数理解析研究所講究録
第1148巻 2000年 106-111

系統樹最節約復元の部分木に関する最小性について

– Some properties of the distortion index on all MPRs –

電気通信大学大学院　宮川 幹平 (Kampei Miyakawa)
東海大学高輪短期大学部　成嶋 弘 (Hiroshi Narushima)

近年の進化生物的研究成果を背景に, 既知の形質情報から最節約原理に従って進化の系統を推定するという問題の数理的定式化とその研究が進められている. 今回, その中でも我々は外点に形質情報が付値された単純無向木構造が与えられたときに, その距離が最小となるような内点への値付けを与えるという第1最節約復元 (Most-Parsimonious Reconstruction; MPR) 問題を扱う. まずは問題の定義を再掲する.

本論文での記法については [3, 6, 7, 10, 11] に従う. $T=(V=V_O\cup V_H, E, \sigma)$ を重み付け関数 $\sigma: V_O\to\Omega$ によって各外点に付値された単純無向木とする. 但し, Ω は線形順序を持つ特性値集合を示している. 以後挙げる例においては簡単のため Ω を非負整数 $\mathbb{N}$ とする. また, V は頂点集合, V_O は外点 (次数1の頂点) 集合, V_H は内点集合, そして E は辺集合をそれぞれ示している. このような構造を我々は**el-tree** と呼んでいる. el-tree T が与えられたとき, $\lambda|V_O$ (V_O に定義域を制限した λ) が σ と等しいような T の各頂点への値付け $\lambda: V\to\Omega$ を T の**復元**と呼ぶ[1]. el-tree T に復元 λ が与えられたとき, 各辺 $e\in E$ に対して**距離** $l(e)$ を $l(e)=|\lambda(u)-\lambda(v)|$, $e=\{u,v\}$ と定義する[2]. またそのとき復元 λ が与えられたときの T の距離を各辺の長さの総和と定義する. 即ち, $L(T|\lambda)=\sum_{e\in E} l(e)$. さらに, T の距離の最小値 $L^*(T)$ を以下のように定義する:

$$L^*(T)=\min\{L(T|\lambda)\,|\,\lambda \text{ is a reconstruction on } T\}.$$

この定義が well-defined であることは容易にわかる. ここで, $L(T|\lambda)=L^*(T)$ となるような復元 λ を特に T の**最節約復元**(*MPR*) と呼ぶ. なお, 一般的に el-tree は一つ以上の MPR を持つことが知られている. そこで, T の MPR 全体の集合を $\mathbf{Rmp}(T)$ と書く. また, 各頂点 u に着目し, 各 MPR がその頂点 u で取り得る値の集合 $\{\lambda(u)|\lambda\in\mathbf{Rmp}(T)\}$ を u の**MPR-set** と呼び, S_u と書く.

与えられた el-tree T について, ある頂点 r を根 (root) と定めることで, rooted el-tree $T^{(r)}$ も定義できる. u の子が v であるとき, $u\to v$ または $u=p(v)$ と書く. なお, 根 r が外点であり, r の子を s としたとき, その rooted el-tree $T^{(r)}$ を特に $T=(T_s, r)$ と書く. また曖昧さを無くすため, 根でない外点を葉 (leaf) と言うことにする. rooted el-tree の任意の頂点 u について u とその子孫からなる部分木を T_u と書くことにする. 詳しい定義については [3, 7] を参照されたい.

I_i ($i\in A$) を Ω 上の閉区間族とし, I_i の全ての端点の中間2点 (median two point) を $\langle x,y\rangle$ としよう. このとき, 閉区間 $[x,y]$ を I_i ($i\in A$) の**中間区間**(*median interval*) と定義し, $med\langle I_i: i\in A\rangle$ と書く. rooted el-tree の各頂点 u(但し根が外点の場合, それを除く) について Ω 上の閉区間 $I(u)$ を以下のように再帰的に与える:

$$I(u)=\begin{cases}[\sigma(u),\sigma(u)] & \text{if } u \text{ is a leaf,}\\ med\langle I(v): u\to v\rangle & \text{otherwise.}\end{cases}$$

[1] 形式的には復元を el-tree 上の内点への値付け関数であると考えても良い
[2] 勿論, Ω 上に適切な演算が定義されているものと仮定している

この閉区間 $I(u)$ を u の**特性区間**, また I を T 上の**特性区間写像**と呼ぶ. これらは第1MPR問題に対する一連の論文のキーコンセプトである.

