并行型 Ramsey 数 DNA 计算模型

许 进[1),2)] 范月科[1)]

[1)](华中科技大学分子生物计算机研究所 武汉 430074)

[2)](北京大学信息科学技术学院高可信软件技术教育部重点实验室 北京 100871)

摘 要 求解 Ramsey 数的困难在于需要搜索的解空间太大,而传统的电子计算机无法在有效的时间和存储空间上进行求解.由于 DNA 计算具有巨大的并行性和高密度存储能力等优点,文中研究了 Ramsey 数的 DNA 计算模型.针对传统的 Ramsey 数 DNA 计算模型存在的 DNA 序列量过多和序列过长的不足,利用 DNA 分子的特性以及生物操作将非解尽可能较早地消除,提出了并行型 Ramsey 数 DNA 计算模型,并以 $R(3,10)$ 为例,给出了具体的求解步骤.

关键词 并行型;DNA 计算;Ramsey 数

中图法分类号 TP301 **DOI 号**: 10.3724/SP.J.1016.2009.02320

The Parallel Type of DNA Computing Model for Solving Ramsey Number Problem

XU Jin[1),2)] FAN Yue-Ke[1)]

[1)](*Institute of Biomolecular Computer, Huazhong University of Science and Technology, Wuhan* 430074)

[2)](*Key Laboratory of High Confidence Software Technologies of Ministry of Education, School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)

Abstract The difficulty of solving the Ramsey number is that the solution space is too large to solve by traditional computer in effective time and storage space. Moreover, for the traditional DNA computing model, lots of oligonuleotides should be designed and generated much longer DNA sequences which are not convenient for bio-operation. This paper proposes a DNA computing model for Ramsey number based on the enormous parallelism and high-density storage capacity of DNA molecules. The advantage of this model is that many false solutions could be deleted as early as possible. Finally, the authors take $R(3,10)$ as an example and give the concrete steps for solving the problem.

Keywords parallel type; DNA computing; Ramsey number

1 引 言

Ramsey 理论[1-2]是组合数学中的一个重要分支,在通信、计算机信息检索和决策学等方面有一系列的具体应用,近年来在与其它数学分支的相互渗透中得到了迅速发展.经典 Ramsey 数是 Ramsey 理论[3-4]的重要研究对象.

DNA 计算发展已有十多年[5-6],它的优势是利用 DNA 分子具有海量的存储能力及生化反应的巨大并行性等特点进行计算.理论分析证明,未来的 DNA 计算机在信息存储和运算速度等方面是电子计算机无法比拟的[7-9].随着分子生物学技术的发展,其研究前景越来越被看好.它打破了传统意义上的计算机概念,为计算机科学的发展提供了新的思路.

收稿日期:2009-11-04.**许 进**,男,1959 年生,教授,博士生导师,研究兴趣为 DNA 计算机、神经网络、遗传算法、图论等. E-mail: jxu@pku.edu.cn. **范月科**,男,1959 年生,博士,研究兴趣为 DNA 计算和 DNA 计算机、图论等. E-mail: yuekefan@sohu.com.

借助于电子计算机在求解经典 Ramsey 数模型的研究,Mckay 等人作出了杰出的贡献[10-11].他们所建立模型的基本思想是利用图的自同构群来删除同构子图,在进一步构造所需的图集合时,一般通过每次增加一个顶点来完成.即使如此,一下子增加的非解也很多,从而导致电子计算机超过一定的范围就无能为力.这也是目前利用电子计算机求解 Ramsey 数停滞不前的根本原因所在.

到目前为止,很多研究者利用 DNA 计算对不少求解组合优化中 NP-完全问题的 DNA 计算模型的图论中的问题展开研究[12-14],DNA 计算用于 Ramsey 数的研究,能够建立有效求解 Ramsey 数的 DNA 计算模型[15-16],为 Ramsey 数的研究提供新的方法和理论支持.

对于 DNA 计算机而言,也存在着下述困难:(1)由于一般需要搜索 Ramsey 数图的顶点数较大,一般为 40 个以上,因此按照位序列方法,需要的位数近 800 个,从而需设计的 DNA 分子数目要 1600 个左右.因而,每个 DNA 链的长度必在 100bp 以上,因此,总链的长度约为 80000bp.这么长的 DNA 序列对当前的生物操作来说,实在是太困难了!(2)若按照位序列模型,需要的长度为 80000bp 的 DNA 序列不同的数目约 2 的 800 次方,这个量显然不可能实现!因此,纯粹采用上述位序列 Ramsey 数 DNA 计算机模型是不可能用于求解 Ramsey 数的.因此我们必须克服这两个问题,方能设计用于求解 Ramsey 的 DNA 计算机.

