J].计算机科学,2002,29(8):109-111.

[3] 刘永革,栗青生.可视化甲骨文输入法的编码与实现[J].计算机工程与应用,2004,40(17):139-140.

[4] 马季兰,杨玉星.基于粘贴模型的图顶点着色问题的 DNA 算法[J].计算机应用,2006(12):2998-3000.

[5] Ravinderjit S B,Chelyapov N,Johnson C,et al.Solution of a 20-variable 3-SAT problem on a DNA computer[J].Science,2002,296:499-502.

[6] 王毅鹏,付宇卓.DNA 计算虚拟生物实验室设计[J].计算机仿真,2006(11):280-283.

[7] 刘书桂,杨芳,陶晋.计算几何在测试计量技术中的应用-求解最小外接圆[J].工程图学学报,2000,21(3):83-89.

[8] 马季兰,杨玉星,孙承意.粘贴 DNA 模型的多级分离技术及其应用[J].计算机工程与设计,2007,28(13):3039-3041.