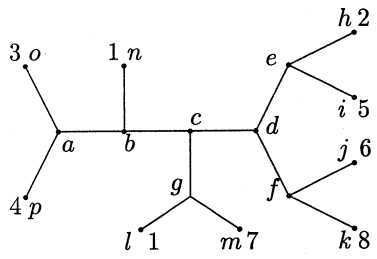


図 1: An undirected el-tree T

λ \ u	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p
λ_1	3	3	3	3	3	6	3	2	5	6	8	1	7	1	3	4
λ_2	3	3	3	4	4	6	3	2	5	6	8	1	7	1	3	4
λ_3	3	3	3	5	5	6	3	2	5	6	8	1	7	1	3	4
λ_4	3	3	4	4	4	6	4	2	5	6	8	1	7	1	3	4
λ_5	3	3	4	5	5	6	4	2	5	6	8	1	7	1	3	4
λ_6	3	3	5	5	5	6	5	2	5	6	8	1	7	1	3	4
λ_7	4	4	4	4	4	6	4	2	5	6	8	1	7	1	3	4
λ_8	4	4	4	5	5	6	4	2	5	6	8	1	7	1	3	4
λ_9	4	4	5	5	5	6	5	2	5	6	8	1	7	1	3	4

表.1 $\mathbf{Rmp}(T)$

与えられた el-tree T に対して, その最短距離 $L^*(T)$, 各頂点 u の MPR-set S_u 等を求める線形時間アルゴリズムが既に得られている ([3, 7]). また, el-tree における MPR の特徴づけも以下の定理 ([3]) によって与えられている:

Theorem A *T を rooted el-tree (T_s, r) とし, λ を T 上の復元とする. λ が T の MPR であるための必要十分条件は各頂点 $u \in V_H$ において, $\lambda(u) \in med\langle[\lambda(p(u)), \lambda(p(u))], I(v) : u \to v\rangle$ (但し I は T 上の特性区間写像) を満たすことである.*

この定理を第1MPR問題の基本定理と呼んでいる. またこれを用いることで el-tree T の全ての MPR を列挙することも可能であるが, $\Omega = \mathbb{N}$ の場合であれば有限ではあるものの一般に MPR は指数個以上存在する. ここで, 進化生物学的観点から導入されたふたつの復元について述べる. これらは外点で根付けされた rooted el-tree について定義されており, ひとつは ACCTRAN 復元と呼ばれ, 根に近いほど形質値の変化を加速する性質を待つ. またもうひとつは DELTRAN 復元と呼ばれ, これは逆に形質値の変化を遅らせるという性質を持つ. これらの進化生物学的な特徴について詳しくは [4, 12] を参照されたい. なお, ACCTRAN 復元を λ_{ACT}, DELTRAN 復元を λ_{DET} とそれぞれ書く. これらは以下のように木の親子関係に関して再帰的に定式化できる ([10, 11]). 与えられた rooted el-tree $T = (T_s, r)$ に対して, その任意の頂点 u において,

$$\begin{aligned} \lambda_{\mathrm{ACT}}(u) &= median\langle \lambda_{\mathrm{ACT}}(p(u)), \min(I(u)), \max(I(u)) \rangle \\ \lambda_{\mathrm{DET}}(u) &= median\langle \lambda_{\mathrm{DET}}(p(u)), \min(S_u), \max(S_u) \rangle. \end{aligned}$$

と定める. 但し $median\langle a, b, c\rangle$ は値 a, b, c の中で中間の値を返す関数とする.

