一种基于 DNA 计算的指定结点路由算法

杨 磊 黄启鑫 李肯立 李仁发
(湖南大学计算机与通信学院 长沙 410082)

摘 要 带指定结点约束的路由问题是一个 NP 难问题,该问题是电信行业路由智能化和交通电力运输等领域的关键问题之一.基于 DNA 计算的高度并行性,文中提出一种将电子计算机与 DNA 计算机相结合的方法求解指定结点路由问题.算法由转化算法 *Transform*()、首末结点搜索切割算法 *FirstEndSearcher*()、转化图结果搜索算法 *DNASearcher*()和结果读取算法 *ResultReader*()共 4 个子算法组成.分析表明:算法的电子计算机部分缩小了问题结点和边的规模,从而使解决问题所需的 DNA 分子链数数量级从 $O((n-2)!)$减少至 $O((m-2)!)(n\geq 2$ 为图中结点数,$m\geq 2$ 为图中指定必经结点数).算法的 DNA 计算机部分采用了有针对性的 DNA 编码新方案,提高了边权值编码的信噪比,通过一系列生物操作,筛选出问题的精确解.和单纯 DNA 超级计算或电子计算机指定结点路由算法相比,文中算法可显著扩大理论上待求解问题的规模.

关键词 DNA 计算;松散指定路由;指定结点路由;MPLS ASON
中图法分类号 TP301 **DOI 号**: 10.3724/SP.J.1016.2009.02373

A Molecular Solution for Routing Satisfying Explicit Node Constraint on DNA-Based Supercomputing

YANG Lei HUANG Qi-Xin LI Ken-Li LI Ren-Fa
(*College of Computer and Communication*, *Hunan University*, *Changsha* 410082)

Abstract Routing algorithms satisfying explicit node constraint is a NP-hard mathematical problem. This problem is not only a obstacle for intelligent routing in telegraphic industry, but also for transportation and power transmission. Based on super parallel-computing of DNA computation, an algorithm that combined the merits of traditional computer and DNA computation is proposed to solve the routing algorithms satisfying explicit node constraint in this paper. The proposed algorithm consists of four sub-algorithms: *Transform*(), *FirstEndSearcher*(), *DNASearcher*(), *ResultReader*(). The theoretic analysis shows that the use of traditional computer part of the algorithm could cut down the amount of nodes and edges sharply so that the corresponding DNA volume strands could decrease from $O((n-2)!)$ to $O((m-2)!)$ where n and m are the amount of nodes and explicit node respectively. A series of biological operations are proposed to search for the accurate solution. In addition, in order to advance the coding-SNR of border-weight and make biological operation feasible, a new DNA-coding rule is also proposed. So that, fast routing algorithms satisfying explicit node constraint will be solved in reasonable time provided that the technology of DNA computing is mature enough in the future.

Keywords DNA-coding; loose explicit node constraint route; explicit node constraint route; MPLS ASON

收稿日期:2009-07-06;最终修改稿收到日期:2009-10-15.本课题得到国家自然科学基金(90715029,60603053)和教育部新世纪优秀人才计划部分资助.**杨 磊**,男,1976 年生,博士研究生,副教授,研究兴趣为 DNA 计算、计算机网络.**黄启鑫**,男,1985 年生,硕士研究生,研究方向为 DNA 计算.**李肯立**,男 1971 年生,教授,博士生导师,研究领域为并行处理、DNA 计算机.E-mail: lkl510@263.net.**李仁发**,男,1957 年生,教授,博士生导师,研究领域为嵌入式系统、DNA 计算等.

1 引 言

随着路由智能化技术的深入发展,带有一个或多个约束条件的路由策略成为关注热点.然而,尽管国内外文献中对于 QoS 约束的研究较为成熟[1-4].但针对指定结点约束路由问题的研究却较少.指定结点路由问题可一般性地定义为带指定结点约束的最短路径问题,它不同于一般的 QoS 约束,它是一种松散指定路由约束[1,5-6].在 MPLS 或 ASON 网络中,由于网络运营商对于业务路径控制的需要、各个网元结点交换能力的差别等原因,在路由计算中指定一个或多个必经结点有着现实的需求.而且,该问题的算法在交通运输、电力输送等领域都有着很高的应用价值.然而,指定结点路由问题是一个 NP 难问题[3],为保证算法的时间效率,目前实际应用中的系统要么必须指定整条路径,要么必须给定业务路径经过所有指定结点的顺序.但是按照给定的结点顺序并不能保证得到最短路径.虽然已有相关文献给出了指定结点路由问题的伪多项式时间算法[1],但算法的时间效率仍无法保证.

1994 年 Adleman[7] 在《Science》杂志上发表了开创性论文,提出了 DNA 计算的概念.之后,DNA 计算和 DNA 计算机的研究迅速发展成为理论计算机科学和数学等学科的研究热点之一[8-13],目前已设计出许多难解问题的 DNA 计算机算法.因此,由于指定结点路由问题的特点,采用具有高度并行性的 DNA 计算机理论上可在多项式时间内得出问题的精确解.然而,DNA 计算目前仍面临很多困难,目前影响 DNA 计算发展的最大障碍是"指数爆炸"问题.

经初步探索,如不进行任何算法优化设计,而只是如典型 DNA 计算机算法那样使用简单穷举方法,则解决本文问题所需的 DNA 分子链数将为图中顶点数的阶乘级,比纯指数数量级更差.

