# 翻转距离星树问题的计算复杂度和近似算法*

朱大铭, 马绍汉, 雷 鹏

(山东大学 计算机科学技术学院,山东 济南 250100)
E-mail: dmzhu@sdu.edu.cn
http://cs.sdu.edu.cn

摘要: 讨论基于基因组翻转距离的星型进化树问题的算法和复杂性.首先证明星树问题是 NP-难解的,再证明该问题不存在绝对近似求解算法,最后给出一个求解星树问题的常数近似算法,近似性能比为 2.

关 键 词: 算法;进化树;基因组;NP-完全性;近似性能比
中图法分类号: TP301 文献标识码: A

根据生物基因组推演物种进化历史是研究物种进化规律十分科学的方法,进化树构造是这一研究的关键.近年来进化树问题吸引了许多计算机科学家的研究兴趣[1,2].星型进化树是进化树的一种最简单形式.构造进化树的基础问题是进化距离计算问题.

基因组翻转距离计算问题首先由 D. Sankoff 等人提出.J. Kececioglu 与 D. Sankoff 以及 V. Bafna 与 P.A. Pevzner 给出翻转距离计算问题十分有效的近似算法[3~6].S. Hannenhalli 与 P.A. Pevzner 的多项式时间算法[7]最终解决了任意两个有向符号序列的翻转距离计算问题.进化树问题的基本含义是根据一组表达基因组的符号序列构造一棵树,使已知序列标记在叶节点上,猜测序列标记在中间节点上,并使总进化距离达到最小.这样就可以推演物种进化历史和进化规律.若只需猜测一个序列而构造星型进化树,则该问题称为星树问题,以下简称为 ST 问题.星树问题是一类形式最简单进化树问题,其价值不仅在于推演物种进化规律,在医药设计、医疗诊断中,也有重要价值.本文首先证明翻转距离星树问题是 NP-难解的,然后讨论星树问题的近似可计算性,一方面证明若 P≠NP,则该问题不存在绝对近似算法,另一方面给出该问题近似性能比为 2 的一个常数近似算法.本文利用了 Hannehalli 等给出的有向符号序列翻转距离计算的有关结果,下面简要介绍.

设 $S=\{0,1,2,\ldots,n\}$为由 $n+1$ 个符号组成的集合,$S$ 的反向符号集合记为 $S^-=\{-0,-1,\ldots,-n\}$.考虑由 $S$ 的所有的有向符号组成的一个排列 $\pi=\pi_0\pi_1\pi_2\ldots\pi_n$,若 $|\pi|=|\pi_0|,|\pi_1|,|\pi_2|,\ldots,|\pi_n|$恰为符号集 $S$ 的一个置换,则 $\pi$称为长度为 $n+1$ 的有向符号序列,并称 $S$ 为序列 $\pi$的符号集,记为 $S=S(\pi)$.其中 $\pi_k\in S\cup S^-,|\pi_k|\in S$.一个有向符号序列实际表示一条染色体,其中 $\pi_k$ 表示一个基因,$|\pi_k|$为该基因的符号,符号前的正负号表达了该基因的方向.有向符号序列 $\pi$在位置 $(i,j)$的一次翻转变异 $\rho$形成一个新的序列:$\pi\cdot\rho(i,j)=\pi_1\ldots\pi_{i-1}-\pi_j\ldots-\pi_i\pi_{j+1}\ldots\pi_n$.设有符号序列 $\pi^1$ 和 $\pi^2$,若存在一组翻转 $\rho_1,\rho_2,\ldots,\rho_k$ 使 $\pi^1\cdot\rho_1\cdot\rho_2\cdot\ldots\cdot\rho_k=\pi^2$,且 $k$ 是最小的,则称 $\pi^1$ 和 $\pi^2$ 的翻转距离为 $k$,记为 $d(\pi^1,\pi^2)=k$.欲计算两序列的翻转距离,首先构造 RD 图 $G_R(V(S),E(\pi^1),E(\pi^2))$.$V(S)$为该图的点集:$V(S)=\{v_{kh},v_{kt}|0\le k\le n\}$,并将 $G_R$ 的顶点根据 $\pi^1$ 由左至右排列为 $v_{01},v_{02},v_{11},v_{12},\ldots v_{n1},v_{n2}$,若 $\pi^1_k$ 方向为正,则 $v_{k1},v_{k2}=v_{kh},v_{kt}$;否则 $v_{k1},v_{k2}=v_{kt},v_{kh}$.$E(\pi^1)=\{e\left(\pi^1_k,\pi^1_{k+1}\right)|0\le k\le$