この2つの復元が MPR であるということは既に示されており, また特に ACCTRAN 復元について本論文の動機付けともなった重要な結果が示されている ([10, 11]).

Theorem B *rooted el-tree $T = (T_s, r)$ 上の ACCTRAN 復元は完全最節約性を持つ, 即ち T の全ての部分木の距離を最小化する唯一の MPR である.*

次に, 復元間の関係を調べる為に T の復元全体の集合に対して幾つかの半順序関係が導入された ([5]). それらのうち, 復元間の通常順序 $\leq$ とは全ての頂点 u において $\lambda(u) \leq \mu(u)$ のとき, $\lambda \leq \mu$ と定義される. この順序が半順序であることは容易にわかる. これを el-tree T の MPR 全体集合 $\mathbf{Rmp}(T)$ に対して導入して得られる半順序集合を通常順序の MPR-poset と呼び, $(\mathbf{Rmp}(T), \leq)$ と書く. この半順序集合の最大元/最小元について以下の結果が得られている ([10]).

Proposition C *T を el-tree とする. $\lambda_{\max}$ ($\lambda_{\min}$) を各頂点 u において $\lambda(u) = \max S_u$ ($\min S_u$) なる T 上の復元とする. このとき, $\lambda_{\max}$ ($\lambda_{\min}$) は $(\boldsymbol{Rmp}(T), \leq)$ の最大元 (最小元) である.*

これら従来の結果を踏まえ, ACCTRAN 復元と DELTRAN 復元の関係を示す新たな命題を以下に与える.

Proposition 1 *el-tree T において, もし (T_s, r) 上で ACCTRAN 復元と DELTRAN 復元が等しくなるような頂点 $r \in V_O$ が存在するならば, $|\mathbf{Rmp}(T)| = 1$ である.*

次に一般に指数個以上存在する MPR について, それを評価する基準として [5] において導入された**distortion index** を以下のように定式化する.

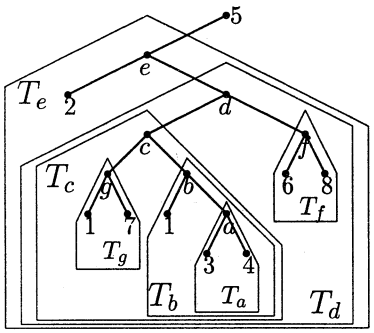
rooted el-tree $T = (T_s, r)$ において, T 上の MPR λ についてその distortion index $I_D(\lambda)$ を

$$I_D(\lambda) = \sum_{u \in V_H} (L(T_u|\lambda_{<u>}) - L^*(T_u)),$$

と定める. 但し $\lambda_{<u>}$ は T の部分木 T_u に制限した λ とする.

任意の el-tree T について, その最小全長 $L^*(T)$ を求めるのには, その頂点数に関して線形時間で十分であるという既知の結果 ([3]) から, 任意の MPRλ の distortion index $I_D(\lambda)$ を求めるのにも頂点数に関して線形時間で十分であるとわかる.

ここで, distortion index に関する例を挙げる. 図.1 に示された el-tree を i で rooting したとき, 図.2 左にあるように, T_a, T_b, T_c, T_d, T_e, T_f, そして T_g がその全ての部分木となる. このとき表.1 に列挙された全ての MPR λ について, それぞれの distortion index の値 $I_D(\lambda)$ は図.2 右の通り求めることができる.



$$I_D(\lambda_3) = \sum_{u \in V_H} (L(T_u|\lambda_{3<u>}) - L^*(T_u))$$
$$= 1+3+9+14+17+2+6$$
$$-(1+3+9+14+15+2+6) = 2$$

$$I_D(\lambda_1) = 0,\ I_D(\lambda_2) = 1,\ I_D(\lambda_4) = 2,$$
$$I_D(\lambda_5) = 3,\ I_D(\lambda_6) = 4,\ I_D(\lambda_7) = 3,$$
$$I_D(\lambda_8) = 4,\ I_D(\lambda_9) = 5$$

図 2: distortion index の計算 in (T_e, i)

ここで定理 B より以下のことは自明である.

