DNA 计算中编码序列的过滤函数研究

王子成,周 康,罗 亮,强小利
WANG Zi-cheng,ZHOU Kang,LUO Liang,QIANG Xiao-li

华中科技大学 控制科学与工程系 信息处理与智能控制重点实验室,武汉 430074
Key Laboratory of Image Processing and Intelligent Control,Department of Control Science and Engineering,Huazhong University of Science and Technology,Wuhan 430074,China

WANG Zi-cheng,ZHOU Kang,LUO Liang,et al.Research of filtering function on coding sequence in DNA computing. Computer Engineering and Applications,2008,44(32):10-11.

Abstract: Authors construct a coding function and bring forward an algorithm useful for DNA words design.By using the filtering function and the algorithm proposed in this paper,authors achieve DNA codeword which satisfies some combinatorial and thermodynamic constraints.The quality of DNA words is greatly increased,and the reliability of DNA computing is also increased.
Key words: DNA computing;DNA coding;coding filtering function

摘 要:构造了用于 DNA 编码序列过滤的函数,并给出了 DNA 序列编码的算法,采用该文设计的过滤函数和算法所得到的 DNA 编码序列,能够满足一定的组合约束条件,并满足一定热力学条件,大大提高了 DNA 编码字的质量,有利于提高 DNA 计算的可靠性。
关键词:DNA 计算;DNA 编码;编码过滤函数
DOI:10.3778/j.issn.1002-8331.2008.32.003 **文章编号:**1002-8331(2008)32-0010-02 **文献标识码:**A **中图分类号:**TP301

1 引言

20 世纪 50 年代,诺贝尔奖获得者 Richard Feynman 首次提出了在分子水平上计算的思想,阐述了分子计算的可能性。1994 年,Adleman 首次提出了 DNA 计算的概念,并将 DNA 计算用于解决 7 个顶点有向 Hamilton 路有向图问题[1]。从此以后,DNA 计算的思想引起了全球范围内许多学者的关注。

1995 年,Lipton 又将 DNA 计算用于解决 3-可满足性问题[2],1997 年 Ouyang 等人提出了用于解决图论中最大团问题的 DNA 计算模型[3],随后,用于解决最大独立集问题[4],最小覆盖问题[5]等的 DNA 计算模型被提出来,2002 年,Ravinderjit 又提出了 20 个变量的 3 可满足性问题的 DNA 计算模式[6]。

作为一种新的计算模式,DNA 计算的出现,是计算科学史上的里程碑事件,在全球范围内,引起了广泛关注。DNA 计算在解决 NP 问题上具有独特的优势,显示出了其强大生命力,相对传统的硅基电子计算机而言,DNA 计算机显示出其无法比拟的优越性:强大的信息储存能力,DNA 作为信息的载体其贮存的容量非常之大;运算速度快,DNA 计算能力具有高度的并行性;DNA 计算耗能低,所消耗的能量只占一台电子计算机完成同样计算所消耗的能量的十亿分之一。

DNA 计算是以 DNA 分子为工具,将实际的问题编码成 DNA 分子,并利用 DNA 分子进行信息的编码和信息处理。DNA 计算的核心思想是 DNA 分子 Watson Crick 模型,它利用了 DNA 计算的特异性杂交性质。

DNA 计算过程可分为三个阶段:(1)编码阶段:即将待解决的问题通过编码,映射为 DNA 分子的集合;(2)计算过程:即生化反应过程,在编码 DNA 分子间进行各种可能的生化反应,即杂交过程,反应结束后,生成解空间的过程;(3)解的检测过程:在计算结果的解空间中,运用分子生物技术,如:PCR(Polymerize Chain Reaction)技术,POA(Parallel Overlap Assembly)技术,分子纯化,电泳,磁珠分离等,借助于这些生物操作手段,从生成的解空间中提取出正确解的过程。

2 DNA 计算中的编码问题

DNA 计算模式中,信息的编码单元定义在字母表{A,C,T,G}上,DNA 分子的杂交反应的核心法则是 Waston-Crick 碱基互补匹配原则,由于杂交反应受实验环境的影响比较明显,在生化反应过程中,两条编码 DNA 序列之间在碱基没有完全匹配的情况下也可以发生,出现假阳性现象,从而形成多种所不希望的二级结构,出现了非解结构,不利于 DNA 计算的正确解的生成。在结果检测阶段,同样在引物之间也会出现不希望的假阳性现象。

