日本オペレーションズ・リサーチ学会
2022年 春季研究発表会

ℓ_2 最小増加超距離木問題に対する k 制限部分木交換近傍に基づく局所探索アルゴリズム

01013123　静岡大学　*安藤和敏　ANDO Kazutoshi
静岡大学　水越雅紀　MIZUKOSHI Masaki

1. はじめに

系統樹とは生物の進化の歴史を表す木であり, 系統樹を推定することは分子系統学における重要な課題である. 本研究では, 求める系統樹として**超距離木**を考える. 超距離木とは根から任意の葉への距離がすべて等しいような枝重み付き根付き木である. 一般性を失わず超距離木は 2 分木であると仮定する. 系統樹の推定に用いられる主な方法の一つに距離法がある. X を考察の対象である生物種の集合とする. 距離法では, まず生物種間の距離を表す行列 $M\colon X\times X\to\mathbb{R}_+$ を求める. その後に以下の問題を解く.

$$\left|\begin{array}{ll}\min & \|D_{(T,l)}-M\|_p\\ \text{s.t.} & (T,l)\text{ は葉集合が }X\text{ であるような超距離木.}\end{array}\right.$$

ここで, 各 $x,y\in X$ に対して, $D_{(T,l)}[x,y]$ は (T,l) における x-y 間の距離であり, $\|\cdot\|_p$ は ℓ_p ノルムを表す. 上の問題は ℓ_p **最良近似超距離木問題**と呼ばれる. また, 入力として与えられた生物種間の相違は実際は真の相違の下限であるという場合がしばしばある. このような状況においては最良近似超距離木問題に対して $M\le D_{(T,l)}$ という制約を追加したものを考える方がより適切である. この問題は ℓ_p **最小増加超距離木問題**と呼ばる. 最良近似超距離木問題と最小増加超距離木問題は, $p=\infty$ のときは多項式時間アルゴリズムが存在するが $p<\infty$ のときは NP-困難である.

$p<\infty$ の場合の ℓ_p 最小増加超距離木問題に対するアルゴリズムとして, 分枝限定法 [4] や近似アルゴリズム [1] がある. また石川他 [3] は, NNI (最小近傍交換) 操作 や SS (部分木交換) 操作と呼ばれる 2 分木の変形操作に基づいてこの問題に対する局所探索アルゴリズムを導入した. 最近, 安藤他 [2] は, NNI 操作と SS 操作を同時に一般化する 2 分木の変形操作である k 制限部分木交換操作 (kSS 操作) を導入し, $p=1$ の場合の ℓ_p 最小増加超距離木問題に対する kSS 操作に基づく局所探索アルゴリズムを提案した.

本研究では, 安藤他 [2] の結果が $p=2$ の場合の ℓ_p 最小増加超距離木問題に対する kSS 操作に基づく局所探索アルゴリズムの 1 反復あたりの計算時間が $p=1$ のときのそれと同じ $\mathrm{O}(n\min\{2^{k+1},n\}k)$ であることを示す. ここで, $n=|X|$ である. さらに, このアルゴリズムの実際的計算量を数値実験によって検証する.

2. 局所探索アルゴリズム

行列 $M\colon X\times X\to\mathbb{R}_+$ は, 任意の $x,y\in X$ に対して $M[x,y]=M[y,x]$ かつ $M[x,x]=0$ を満たすとき, X **上の相違行列**と呼ばれる. 根付き 2 分木 $T=(V,E)$ は, その葉集合が X と等しいときに X **上の根付き 2 分木**と呼ばれる. 本研究では根付き 2 分木を全ての枝が根から葉に向かって向き付けられた有向木とみなす. X 上の根付き木 T の点への重み付け $h\colon V\to\mathbb{R}_+$ が, 任意の $v\in X$ に対して $h(v)=0$ かつ, 任意の $(v,w)\in E$ に対して $h(w)\le h(v)$ を満たすとき, (T,h) を**単調点重み付き根付き木**と呼ぶ. 任意の単調点重み付き根付き木 (T,h) に対して, $D_{(T,h)}[x,y]=2h(\mathrm{lca}(x,y))$ と定義する. ここで, $\mathrm{lca}(x,y)$ は T 中の x と y の最小共通祖先である. ℓ_p 最小増加超距離木問題は, 以下の問題と等価である.

$$\left|\begin{array}{ll}\min & \|D_{(T,h)}-M\|_p\\ \text{s.t.} & (T,h)\text{ は }M\le D_{(T,h)}\text{ であるような}\\ & X\text{ 上の単調点重み付き木.}\end{array}\right.\qquad(1)$$

X 上の根付き 2 分木 T を固定して, (T,h) が (1) の最適解となるような点重み関数 $h\colon V\to\mathbb{R}_+$ を見出す問題は **MUTT 問題**と呼ばれる. M と T に対する MUTT 問題の最適解を $h(M,T)$ によって表し, $f_p(T)=\|D_{(T,h(M,T))}-M\|_p$ とする. アルゴリズム 1 は ℓ_p 最小増加超距離木問題に対する局所探索アルゴリズムである [3].

アルゴリズム 1 中の $\mathcal{N}(T)$ として以下で定義する近傍を考える. $1\le k\le n$ とする. 根付き 2 分木 T に対して, $e_1=(v_1,w_1)$ と $e_2=(v_2,w_2)$ を $v_2\ne v_1$ かつ w_1 と w_2 が子孫関係になく, v_1 から v_2 の道の長さが k 以下であるような 2 つの枝とする. T から枝 e_1,e_2 を削除した後 $(v_1,w_2),(v_2,w_1)$ を挿入する操作は, e_1 と e_2 による k **制限部分木交換操作** (kSS 操作) と呼ばれる [2]. 1SS 操作は NNI 操作であり, nSS 操作は SS 操

アルゴリズム 1: 局所探索アルゴリズム.