Corollary 1 *rooted el-tree* (T_s, r) において, *ACCTRAN* 復元は

$$I_D(\lambda_{\mathrm{ACT}}) = \min_{\lambda \in \mathbf{Rmp}(T)} I_D(\lambda) = 0$$

なる唯一の *MPR* である.

それでは逆に I_D の最大値についてはどうであろうか. 任意の MPR の distortion index を得るには線形時間で十分でも, 一般に MPR は指数個以上存在するため単純にそれを列挙して求めることは効率が悪い. ここで distortion index の定義式を以下の様に変形できることを示す.

Lemma 1 *rooted el-tree* (T_s, r) において, 任意の *MPR* λ の *distortion index* は

$$I_D(\lambda) = \sum_{u \in V \setminus V_C} |\lambda(u) - \lambda_{\mathrm{ACT}}(u)|,$$

である. 但し $V_C = \{u \in V | S_{p(u)} \subseteq I(u)\}$.

この補題から

$$I_D(\lambda) \leq \sum_{u \in V_H \setminus V_C} (\max S_u - \min S_u)$$

が成り立つことは容易にわかる. 次に, 各頂点 $u \in V \setminus \{r\}$ に対して関数 $f_u : S_u \to \Omega$ を

- $u \in V_O \setminus \{r\}$ のとき, S_u の任意の値 x(即ち $\sigma(u)$) に対して, $f_u(x) = 0$.
- $u \in V_C$ のとき, S_u の任意の値 x に対して

$$f_u(x) = \sum_{u \to v} \max_{y \in S_v | x} f_v(y).$$

- $u \in V_H \setminus V_C$ のとき, S_u の任意の値 x に対して

$$f_u(x) = \sum_{u \to v} \max_{y \in S_v | x} f_v(y) + |\lambda_{\mathrm{ACT}}(u) - x|.$$

の様に再帰的に定義する. このとき Lemma 1 より以下のことがわかる. なお, T の部分木 T_u に属する頂点集合を $V(T_u)$ と書く.

Lemma 2 T を *rooted el-tree* (T_s, r), λ を T の任意の *MPR* とする. このとき各頂点 u において,

$$\sum_{v \in V(T_u) \setminus V_C} |\lambda_{\mathrm{ACT}}(v) - \lambda(v)| \leq f_u(\lambda(u))$$

が成り立つ.

これらの補題をさらに進めると以下のような定理が得られる.

Theorem 1 T を *rooted el-tree* (T_s, r), λ を T の任意の *MPR* とする. このとき, $I_D(\lambda) = \max_{\mu \in \mathbf{Rmp}(T)} I_D(\mu)$ であるための必要十分条件は, 任意の頂点 $u \in V \setminus \{r\}$ において, $\lambda(u) \in \{x | f_u(x) = \max_{y \in S_u | \lambda(p(u))} f_u(y)\}$ を満たすことである.

この定理を用いて実際に distortion index が最大となる MPR やその最大値を求める為の計算量は f_u を求める為の計算量に大きく依存する. しかし f_u を定義どおりに求めるのは非常に効率が悪い. ここで, 実際に必要となるのは各頂点 u において, 任意の値 $x \in S_{p(u)}$ に対して $\max_{y \in S_u | x} f_u(y) = f_u(z)$ なる値 $z \in S_u | x$ である. このことを踏まえ, distortion index が最大となる MPR やその最大値に関する以下の結果が得られた.

Theorem 2 *rooted el-tree* (T_s, r) において, 以下のことが成立する.