为此,为既利用 DNA 计算在求解难解问题上的优势,又可有效避免指定结点路由问题中出现的 DNA 容量空间指数爆炸问题,本文将电子计算机与 DNA 计算机相结合,以发挥各自的特性和优势.首先利用电子计算机对问题进行优化变形,缩小问题结点和边的规模,从而将待解决问题所需的 DNA 分子链数减少.然后提出一种带权图中边和结点的编码方案,并设计出问题求解的 DNA 计算机算法来求得问题的最终解.

本文第 2 节介绍指定结点路由问题;第 3 节是该问题的算法,包括算法的电子计算机部分和 DNA 计算机部分;第 4 节给出算法性能分析结果;算法的 DNA 编码方案则在第 5 节给出;最后是本文总结和今后工作的研究方向.

2 指定结点路由问题描述

用三元组表示问题中的带权无向连通图 G,$G=\{V(G),U(G),E(G)\}$.其中 $V(G)=\{v_1,v_2,\cdots,v_n\}$ 为图 G 中的结点集合,$U(G)=\{u_1,u_2,\cdots,u_m\}$ 为图 G 中的指定必经结点集合(包括首末结点).$E(G)=\{e_{12},e_{21},\cdots,e_{ij}\}$ 为图 G 中的边集合,用字母 n,m 和 q 分别表示图中的结点数,指定必经结点数和边数.其中 $n\geq 1$,$m\geq 1$,$q\geq 0$,e_{ij} 表示 v_i 到 v_j 的边,边权值为 $W(v_i,v_j)$.由于路由计算的特点,如果 v_i,v_j 之间存在边,均记 $W(v_i,v_j)=1$(不为 1 同样适用),如果 v_i,v_j 之间不存在链路,$W(v_i,v_j)=+\infty$.

定义 1. 从图 G 中结点 v_i 到结点 v_j 的一条路径记为 $PATH_k(v_i,v_j)$,由于 v_i 到 v_j 的路径可能有多条,k 标示路径的编号,该条路径的长度为

$$W(PATH_k(v_i,v_j))=\sum_{(u,v)\in PATH_k(v_i,v_j)}W(u,v),$$

公式中的 u,v 是该路径上的结点.

定义 2. 从图 G 中结点 v_i 到结点 v_j 的最短路径记为 $S(v_i,v_j)$,该条路径为 v_i 到 v_j 所有路径中长度最短的,即

$$W(S(v_i,v_j))=\min(W(PATH_k(v_i,v_j))).$$

定义 3. 指定结点路由问题是指在带权无向图 G 中,计算一条从首结点 v_1 到末结点 v_n 的路径 $P(v_1,v_n)$,路径必须经过 $U(G)$ 中所有结点,并且在满足上述条件的路径中 $W(P(v_1,v_n))$ 最小.

为讨论方便,本文只对简单路径经过指定结点的情况进行讨论.为讨论方便,给出如图 1 所示的示例.

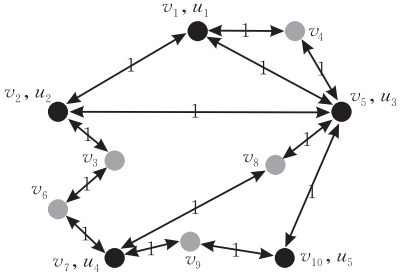

图 1 指定结点路由问题网元组网图

图 1 中的结点数为 $n=10$,其中起始结点为 v_1,

末结点为 v_{10},黑色结点为指定必经结点,以 u_i 重新命名.灰色结点为非必经结点,边权值均为 1.

3　指定结点路由问题算法

3.1　算法电子计算机部分

DNA 计算机通常采用穷举搜索算法,如果直接对本文问题进行 DNA 编码并执行生物操作,将产生大量伪解 DNA 链,这无疑会限制 DNA 计算机所能计算的问题规模.而且,经我们研究,图中的部分结点和边是冗余的,删除这些结点和边并将原图进行转化后并不影响原问题结果输出.本文设计出一种算法,在不改变原图输出结果的前提下,借助电子计算机在多项式时间内将原图 G 进行转化.

3.1.1　电子计算机部分算法实现

本算法将原图 G 进行转化,生成与图 G 相对应的转化图 G'.图 G' 中的结点与图 G 中指定结点形成一一映射关系,即图 G 中结点 u_i 对应于图 G' 中结点 u_i',图 G' 中两结点间的每条边对应图 G 中相应两结点间的边或一条路径.转化算法的伪代码表示如下.

算法 1.　指定结点路由转化算法.

```
Procedure Transform()
 For i=1 to m−1
      For j=i to m
  (1)   If(Exist(e_ij) in graph G)
  (1a)     Construct(e'_ij) in graph G'
  (1b)     W(S(u_i,u_j))=1
  (2)   Else
  (2a)     S(u_i,u_j)=Dijkstra(u_i,u_j)
  (2b)     path_length=W(Dijkstra(u_i,u_j))
  (3)      If(u_k∈S(u_i,u_j),k≠i,j)
  (3a)        W(S(u_i,u_j))=+∞
           Else
  (3b)        Construct(e'_ij) in graph G'
  (3c)        W(S(u_i,u_j))=path_length
           EndIf
        EndIf
      EndFor
 EndFor
EndProcedure
```

引理 1.　经算法 1 转化后图 G' 删除了原图 G 中的非必经结点和冗余边.