* 收稿日期: 2000-07-10; 修改日期: 2001-03-01
基金项目: 国家自然科学基金资助项目(60073042);国家教育部青年教师基金资助项目(y66053;060602);山东省中青年科学家奖励基金资助项目(01bs03)

作者简介: 朱大铭(1964－),男,山东济南人,博士,教授,主要研究领域为算法分析与设计,神经网络;马绍汉(1938－),男,山东济南人,教授,博士生导师,主要研究领域为算法分析与设计,人工智能;雷鹏(1964－),男,山东济南人,工程师,主要研究领域为算法分析与设计.

$n-1\}$为黑色边集;$E(\pi^2)=\{e(\pi_k^2,\pi_{k+1}^2)|0\le k\le n-1\}$为灰色边集,设$|\pi_k^t|=i$, $|\pi_{k+1}^t|=j,0\le i,j\le n,i\ne j,t=1,2$,则有:(a) 若$\pi_k^t=i$, $\pi_{k+1}^t=j$,则 $e(\pi_k^t,\pi_{k+1}^t)=(v_{it},v_{jh})$;(b) 若$\pi_k^t=i$, $\pi_{k+1}^t=-j$,则 $e(\pi_k^t,\pi_{k+1}^t)=(v_{it},v_{jt})$;(c) 若$\pi_k^t=-i$, $\pi_{k+1}^t=j$,则$e(\pi_k^t,\pi_{k+1}^t)=(v_{ih},v_{jh})$;(d) 若$\pi_k^t=-i$, $\pi_{k+1}^t=-j$,则$e(\pi_k^t,\pi_{k+1}^t)=(v_{ih},v_{jt})$.$G_R(V(S),E(\pi^1),E(\pi^2))$简称为$\pi^1,\pi^2$的 RD 图.

例如,设$\pi^1$=01-3245,$\pi^2$=0-1-4235,则$G_R$如图 1 所示.对于有向符号序列$\pi^1,\pi^2$,$G_R(V,E(\pi^1),E(\pi^2))$总由若干互不连通的圈组成,每个圈的两色边间隔相连,将$G_R$的所有圈记为$C(G_R)=\{c_1,c_2,\ldots,c_k\}$.一条黑边按照端点在$G_R$中的排列顺序可分为左端和右端,同样一条灰边的两端也可分为左端和右端.若一条灰边连接两条黑边的同一端,则该灰边称为有序边,否则称为无序边.至少包含一条有序边的$G_R$的圈称为有序圈,否则称为无序圈.若有两条灰边$e_1$和$e_2$使$e_2$的一端位于$e_1$的两端之间,$e_2$的另一端位于$e_1$两端外面,则称灰边$e_1$和$e_2$是相交的;若相交的两条灰边分别属于$G_R$的两个圈$c_1$和$c_2$,则称$c_1$和$c_2$是相交的.设$C$是$C(G_R)$的子集,若(1) 对任意$c_i\in C$,总有$c_j\in C$与$c_i$相交;(2) 不存在$c_k\in C(G_R)\backslash C$与$C$中圈相交;(3) 不存在$C$的子集满足(1)和(2),则称$C$中圈形成的$G_R$的子图为$G_R$的一个 C_图(component).若一个 C_图$G_{RC1}$的所有顶点均位于另一个 C_图$G_{RC2}$的两个顶点$v_1$和$v_2$之间,则称$G_{RC2}$覆盖$G_{RC1}$,设有$k(\ge 2)$个 C_图$G_{RC1},G_{RC2},\ldots,G_{RCk}$被 C_图$G_{RC}$覆盖,若对任意$v\in V(G_{RC}),G_{RC1},G_{RC2},\ldots,G_{RCk}$的所有顶点均位于$v$的左端或右端,则称$G_{RC}$严格覆盖$G_{RCj},1\le j\le k$.若$G_{RC1}$的所有顶点均位于$G_{RC2}$的任一顶点的左边或右边,则称$G_{RC1}$与$G_{RC2}$并列.若一个 C_图中的所有圈均为无序的,则称该 C_图为无序的.若一个无序 C_图$G_{RC1}$不覆盖其他任何无序 C_图,则称$G_{RC1}$为一个 H_图(hurdle);另外,若一个无序 C_图$G_{RC1}$在$G_R$中严格覆盖其他所有无序 C_图,则$G_{RC1}$也称为一个 H_图.设$G_H$是一个 H_图,若$G_H$被一个无序 C_图$G_C$严格覆盖,但$G_C$不是 H_图,且$G_C$不覆盖其他 H_图,则称$G_H$是一个强 H_图.若$G_R$中的 H_图均为强 H_图,且强 H_图个数为奇数,则称$G_R$含有一个 F_图(fortress),任意 RD 图最多只有一个 F_图.设$G_R(V(S),E(\pi^1),E(\pi^2))$共有$b$条黑边,$c$个圈,$h$个 H_图,$f$个 F_图,则$\pi^1$和$\pi^2$之间的翻转距离为$d(\pi^1,\pi^2)=b-c+h+f$.

