数理解析研究所講究録 1205 巻 2001 年 125-130

系統樹最節約復元の部分木に関する最小性について II

– Some properties of the distortion index on all MPRs II –

宮川 幹平 (Kampei Miyakawa)[a], 成嶋 弘 (Hiroshi Narushima)[b]
[a] 電気通信大学大学院情報工学専攻
(*Course in Computer Science and Information Mathematics, Graduate School of Electro-Communications, The University of Electro-Communications*)
[b] 東海大学短期大学部情報・ネットワーク学科
(*Department of Information and Network, Tokai University Junior College*)

近年の進化生物学的研究成果を背景に，既知の形質情報から最節約原理にもとづき，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．本論文で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 最節約復元問題 (The First Most-Parsimonious Reconstruction Problem) を扱う．

本論文での記法については [2, 5, 7, 8, 9] に従う．$T = (V = V_O \cup V_H, E, \sigma)$ を付値関数 $\sigma : V_O \to \Omega$ 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す．以後に示す例においては 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，E は辺集合をそれぞれ示す．このような構造を我々は **el-tree** と呼ぶ．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．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** $l(e|\lambda)$ を $l(e|\lambda) = |\lambda(u) - \lambda(v)|$, $e = \{u, v\}$ と定義する．また復元 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:

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

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 T 上の**最節約復元**(MPR) と呼び，T 上の MPR 全体の集合を $\mathbf{Rmp}(T)$ と書く．また，各 MPR が頂点 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．

例として，図 1 にあるような el-tree が与えられたとき，その MPR 全体集合は表 1 のようになる．

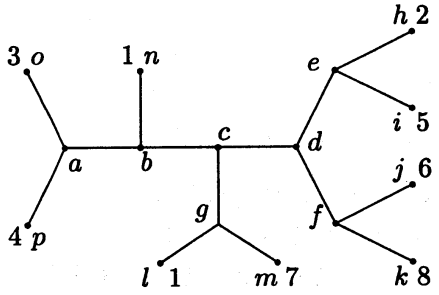

図 1: An undirected el-tree T

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について，ある頂点rを根(root)と定めることで，根付きel-tree $T^{(r)}$も自然に定義できる．uの子がvであるとき，$u \to v$または$u = p(v)$と書く．もし$u_i \to u_{i+1}$ $(i \in [n-1])$なる頂点列$u = u_1, u_2, \cdots, u_n = v$が存在するならば，$u$は$v$の祖先であると呼び，$u \overset{*}{\to} v$と書く．根付きel-tree T上の頂点uに対して，部分木T_uとは頂点部分集合$\{v | u \overset{*}{\to} v\}$によって誘導される$T$の部分グラフとする．なお，根$r$が外点であり，$r$の子を$s$としたとき，その根付きel-tree $T^{(r)}$を特に$T = (T_s, r)$と書く．このT_sを根付きel-tree Tのbodyと呼ぶ．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．詳しくは[2, 7]を参照されたい．

I_i $(i \in A)$をΩ上の閉区間族としたとき，I_iのすべての端点の中間2点（median two point）を$\mathrm{med2}\langle I_i : i \in A\rangle$と書くことにする．このとき，中間2点を端点とする閉区間を閉区間族I_i $(i \in A)$の**中間区間**(*median interval*)と定義し，$\mathrm{med}\langle I_i : i \in A\rangle$と書く．また，同様に閉区間族$I_i$ $(i \in A)$のすべての端点の中間4点（median four point）を$\mathrm{med4}\langle I_i : i \in A\rangle$と書くことにする．根付きel-treeのbodyに属する各頂点u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:

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

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．また，el-treeにおけるMPRの特徴づけも次の定理によって与えられている[2].

