

系統樹最節約復元束の完備分配性について

– Lattice-theoretic properties of MPR-posets in phylogeny –

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

On the background of biology such as taxonomy, cladistics and phylogeny, the principle of maximum parsimony also called Wagner Parsimony has been mathematically formulated and then a mathematical and algorithmic theory has been developing[2-12]. The principle assumes that the total amount of evolutionary changes is globally minimized. The problems and related problems on this minimization have been called the Most-Parsimonious Reconstruction (abbreviated to MPR) Problems in phylogeny.

In Narushima [8], the MPR Problems are classified into two kinds of topics. One is called the First MPR Problem “given a phylogenetic tree with the external nodes (which express the operational taxonomic units) of which characters are stated, find an assignment of character-states to all internal nodes (which express the hypothetical taxonomic units) of the tree, so as to minimize the length (the total amount of evolutionary changes) of the tree.” This is also known as “the character-state minimization.” The other is the Second MPR Problem (or the Wagner Tree Problem) “given a set of operational taxonomic units of which characters are stated, find a phylogenetic tree with the set as the external nodes, and simultaneously an assignment of character-states to all internal nodes of the tree, so as to minimize the length of the tree.” Such an optimal phylogenetic tree is called a Wagner tree. It is well-known that the latter problem is strongly related to the Steiner Problem in Phylogeny (SPP) and the Rectilinear Steiner Tree (RST) problem etc. Both problems described above originate in Farris [2]. In this paper, we discuss the former problem under linearly ordered character-states, that is, in the framework based on the method of Hanazawa, Narushima and Minaka[3].

We use the notation in [3, 6, 9, 10]. In this paper, let the set Ω of linearly ordered character-states be the set \mathbf{R} of real numbers unless otherwise stated, because we discuss the completeness of posets, introduced later. Let Ω^n denote the n -dimensional Cartesian product of Ω . Let $T = (V, E, \sigma)$ be any undirected simple tree whose endnodes are evaluated by a weight function $\sigma : V_O \rightarrow \Omega^n$, which is called a *multi-character state function*, where V is the set of nodes, V_O is the set of endnodes, V_H is the set of internal nodes, and E is the set of branches. Note that $V_O \cup V_H = V$ and $V_O \cap V_H = \emptyset$. We call this tree an *el-tree*. For an el-tree T , we define an assignment $\lambda : V \rightarrow \Omega^n$ such that $\lambda|_{V_O}$ (the restriction of λ to V_O) = σ , where $\lambda(u)$ is called a *state* of u under λ . This assignment is called a *reconstruction* on an el-tree T . We denote the restriction of λ -range to the i -th component of Ω^n ($1 \leq i \leq n$) by λ_i . For each branch e in E of an el-tree T with a reconstruction λ , we define the *length* $l(e)$ of branch $e = \{u, v\}$ by $\sum_{i=1}^n |\lambda_i(u) - \lambda_i(v)|$, which is said to be the Manhattan distance or the rectilinear distance. The *length* $L(T|\lambda)$ of an el-tree T under the reconstruction λ is the sum of the lengths of the branches. That is, $L(T|\lambda) = \sum_{e \in E} l(e)$. Then we define the minimum length $L^*(T)$ of T by

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

Note that $L^*(T)$ is well-defined. A reconstruction λ such that $L(T|\lambda) = L^*(T)$ is called a *most-parsimonious reconstruction* (abbreviated to MPR) on T . Generally, an el-tree T has more than one MPR. The following is one of the key concepts in the subject. The set $\{\lambda(u) \mid \lambda \text{ is an MPR on } T\}$ of states is called the *MPR-set* of a node u and written as S_u .

We here note the following important fact. Considering the definitin of $L(T|\lambda)$, that is,

$$L(T|\lambda) = \sum_{e \in E} l(e) = \sum_{i=1}^n \sum_{\{u,v\} \in E} |\lambda_i(u) - \lambda_i(v)|,$$

we see that this minimization allows us to treat each component (character) independently. Indeed, this independence among characters is a crucial assumption of our method. So, hereafter, we treat only the single-character case for an el-tree.

For a given el-tree $T = (V, E, \sigma)$, we define a *rooted el-tree* $T^{(r)}$ rooted at any element r in V . The rooted el-tree $T^{(r)}$ is simply written T if it is understood. The parent-child relation $\{u, v\}$ in E on a rooted el-tree T is denoted by $u \rightarrow v$ or $p(v) = u$, which means u is a *parent* of v (or v is a *child* of u). For each u and v in V , u is called an *ancestor* of v , written $u \xrightarrow{*} v$, if there is a sequence of nodes $u = u_1, u_2, \dots, u_n = v$ in V such that $u_i \rightarrow u_{i+1}$ ($1 \leq i \leq n-1$). In a rooted el-tree, there is only one node without a parent, which is called the *root*, and a node without a child is called a *leaf*. For each u in V , we denote a subtree of a rooted el-tree T induced from a subset $\{u\} \cup \{v \in V \mid u \xrightarrow{*} v\}$ of V by T_u , where u is the root. If r is an endnode, i.e., $r \in V_O$ and s is its unique child, we denote the rooted el-tree $T^{(r)}$ by (T_s, r) . In this case, the subtree T_s is called the *body* of the tree $T^{(r)}$; otherwise, i.e., if $r \in V_H$, the body of $T^{(r)}$ is $T^{(r)}$ itself.