仔细分析导致上述问题的根本原因,核心问题只有一点,就是大量的非解!因此,如何巧妙地利用 DNA 分子的特性以及生物操作的优势来尽可能把非解"消除"在"萌芽状态",应该是 DNA 计算机设计中的一个关键性的问题.在文献[15－16]的基础上,本文提出了并行型 Ramsey 数 DNA 计算机模型,其基本思想是将位序列进行分段、然后分别删除非解,再逐步逐位进行合并.具体步骤如下:

1. 对 L_q 序列进行分段;

2. 合成每个段位的可能构成的初始解空间;

3. 从每个段位的初始解空间中删除所有可能的非解;

4. 依次相继合成每个段位:首先合成第一与第二段位,并删除构成的所有可能的非解;然后合成新段位与第三个段位,并删除所有的非解,如此类推,直到合成最后一个段位,并删除所有的非解.最后所得到的解即为问题的最终解;

5. 检测最终解.

其中,有关记号与概念见文献[15-16].

在本文最后,我们也将以求解 $R(3,10)$ 为例来说明并行型 Ramsey 数 DNA 计算机模型.

2　求解 Ramsey 数的基本方法思路

求解 Ramsey 数,特别是经典 Ramsey 数问题是当今组合数学乃至整个数学界最困难的问题之一.Ramsey 数已经成为研究的热点和难点.

关于求解 Ramsey 数的基本方法思路,我们将以求解 Ramsey 数 $R(3,10)$ 的方法给予说明.

业已知道,$40\leq R(3,10)\leq 43$.若在 40 个顶点集合没有 Ramsey 图,即在 G_{40} 没有一个既不含 K_3,也不含 N_{10} 的 Ramsey 图,则 $R(3,10)=40$;相反,若在 G_{40} 存在 Ramsey 图,需要继续考察图集 G_{41} 中来寻找 Ramsey 图,若不存在,则说明 $R(3,10)=41$;否则,说明 $R(3,10)\geq 41$;于是图集 G_{42} 中来寻找 Ramsey 图,若不存在,则说明 $R(3,10)=42$;否则,说明 $R(3,10)=43$.这就是一种通过搜索方法来求解 Ramsey 数的思路.在这里,我们虽然仅对 Ramsey 数 $R(3,10)$ 给予说明,但对求解其它 Ramsey 数的方法完全相同.

3　并行型 Ramsey 数 DNA 计算模型

本节将给出并行型 Ramsey 数 DNA 计算模型算法的具体方法步骤,也就是给出第 1 节里 5 个步骤,以 G_{40} 为例,给出具体生化操作的方法步骤分析等.细言之,我们首先给出了序列的分段及编码,给出了初始解空间的构建;然后依次相继合成每个段位,删除构成的所有可能的非解;最后,检测最终解.在文章中通过巧妙利用 DNA 分子的特性以及生物操作的优势尽可能把非解消除在"萌芽状态",使得 DNA 编码的数量和长度都得到了很大的减少,更加有利于 DNA 计算生物实验的操作.

3.1　序列分段

截至目前,Ramsey 数 $R(3,10)$ 的界仍是 $40\leq R(3,10)\leq 43$.按照上面求解 Ramsey 的思路,在顶点数为 40 的图集 G_{40} 中,对应的序列 L_q 的长度为 780,因此,可分解成 52 个长度为 15 位的段位.

这里要强调说明的是,长度可以不等,但至多为两种为宜.

分段的原则是:① 位数一般在 15 位,这样易于编码及生化操作;② 一般取相等长度为宜.

3.2　初始解空间构建

在确定好编码之后,按照一定的顺序连接,就可

以形成所有可能解的 DNA 链序列，这些被合成所有的 DNA 序列被称为库链，它们的集合被称为存储库或初始解空间.