证明.　算法 1 通过两层循环实现.对于图 G 中任意两个必经结点 u_i,u_j,步(1)首先用 $Exist()$ 函数判断这两个结点间是否有边相连,如果存在边,则子步(1a)在转化后图 G' 中对 u_i',u_j' 之间设置边,权值为 1.如果不存在边,则进入步(2)进行处理.子步(2a)通过 $Dijkstra(u_i,u_j)$ 函数求出 u_i 和 u_j 间的最短路径($Dijkstra(u_i,u_j)$ 函数为求两结点间最短路径的 Dijkstra 算法函数实现),记为 $S(u_i,u_j)$.子步(2b)用 $W(Dijkstra(u_i,u_j))$ 函数求出该路径的长度,赋给变量 $path\_length$,标示图 G' 中相应两结点间边的长度.

例如,对图 1 中所有结点执行步(1)、(2)后得到的结果如图 2 所示.

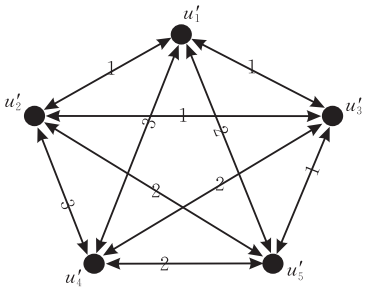

图 2　删除冗余结点转化图

图 2 中只含图 1 中的指定结点,任意两结点间的边长度为图 G 中对应两结点间的最短路径长度.

经步(1)、(2)处理后的图中仅含有原图 G 中的指定结点,但仍含有冗余边,算法 1 通过步(3)将冗余边去除.首先,判断 $S(u_i,u_j)$ 中是否包含其它指定必经结点,如果包含则图 G' 中 u_i',u_j' 间不需要设置边,子步(3a)将 u_i' 和 u_j' 间的边权值置为 $+\infty$,生物操作时不进行 DNA 编码.否则,子步(3b)在 u_i' 和 u_j' 间设边,子步(3c)将边权值置为 $Dijkstra(u_i,u_j)$ 函数求出的 $path\_length$ 值.

例如,图 2 中所有结点经步(3)处理后得到的结果如图 3 所示.

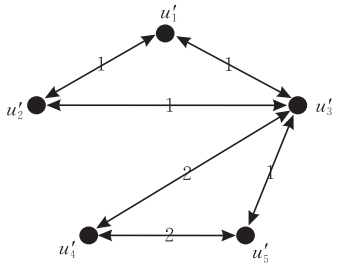

图 3　删除冗余边转化图

图 3 中删除了图 2 的冗余边,例如,图 2 中结点 u_2' 到 u_4' 的最短路径 $P(u_2',u_4')$ 为 $u_2'\rightarrow u_3'\rightarrow u_4'$,路径中包含结点 u_3',所以边 $u_2'\rightarrow u_4'$ 为冗余边,应该删去.以下证明这一步骤的正确性:假设路径 $S(u_i',u_j')$ 中含其它必经结点为 u_{i+1}',u_{i+2}',…,u_{j-1}',$S(u_i',u_j')$ 经过

其它必经结点的顺序为 $u'_{i+1}, u'_{i+2}, \cdots, u'_{j-1}$，本算法执行结束后可以保证图 G' 中的 $u'_i \rightarrow u'_{i+1}$，$u'_{i+1} \rightarrow u'_{i+2} \cdots u'_{j-1} \rightarrow u'_j$ 间均有边直接相连，且这些边连接成的路径长度即为 $u'_i \rightarrow u'_j$ 最短路径长度（由步(1c)保证），所以 u'_i 和 u'_j 间已经有权值最小的路径相连，记为 $S(u'_i, u'_j)$，不需要再设边.

综上所述，经过算法 1，原图 G 转化为图 G'，且图 G' 中只包含图 G 中的必经结点，图 G' 中任意边的权值等于图 G 中相应两结点间的最短路径值，去除了冗余边.

3.1.2 转化图输出结果分析

定理 1. 转化后图 G' 与原图 G 对指定结点路由问题同解.

证明. 证明可分为两个部分，首先证明图 G' 中经过所有结点的最短路径 $S(u'_1, u'_m)$ 可以映射为图 G 中经过所有指定结点的一条路径 $S(u_1, u_m)$. 根据算法 1 可以给出图 G' 中边和图 G 中对应路径的映射关系表，例如，文中图 1 与图 3 的映射关系表如表 1 所示. 不妨设路径 $S(u'_1, u'_n)$ 在图 G' 中经过各结点的顺序是 $u'_1, u'_2, \cdots, u'_m$，$S(u'_1, u'_n)$ 上的每一条边都在图 G 中唯一对应一条路径，则可以在图 G 中根据映射关系表映射出一条以同样顺序经过指定结点 $u_1, u_2, \cdots, u_m$ 的路径，设为 $S(u_1, u_m)$，而且根据映射关系表即可证明两条路径长度相等.

