

基于闭环 DNA 的指派问题算法^{*}

周 康^{1,2} 同小军^{1,2} 许 进²

(武汉工业学院数理科学系 武汉 430023)¹ (华中科技大学控制科学与工程系 武汉 430074)²

摘 要 给出了闭环 DNA 计算模型及其生化实验。用闭环 DNA 计算模型设计出了指派问题的 DNA 算法。首先对决策变量进行二维 DNA 编码来存放决策变量和效益值,然后通过有目的的终止技术和删除实验得到指派问题的全部可行解,最后通过电泳实验和检测实验获得最优指派问题的最优解。举例说明了算法的可行性。最后,为减少 DNA 编码数量和缩短 DNA 编码的码长,讨论了算法的两种改进方法。

关键词 闭环 DNA 计算模型,指派问题,删除实验,有目的的终止技术

An Algorithm of Assignment Problem Based on Closed Circle DNA

ZHOU Kang^{1,2} TONG Xiao-Jun^{1,2} XU Jin²

(Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan 430023)¹

(Department of Control Science and Engineer, Huazhong University of Science and Technology, Wuhan 430074)²

Abstract In this paper, a model of closed circle DNA computing and its bio-chemistry experiments are brought forward. An algorithm based on closed circle DNA is designed for assignment problem. In the algorithm, first planar DNA encoding for each decision-making variable is encoded to deposit each decision-making variable and benefit value. Then all feasible solutions are gained with purposive ending technology and delete experiment. And then all optimization solutions are obtained with electrophoresis experiment and detect experiment. Feasibility of this algorithm is explained by an example. Finally two improving methods of this algorithm are discussed in order to decrease number of DNA encodings and to shorten encoding length of DNA encodings.

Keywords Model of closed circle DNA computing, Assignment problem, Delete experiment, Checking experiment

从 1994 年 Adleman^[1] 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 计算发展迅速^[2~8],各种新的 DNA 计算模型不断涌现。2006 年,周康、许进等人^[5,6] 提出了闭环 DNA 模型。该模型的闭环 DNA 分子的结构灵活性的特点已经解决了多个 NP-完全问题。因此,根据解决问题的需要选择不同结构的闭环 DNA 分子成为对闭环 DNA 模型研究的一个热点。

以往的初始闭环 DNA^[5] 为每个位置上有单一的等长 DNA 编码。为了存储更多的信息,本文首次在每个位置上采用二维 DNA 编码构成不等长的闭环 DNA 来存储指派问题的解的信息,推广了闭环 DNA 模型。另外,推广的闭环 DNA 模型的删除实验是通过将不满足要求的闭环 DNA 变成不可环化的 DNA 片段不可逆转地被删除,而基本闭环 DNA 模型的删除实验只是通过删除外源 DNA 来调节闭环 DNA 的长度。因此,推广的闭环 DNA 模型与以往的闭环 DNA 模型有本质区别。本文正是用推广的闭环 DNA 模型解决了指派问题。

1 闭环 DNA 计算模型

基本的闭环 DNA 由若干个限制性内切酶的识别序列在识别位点处首尾相连而成。设闭环 DNA 分子中有 $n+1$ 个位置存放识别序列,它们分别为 p_i ($i=0,1,2,\dots,n$),每个 p_i

的结构为 $s_i t_i$,其中 s_i 和 t_i 之间是限制性内切酶的识别位点。本文使用的闭环 DNA 计算模型以推广的闭环 DNA 为计算对象,推广的闭环 DNA 由若干个 DNA 序列首尾相连而成。其每个位置^[5]上的 DNA 序列的结构为 $p_i q_i$ ($i=1,2,\dots,n$),其中 p_i 为识别序列, q_i 为线性 DNA。开环的位置选择在 p_0 处。在 p_0 处开环后的线性 DNA 中 5' 端规定为闭环 DNA 的首端,其识别位点的位置为 1,另一端为末端,那么每个识别位点均对应一个位置,我们称第 k 个识别位点为 k 位识别位点,相应的限制酶为 k 位限制酶。