Theorem A *Tを根付きel-tree (T_s, r)とし，λをT上の復元とする．λがT上の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，$\lambda(u) \in med\langle [\lambda(p(u)), \lambda(p(u))], I(v) : u \to v\rangle$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のすべてのMPRを列挙することも可能であるが，$\Omega = \mathbf{N}$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．なお，以下では任意の$x \in S_{p(u)}$に対して$\lambda(u) \in \mathrm{med}\langle [x, x], I(v) : u \to v\rangle$を$S_u|x$と書く．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根付き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（即ち進化系統上の共通祖先）に近いほど形質値の変化を加速させる性質を待つ．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意味について詳しくは[1, 3, 10]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[8, 9]．与えられた根付きel-tree $T = (T_s, r)$に対して，その任意の頂点uにおいて，

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

と定める．但し$\mathrm{median}\langle a, b, c\rangle$は値$a, b, c$の中で2番目に大きい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いる[6, 8]．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も示されており[8]，以下に引用する．

Theorem B *根付きel-tree $T = (T_s, r)$上の$ACCTRAN$復元は完全最節約性を持つ，即ちTのすべての部分木T_u $(u \in V)$の距離を最小化する唯一のMPRである.*

λ \ u	a	b	c	d	e	f	g	$D_I(\lambda)$
λ_1	0	0	0	0	2	0	0	2
λ_2	0	0	0	0	1	0	0	1
λ_3	0	0	0	0	0	0	0	0
λ_4	0	0	1	0	1	0	0	2
λ_5	0	0	1	0	0	0	0	1
λ_6	0	0	0	0	2	0	0	2
λ_7	0	1	1	0	1	0	0	3
λ_8	0	1	1	0	0	0	0	2
λ_9	0	1	2	0	0	0	0	3

表 2: distortion sequence 一覧

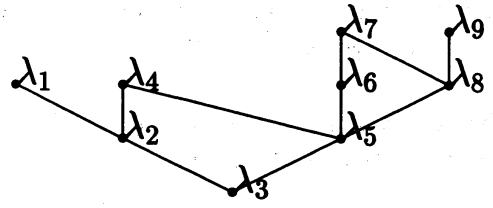

図 2: distortion MPR-poset の例

ここで，新たにいくつかの記法を導入する．T を根付き el-tree (T_s, r) としたとき，$V \setminus \{r\}$ の部分集合 V_C，V_G，そして V_L をそれぞれ $S_{p(v)} \subseteq I(v)$ なる頂点 v の集合，$\min S_{p(v)} < \min I(v)$ なる頂点 v の集合，そして $\max I(v) < \max S_{p(v)}$ なる頂点 v の集合と定義する．このときこれらが互いに素であることは直ちに導かれる．また，任意の頂点 u の子 v の集合を $V(u)$ と書き，$V(u) \cap V_J = V_J(u)$, $(J = \{L, C, G\})$ とする．任意の 2 頂点 $u \overset{*}{\to} v$ について，$u \overset{*}{\to} w \overset{*}{\to} v$ なる $\forall w \neq u, v$ が $w \in V_C$ であるようなとき，$u \overset{C}{\to} v$ と書くことにする．

三中 [4] は，与えられた el-tree T を外点で根付け $(T = (T_s, r))$ したときに，各 MPR λ が T の各部分木の全長をどれだけ最小化しているかに着目した評価基準として **distortion index** $D_I(\lambda)$ を導入した．我々はその概念をより精密に表現するために，各 MPR λ に対して長さ $|V|$ の列 $D_S(\lambda)$ を割り当てる．この列の各要素は T の各頂点 u に対応し，部分木 T_u の最小長と λ を T_u に割り当てたときの全長の差を格納する．即ち，$D_S : \mathbf{Rmp}(T) \to \Omega^V$ を

$$D_S(\lambda)(u) = L(T_u|\lambda_{\langle u \rangle}) - L^*(T_u),$$

と定義する．但し，$\lambda_{\langle u \rangle}$ は T の部分木 T_u に定義域を制限した λ とする．この列を MPR λ の **distortion sequence** と呼ぶ．また，この列の要素の総和が前述の distortion index に対応する．即ち $\sum_{u \in V} D_S(\lambda)(u) = D_I(\lambda)$．次に，任意の MPR λ, μ に対して $\lambda \leq_D \mu$ であることを $\forall u \in V, D_S(\lambda)(u) \leq D_S(\mu)(u)$ と定義すると MPR 全体集合上の半順序となる．そこで，半順序集合 $(\mathbf{Rmp}(T), \leq_D)$ を根付き el-tree T の **distortion MPR-poset** と呼ぶことにする．