表 1 原图与转化图映射关系表

图 G' 中边	图 G 中对应路径	权值
$(u'_1 \rightarrow u'_2)$	$(v_1 \rightarrow v_2)$	1
$(u'_1 \rightarrow u'_3)$	$(v_1 \rightarrow v_5)$	1
$(u'_2 \rightarrow u'_3)$	$(v_2 \rightarrow v_5)$	1
$(u'_3 \rightarrow u'_4)$	$(v_5 \rightarrow v_8 \rightarrow v_7)$	2
$(u'_3 \rightarrow u'_5)$	$(v_3 \rightarrow v_5)$	1
$(u'_4 \rightarrow u'_5)$	$(v_7 \rightarrow v_9 \rightarrow v_{10})$	2

然后证明 $S(u_1, u_m)$ 为图 G 中经过所有指定结点的最短路径即可. 不妨设路径长度 $W(S'(u'_1, u'_m)) = W(S(u_1, u_m)) = k$. 利用反证法，假设图 G 中另存在经过所有必经结点的路径 $Q(u_1, u_m)$ 且权值 $W(Q(u_1, u_m)) < k$，可以根据 $Q(u_1, u_m)$ 经过指定必经结点的顺序在图 G' 中找到一条相应路径 $Q'(u'_1, u'_m)$，如果 $W(Q'(u'_1, u'_m)) < k$，这与 $P(u'_1, u'_m)$ 是图 G' 中最短路径的假设矛盾. 如果 $W(Q'(u'_1, u'_m)) \geq k$，由于 $Q(u_1, u_m)$ 和 $Q'(u'_1, u'_m)$ 经过必经结点的顺序是一样的，则在 $Q(u_1, u_m)$ 中至少存在两个结点 u_i，u_j，使得这两个结点之间的最短路径长度 $W(S(u_i, u_j))$ 小于 $Q'(u'_1, u'_m)$ 中对应两结点 u'_i, u'_j 间边的权值 $W(u'_i, u'_j)$，这与图 G' 中任意两结点间边的权值是图 G 中对应两结点间最短路径的前提矛盾.

3.2 算法 DNA 计算机部分

本算法将经过电子计算机转化后的图进行 DNA 计算，从而得出指定结点路由问题的最终解.

3.2.1 DNA 计算模型

DNA 计算解决问题的基本思想是[7]：利用 DNA 特殊的双螺旋结构和碱基互补配对原则对问题进行编码，把要运算的对象映射成 DNA 分子链，在含 DNA 溶液的试管里，在生物酶的作用下，生成各种数据池（datapool），然后按照一定的规则将原始问题的数据运算高度并行地映射成 DNA 分子链的可控的生化过程. 最后，利用分子生物操作 PCR、POA、超声波降解、亲和层析、克隆、诱变、分子纯化、电泳、磁珠分离等，破获运算结果. DNA 分子生物计算的研究包括[9]：DNA 计算模型、使用具体模型求解问题时的 DNA 计算算法和包括编码、分子合成、输入输出技术等在内的 DNA 计算实现技术等 3 个方面. 自 Head 首先提出 splicing 模型后，已有多种 DNA 计算模型被提出，而且仍在发展中，目前 DNA 计算常用的模型有剪接系统（splicing systems）[10]、粘贴模型（sticker）[11-12]、表面计算模型（surface-based）[13]、基于质粒（plasmids）的 DNA 计算模型[14]、禁止-允许模型（forbidding-enforcing）[15]等. 根据本文问题特点，结合 Adleman-Lipton 模型[7]，并增加 5′-限制性切割（5′-cut）、3′-限制性切割（3′-cut）和凝胶电泳（gel electrophoresis）操作.

(1) 5′-限制性切割（5′-cut）. 给定试管 P 和一个短的 DNA 链 Z，该操作使用限制性内切核酸酶对 P 中含 Z 的 DNA 双链从 Z 的 5′-端进行平头切割，使其断链.

(2) 3′-限制性切割（3′-cut）. 给定试管 P 和一个短的 DNA 链 Z，该操作使用限制性内切核酸酶对 P 中含 Z 的 DNA 双链从 Z 的 3′-端进行平头切割，使其断链.

(3) 凝胶电泳（gel electrophoresis）. 给定试管 P，该操作将试管中的 DNA 链按照分子大小进行分离.

3.2.2 DNA 计算机算法实现

本文 DNA 算法的基本思想是：将图 G' 中的每个结点 u_i、每条边 e_{ij} 和边权值的逆补 $\overline{e_{ij}}$ 分别进行 DNA 编码，将所有 DNA 分子链置入试管溶液中充分混合杂交，产生所有可能路径的 DNA 双链. 然后通过设计出的生物操作步骤，逐步分离出所有以 u_1

为首的结点,u_m为末结点的 DNA 链,为保证每条路径都经过所有指定必经结点,还应去除包含结点不完全的 DNA 链,然后通过凝胶电泳法对结果 DNA 分子链进行筛选,由于越短的 DNA 分子链在电泳过程中迁移得越快,即可得到表示符合指定结点约束的最短路径 DNA 双链,再经过 DNA 序列解析过程,得出该解路径的核苷酸序列,根据图 G' 中边和结点与 DNA 编码的对照表得出路径序列,最后查找图 G' 中边和图 G 中路径的映射表,得出图 G 中指定结点最短路径.

(1) 首末结点搜索切割算法

在含所有可能路径 DNA 双链的试管 T_0 中,必须排除首结点或末结点不在 DNA 链两端的 DNA 链. 因此,应通过相应的生物操作保证试管 T_0 中的 DNA 链以 u_1' 为首结点,u_m' 为末结点.

算法 2.　首末结点搜索切割算法.