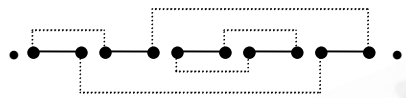

Fig.1 RD graph
图 1 RD 图

## 1 星树问题是 NP_完全的

考虑建立在翻转距离计算基础上的星树问题 ST,ST 问题的判定形式如下:

实例:一组长度为$n$的有向染色体符号序列$\pi^1,\pi^2,\ldots,\pi^m$,正整数$M$.

询问:是否存在一个有向符号序列$\pi^*$,使$\sum_{i=1}^m d(\pi^i,\pi^*)\le M$.

已知 3SAT 问题是 NP_完全的[8].3SAT 问题形式化描述如下:

实例:(1) 布尔变量集合:$X=\{x_1,x_2,\ldots,x_n\}$及其反变量集合$\overline{X}=\{\overline{x}_1,\overline{x}_2,...,\overline{x}_n\}$;(2) 项集合:$C=\{c_1,c_2,\ldots,c_m\}$,其中$c_i=\{c(x_{i_1}),c(x_{i_2}),c(x_{i_3})\}$,$1\le i_1,i_2,i_3\le n$,$c(x_k)\in\{x_k,\overline{x}_k\}$,$1\le k\le n$.

询问:是否存在$X$的真值指派使$X\to\{0,1\}$使$C=\bigwedge_{i=1}^m(c(x_{i_1})\vee c(x_{i_2})\vee c(x_{i_3}))$为真或$C$满足.

定义 **1.** 在 3SAT 的实例中,设$x_i$是任一布尔变量,$x_i$和$\overline{x}_i$分别称为布尔变量$x_i$的正变量和负变量,并将$x_I$有关的项集合记为$C(x_i)=\{c_k|x_i\in c_k,1\le k\le m\}$,$C(\overline{x}_i)=\{c_k|\overline{x}_i\in c_k,1\le k\le m\}$.

若 3SAT 实例满足:(1) $m$为偶数;(2) 对任意布尔变量$x_i$,若$C(x_i)\ne\varnothing\wedge C(\overline{x}_i)\ne\varnothing$,则$|C(x_i)|=|C(\overline{x}_i)|$,则相应的 3SAT 问题称为 E3SAT 问题.

定理 **1.** E3SAT 是 NP-完全的.

证明:将 3SAT 问题规约到 E3SAT.设 3SAT 的实例为$X_{3SAT}$和$C_{3SAT}$,构造 E3SAT 的实例如下:

$$X_{E3SAT}=X_{3SAT}\cup X_E\cup X_N,C_{E3SAT}=C_{3SAT}\cup C_E\cup C_N,X_E=\bigcup_{i=1}^n X_{E_i},C_E=\bigcup_{i=1}^n C_{E_i}. \quad (1)$$

考察布尔变量$x_i\in X_{3SAT},1\le i\le n$,若$C_{3SAT}(x_i)=\varnothing$或$C_{3SAT}(\overline{x}_i)=\varnothing$,或$|C_{3SAT}(x_i)|=|C_{3SAT}(\overline{x}_i)|$,则$X_{E_i}=\varnothing,C_{E_i}=\varnothing$.若

$C_{3SAT}(x_i)\neq\varnothing\wedge C_{3SAT}(\bar{x}_i)\neq\varnothing\wedge|C_{3SAT}(x_i)|\neq|C_{3SAT}(\bar{x}_i)|$,则如下构造 $X_{E_i}$ 和 $C_{E_i}$:

(a) 若 $|C_{3SAT}(x_i)|>|C_{3SAT}(\bar{x}_i)|$,设 $k=|C_{3SAT}(x_i)|-|C_{3SAT}(\bar{x}_i)|$,则 $X_{E_i}=\{y_{i1},y_{i2},\ldots,y_{i2k-1},y_{i2k}\}$; $C_{E_i}=\{\{\bar{x}_i,y_{i1},y_{i2}\},\{\bar{x}_i,y_{i3},y_{i4}\},\ldots,\{\bar{x}_i,y_{i2k-1},y_{i2k}\}\}$;

(b) 若 $|C_{3SAT}(x_i)|<|C_{3SAT}(\bar{x}_i)|$,设 $k=|C_{3SAT}(\bar{x}_i)|-|C_{3SAT}(x_i)|$,则 $X_{E_i}=\{y_{i1},y_{i2},\ldots,y_{i2k-1},y_{i2k}\}$; $C_{E_i}=\{\{x_i,y_{i1},y_{i2}\},\{x_i,y_{i3},y_{i4}\},\ldots,\{x_i,y_{i2k-1},y_{i2k}\}\}$.

增加 $X_N$ 和 $C_N$ 使 $C_{E3SAT}$ 中恰有偶数个项:若 $|C_{3SAT}\cup C_E|$ 为奇数,则 $X_N=\{x_{N1},x_{N2},x_{N3}\}$,$C_N=\{\{x_{N1},x_{N2},x_{N3}\}\}$;若 $|C_{3SAT}\cup C_E|$ 为偶数,则 $X_N=\varnothing$,$C_N=\varnothing$.以上构造的 $X_{E3SAT}$ 和 $C_{E3SAT}$ 显然满足 E3SAT 问题的条件,且该规约可在 $O(mn)$时间内完成.

若 3SAT 实例存在使 $C_{3SAT}$ 满足的 $X_{3SAT}$ 的真值指派,则令 $X_E$ 和 $X_N$ 中的布尔变量取值为‘1’,所有 E3SAT 中的项必满足.若 E3SAT 实例存在使 $C_{E3SAT}$ 满足的 $X_{E3SAT}$ 的真值指派,则其中 $X_{3SAT}$ 的赋值就可使 $C_{3SAT}$ 的所有项满足.故 E3SAT 问题是 NP-完全的. □

在下文的叙述中,直接采用整数“变量表达式”表示该表达式所代表的整数“符号”.

定理 **2**. ST 问题是 NP-完全的.

证明:将 E3SAT 问题规约到 ST.设 $X=\{x_1,x_2,\ldots,x_n\}$,$C=\{c_1,c_2,\ldots,c_m\}$是 E3SAT 的实例,欲构造的整数符号序列分为两部分,每条序列长度均为 $2n+2$,第 1 部分有 $m$ 条序列:$\pi^1,\pi^2,\ldots,\pi^m$.$m$ 条序列的中间符号及其方向根据 E3SAT 实例的 $m$ 个项决定.设 $c_k=\{c(x_{k_1}),c(x_{k_2}),c(x_{k_3})\}$,$c(\bullet)\in\{\bullet,\bar{\bullet}\}$,$1\le k_1<k_2<k_3\le n$,$1\le k\le m$,则 $\pi^k=\pi_0^k\,\pi^{k0}\pi^{k1}\pi_{2k_1-1}^k\,\pi^{k2}.\pi_{2k_2-1}^k\,\pi^{k3}\,\pi_{2k_3-1}^k\,\pi_{2n+1}^k$,其中

$$\pi_0^k=0,\quad \pi_{2n+1}^k=2n+1, \tag{2}$$

$$\pi_{2j-1}^k=\begin{cases}2j-1 & c(x_j)=x_j\\ -2j+1 & c(x_j)=\bar{x}_j\end{cases},\quad j\in\{k_1,k_2,k_3\}. \tag{3}$$