因此,将实际的问题映射为 DNA 计算模型后,如何有效快速地对 DNA 序列进行编码,而又能够保证随后的 DNA 计算的过程中的生化反应能够顺利可靠进行,保证 DNA 计算的可靠性,以利于生成正确的解空间,并有助于最终正确解的检测过程,是 DNA 计算中非常重要的问题,也是一个重点和难点问题。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60533010,No.60574041,No.60674106)。
作者简介:王子成(1976-),男,博士研究生,目前从事图论与智能计算,DNA 计算等领域的研究工作;周康,罗亮,强小利,博士研究生。
收稿日期:2008-06-10 **修回日期:**2008-08-27

研究 DNA 序列编码的目的就是希望经过编码的信息单元，即每一条编码的 DNA 序列在生化反应的过程中能够尽可能地被唯一识别,以保证在生化反应过程中,完全互补的两条编码 DNA 序列之间能够产生特异性杂交,避免非互补的 DNA 单链间的非特异性杂交,以促使生化反应尽量按照实验所设计的方案进行,以便提高 DNA 计算的可靠性与正确性。

以往有关 DNA 序列编码问题的研究从组合约束和热力学约束的角度对 DNA 编码问题进行了研究，保证了编码序列中的 G、C 含量，也降低了编码序列之间的相似度，有助于提高 DNA 编码序列的可靠性,提高对 DNA 计算的可靠性。

本文对 DNA 编码问题进行了研究,将编码 DNA 序列映射为实数空间上的向量,设计了用于 DNA 编码序列过滤的函数，以便最终所得到的 DNA 编码序列具有更高的质量，经过滤函数过滤后的 DNA 编码序列内的碱基分布更均匀，最终所得到的 DNA 编码序列既能满足组合约束条件,又能满足编码 DNA 序列的热力学条件,使其具有更高的热力学稳定性,有助于避免生化反应过程中的假阳性现象,提高了 DNA 编码序列的质量。

3　过滤函数设计

DNA 计算的生化反应过程中,互补的两条 DNA 单链之间可以经过杂交反应形成 DNA 双链,其中完全互补的 DNA 单链间的杂交反应称为特异性杂交，这是 DNA 计算中所希望出现的杂交反应，经特异性杂交反应所形成的 DNA 双链处于较低的能态,具有稳定的热力学结构。相对而言,部分互补的 DNA 单链间也可以产生杂交现象,从而在生化反应过程中会出现所不希望的假阳性现象,这是 DNA 计算中应该力求避免的反应,由假阳性杂交反应而形成的 DNA 双链处于部分匹配的状态,其自由能变化比较低,所形成的 DNA 双链处于较高的能态,其具有很不稳定的热力学结构。

DNA 分子存在于所有的细胞有机体中,DNA 分子由两条方向相反且彼此互补的 DNA 单链序列相互缠绕而成，经过特异性杂交而形成稳定的双螺旋结构，每条 DNA 双链由嘌呤碱(A,G)和嘧啶碱(T,C)经过磷酸二酯键的结合组成,根据 Watson-Crick 碱基互补原则,碱基 A 与 T 之间相互匹配,C 与 G 之间相互匹配。

定义函数,

$$f(x)=\begin{cases}-2, x=A\\-2, x=T\\-3, x=G\\-3, x=C\end{cases}$$

式中,变量 x 为任意一个碱基,取值范围为字母表$\{A,T,G,C\}$。

定义函数:

$$E(x_i,x_{i+1})=f(x_i)f(x_{i+1})=\begin{cases}4,(x_i,x_{i+1})\in\{(A,T),(T,A),(A,A),(T,T)\}\\9,(x_i,x_{i+1})\in\{(G,G),(G,C),(C,G),(C,C)\}\\6,(x_i,x_{i+1})\in\{(A,G),(A,C),(T,G),(T,C),\\\qquad(G,A),(C,A),(G,T),(C,T)\}\end{cases},i=1,2,\cdots,n-1$$