Procedure $FirstEndSearcher(T_0)$

(1) $T_1, T_2, T_3, T_4, T_5 = \varnothing$
(2) $Amplify(T_0, T_1)$
(3) $T_2 = +(T_1, u_1')$
(4) $T_3 = +(T_2, u_m')$
(5) $Discard(T_1), Discard(T_2)$
(6) If ($Detect(T_3)$="yes") then
(6a) $5'\text{-}Cut(T_3, u_1')$
(6b) $3'\text{-}Cut(T_3, u_m')$
Else
(6c) break Stop;
EndIf
(7) $T_4 = +(T_3, u_1')$
(8) $T_5 = +(T_4, u_m')$
(9) $Discard(T_3), Discard(T_4)$
(10) $Amplify(T_5, T_0)$
EndProcedure

引理 2.　算法 $FirstEndSearcher(T_0)$ 可以生成所有以 u_1' 为起点、u_m' 为终点的路径.

证明.　步(1)准备 5 支空试管 T_1, T_2, T_3, T_4, T_5,步(2)将 T_0 中的溶液复制到 T_1 中,并将 T_0 置空. 步(3)从 T_1 中抽取出所有包含 u_1' 的 DNA 链并置入 T_2 中. 步(4)从 T_2 中抽取出所有包含 u_m 的 DNA 链并置入试管 T_3 中,此时试管 T_3 中所有 DNA 链都包含 u_1' 和 u_m',但由于生物操作中可能产生的杂交错误,不能保证 u_1' 和 u_m' 都位于 DNA 链两端. 步(5)将试管 T_1 和 T_2 中的溶液清除掉. 步(6)判断试管 T_3 中是否还有 DNA 分子链,如果没有,说明试管中不存在问题的解,算法终止. 否则,就进行首末结点切割操作,步(6a)使用 5′-限制性切割操作将 T_3 中的 DNA 双链从结点 u_1' 的 5′-端平头切割,步(6b)使用 3′-限制性切割操作将 T_3 中 DNA 双链从结点 u_m' 的 3′-端平头切割,则 T_3 中首末结点 u_1' 和 u_m' 要么均在链两端,要么切割出来的 DNA 链中不包含 u_1' 和 u_m'. 步(7)和步(8)将再次执行筛选含有结点 u_1' 和 u_m' 的 DNA 链的操作,目的是将切割后不含 u_1' 和 u_m' 的 DNA 链去除. 步(9)将试管 T_3 和 T_4 中的溶液清除掉. 步(10)将试管 T_5 复制到试管 T_0 中. 因此,试管 T_0 包含所有以 u_1 为起点、u_m 为终点的可能路径,得证.　　证毕.

算法 2 执行两次复制操作、四次抽取操作、四次丢弃操作、一次检测操作,所用试管数为 6.

(2) 转化后图结果 DNA 链搜索

执行算法 2 后试管 T_0 中的 DNA 双链仍可能存在包含结点不完整的路径. 应设计搜索算法搜索出路径中包含图 G' 中所有结点的 DNA 双链.

算法 3.　转化图结果搜索算法.

Procedure $DNASearcher(T_0, m)$

(1) $Amplify(T_0, T_2)$
(2) For $i=2$ to $(m-1)$
(2a) $T_{i+1} = +(T_i, u_i')$
(2b) $Discard(T_i)$
End For
(3) If ($Detect(T_m)$="yes") then
(3a) $Amplify(T_m, T_0)$
(3b) $resultDNA = Gel(T_0)$
Else
(3c) break Stop;
EndIf
EndProcedure

引理 3.　算法 $DNASearcher(T_0, m)$ 可筛选出图 G' 中经过所有结点的最短路径 DNA 链.

证明.　步(1)将试管 T_0 复制到 T_2 中,并将 T_0 置空. 步(2)通过 $m-2$ 次循环删除试管中包含结点不完整的 DNA 双链. 具体算法为:步(2a)将试管 T_i 中包含结点 u_i 的所有 DNA 双链复制到试管 T_{i+1} 中. 该步骤的生物操作方法为:先将试管中所有 DNA 分子双链加热解链为单链[16],在试管壁上粘贴结点 u_i 的引物,充分反应后含有 u_i 的 DNA 链和引物结合粘贴在试管壁上,倒掉试管中的溶液并将试管壁上的 DNA 分子链提取出即可. 步(2b)将不需要的试管丢弃. 步(3)判断试管 T_m 中是否还存在 DNA 链,如果不存在,则试管中不包含问题的解 DNA 链,算法终止. 否则,就执行下面算法输出结

果,步(3a)将试管 T_m 中的溶液复制到试管 T_0. 步(3b)通过凝胶电泳法得到表示图 G' 中经过所有结点的最短路径的 DNA 链,记为 *resultDNA*,得证. 证毕.

算法 3 执行两次复制操作、$m-2$ 次抽取操作、$m-2$ 次丢弃操作、一次检测操作、一次凝胶电泳操作,所用试管数为 m.

(3)指定结点路由结果输出

通过算法 3 得到转化后图的指定结点最短路由 DNA 链,还需要对 DNA 序列进行读取并将读取结果转化成原图 G 中的路径.

算法 4. 结果读取算法.

Procedure $ResultReader(T_0)$

(1) $T_1=\varnothing$

(2) $T_1=FirstEndSearcher(T_0)$

(3) $resultDNA=DNASearcher(T_1,m)$

(4) $result\_strand=Read(resultDNA)$

(5) $Path(G')=StrandToPath(result\_strand)$

(6) $Path(G)=PathExchange(Path(G'))$

EndProcedure

引理 4. 算法 $ResultReader(T_0)$ 可读出原图 G 中指定结点路由问题最终解.