図 1 にある el-tree を外点 h で根付けしたときの根付き el-tree $T = (T_e, h)$ において，各 MPR に対する distortion sequence は定義より表 2 のように得られる．このとき，T の distortion MPR-poset のハッセ図は図 2 のように得られる．なお，$\lambda_3 = \lambda_{\mathrm{ACT}}$ である．

定理 B より以下の系が直ちに得られる．

Corollary 1 根付き *el-tree* (T_s, r) において，*ACCTRAN* 復元 λ_{ACT} は ***distortion MPR-poset*** における最小元であり，かつ $D_S(\lambda_{\mathrm{ACT}}) = \mathbf{0}$ である．即ち $D_I(\lambda_{\mathrm{ACT}}) = 0$.

一方，例にあるように distortion MPR-poset が常に最大元を持つとは限らない．distortion MPR-poset の束論的特徴づけとして，以下を得た．

Theorem 1 T を根付き *el-tree* (T_s, r) とする．このとき T の ***distortion MPR-poset*** は下に完備な半束である．

Corollary 2 T を根付き *el-tree* (T_s, r) とする．このとき，***distortion MPR-poset*** 上の任意の閉区間 $[\lambda, \mu]$ は完備分配束である．

Theorem 2 T を根付き *el-tree* (T_s, r) とする．T の復元 λ が ***distortion MPR-poset*** の極大元であるための必要十分条件は任意の内点 u について，

$$\lambda(u) \in \begin{cases} \left[\max\{\min S_u, \lambda(p(u))\}, \max\{\min S_u, \lambda(p(u)), \max\limits_{u \xrightarrow{C} v \in V_L} (\max S_v)\}\right] & (u \in V_L), \\ \left[\min\{\max S_u, \lambda(p(u)), \min\limits_{u \xrightarrow{C} v \in V_G} (\min S_v)\}, \min\{\min S_u, \lambda(p(u))\}\right] & (u \in V_G), \end{cases}$$

$$\lambda(u) = \lambda(p(u)) \quad (u \in V_C),$$

を満たすことである．

定理 2 と DELTRAN 復元の定義から以下の系が直ちに得られる．

Corollary 3 根付き *el-tree* $T = (T_s, r)$ において，*DELTRAN* 復元は T の ***distortion MPR-poset*** の極大元である．

これらの結果から，$\Omega = \mathbf{N}$ である場合，distortion MPR-poset において，任意の MPR λ に対して鎖 $C(\lambda_{\mathrm{ACT}}, \lambda)$ の長さ（即ち分配束 $([\lambda_{\mathrm{ACT}}, \lambda], \leq_D)$ の高さ）が λ の distortion index と等しいことはその定義から直ちに導かれる．しかし distortion MPR-poset の極大元は一般に指数個以上存在し，かつ極大元と最小元から誘導される鎖すべてが最大長を持つとは限らないため，この distortion index を最大にする MPR 及び distortion index の最大値を決定する問題は自明でない．よって以下では [11] で述べられた distortion index が最大であるような MPR の決定について，その具体的なアルゴリズムを示す．

T を根付き el-tree (T_s, r) としたとき，T の body 上の頂点 u に対して，Ω 上の閉区間 $I_D(u)$[1]，多重集合 $M_L(u), M_G(u)$，及びフラグ変数 $stop(u)$ を以下のようにボトムアップ (leaf $\to$ root) に定める．このとき，λ を T 上の MPR とすると任意の $u \notin V_C$ に対して $\lambda(u) \in I_D(u)$ であることが $\sum_{v \in V(T_u)} D_S(\lambda)(v)$ が最大となるための必要条件となる（証明は省略）．また $M_L(u), M_G(u)$ は $I_D(u)$ を決定するために用いる．