初始解空间的构建需要围绕着如下几个步骤进行：

(1) 确定子图集的顶点数 p，从而得到了所有可能的 DNA 序列的片断数，并在此基础上确定每个 DNA 片断的长度；

为了实现并行处理的思想，在计算时采用并行处理的方法，需要进行预先分析，进行合理划分，确定图集的顶点数 p. 为了便于操作和实现并行处理，我们提出了几个划分原则：① 图集标定的起始点、终点的度数应该尽可能大，尤其是子图标定的起始点和终点的度数之和应该尽可能大；② 对于一个确定的子图，顶点排序是很重要的，排序的原则是：排序相邻的顶点之间在图中尽可能有边相连，当然，最好是排序后的顶点序列在图中是一个 Hamilton 圈；③ 子图的规模原则：一般子图的规模在 10～20 个顶点之间为宜.

(2) 编码及探针设计. 编码生成所有 DNA 片断，如 G_{40} 共需要 1560 个编码，并设计相应的探针，通过合成所需要的 DNA 序列，并用一定量的 T4 连接酶与缓冲液等；

DNA 计算模型的研究中，任何模型遇到的首要问题都是 DNA 计算的编码问题，编码的合理与否将直接关系到模型能否成功地被生化实验验证. 一般地，在 DNA 计算编码问题的研究中，主要有两个指标：编码的数量和质量. 编码的质量越高，计算的可靠性越高；编码的数量越大，解决应用问题的规模也就越大. 但是，这两个指标是相互矛盾的. 编码质量越高，编码的数量反而越小.

在序列分段的基础上，和我们已有的编码方法来看，由于每位需要用两个 DNA 片断来分别表示该位是红色和蓝色，因此，780 位共需要的 DNA 片断的数目共有 1560 个. 因此，为了得到足够多的，又能够保证生化反应顺利进行的编码，我们对编码条件及部分参数进行了调整. 每个片断的长度需要用约为 30bp 的碱基序列来表示，因此，最终的碱基个数为 23400bp.

不失一般性，我们以第一个段位为例给予说明，且假定位数为 15，该段位对应的编码如下所示：

$$\begin{array}{ccccccc} 1 & 2 & 3 & \cdots & 15 \\ r_1 & r_2 & r_3 & \cdots & r_{15} \\ b_1 & b_2 & b_3 & \cdots & b_{15} \end{array}$$

探针是为建立非枚举的、尽可能小的初始解空间而设计的. 探针的建立方法具体如下：

这里，我们令 $\overline{x_i x_{i+1}}$，$x \in \{r, b\}$，$i = 1, 2, \cdots, 15$，记为对应于每两个相邻的位 i 与 $i+1$ 之间的探针，则这样的探针共有 4 条，分别记作：

$$\overline{r_i r_{i+1}}, \overline{r_i b_{i+1}}, \overline{b_i r_{i+1}}, \overline{b_i b_{i+1}},$$

其中，每个探针由两个 DNA 片段构成. 探针的设计方法是：将代表第 i 位的 DNA 片段的后半部分和第 $i+1$ 位的 DNA 序列的前半部分的寡聚核苷酸序列合并，然后再对该序列取其 Watson-Crick 互补序列，这个互补的序列即为探针. 例如，顶点 1 与顶点 2 之间探针设计如下：根据顶点 1 和 2 的着色集合，若

$r_1 = 5'$-GAATTGGATATATGGAATCCCCGGT-ATATATAGCGCGGTTGCCTCCTCA-$3'$,

$r_2 = 5'$-TTTTAAAAGGGCCCCATATATAACC-GGATTTTAACGCGCGGGAAATTAT-$3'$.

则有

$r_1 r_2 = 5'$-ATATATAGCGCGGTTGCCTCCTCATTTTAAAAGGGCCCCATATATAACC-$3'$,

$\overline{r_1 r_2} = 5'$-ggTTATATATggggCCCTTTTAAAATgAggAggCAACCgCgCTATATAT-$3'$.

按照上面的探针生成方法，第一段位总共需要构建 56 个探针，其相应的形式化的探针给出如下：

$$\begin{array}{cccc} \overline{r_1 r_2}, & \overline{r_1 b_2}, & \overline{b_1 r_2}, & \overline{b_1 b_2} \\ \overline{r_2 r_3}, & \overline{r_2 b_3}, & \overline{b_2 r_3}, & \overline{b_2 b_3} \\ \overline{r_3 r_4}, & \overline{r_3 b_4}, & \overline{b_3 r_4}, & \overline{b_3 b_4} \\ \overline{r_4 r_5}, & \overline{r_4 b_5}, & \overline{b_4 r_5}, & \overline{b_4 b_5} \\ \overline{r_5 r_6}, & \overline{r_5 b_6}, & \overline{b_5 r_6}, & \overline{b_5 b_6} \\ \vdots & \vdots & \vdots & \vdots \\ \overline{r_{14} r_{15}}, & \overline{r_{14} b_{15}}, & \overline{b_{14} r_{15}}, & \overline{b_{14} b_{15}} \end{array}$$

所设计好的编码及相应的探针序列都通过生物公司进行合成.