证明. 步(1)首先给出一支空试管 T_1,步(2)将所有以 u_1' 为首结点且 u_m' 为末结点的 DNA 链复制到试管 T_1 中. 步(3)得出图 G' 中指定结点最短路径的结果 DNA 链,记为 *resultDNA*. 步(4)读出 *resultDNA* 的碱基排列顺序,记为 *result_strand*. 步(5)通过 *StrandToPath*()函数将碱基排列顺序解码为图 G' 中的路径,记为 $Path(G')$. 步(6)通过查找转化图与原图映射关系表,得出图 G 中指定结点最短路径,记为 $Path(G)$,即为最终解.

4 指定结点路由问题算法性能分析

4.1 计算机算法复杂度分析

由于Dijkstra算法的时间复杂度是$(n+q)\log(n)$,判断和赋值操作时间忽略不计,在每次循环都需要执行 Dijkstra 算法的情况下,算法 1 共需要执行 $m^2/2$ 次循环,算法时间复杂度为 $O((n+q)\log(n)m^2/2)$. 应该指出的是,Dijkstra 算法是本文计算机算法中最耗时的部分. 为避免每次循环都执行该算法,算法 1 中加入了对于图 G 中两个指定结点 u_i 和 u_j 间是否存在边的判断,如果存在边,即可直接得出图 G' 中对应结点 u_i' 和 u_j' 间的最短路径及长度,不需要使用 Dijkstra 算法. 在实际应用中,随着图中指定结点的增多,两指定结点间存在边的概率也随之增加,因此,一般情况下执行 Dijkstra 算法少于 $m^2/2$ 次,尤其在两种极端情况下(指定结点数为 0 或 n)均不需使用 Dijkstra 算法. 所以,$O((n+q)\log(n)m^2/2)$的多项式时间复杂度为算法 1 的最差情况,可以通过电子计算机完成.

4.2 DNA 计算机复杂性分析

对于本文中问题,如果不经电子计算机算法转化而直接使用 DNA 计算机计算,则需使用 DNA 分子链穷举出原图 G 中的所有路径,然后在这些路径 DNA 链中筛选出解路径. 因此,生物操作中将会出现大量伪解 DNA 链.

定理 2. 指定结点路由问题直接使用 DNA 计算机计算所需 DNA 分子链数量级为 $O((n-2)!)$,其中 n 为图 G 的结点总数.

证明. 对图中任意两个结点间都存在边的情况讨论,每条结果 DNA 链对应一条路径,所以仅需证明结点数为 n 的图中最多存在$(n-2)!$条可能路径即可. 利用数学归纳法证明定理对于 $n=2$ 的情况成立. 假设该定理对于 $n-1$ 个结点的图也成立,讨论包含 n 个结点的图. 图中含非首末结点 $n-2$ 个,从首结点到非首末结点的任一结点 v_i 都有 k 条不同路径 $PATH_k(v_1,v_i)$,根据假设,对每一个结点 v_i 的 k 取值都为$(n-3)!$,即存在$(n-3)!$条从 v_1 到 v_i 的不同路径,而且,由于图中任意两个结点间都存在边,这些非首末结点均可以与末结点相连形成一条连接首末结点的新路径,因此连接首末结点的路径至多有$(n-2)((n-3)!)$条,即$(n-2)!$条. 证毕.

需要指出的是,减少问题原图中的边数也会对结果 DNA 链的规模产生影响. 因为在连通图中,如果一条边 e_{ij} 属于从首结点到末结点的一条或多条路径中,将该边删除则可将所有经过该边的路径切断. 如果这条边不在任何一条连接首末结点的路径上,则删除该边不对 DNA 链数产生影响.

定理 3. 指定结点路由问题经算法 1 转化后,解决问题所需 DNA 分子链数为 $O((m-2)!)$,其中 m 为图 G 的必经结点总数.

证明. 经过算法 1 转化后图中的结点数从 n 减少为 m,由定理 2 中的证明可知,解决含 m 个结点的图的指定结点路由问题至多需要$(m-2)!$个 DNA 分子链. 而且,只要原图 G 中的两指定结点间最短路径上有其它指定结点,则不需要在这两个结点间设边,一般情况下,通过算法 1 可以明显减少图中边数,从而进一步去除了原图中的部分伪解路径,

使生物操作所需的 DNA 链数降至更低.特殊情况下(例如星型网络中心结点为非指定结点,其余结点均为指定结点,则转化后图中每两个指定结点间均要设边),转化后图 G' 中边数可能会大于原图 G 的边数,但并不会增加生物操作的 DNA 分子链数.容易证明,根据图 G' 中边和图 G 中对应路径的映射关系表可知,转化后图 G' 中每条边均是原图 G 中的已有边或者已有边的组合,因此,即使转化后边数增加,仍不会比原图多出任何新边或新路径,不会增加生物操作的 DNA 链数.

综上所述,通过电子计算机的转化处理,解决指定结点路由问题所需的 DNA 分子链数至少从 $O((n-2)!)$ 减少至 $O((m-2)!)$,并随边的减少程度而进一步减小,但总的 DNA 链数即空间复杂度的上界保持不变.