又 $\pi^{k0},\pi^{k1},\pi^{k2}$ 和 $\pi^{k3}$ 均为有向符号子序列,或称为有向符号段.先给出其每个位置上的符号:

$$\pi^{k0}=\pi_1^{k0}\ldots\pi_{2k_1-2}^{k0}. \tag{4}$$

$$\pi_{2j}^{k0}=2j,\ 1\le j\le k_1-1;\quad |\pi_{2j-1}^{k0}|=2j-1,\ 1\le j\le k_1-1. \tag{5}$$

$$\pi^{k1}=\pi_{2k_1}^{k1}\ldots\pi_{2k_2-2}^{k1}. \tag{6}$$

$$\pi_{2j}^{k1}=2j,\ k_1\le j\le k_2-1;\quad |\pi_{2j-1}^{k1}|=2j-1,\ k_1+1\le j\le k_2-1. \tag{7}$$

$$\pi^{k2}=\pi_{2k_2}^{k2}\ldots\pi_{2k_3-2}^{k2}. \tag{8}$$

$$\pi_{2j}^{k2}=2j,\ k_2\le j\le k_3-1;\quad |\pi_{2j-1}^{k2}|=2j-1,\ k_2+1\le j\le k_3-1. \tag{9}$$

$$\pi^{k3}=\pi_{2k_3}^{k3}\ldots\pi_{2n}^{k3}. \tag{10}$$

$$\pi_{2j}^{k3}=2j,\ k_3\le j\le n;\quad |\pi_{2j-1}^{k3}|=2j-1,\ k_3+1\le j\le n. \tag{11}$$

由符号序列的构造方法可知,绝对值为 $2j-1$ 的符号对应于 E3SAT 实例的布尔变量 $x_j$.在前面的符号序列构造过程中,每条序列还剩 $n-3$ 个符号尚未确定方向,下面给出其方向.设 $S=\{\pi_{2j-1}^{k\bullet}\mid 1\le k\le m, x_j,\bar{x}_j\notin c_k, j\text{ 固定}\}$为 $x_j$ 在 $m$ 条序列中对应的未定方向符号集合,$|S|=m-|C(x_j)|-|C(\bar{x}_j)|$,若 $C(\bar{x}_j)=\varnothing$,则 $S$ 中每个符号均取负向:$\pi_{2j-1}^{k\bullet}=-2j+1$;若 $C(x_j)=\varnothing$,则 $S$ 中每个符号均取正向:$\pi_{2j-1}^{k\bullet}=2j-1$.不然必有 $|C(\bar{x}_j)|=|C(x_j)|$,因 $m$ 为偶数,所以 $|S|=m-2|C(x_j)|$为偶数.随机在 $S$ 中取 $|S|/2$ 个元素取为正向,另外 $|S|/2$ 个元素取为负向.

因 $m$ 为偶数,所以 $m(n-3)/2+6m$ 为整数,取 $K$ 为不小于 $m(n-3)/2+6m$ 的最小偶数,构造 ST 实例的第 2 部分符号序列,共 $K$ 条:$\pi^{m+1},\pi^{m+2},\ldots,\pi^{m+K}$,其中

$$\pi^k=\pi_0^k\pi_1^k\ldots\pi_{2n}^k\pi_{2n+1}^k,\ m+1\le k\le m+K. \tag{12}$$

$$\pi_0^k=0,\quad \pi_{2n+1}^k=2n+1,\ m+1\le k\le m+K. \tag{13}$$

$$\pi^k_{2j}=2j,\ 1\le j\le n,\ m+1\le k\le m+K. \tag{14}$$

$$\pi^k_{2j-1}=\begin{cases}2j-1 & 1\le j\le n,\ m+1\le k\le K/2\\ -2j+1 & 1\le j\le n,\ K/2+1\le k\le m+K\end{cases}. \tag{15}$$

取 $M=Kn/2+m(n_1-3)/2+6m$,其中 $n_1=n-|\{x_i|C(x_i)=\varnothing\vee C(\bar{x}_i)=\varnothing,1\le i\le n\}|$.根据 E3SAT 实例构造 ST 实例的最坏复杂度为 O($Kn$),显然是多项式的.