(3) 合成解空间. 分别合成第一段位、第二段位，一直到最后一个段位的初始解空间.

一旦编码和探针合成好之后，就采用一定的生物技术手段，按照一定的顺序，利用探针将代表每个位的寡聚核苷酸编码连接形成解空间. 通常，合成的方法是：分别以代表每个段位的首位与末位的两个 DNA 片断为引物，并加入必须的探针与其它试剂等进行 PCR 扩增即可.

我们在合成解空间的时候，通过依次采用首先对 $5'$端磷酸化、退火、连接反应和 PCR 反应. 通过前

3 步反应,可以使得代表每个位的寡聚核苷酸编码连接形成一定长度的 DNA 链,此时为单链 DNA. 以这些 DNA 链为模板,分别以$\langle r_1,\overline{r_{15}}\rangle$、$\langle r_1,\overline{b_{15}}\rangle$、$\langle b_1,\overline{r_{15}}\rangle$、$\langle b_1,\overline{b_{15}}\rangle$为引物对,进行 PCR 扩增,得到双链 DNA,这些 DNA 链所构成的集合即为解空间. 此初始解空间应该包含有 2^{15}种 DNA 链.

最后采用 PCR 反应对所合成的解空间的完备性进行检测. 将所得到的 DNA 链作为模板,分别用$\langle r_1,\overline{x_i}\rangle$和$\langle b_1,\overline{x_i}\rangle$为引物对,其中 $x\in\{r,b\}$,$i=2,3,\cdots,15$,检测库链是否完备.

这里我们是以第一段位为例来阐述解空间的构建方法的,对于其它段位来讲,合成解空间的方法完全一致. 可在同时并行地对所有的段位进行解空间的合成,实现并行处理.

3.3　从初始解空间中删除所有的非解

在按照上述方法所构建的解空间中,还存在有大量的代表非解的 DNA 链,即这些 DNA 链所代表的子图中要么含有 K_3子图,要么含有 N_{10}子图. 因此,我们必须把这些非解的 DNA 链全部删除. 具体的操作方法是利用酶切反应来删除非解,这里我们仅以删除第一段位中的 K_3为例,对该方法在此模型中的应用给予说明.

第 1 步. 首先将得到代表第一段位解空间的 DNA 链平均分成 3 份,分别记为 T_1,T_2和 T_3,然后给相应的试管中加入相应的限制性内切酶,切断对应位上为 1 的 DNA 序列.

第 2 步. 采用琼脂糖凝胶电泳技术,将 T_1,T_2和 T_3中代表非解的被切断的 DNA 链,分离并回收剩余 DNA 链.

第 3 步. 将 T_1,T_2和 T_3中的剩余的 DNA 链合并至 T,则这些溶液中不含有 3 条边同时为 1 的 DNA 序列.

第 4 步. 将 T 重新分成 3 份,并继续标记为 T_1,T_2和 T_3,重复操作第 1 步～第 3 步,直到删除完所有的 K_3.

对于删除含有 N_{10}的非解来讲,操作方法与上述的方法完全一致,只是需要删除的是同时为 0 的情况.

3.4　解的检测

经过运算子系统后,所剩下的 DNA 序列就代表可能的 Ramsey 图. 我们利用 PCR 或者是 DNA 测序技术将所剩下的 DNA 链进行测序.

4　结　论

Ramsey 数的求解一直是组合优化问题研究中的难点和热点,广泛地被各国科学家所研究. 传统的电子计算机搜索解空间的方法在时间和存储空间上也显得无能为力. 发展新的计算工具求解 Ramsey 数显得尤其重要,而 DNA 计算由于具有巨大的并行性和高密度存储能力等优点,为我们提供了一种新的计算工具. 本文提出了 Ramsey 数的 DNA 计算模型;通过巧妙利用 DNA 分子的生物特性和生物操作,将大量非解消除在"萌芽"阶段,从而大大减少了 DNA 编码序列的长度和编码的数量,能有效地提高 DNA 生物实验操作和可靠性.