$T \leftarrow$ 任意の根付き 2 分木;
do
  $\bar{T} \leftarrow \mathrm{argmin}\{f_p(T') \mid T' \in \mathcal{N}(T)\}$;
  if $f_p(\bar{T}) < f_p(T)$ **then** $T \leftarrow \bar{T}$;
while T は局所最適でない;

作である. T から 1 回の kSS 操作によって得られる 2 分木の集合を T の k **制限部分木交換近傍** (kSS 近傍) と呼ぶ. kSS 近傍のサイズは $\mathrm{O}(\min\{2^{k+1}, n\}n)$ であり, MUTT 問題を解く $\mathrm{O}(n^2)$ 時間アルゴリズムが存在するため [4], アルゴリズム 1 の 1 反復あたりの計算時間は $\mathrm{O}(\min\{2^{k+1}, n\}n^3)$ である.

3. $p = 2$ の場合のアルゴリズムの高速化

M を X 上の任意の相違行列, $T = (V, E)$ を X 上の根付き 2 分木とする. (M, T) に対する**総和行列** $H\colon V \times V \to \mathbb{R}_+$ を

$$H[v, w] = \sum_{\substack{x \in L(T_w) - L(T_v), \\ y \in L(T_v)}} M[x, y] \tag{2}$$

で定義する. ここで, T_v は v を根とする T の部分木を表わし, $L(T_v)$ は T_v の葉集合を表す.

T' を T に対する $e_1 = (v_1, w_1)$ と $e_2 = (v_2, w_2)$ による kSS 操作で得られる 2 分木とし, P を v_1 から v_2 への T 中の道とする.

補題 3.1 T の任意の内部点 v に対して, T における v の子を v_+, v_-, T' における v の子を v'_+, v'_-, (M, T') に対する総和行列を H', $h = h(M, T)$, $h' = h(M, T')$ とする. このときが成り立つ.

$$\begin{aligned} f_p(T') - f_p(T) = 8 \sum_{v \in V(P)} & (|L(T'_{v'_+})||L(T'_{v'_-})|h'(v)^2 \\ & - |L(T_{v_+})||L(T_{v_-})|h(v)^2 \\ & - H'[v'_+, v'_-]h'(v) + H[v_+, v_-]h(v)). \end{aligned}$$

(M, T) に対する**最大値行列** $K\colon V \times V \to \mathbb{R}_+$ を

$$K[v, w] = \begin{cases} \max\limits_{\substack{x \in L(T_{w_+}) - L(T_v), \\ y \in L(T_{w_-}) - L(T_v)}} M[x, y]/2 & \text{if } w \text{ が } v \text{ の祖先}, \\ 0 & \text{if } v \text{ が } w \text{ の祖先}, \\ \max\limits_{\substack{x \in L(T_v), \\ y \in L(T_w)}} M[x, y]/2 & \text{otherwise} \end{cases}$$

で定義する.

補題 3.2 (M, T) に対する総和行列 H, 最大値行列 K, $h(M, T)$ および $|L(T_v)|$ $(v \in V)$ が与えられているとする. このとき, $f_p(T')$ は $\mathrm{O}(k)$ 時間で計算できる.

定理 3.3 $p = 2$ かつ $\mathcal{N}(T)$ を T の kSS 近傍とするとき, アルゴリズム 1 の 1 反復あたりの計算時間は $\mathrm{O}(n \min\{2^{k+1}, n\}k)$ である.

4. 数値実験

局所探索アルゴリズム (アルゴリズム 1) の実行時間を数値実験によって計測した. 表 1 は $n = 100, 200, 300, 400, 500; k = 1, 5, 10, 50, 100, n$ のそれぞれに対して 10 個のランダムに生成した相違行列を入力としてアルゴリズムを実行したときの 1 反復あたりの計算時間の平均である. 実験結果は k が n に対して小さいときはほぼ理論的計算時間と一致している. k の増加にしたがって計算時間のオーダはゆるやかに増加するが $k = n$ のときでさえも $n^{2.70}$ 程度であった.

表 1: 1 反復あたりの平均計算時間 [ms].

k \ n	100	200	300	400	500
1	0.05	0.10	0.15	0.23	0.32
5	0.68	1.75	2.32	3.66	5.80
10	2.48	7.08	12.82	22.39	31.53
50	5.90	37.60	121.77	259.13	427.31
100	6.09	38.07	123.14	274.29	539.73
n	6.09	35.16	98.19	222.89	513.44

謝辞

本研究は JSPS 科研費 18K11180 の助成を受けたものである.

参考文献

[1] N. Ailon et al.: Fitting tree metrics: hierarchical clustering and phylogeny. *SIAM Journal on Computing* **40** (2011) 1275–1291.

[2] 安藤和敏, 水越雅紀: 最小増加超距離木問題に対する k 制限部分木交換近傍に基づく局所探索アルゴリズム. 日本オペレーションズ・リサーチ学会 2021 年春季研究発表会アブストラクト集 (2021) 2-B-9.

[3] 石川累, 安藤和敏: 最小増加超距離木問題に対する局所探索アルゴリズム. 数理解析研究所講究録 **2027** (2017) 15-29.

[4] B. Y. Wu et al.: Approximation and exact algorithms for constructing minimum ultrametric trees from distance matrices. *Journal of Combinatorial Optimization* **3** (1999) 199–211.