5　DNA 编码方案

将问题原图经电子计算机转化后,要使用 DNA 计算机进行生物实验,必须对 G' 中的结点和边的权值进行 DNA 编码.权值编码是 DNA 计算中的难点之一.Lee[17] 等提出采用权值定长编码然后通过解链温度不同得出最短路径的方法,然而这种方法生物可行性不高.韩爱丽[18] 等提出将带权边转化为结点后利用边和结点的逆补形成稳定 DNA 双链的方案,然而,若边权值为 1,该方案仅用 1 个核苷酸编码,这必定会在实际生物操作中出现误码问题,而且,该方案没有考虑碱基的排列方式对编码误码率的影响.本文采用变长编码,而且,为了能够准确地得出最短路径长度,权值编码应该采用信噪比高的方案.Braich[19] 等提出了一组高信噪比的编码规则,并用生物实验加以验证.朱翔鸥[20] 等根据 Braich 的编码规则给出了搜索 DNA 编码序列的算法.针对本文中问题的特点,结合韩爱丽和 Braich 的编码规则,我们提出本文的 DNA 编码规则,如下所示:

(1)首末结点用 10 个核苷酸的 DNA 单链编码,非首末结点用 20 个核苷酸的 DNA 单链编码,方向为 5′到 3′.

(2)使用含 $10+n$ 个核苷酸的 DNA 单链为边的权值 $n(n>0$ 为自然数,$n\neq+\infty)$ 编码,权值为 $+\infty$ 的边不进行编码.

(3)权值编码采用三字母表 $\Sigma=\{A,T,C\}$,且其碱基 C 的含量约为 1/3;

(4)所有的权值编码序列中不含 5 个或 5 个以上连续相同的碱基;

(5)相同权值使用不同的编码序列,任两个编码序列的海明距离 $H(x_i,x_j)\geq d_h$,d_h 为最小海明距离,d_h 的取值视具体问题而定.

(6)边 $e_{ij}(u'_i\rightarrow u'_j)$ 的编码分为 3 段.第 1 段是结点 u'_i 的 3′→5′方向 10 个碱基片段的逆补(如果 u_i 为首结点,则是整个 u'_i 的逆补);第 2 段为权值编码;第 3 段是结点 u'_j 的 5′→3′方向 10 个碱基片段的逆补(如果 u'_j 是末结点,它是整个 u'_j 的逆补).这样的编码设计是为了在边两端形成粘性末端,与边两端的结点形成 DNA 双链,保证结果 DNA 链的稳定性.

(7)为保证形成 DNA 双链稳定结构,还应给出边 e_{ij} 权值编码的逆补 $\overline{e_{ij}}$,必须强调的是,文中 $\overline{e_{ij}}$ 代表边 e_{ij} 权值的逆补而非整个边 e_{ij} 编码的逆补.

例如,按照以上编码规则,给出对于图 2 中结点的编码,见表 2.

表 2　转化图中结点 DNA 编码表

图 G' 中结点	结点编码
u'_1	CTGATTGTCG
u'_2	ACTGTAGTCCAGTAAGCCTT
u'_3	ATCGGTAGCTAGTTACCGTA
u'_4	TACCAGCTTGTGGCTATAGG
u'_5	AACTTCACTA

表 2 中首末结点编码为 10 个核苷酸单链,非首末结点为 20 个核苷酸单链.

按照以上编码规则,取 $d_h=6$,给出对于图 2 中边权值编码,见表 3.

表 3　转化图中权值 DNA 编码表

图 G' 中边	边权值编码	权值
$(u'_1\rightarrow u'_2)$	ACATCACATAC	1
$(u'_1\rightarrow u'_3)$	ACTCTACTCAA	1
$(u'_2\rightarrow u'_3)$	TATCTTACCCT	1
$(u'_3\rightarrow u'_4)$	CCTTTTAATCCA	2
$(u'_3\rightarrow u'_5)$	CTCATTCAATC	1
$(u'_4\rightarrow u'_5)$	TCATAATTCCCA	2

根据表 3,可以得到图 G' 中 $u'_3\rightarrow u'_4$ 的权值编码,与表 2 中结点 u_3 和 u_4 的 DNA 编码组合即为边 $u'_3\rightarrow u'_4$ 的编码方案,见表 4.

表 4　边 DNA 双链编码实例

边 $u'_3\rightarrow u'_4$ 编码方案
e'_{34}: TCAATGGCATCCTTTTAATCCAATGGTCGAAC
$\overline{e'_{34}}$: GGAAAATTAGG
e'_{34}: TCAATGGCATCCTTTTAATCCAATGGTCGAAC
$\overline{e'_{34}}$: GGAAAATTAGG
$u'_3\rightarrow u'_4$: TCAATGGCATCC……CCAATGGTCGAAC
AGTTACCGTAGG……GGGTACCAGCTTG

6 结 论

本文通过将电子计算机与 DNA 计算机结合，给出了指定结点路由问题的多项式时间算法.算法的电子计算机部分通过转化算法 $Transform()$将图中的结点规模由 n 降至 m，并删除图中的冗余边.电子计算机部分时间复杂度为$O((n+q)\log(n)m^2/2)$.算法的 DNA 计算机部分采用误码率较低的编码方案对边权值进行编码，提高了编码的信噪比.

通过在预处理阶段引入经典的电子计算机算法减少结点规模，从而使 DNA 分子生物算法在多项式操作复杂性条件下，使得算法的 DNA 药物容量从直接穷举法的$O((n-2)!)$显著减少为$O((m-2)!)$.因此，本算法在保证了利用 DNA 计算海量的并行性的同时，一定程度上避免了 DNA 计算机算法出现的“指数爆炸”问题，从而大大扩大待求解的指定结点路由问题的规模.