(→)若 E3SAT 实例存在 $X$ 的真值指派 $a:X\to\{0,1\}$使所有项满足.不失一般性假设:若 $C(x_j)=\varnothing$,则 $a(x_j)=0$;若 $C(\bar{x}_j)=\varnothing$,则 $a(x_j)=1$.如下确定一条符号序列:$\pi^*=\pi^*_0\pi^*_1\ldots\pi^*_{2n+1}$.

$$\pi^*_0=0,\ \ \pi^*_{2n+1}=2n+1,\ \ \pi^*_{2j}=2j,\ 1\le j\le n. \tag{16}$$

$$\pi^*_{2j-1}=\begin{cases}-2j+1 & a(x_j)=1\\ +2j-1 & a(x_j)=0\end{cases},\quad 1\le j\le n. \tag{17}$$

考察$\pi^*$与第 1 部分序列中任意一条序列$\pi^k$的翻转距离.如图 2 所示,在图 $G_R(V,E(\pi^*),E(\pi^k))$中不可能形成强 H_图,因此无 F_图,即 $f=0$.在图中将顶点$(2j-1)t$ 与$(2j-1)h$ 并为一点,真值指派 $a(x_{k_1}),a(x_{k_2}),a(x_{k_3})$确定了 $\pi^*_{2k_1-1}$,$\pi^*_{2k_2-1}$和 $\pi^*_{2k_3-1}$的方向,因此确定了$(2j-1)t$ 与$(2j-1)h$ 的排列顺序,$j\in\{k_1,k_2,k_3\}$.不论 3 个符号的方向如何,点集$\{(2j-1)h,(2j-1)t|j=k_1,k_2,k_3\}$ 中的点总处在同一个 7 条黑边的圈 $C^{7b}$ 中.因为 $X$ 的真值指派使 $c(x_{k_1})\vee c(x_{k_2})\vee c(x_{k_3})=1$,不妨设 $c(x_{k_1})=1$.若 $c(x_{k_1})=x_{k_1}$,则 $\pi^k_{2k_1-1}=2k_1-1$, $\pi^*_{2k_1-1}=-2k_1+1$,此时灰边($v_{(2k_1-1)h}$,$v_{(2k_2-2)t}$)是有序边,因此 $C^{7b}$ 是有序圈;同理可证,当 $c(x_{k_1})=\bar{x}_{k_1}$ 时,$C^{7b}$ 也为有序圈,所以 $h=0$.又因为当 $C(x_i)=\varnothing$或 $C(\bar{x}_i)=\varnothing$时,必有 $d(\pi^*_{2i-1},\pi^k_{2i-1})=0$,故由以上分析不难得到

$$\sum_{k=1}^{m+K}d(\pi^*,\pi^k)=Kn/2+m(n_1-3)/2+6m\le M. \tag{18}$$

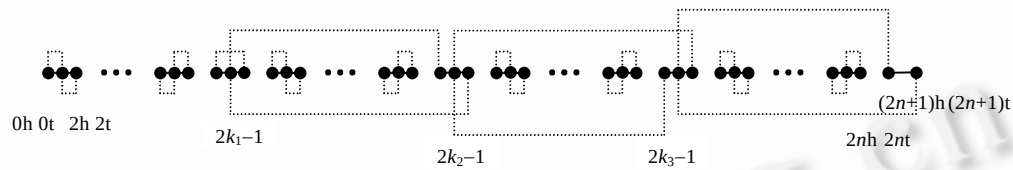


Fig.2 RD graph based on $\pi^*$ and $\pi^k$

图 2 由$\pi^*$和$\pi^k$得到的 RD 图

(←)若 ST 实例存在$\pi^*$使 $\sum_{k=1}^{m+K}d(\pi^*,\pi^k)\le M$,则$\pi^*$的符号顺序一定是 $\pi^*_{2k}=2k$,$|\pi^*_{2k+1}|=2k+1$,$0\le k\le n$.这是因为否则对于 $i,j,m+1\le i\le m+K/2,m+K/2+1\le j\le m+K$,必有 $d(\pi^*,\pi^i)+d(\pi^*,\pi^j)\ge n+2$,因此易证$\pi^*$与两部分符号序列翻转距离之和满足:

$$\sum_{k=1}^{m}d(\pi^*,\pi^k)\le 6m+m(n_1-3)/2,\quad \sum_{k=m+1}^{m+K}d(\pi^*,\pi^k)\le Kn/2. \tag{19}$$

由$\pi^*$和$\pi^k(1\le k\le m)$形成的 RD 图 $G_R(V,E(\pi^*),E(\pi^k))$仅有一个 7 条黑边的最大圈,且该圈是有序圈.因此,只需如下确定 $X$ 的真值指派:

$$A(x_i)=\begin{cases}1, & \text{if } \pi^*_{2i-1}=-2i+1\\ 0, & \text{if } \pi^*_{2i-1}=2i-1\end{cases},\quad 1\le i\le n, \tag{20}$$

就可使 $m$ 个项全部满足. □

## 2 星树问题的近似算法

下面讨论 ST 问题的近似算法.ST 问题的优化形式为:给定一组长度为 $n+1$ 的有向染色体符号序列$\pi^1,\pi^2,\ldots,\pi^m$,求$\pi^*$使 $\sum_{i=1}^n d(\pi^i,\pi^*)=\min\{\sum_{i=1}^n d(\pi^i,\pi^\forall)\}$.