构造实空间的向量 $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_{n-1})$,其中 $\alpha_i=E(x_i,x_{i+1})$,$i=1,2,\cdots,n-1$。

借助于函数 $f(x)$和 $E(x_i,x_{i+1})$,$i=1,2,\cdots,n-1$,可以将长度为 n 的编码 DNA 序列 $X=(x_1,x_2,\cdots,x_n)$,其中 x_i 定义在字母表$\{A,T,G,C\}$上,映射为实空间中的 $n-1$ 维实向量 $\boldsymbol{\alpha}$。如:DNA 编码序列 5′-ATGGAGCTTC-3′可以映射为向量 $\boldsymbol{\alpha}=(4,6,9,6,6,9,6,4,6)$,编码 DNA 序列的补链 3′-TACCTCGAAT-5′也可以映射为向量 $\boldsymbol{\alpha}=(4,6,9,6,6,9,6,4,6)$,其反链 3′-CTTCGAGGTA-5′可以映射为向量 $\boldsymbol{\alpha}'=(4,6,9,6,6,9,6,4,6)$。

定义函数 $M(\alpha,\beta)=\sum_{i=1}^{n-1}E(x_i,x_{i+1}),x_i,x_{i+1}\in\{A,T,G,C\},i=1,2,\cdots,n-1$。式中 α、β 为 DNA 编码序列的集合中任意不同的两条 DNA 编码序列所对应的实空间中的向量。

根据最近邻模型,在 DNA 计算中单链 DNA 进行杂交反应时,完全互补的 DNA 两条单链 DNA 之间会发生特异性杂交反应,杂交反应前后 DNA 编码序列的自由能变化比较大,从而有利于形成具有热力学稳定性结构的 DNA 双链。

借助于函数 $M(\alpha,\beta)$,设计算法如下,采用该算法可以得到满足一定条件的 DNA 编码序列，提高了 DNA 编码字的质量,又助于 DNA 计算的顺利进行。

4　算法设计

若待设计的 DNA 编码序列的长度为 n，编码字的设计算法如下:

步骤 1　随机产生长度为 n 的 DNA 序列集合 S,共 4^n 条;

步骤 2　计算 DNA 编码序列集合 S 中的任意两条 DNA 编码序列之间的 $M(\alpha,\beta)$,$M(\alpha,\beta)=\sum_{i=1}^{n-1}E(x_i,x_{i+1}),x_i,x_{i+1}\in\{A,T,G,C\}$,$i=1,2,\cdots,n-1$,保留 $M(\alpha,\beta)$值在$[6n,7(n-1)]$之间的序列;

步骤 3　选择 DNA 编码序列集合中的任意一段长度为 5 的子序列,计算相应的 $E(x_i,x_{i+1})$,以保证其取值范围在[22,42]之间;

步骤 4　检测集合 S 中的 DNA 编码序列,删除含有特定子序列的 DNA 编码序列;

步骤 5　检测 S 中的 DNA 编码序列，以保证编码 DNA 序列间的汉明距离大于 $n/2$;

步骤 6　删除集合 S 中的 DNA 编码序列中含有回文子序以及能够形成发卡结构的 DNA 编码序列;

步骤 7　确定最终的 DNA 编码字集合 S。

5　实例

本文采用上述算法，构造了长度为 10，汉明距离为 6 的 DNA 编码序列,最小滑动错配长度为 4,最小相同子序列为 3,同时保证编码 DNA 序列中没有出现特定的子序列和回文现象,共得到 DNA 编码序列 7 条,见表 1 所示。

表 1　n=10 的编码 DNA 序列

DNA 序列	DNA 序列
AACGAAGACA	ACCTATCCAG
AACTAAGGAC	ACGCACTCAC
AAGTAATGCC	AGAGCGACTG
AATACACCGA	

6　结果与讨论

在 DNA 计算中,根据 Watson-Crick 碱基互补匹配原则,两条能够形成特异性杂交的 DNA 单链间，在经过特异性杂交后形成的 DNA 双链的过程中具有较大的自由能变化，从而使杂

(下转 21 页)