参 考 文 献

[1] Gao B, Yang Y B, Chen C. Implementing a Constrained-Based Shortest Path First Algorithm in Intelligent Optical Networks, USA: White paper. Mahi Networks, Inc, 2003

[2] Mieghem P V, Kuipers F A. Concepts of exact QoS routing algorithms. ACM/IEEE Transactions on Nerworking, 2004, 12(5): 851-864

[3] Liu G, Ramakrishnan K G. An algorithm for finding K shortest paths subject to multiple constraints//Proceedings of the IEEE INFOCOM01. Anchorage, AK, USA, 2001, 2: 743-749

[4] Ariel Orda, Alexander Sprintson, Er Sprintson. Efficent algorithms for computing disj oint QoS path//Proceedings of the IEEE INFOCOM04. Hong Kong, 2004, 1: 727-738

[5] Mannie E. Generalized multi-protocol label switching (GMPLS) architecture. Internet RFC 3945, 2004.10

[6] Jamoussi B. Constraint-Based LSP setup using LDP. Internet RFC 3221, 2002-01

[7] Adleman L. Molecular computation of solutions to combinatorial problems. Science, 1994, 266(5187): 1021-1024

[8] Braich R S, Nickolas Chelyapov, Cliff Johnson, Paul W K Rothemund, Leonard Adleman. Solution of a 20-variable 3-SAT problem on a DNA computer. Science, 2002, 296 (19): 499-502

[9] Bach E, Condon A, Glaser E, Tanguay C. DNA models and algorithms for NP-complete problems. Journal of Computer and System Science, 1998, 57(2): 172-186

[10] Li Ken-Li, Zou Shu-Ting, Xu Jin. Fast parallel molecular algorithms for DNA-based computation: Solving the elliptic curve discrete logarithm problem over $GF(2^n)$. Journal of Biomedicine and Biotechnology, 2008, (1): 1-10

[11] Leandro N C. Fundamentals of natural computing: An overview. Physics of Life Reviews, 2007, 4(1): 1-36

[12] Yeh C W, Chu C P. Molecular verification of rule-based systems based on DNA computation. IEEE Transactions on Knowledge and Data Engineering, 2008, 20(7): 965-975

[13] Chang W L, Ho M S, Guo M. Molecular solutions for the subset-sum problem on DNA-based supercomputing. BioSystems, 2004, 73(2): 117-130

[14] Chang W L, Guo M, Michael H. Fast parallel molecular solutions for DNA-based supercomputing: Factoring integers. IEEE Transactions on Nanobioscience, 2005, 4(2): 149-163

[15] Martínez-Pérez I M, Zimmermann K H. Parallel bioinspired algorithms for NP complete graph problems. Journal of Parallel and Distributed Computing, 2009, 69(3): 221-229

[16] Xu Jin, Zhang Lei. DNA computer principle, advances and difficulties (I): Biological computing system and its applications to graph theory. Chinese Journal of Computers, 2003, 26(1): 1-11(in Chinese)
(许进，张雷. DNA 计算机原理、进展及难点(I)：生物计算机系统及其在图论中的应用. 计算机学报，2003，26(1)：1-11)

[17] Lee J Y, Shin S Y, Park T H et al. Solving traveling salesman problems with DNA molecules encoding numerical values. BioSystems, 2004, 78(1-3): 39-47

[18] Han Ai-Li, Zhu Da-Ming. DNA computing model based on a new scheme of encoding weight for Chinese postman problem. Journal of Computer Research and Development, 2007, 44(6): 1053-1062(in Chinese)
(韩爱丽，朱大明. 基于一种新的边权编码方案的中国邮递员问题的 DNA 计算模型 . 计算机研究与发展，2007，44(6)：1053-1062)

[19] Braich Ravinderjit S, Johnson Cliff, Rothemund Paul W K, Adleman Leonard M. Solution of a satisfiability problem on a gel-based DNA computer//Proceedings of the 6th International Workshop on DNA-Based Computers. London: Springer-Verlag, 2001: 27-42

[20] Zhu Xiang-Ou, Liu Wen-Bing, Sun Chuan. Research on the DNA words and algorithm. Acta Electronica Sinica, 2006, 34(7): 1169-1174(in Chinese)
(朱翔鸥，刘文斌，孙川. DNA 计算编码研究及其算法. 电子学报，2006，34(7)：1169-1174)

YANG Lei, born in 1976, Ph. D. candidate, associate professor. His main research interests include DNA computing and computer networks.

HUANG Qi-Xin, born in 1985, M. S. candidate. His main research interests focus on DNA computing.

LI Ken-Li, born in 1971, Ph. D., professor. His main research interests include parallel processing and DNA computing.

LI Ren-Fa, born in 1957, Ph. D., professor, Ph. D. supervisor. His main research interests include embedded computing and DNA computing.

Background

This research is supported by the National Natural Science Foundation of China under grants (90715029, 60603053), the Cultivation Fund of the Key. Scientific and Technical Innovation Project, Ministry of Edacation of China (Grant No. 708066), the Program for New Century Excellent Talents in University, NCET): Research on a scalable DNA computer model, the Theory, Model and Method of DNA Computer etc. The projects mainly focus on DNA computer models for processing some hard problems, including encoding DNA sequences, synthesizing DNA molecules, setting up the model, detecting solutions, etc. Our research group has been working on many aspects of DNA computing since 1996 and have published a monograph and more than 50 papers on DNA computing and DNA computer. In this paper, the authors proposed an improved sticker model and DNA algorithm for the problem of routing satisfying explicit node constraint.