定理 **3**. 若 P≠NP,则对 ST 问题的实例$\Pi$,不存在近似算法 $A$ 使 $A(\Pi)-OPT(\Pi)\leq k$,$k$ 为常数.

证明:只需将$\Pi$的 $m$ 条序列分别复制 $K+1$ 份得到新实例$\Pi_{new}$.易证若存在算法 $A$ 使 $A(\Pi_{new})-OPT(\Pi_{new})\leq k$,则可给出实例$\Pi$的多项式算法,与 P≠NP 矛盾. □

下面给出一个近似性能比为 2 的 ST 问题近似算法.算法描述如下:

设$\pi^1,\pi^2,\ldots,\pi^m$ 为给定 $m$ 条有向符号序列,则算法 $A(\pi^1,\pi^2,\ldots,\pi^m)$选择$\pi^1,\pi^2,\ldots,\pi^m$ 中一个序列$\pi^k$ 满足:$\sum_{i=1}^{m}d(\pi^k,\pi^i)=\min\{\sum_{i=1}^{m}d(\pi^j,\pi^i)\,|1\leq j\leq m\}$,算法 $A$ 的输出为$\pi^k$和$\sum_{i=1}^{m}d(\pi^k,\pi^i)$.

定理 **4**. 对任意 ST 问题实例$\Pi$,设算法 $A$ 求得解为 $A(\Pi)=d^A$,则 $A(\Pi)/OPT(\Pi)\leq 2$.

证明:设$\Pi$的 $m$ 个有向符号序列为:$\pi^1,\pi^2,\ldots,\pi^m$.记 $d_{ij}=d(\pi^i,\pi^j)$,显然 $d_{ij}=d_{ji}$.设$\pi^*$满足$\sum_{i=1}^{m}d(\pi^*,\pi^i)=\min\{\sum_{i=1}^{m}d(\pi^\forall,\pi^i)\ |S(\pi^\forall)=S(\pi^1)\}=OPT(\Pi)$.根据三角不等式有 $d(\pi^*,\pi^i)+d(\pi^*,\pi^j)\geq d_{ij}$.故,$2(m-1)\sum_{i=1}^{m}d(\pi^*,\pi^i)\geq\sum_{i=1}^{m}\sum_{j=1}^{m}d_{ij}$.由算法 $A$ 求得的$\pi^A$满足:$d^A=\sum_{i=1}^{m}d(\pi^A,\pi^i)\leq\sum_{i=1}^{m}d(\pi^k,\pi^i)$,$1\leq k\leq m$.由此易得$\sum_{i=1}^{m}\sum_{j=1}^{m}d_{ij}\geq m\sum_{i=1}^{m}d(\pi^A,\pi^i)=md^A$.所以,

$$\frac{A(\Pi)}{OPT(\Pi)}=\frac{d^A}{\sum_{i=1}^{m}d(\pi^*,\pi^i)}\leq 2\left(1-\frac{1}{m}\right)\leq 2. \tag{21}$$

由 Hannenhalli 等人给出的方法可知[7],完成长度为 $n$ 的两个有向符号序列的翻转距离计算,最坏时间复杂度为 $O(n^5)$.由于算法 $A$ 共调用该算法 $m^2$ 次,因此其最坏时间复杂度为 $O(m^2n^5)$.

**References**:


[1] Wang, Lu-sheng, Jiang Tao. Approximation algorithms for tree alignment with a given phylogeny. Algorithmica, 1996, 16(3):302~315.

[2] Gusfield, D. Efficient algorithms for inferring evolutionary trees. Networks, 1991,21(1):19~28.

[3] Kececioglu, J., Sankoff, D. Exact and approximation algorithms for the reversal distance between two permutations. Algorithmica, 1995,13(1):180~210.

[4] Kececioglu, J., Sankoff, D. Efficient bounds for oriented chromosome inversion distance. In: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching. Lecture Notes in Computer Science 807, 1994. 307~325.