- u が leaf の場合，$I_D(u) = [\sigma(u), \sigma(u)]$, $M_L(u) = M_G(u) = \emptyset$, $stop(u) = \mathrm{On}$.
- $u \in V_H \cap V_C$ の場合，$M_L(u) = M_G(u) = \emptyset$, $stop(u) = \mathrm{Off}$.
- $u \in V_H \setminus V_C$ の場合，

  ⟨1⟩ $M_L(u), M_G(u)$ を下記のように定める:

$$M_L(u) = \begin{cases} \{\max S_u\} & u \in V_L, \\ \emptyset & u \in V_G. \end{cases}, M_G(u) = \begin{cases} \emptyset & u \in V_L, \\ \{\min S_u\} & u \in V_G. \end{cases}.$$

[1] $u \in V_H \cap V_C$ なる頂点 u のについては未定義とする．

⟨2⟩ $u \xrightarrow{C} v \notin V_C$ なる頂点の中で，$M_L(v) \neq \emptyset$ かつ $M_G(w) \neq \emptyset$ であるような頂点 v, w が存在するか調べる．

◇ 存在しない場合，$stop(u) = \text{off}$ かつ

$$I_D(u) = \begin{cases} [\max S_u, \max S_u] & (u \in V_L), \\ [\min S_u, \min S_u] & (u \in V_G). \end{cases}$$

と定める．

◇ 存在した場合，$stop(u) = \text{on}$ とし，多重集合 $\hat{M}_L(u), \hat{M}_G(u)$ を以下のように定める:

$$\begin{aligned} \hat{M}_L(u) &= \{M_L(t) | u \xrightarrow{s} v\}, \\ \hat{M}_G(u) &= \{M_G(t) | u \xrightarrow{s} v\}, \end{aligned}$$

但し，$u \xrightarrow{s} v$ を $u \xrightarrow{*} v$ でありかつ $u \xrightarrow{*} w \xrightarrow{*} v$ なる頂点 $w \neq u, v$ について $stop(w) = \text{off}$ を満たすことと定義する．

ここで，$I_D(u) = [Z_{|\hat{M}_L(u)|}, Z_{|\hat{M}_L(u)|+1}] \cap S_u$ と定める．但し，多重集合 $\hat{M}_L(u) \cup \hat{M}_G(u)$ の中で，i 番目に小さい要素を Z_i と書くとする．また

$$\begin{cases} \text{if } u \in V_L \text{ then} & M_L(u) = \{Z_1, Z_2, \cdots, Z_{|\hat{M}_L(u)|}\}, \\ \text{if } u \in V_G \text{ then} & M_G(u) = \{Z_{|\hat{M}_L(u)|+1}, Z_{|\hat{M}_L(u)|+2}, \cdots, Z_{|\hat{M}_L(u)|+|\hat{M}_L(u)|}\}, \\ \text{if } u \in V_C \text{ then} & M_L(u) = \hat{M}_L(u),\ M_G(u) = \hat{M}_G(u). \end{cases}$$

とする．

次はトップダウン (root → leaf) に，T 上の復元 λ を以下のように決定する．

$$\begin{cases} \lambda(u) \in [\alpha_u, \beta_u] & u \in V_L \\ \lambda(u) = \lambda(p(u)) & u \in V_C \\ \lambda(u) \in [\gamma_u, \delta_u] & u \in V_G \end{cases}$$

但し，$\text{med4}\langle I_D(u), [\lambda(p(u)), \lambda(p(u))]\rangle = \langle \alpha_u, \beta_u, \gamma_u, \delta_u \rangle$ とする．

なお証明は省略するが，このアルゴリズムは根付き el-tree $T = (T_s, r)$ について，distortion index を最大化する T 上の MPR を全て列挙している．このアルゴリズムの正当性及び計算量に関して詳しくは [12] を参照されたい．