推广的闭环 DNA 模型为以推广的闭环 DNA 为物质基础,有删除、扩增、分离、合并、电泳、检测等生化实验的 DNA 计算模型。本文还使用了有目的的终止技术。

删除实验:在多种限制性内切酶作用下,除掉混合物中含有两种及以上的对应该识别位点的所有闭环 DNA。其生化操作过程:首先在多种限制性内切酶作用下,闭环 DNA 分子变成线性;然后,经过高效液相层析除去限制性内切酶,并在 T_4 DNA 连接酶作用下,部分 DNA 重新闭环;最后,经过凝胶电泳除去线性 DNA 和 T_4 DNA 连接酶。

分离操作:将闭环 DNA 混合物均匀、等量地分为两份闭环 DNA 混合物。

合并操作:将两种闭环 DNA 混合物均匀、充分地混合得到新的闭环 DNA 混合物。

^{*} 国家自然科学基金项目资助(编号:60403002);湖北省自然科学基金项目资助(编号:2005ABA233);湖北省优秀中青年科技创新团队计划项目资助;浙江省自然科学基金项目资助(编号:ZJNSF-Y105654)。周 康 副教授,博士生,主要研究方向为 DNA 计算、运筹学、现代优化算法、系统稳定性。

凝胶电泳:通过电泳反应将 DNA 分子按照链长进行分离。

聚合酶链反应(PCR 扩增):在 *Taq* DNA 聚合酶作用下,选择 3 个温度进行热循环,从而使 DNA 分子扩增。

检测实验:利用分子信标技术,通过标记和去标记实验检测 DNA 序列的过程。

有目的的终止技术^[3]:以杂交实验为核心利用芯片技术和保护基技术合成满足一定要求的线性 DNA 的合成技术。

2 指派问题及其数学模型

指派问题^[9]的提法是:某单位需完成 n 项任务,恰好有 n 个人可承担这些任务。各人完成任务不同,效率也不同。那么,如何进行指派,使总效率最高。

设决策变量为 x_{ij} ($i, j=1, 2, \dots, n$),当指派第 i 个人去完成第 j 项任务时, $x_{ij}=1$; 否则, $x_{ij}=0$ 。设第 i 个人完成第 j 项任务的效率(时间)为 c_{ij} ($i, j=1, 2, \dots, n$), $c_{ij} > 0$ 。如果要指派后的总时间最小,则最优指派问题的数学模型 1 为

$$\begin{aligned} \text{Min } z &= \sum_{i=1}^n \sum_{j=1}^n c_{ij} x_{ij} \\ \text{s. t. } &\begin{cases} \sum_{j=1}^n x_{ij} = 1 & i=1, 2, \dots, n \\ \sum_{i=1}^n x_{ij} = 1 & j=1, 2, \dots, n \\ x_{ij} = 1 \text{ or } 0 \end{cases} \end{aligned} \quad (1)$$

从模型 1 的约束条件的结构不难看出:设决策变量所组成的矩阵(简称:决策变量矩阵)为 $A=(x_{ij})_n$,则最优指派问题的解 x_{ij} ($i, j=1, 2, \dots, n$) 为可行解的充分必要条件为: A 中恰有 n 个数为 1,第 j ($j=1, 2, \dots, n$) 列恰有 1 个数为 1,第 i ($i=1, 2, \dots, n$) 行恰有 1 个数为 1。

算法分析:首先通过编码及 DNA 的合成技术产生满足行要求的闭环 DNA 序列,再通过批删除实验满足列要求,得到可行解。为了得到最优解将效率值编入 DNA 序列中,因此,采用二维的 DNA 编码。

3 DNA 计算的算法

3.1 DNA 计算的算法

设模型 1 中 $c_{ij} > 0$ 且为整数值。则指派问题的闭环 DNA 计算算法见图 1,具体步骤为:

步骤 1 对每个决策变量采用 2 维编码,生成决策变量组、夹板组、探针组。

步骤 2 通过有目的的终止技术^[3],合成由 n 个决策变量形成的闭环 DNA 的集合,构成满足行要求的初始解空间。

步骤 3 通过删除实验,生成可行解空间。

步骤 4 通过电泳实验,产生最优解。

步骤 5 检测实验结果。

3.2 DNA 计算算法的实现

对步骤 1,算法的生化实现过程为:

1. 产生三组 DNA 编码

决策变量组:

$3'-p_{ij}q_{ij}-5', 3'-p_{0i}p_{1j}q_{1j}-5' (i=2, \dots, n, j=1, 2, \dots, n);$