- T の任意の *MPR* λ について, $I_D(\lambda) \leq \max_{y \in S_s \cap \sigma(V)} f_s(y)$.
- 任意の頂点 $u \in V \setminus \{r\}$ において,

$$\lambda(u) \in \{x | f_u(x) = \max_{y \in S_u | \lambda(p(u)) \cap \sigma(V)} f_u(y)\}$$

が成り立つような T の *MPR* λ が存在し, $I_D(\lambda) = \max_{y \in S_s \cap \sigma(V)} f_s(y)$.

Corollary 2 *rooted el-tree* (T_s, r) において, その *distortion index* の最大値及び最大値を取るようなある *MPR* を実際に構築するのには頂点数 n について $O(n^2)$ で十分である.

また, Theorem 2, Corollary 2 で示した結果を用いて, distortion index に関する興味深い結果が得られた.

Theorem 3 *rooted el-tree* (T_s, r) において, 任意の値 $\omega \in \Omega$ $(0 \leq \omega \leq \max_{\mu \in \mathbf{Rmp}(T)} I_D(\mu))$ について, $I_D(\lambda) = \omega$ となるような T の *MPR* λ が存在し, T の頂点数 n について $O(n^2)$ で実際に構築できる.

最後に, 前述した ACCTRAN 復元と DELTRAN 復元について, distortion index の最大値を取る MPR との関わりを示す結果を与える.

Theorem 4 *rooted el-tree* (T_s, r) において, *ACCTRAN* 復元が $(\boldsymbol{Rmp}(T), \leq)$ の最大元または最小元であるとき, *DELTRAN* 復元は *distortion index* の最大値を取る唯一の *MPR* である.

一般に, この定理の逆は成り立たない.

参考文献

[1] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, Time bounds for selection, JCSS 7 (1973) 448-461.

[2] J. M. Farris, Methods for computing Wagner trees, Systematic Zoology 19 (1970) 83-92.

[3] M. Hanazawa, H. Narushima and N. Minaka, Generating most parsimonious reconstructions on a tree: a generalization of the Farris-Swofford-Maddison method, Discrete Applied Mathematics 56 (1995) 245-265.

[4] N. Minaka, Parsimony, phylogeny and discrete mathematics: combinatorial problems in phylogenetic systematics (in Japanese: with English summary), Natural History Research, Chiba Prefectural Museum and Institute, Vol.2 No.2 (1993) 83 - 98.

[5] N. Minaka, Algebraic properties of the most parsimonious reconstructions of the hypothetical ancestors on a given tree, Forma 8 (1993) 277-296.

[6] K. Miyakawa and H. Narushima, Lattice-theoretic properties of MPR-posets in phylogeny, preprint.

[7] H. Narushima and M. Hanazawa, A more efficient algorithm for MPR problems in phylogeny, Discrete Applied Mathematics 80 (1997) 231-238.

[8] H. Narushima and N. Misheva, On a role of the MPR-poset of most-parsimonious reconstructions in phylogenetic analysis – A combinatorial optimization problem in phylogeny –, in: W. Y. C. Chen, D. Z. Du, D. F. Hsu, H. Y. Hap (Eds.), Proc. The International Symposium on Combinatoircs and Applications 28–30, June, 1996, Tianjin, P.R.China, pp. 306–313.

[9] H. Narushima, On globally optimal reconstructions of phylogenetic trees (in Japanese: with a part in English), RIMS Koukyuuroku 992 "Computation Theory and Its applications" (Kyoto Univ., May, 1997), pp. 5–11.

[10] H. Narushima and N. Misheva, On characteristics of ancestral character-state reconstructions under the accelerated transformation optimization, preprint.

[11] H. Narushima, On extremal properties of ACCTRAN reconstructions in phylogeny, preprint.

[12] D. L. Swofford and W. P. Maddison, Reconstructing ancestral character states under Wagner parsimony, Mathematical Biosciences 87 (1987) 199-229.

[13] D. L. Swofford, W. P. Maddison, Parsimony, character-state reconstructions, and evolutionary inferences, in: R. L. Mayden (Ed.), Systematics, Historical Ecology, and North American Freshwater Fishes, Stanford Univ. Press, California, 1992, pp. 186–223.