本文建立的这种并行型求解 Ramsey 数 DNA 计算模型,其基本思想是将一个给定图所转换的位序列进行分段;然后分别按照各个小段删除非解,再逐步逐位进行合并,并以 $R(3,10)$为例,给出了具体的求解步骤.

参　考　文　献

[1] Burr S A. Determining generalized Ramsey numbers is NP-hard. Ars Combinatoria, 1984, 17: 21-25

[2] Wang Qing-Xian. Ramsey number of applications in computer science. Journal of University of Information Technology, 1997, 16(1): 1-5(in Chinese)
(王清贤. Ramsey 数在计算机科学中的应用. 信息工程学院学报, 1997, 16(1): 1-5)

[3] Radziszowski S P. Small Ramsey numbers. Electronic Journal of Combinatorics Dynamical Survey, 2006, 11: 3-36

[4] Tse K K. On the Ramsey number of the quadrilateral versus the book and the wheel. Australasian Journal of Combinatorics, 2003, 27: 163-167

[5] Adleman L M. Molecular computation of solution to combinatorial problems. Science, 1994, 266: 1021-1024

[6] Lipton R J. DNA solution of hard computation problems. Science, 1995, 268(4): 542-545

[7] Ouyang Q, Kaplan P D, Liu S et al. DNA solution of the maximal clique problem. Science, 1997, 278(17): 446-449

[8] Liu Q, Wang L, Frutos A et al. DNA computing on surfaces. Nature, 2000, 403: 175-179

[9] Braich R S, Chelyapov N, Johnson C et al. Solution of a 20-variable 3-SAT problem on a DNA Computer. Science, 2002, 296(5567): 430-434

[10] Mckay B D, Radziszowski S P. $R(4,5)=25$. Journal of Graph Theory, 1995, 19(3): 309-322

[11] Mckay B D, Zhang K M. The value of the Ramsey number $R(3,8)$. Journal of Graph Theory, 1992, 16(1): 99-105

[12] Pan L Q, Liu Y C et al. A surface based DNA algorithm for the maximal clique problem. Chinese Journal of Electronics, 2002, 11(4): 469-471

[13] Liu W B, Xu J. A DNA algorithm for the graph coloring problem. Journal of Chemical Information and Computers, 2002, 42(5): 1176-1178

[14] Liu Y C, Pan L Q, Xu J et al. DNA solution of graph coloring problem. Journal of Chemical Information and Computer Sciences, 2002, 42(3): 529-534

[15] Xu Jin, Fan Yue-Ke. Classical Ramsey number DNA computing model (I): Add-Bit-Sequence model. Chinese Journal of Computers, 2008, 31(12): 2073-2080(in Chinese)
(许进, 范月科. 经典 Ramsey 数 DNA 计算模型(I): 位序列计算模型. 计算机学报, 2008, 31(12): 2073-2080)

[16] Xu Jin, Fan Yue-Ke. Classical Ramsey number DNA computing model (Ⅱ): Add-Bit-Sequence DNA computing model. Chinese Journal of Computers, 2008, 31(12): 2081-2089 (in Chinese)
(许进, 范月科. 经典 Ramsey 数 DNA 计算模型(Ⅱ): 基于位序列的 DNA 计算模型. 计算机学报, 2008, 31(12): 2081-2089)

XU Jin, born in 1959, professor, Ph. D. supervisor. His research interests include DNA computing and DNA computer, neural networks, genetic algorithms, graph theory etc.

FAN Yue-Ke, born in 1959, Ph. D.. His research interest include DNA computing and graph theory.

Background

Classical Ramsey number problem belongs to NP complete problem, which will spend exponential time in solving by traditional electronics computer. It is necessary to study new computation methods because traditional electronics computer faces with greatly difficulty in solving NP complete problem. DNA computing posssesses hign parallelism in data and higher storage capacity than normal systems. Hence, in theory, it is feasible to solve NP complete problem with DNA computing. Up to now, many accomplishments have been achieved to improve its performance and increase its reliability.

In this paper, a parallel type DNA computing model is proposed and used to solve the classicla Ramsey number problem based on the enormous parallelism and high-density storage capacity of DNA molecules. The basic idea is to construct a graph by bit by site. Here, the authors take as an example and give the concrete steps for solving the problem. The advantage of this model is that many false solution could be deleted as soon as possible.