夹板组:

$5'-q'_{ki}p'_{(k+1)j}-3',$

$5'-q'_{ni}p'_{0j}p'_{1j}-3' (k=1, 2, \dots, n-1, i, j=1, 2, \dots, n);$

探针组:

$5'-q'_{ij}-3' (i, j=1, 2, \dots, n)。$

其中,决策变量 x_{ij} 的 DNA 编码 $p_{ij}q_{ij}$ 由两部分组成, p_{ij} 为识别序列,其长度为 6; q_{ij} 存储 c_{ij} 的值,其长度为 $15 \times c_{ij}$, p_0 为开环识别序列。夹板组中 p'_{ij} 、 q'_{ij} 为 p_{ij} 、 q_{ij} 的补序列。探针组用于标记实验,探针 q'_{ij} 称为第 (i, j) 号探针,并进行荧光素标记。

2. 合成三组 DNA 编码组的 DNA 分子并进行 PCR 扩增,得到足量的三组 DNA 混合物。

对步骤 2,算法的生化实现过程为:

1. 将生物素分子连接到 $p_0p_{1j}q_{1j}$ 的 3' 端,再将 DNA 片段固定在经过修饰后的表面上,制成 DNA 芯片;对 $q'_{ni}p'_{0j}p'_{1j}$ 的 3' 端以及 $p_{ni}q_{ni}$ 和 $q'_{1i}p'_{2j}$ 的 5' 端 ($j=1, 2, \dots, n$) 进行修饰保护,使其失去化学活性。

2. 取过量的决策变量 DNA 分子 $p_{ij}q_{ij}$ 、过量的夹板 DNA 分子 $q'_{ki}p'_{(k+1)j}$ 、 $q'_{ni}p'_{0j}p'_{1j}$ 和经过处理的 $q'_{ni}p'_{0j}p'_{1j}$ 、 $p_{ni}q_{ni}$ 和 $q'_{1i}p'_{2j}$ 与 DNA 芯片充分混合后,将混合物由 94°C 缓慢冷却至 68°C ,发生充分的杂交反应。

3. 取出 DNA 芯片,用 68°C 缓冲溶液冲洗干净后称重。

4. 将温度升至 94°C ,继续如 2 的杂交实验,取出 DNA 芯片,冲洗干净后称重。

5. 如果重量有增加,返回 4;如果重量没有变化,则从芯片上取出双链 DNA,并去掉双链 DNA 两端修饰部分,再进行凝胶电泳实验,除去杂质和过反应 DNA 分子的 DNA 分子,即保留链长在区间 $I=[6 \times (n+1) + 15 \times \sum_{i=1}^n (\min_{j=1}^n c_{ij}), 6 \times (n+1) + 15 \times \sum_{i=1}^n (\max_{j=1}^n c_{ij})]$ 内的 DNA 分子。

6. 在 T_4 DNA 连接酶作用下,将混合物中线性 DNA 分子变为闭环 DNA 分子,并经过凝胶电泳实验,除去杂质、线性 DNA 和链长超过 $6 \times (n+1) + 15 \times \sum_{i=1}^n (\max_{j=1}^n c_{ij})$ 的 DNA 分子,得到由闭环 DNA 组成的混合物 H 。其中每个闭环 DNA 分子具有结构: $p_0p_{1j1}q_{1j1}p_{2j2}q_{2j2} \dots p_{njn}q_{njn} (j1, j2, \dots, jn=1, 2, \dots, n)。$

混合物 H 中的 DNA 闭环 DNA 分子 $p_0p_{1j1}q_{1j1}p_{2j2}q_{2j2} \dots p_{njn}q_{njn}$ 的含义是:闭环 DNA 分子中出现了 DNA 片段 p_{kjk} q_{kjk} ,表示决策变量 $x_{kjk}=1$;没有在闭环 DNA 分子中出现的 DNA 片段 $p_{ij}q_{ij}$,表示决策变量 $x_{ij}=0$ 。因此,该闭环 DNA 分子代表决策变量 $x_{1j1}=x_{2j2}=\dots=x_{njn}=1$,其余决策变量为 0 的指派问题一个可能解(注意:可能解满足行要求);闭环 DNA 分子的长度 $6 \times (n+1) + 15 \times \sum_{k=1}^n p_{kjk}$ 代表该可能解的目标函数值。