最後に，図 3 の根付き el-tree に対して distortion index が最大となる MPR λ を構築する例を挙げる．なお，特性区間を定義どおり求めると $a, b, c \in V_L$, $e, f \in V_G$ であるとわかる．図 1 の el-tree T を外点 h で根付けした根付け el-tree (T_e, h) に対して，頂点 e の全ての子孫 v における $I_D(v), M_L(v), M_G(v)$ が前述のアルゴリズムに従って図 3 にあるように定義済みであるとする（$M_L(v), M_G(v)$ の記入の無い場合は空集合とする．また図内の閉区間は $I_D(v)$ を示し，記入が無い場合は未定義とする．また $stop = \text{on}$ なる頂点は二重丸で，$V_H \cap V_C$ なる頂点は方形で表している．）．$e \xrightarrow{C} v \notin V_C$ なる頂点は c, e, f, i であるが，$M_L(c), M_G(f)$ はともに空集合でない．ここで $\hat{M}_L(d) = \{4, 5\}, \hat{M}_G(e) = \{3, 6\}$ となるから，$I_D(e) = [4, 5]$ である．$e \in V_G$ であるから，$M_L(e) = \emptyset, M_G(e) = \{5, 5\}$ となる．また λ の決定においては，図 4 にあるように $\text{med4}\langle I_D(e), [\lambda(h), \lambda(h)]\rangle = \langle 2, 2, 4, 5 \rangle$ となるから，$e \in V_G$ より $\lambda(e) \in [4, 5]$. $\lambda(e) = 4$ とすると，$d \in V_C$ であるから $\lambda(d) = \lambda(e) = 4$. 以下，再帰的に λ は $\lambda(c) = \lambda(g) = \lambda(b) = \lambda(a) = 4$, $\lambda(f) = 6$ となって表 1, 2 より確かに $\lambda = \lambda_7$ の distortion index は最大 (3) である．$\lambda(e) = 5$ とした場合も同様に $\lambda = \lambda_9$ を得て，このときも確かにその distortion index は最大である．

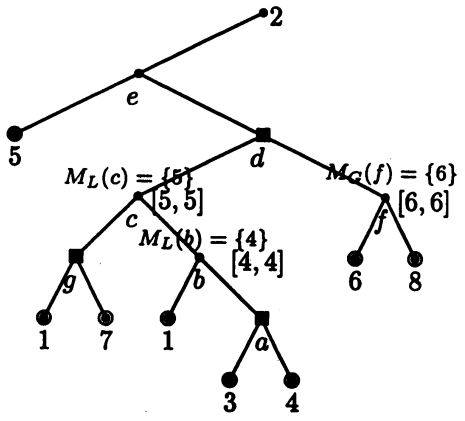

図 3: $I_D(u)$ の構築

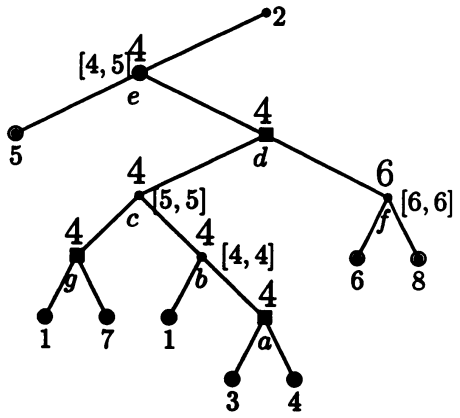

図 4: $D_I(\lambda)$ が最大となる MPR λ の構築例

参考文献

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

[2] 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.

[3] 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.

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

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

[6] K. Miyakawa and H. Narushima, On mathematical properties of ancestral character-state reconstructions under the delayed transformation optimization, 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 characteristics of ancestral character-state reconstructions under the accelerated transformation optimization, preprint.

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

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

[11] 宮川 幹平, 成嶋 弘, 系統樹最節約復元の部分木に関する最小性について, 京都大学数理解析研究所講究録 「計算機科学の基礎理論 : 21 世紀の計算パラダイムを目指して」 No.1148 (2000) 106–111.

[12] 宮川 幹平, 成嶋 弘, 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/sequence on all MPRs, preprint.