[5] Bafna, V., Pevzner, P.A. Sorting by reversals: genome rearrangements in plants organelles and evolutionary history of X chromosome. Molecular Biology and Evolution, 1995,12:239~246.

[6] Kececioglu, J.D., Myers, E.W. Combinatorial algorithms for DNA sequence assembly. Algorithmica, 1995,13(1):7~15.

[7] Hannenhalli, S., Pevzner, P.A. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In: Proceedings of the 27th Annual ACM Symposium on Theory of Computing (STOC'95). ACM press, 1995. 178~189.

[8] Ma, Shao-han, Designation and Analysis of Algorithms. Ji'nan: Shandong University Press, 1992 (in Chinese).


附中文参考文献:


[8] 马绍汉.算法分析与设计.济南:山东大学出版社,1992.

# Computational Complexity and an Approximation Algorithm for Star-Tree Phylogeny Problem with Reversal Distance*

ZHU Da-ming, MA Shao-han, LEI Peng

(*School of Computer Science and Technology, Shandong University, Ji'nan* 250100, *China*)

E-mail: dmzhu@sdu.edu.cn

http://cs.sdu.edu.cn

**Abstract:** In this paper, the algorithms and the computational complexity of Star-Tree phylogeny problem are studied. The Star-Tree phylogeny problem is proved to be NP-complete first. Then it is proved that there is no absolute approximation algorithm for this problem. At last, a polynomial approximation algorithm of ratio 2 is presented to compute the Star-Tree phylogeny problem.

**Key words:** algorithm; phylogeny; genome; NP-completeness; approximation ratio

* Received July 10, 2000; accepted March 1, 2001

Supported by the National Natural Science Foundation of China under Grant No.60073042; the Chinese Educational Branch Foundation for Youth Teacher under Grant Nos.y66053, 060602; the Middle Age or Youth Award Foundation of Shandong Province of China under Grant No.01bs03

## 敬告作者

《软件学报》创刊以来,蒙国内外学术界厚爱,收到许多高质量的稿件,其中不少在发表后读者反映良好,认为本刊保持了较高的学术水平.但也有一些稿件因不符合本刊的要求而未能通过审稿.为了帮助广大作者尽快地把他们的优秀研究成果发表在我刊上,特此列举一些审稿过程中经常遇到的问题,请作者投稿时尽量予以避免,以利大作的发表.

1. 读书偶有所得,即匆忙成文,未曾注意该领域或该研究课题国内外近年来的发展情况,不引用和不比较最近文献中的同类结果,有的甚至完全不列参考文献.

2. 做了一个软件系统,详尽描述该系统的各个方面,如像工作报告,但采用的基本上是成熟技术,未与国内外同类系统比较,没有指出该系统在技术上哪几点比别人先进,为什么先进.一般来说,技术上没有创新的软件系统是没有发表价值的.

3. 提出一个新的算法,认为该算法优越,但既未从数学上证明比现有的其他算法好(例如降低复杂性),也没有用实验数据来进行对比,难以令人信服.

4. 提出一个大型软件系统的总体设想,但很粗糙,而且还没有(哪怕是部分的)实现,很难证明该设想是现实的、可行的、先进的.

5. 介绍一个现有的软件开发方法,或一个现有软件产品的结构(非作者本人开发,往往是引进的,或公司产品),甚至某一软件的使用方法.本刊不登载高级科普文章,不支持在论文中引进广告色彩.

6. 提出对软件开发或软件产业的某种观点,泛泛而论,技术含量少.本刊目前暂不开办软件论坛,只发表学术文章,但也欢迎材料丰富,反映现代软件理论或技术发展,并含有作者精辟见解的某一领域的综述文章.

7. 介绍作者做的把软件技术应用于某个领域的工作,但其中软件技术含量太少,甚至微不足道,大部分内容是其他专业领域的技术细节,这类文章宜改投其他专业刊物.

8. 其主要内容已经在其他正式学术刊物上或在正式出版物中发表过的文章,一稿多投的文章,经退稿后未作本质修改换名重投的文章.

本刊热情欢迎国内外科技界对《软件学报》踊跃投稿.为了和大家一起办好本刊,特提出以上各点敬告作者.并且欢迎广大作者和读者对本刊的各个方面,尤其是对论文的质量多多提出批评建议.