对步骤 3,算法的生化实现过程为:

1. 置 $k=1, H_{k-1}=H。$

2. 取出识别序列 p_{1k} 、 p_{2k} 、 $\dots$ 、 p_{nk} 对应的限制性内切酶,组成集合 T_k 。混合物 H_{k-1} 与 T_k 混合,在限制性内切酶集合 T_k 作用下发生酶切反应。经过凝胶电泳实验,除去限制性内切酶、剪切下来的 DNA 片段。在 T_4 DNA 连接酶作用下,开环的 DNA 形成新的闭环 DNA,经过凝胶电泳实验保留链长在区间 I 内的闭环 DNA 分子,得到新的闭环 DNA 混合物 H_k (注意:如果闭环 DNA 分子被剪切成两个或两个以上部分的线性 DNA 片段,则在环化时每个部分不能变成闭环 DNA 分子,且电泳速度快,即不满足列要求的可能解被淘汰。)

3. 如果 $k=n$,则得到混合物 H_n ,转入步骤 4;否则,置 $k=k+1$,转入 2。

混合物 H_n 中每一条闭环 DNA 代表指派问题的一个可

行解;闭环 DNA 序列的链长代表该可行解的目标函数值。

对步骤 4,算法的生化实现过程为:

1. 识别序列 p_0 对应的限制性内切酶与混合物 H_n 混合,在该限制性内切酶作用下发生酶切反应。

2. 对混合物 H_n 中开环的 DNA 分子,以聚丙烯酰胺为支持物,进行凝胶电泳。

3. 在紫外光照射下,将胶上 DNA 分子长度最小的区带切割下来,对切割部分继续电泳使胶上 DNA 分子释放到缓冲溶液中。得到双链 DNA 的混合物 R 。

混合物 R 中每一条线性 DNA 链代表指派问题的一个最优解。

对步骤 5,算法的生化实现过程为:

1. 将线性 DNA 混合物 R 的温度升到 95°C ,发生变性反应,再将混合物骤然冷却至 68°C ,形成单链的 DNA 分子混合物。再对混合物 R 中的 DNA 分子用“-SH”基取代 $5'$ 的修饰保护基并固定在硅片表面上,制成 DNA 芯片。置 $i=1, j=1$ 。

2. 取第 (i, j) 号标记探针与 DNA 芯片充分混合后,温度由 94°C 缓慢冷却至 68°C ,进行杂交反应,并用 68°C 的缓冲溶液冲掉未杂交的探针分子。用激光共聚焦显微镜扫描 DNA 芯片对杂交结果进行检测,并记录下 DNA 芯片上的亮点。有亮点处 DNA 分子代表决策变量 $x_{ij}=1$;没有亮点的代表决策变量 $x_{ij}=0$ 。

3. 用 94°C 的缓冲液与 DNA 芯片混合均匀,发生变性反应,并用 94°C 的缓冲液冲洗掉 DNA 芯片上全部探针。

4. 如果 $j=n$,则置 $j=1$,转入 5;否则,置 $j=j+1$,转入 2。

5. 如果 $i=n$,则检测结束,转入 6;否则,置 $i=i+1$,转入 2。

6. 对 DNA 芯片上每个点所做的记录进行分析,可以得出最优指派问题的全部最优解。

4 算例

为说明 DNA 计算求解指派问题的可行性下面举一个 3×3 指派问题的例子加以验证。

表 1 3×3 指派问题效率 c_{ij} 数据一览表

任务 人数	任务		
	1	2	3
1	2	2	1
2	3	1	2
3	2	3	1

效率值为表 1 的指派问题计算过程为:

(1) p_{ij} 编码及相应的限制性内切酶见表 2。

表 2 识别序列及相应的限制性内切酶一览表

p_{ij}	1	2	3
1	AACGTT <i>Acl</i> I	CTGAAG <i>Acu</i> I	ACCGGT <i>Age</i> I
2	GGGCCC <i>Apa</i> I	GTGCAC <i>Apa</i> L I	ATTAAT <i>Ase</i> I
3	CCTAGG <i>Avr</i> II	ATGCAT <i>Bfr</i> B I	AGATCT <i>Bgl</i> II

p_0 的识别序列 GGTCTC;内切酶为 *Bsa* I。

(2) q_{ij} 的编码为:

q_{11} :AAACTATCTATCTCG-CCAATACTTATACCT
 q_{12} :AAACTTCATACTTCC-AAACTCATTATCCCT
 q_{13} :AACTAACCCATTCTT
 q_{21} :AACTATACTATCCTG-CCTCATCCTAAATTA
 -AACTACTCTATTACC
 q_{22} :AACTACCTTAACTCT
 q_{23} :ACTAATACTACTTCC-ACTAACCACACTTTT
 q_{31} :TTTACACTTAACACC-TTCCTATTAAACCAC
 q_{32} :TTCCTTTCCAACAAA-TTCCTACCAAATAT
 -CATTTTATCACACAC
 q_{33} :CTAAACCTTATTACC

3. 经过有目的的终止技术和删除实验后闭环 DNA 混合物组成为:

$p_0 p_{11} q_{11} p_{22} q_{22} p_{33} q_{33}$ 、 $p_0 p_{11} q_{11} p_{23} q_{23} p_{32} q_{32}$ 、
 $p_0 p_{12} q_{12} p_{21} q_{21} p_{33} q_{33}$ 、 $p_0 p_{12} q_{12} p_{23} q_{23} p_{31} q_{31}$ 、
 $p_0 p_{13} q_{13} p_{21} q_{21} p_{32} q_{32}$ 、 $p_0 p_{13} q_{13} p_{22} q_{22} p_{31} q_{31}$ 。

4. 经过酶切反应和凝胶电泳实验以后,链长最短的 DNA 序列为: $p_0 p_{11} q_{11} p_{22} q_{22} p_{33} q_{33}$ 。

5. 将上述单链 DNA 序列 $p_0 p_{11} q_{11} p_{22} q_{22} p_{33} q_{33}$ 固定在硅片表面上,制成 DNA 芯片。进行标记实验,得到最优解为 $(1, 0, 0, 0, 1, 0, 0, 0, 1)$ 。

结束语 本算法可以在以下两方面进行改进。

设 p_{ij} 的链长为 6; q_{ij} 的链长为 $m \times c_{ij}$,则闭环 DNA 的链长不低于 $6 \times (n+1) + m \times n \times \min(c_{ij})$ 。发生酶切反应的线性 DNA 的链长不超过 $6 \times n + m \times (n-1) \times \max(c_{ij})$,因此当 $m \times (n-1) \times \max(c_{ij}) - m \times n \times \min(c_{ij}) < 6$ 时,只需要取 $p_{ij} = p_j$ 以及通过电泳实验除去小于 $6 \times (n+1) + m \times n \times \min(c_{ij})$ 的线性 DNA 即可。这样,限制性内切酶的规模由原来的 n^2 个变为 n 个。

算法开始首先处理系数矩阵,使系数矩阵产生更多的零。则对 q_{ij} 编码的链长可以缩短,从而减少编码的规模。

参考文献

- Adleman L M. Molecular computation of solutions to combinatorial problems [J]. Science, 1994, 266(11):1021~1024
- Zhou Kang, Tong Xiao-Jun, Xu Jin. An Algorithm of Sticker DNA Chip Model on Making Spanning Tree Problem [J]. In: Proc. eedings of the Fifth International Conference on Machine Learning and Cybernetics, 2006. 4287~4292
- 周康,同小军,许进. 路序问题基于表面的 DNA 算法[J]. 华中科技大学学报(自然科学版), 2005, 33(8):100~103
- 周康,许进. 最小顶点覆盖问题的闭环 DNA 算法[J]. 计算机工程与应用, 2006, 42(20):7~9
- 周康,王延峰,刘文斌,许进. 基于闭环 DNA 的边着色问题 DNA 算法[J]. 华中科技大学学报(自然科学版), 2006, 34(9):25~28
- Zhou Kang, Gao Zunhai, Xu Jin. An Algorithm of DNA Computing on 0-1 Planning Problem [J]. Advances in Systems Science and Applications, 2005, 5(4):587~593
- Paun G,等. DNA 计算一种新的计算模式[M]. 北京:清华大学出版社, 2004
- 钱颂迪,田丰,胡运权,等. 运筹学. 修订版. [M]. 北京:清华大学出版社, 1990