We denote the set $\{1, 2, \dots, n\}$ of n elements by $[n]$. Let a_i ($i \in [2n]$) be any elements in Ω , and be sorted in ascending order as follows:

$$x_1 \leq x_2 \leq \dots \leq x_n \leq x_{n+1} \leq \dots \leq x_{2n}.$$

Then we call x_n and x_{n+1} the *median two points* of the numbers a_i ($i \in [2n]$), and denote $\langle x_n, x_{n+1} \rangle$ by

$$\text{med2}\langle a_1, a_2, \dots, a_{2n} \rangle \text{ or } \text{med2}\langle a_i : i \in [2n] \rangle.$$

Let I_i ($i \in [m]$) be any family of closed intervals in Ω . Then we denote the median two points of all the endpoints of I_i ($i \in [m]$) by

$$\text{med2}\langle I_1, I_2, \dots, I_m \rangle \text{ or } \text{med2}\langle I_i : i \in [m] \rangle.$$

Let $\text{med2}\langle I_i : i \in [m] \rangle = \langle x, y \rangle$. Then we call the closed interval $[x, y]$ in Ω the *median interval* of I_i ($i \in [m]$), which is the key concept in a series of our papers, and denote it by

$$\text{med}\langle I_1, I_2, \dots, I_m \rangle \text{ or } \text{med}\langle I_i : i \in [m] \rangle.$$

For each node u in the body of a rooted el-tree T , we assign a closed interval $I(u)$ of Ω recursively as follows:

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

This $I(u)$ is called the *characteristic interval* of a node u , and so is I the *characteristic interval map* on T .

We now restate some previous results which are particularly related to new results stated later. The following is a qualitative expression of Theorem 1 (Theorem 3(ii)) in [3], which shows the necessary and sufficient condition for a reconstruction on T to be an MPR on T . This theorem may be said to be the fundamental theorem on the first MPR problem.

Theorem A *Let T be a rooted el-tree (T_s, r) and λ be a reconstruction on T . λ is an MPR on T if and only if for any $u \in V_H$, $\lambda(u) \in \text{med}\langle [\lambda(p(u)), \lambda(p(u))], I(v) : u \rightarrow v \rangle$, where I is the characteristic interval map on T .*

By using Theorem A, we can recursively obtain all MPRs on a given el-tree T . For details see [3, 6]. Then we denote the set of MPRs on an el-tree T by $\mathbf{Rmp}(T)$. The following is Corollary 5 in [3], which gives a characterization for each MPR-set.

Corollary B *Let u be any internal node of an el-tree T . Let I be the characteristic interval map on a rooted el-tree $T^{(u)}$. Then $I(u)$ is the MPR-set S_u .*

From Theorem A, we see that $\text{med}([\lambda(p(u)), \lambda(p(u))], I(v) : u \rightarrow v)$ is the MPR-set of node u under the restriction that an element $\lambda(p(u))$ in $S_{p(u)}$ has been assigned to u 's parent $p(u)$. We denote this subset of the MPR-set S_u by $S_u | x$. That is,

$$S_u | x = \text{med}([x, x], I(v) : u \rightarrow v),$$

where x is an element in $S_{p(u)}$.

The following is Theorem 1 in [6], which gives a recursive characterization for each MPR-set.

Theorem C *Let T be a rooted el-tree (T_s, r) . Then each MPR-set S_u for each internal node u of T is recursively decided by*

$$S_u = [\min(S_u | \min(S_{p(u)})), \max(S_u | \max(S_{p(u)}))].$$

We here show some examples for illustrating our previous results. In examples showed in this paper, Ω is restricted to the set \mathbf{N} of integers. An el-tree T hereafter used, is shown in Fig.1. All MPRs on T are recursively generated by the algorithm based on Theorem A and shown in Table 1. Each MPR-set S_u is also obtained recursively by the Hanazawa–Narushima algorithm based on Theorem C and shown in Fig.2. For details on the computational complexity of the algorithms, see [3, 6].

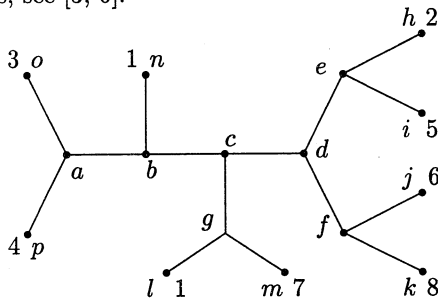


Fig. 1: An undirected el-tree T

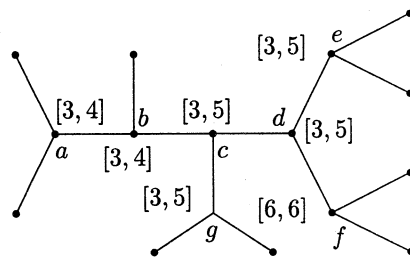


Fig. 2: All MPR-sets on T

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

Table 1: $\mathbf{Rmp}(T)$

Since the minimization of a reconstruction $\lambda : V \rightarrow \Omega$ on an el-tree $T = (V, E, \sigma)$ is our center of interest, it is sufficient for us to consider the range of λ as the closed interval $[\min \sigma, \max \sigma]$

(written as Δ) of Ω . Therefore, we may think of the set $\{\lambda : V \rightarrow \Delta\}$ of reconstructions on T as the general framework of our subject. Let $\mathbf{Rec}(T)$ denote the set $\{\lambda : V \rightarrow \Delta\}$. Then the usual ordering $\lambda \leq \mu$ on $\mathbf{Rec}(T)$ is defined by $\lambda(u) \leq \mu(u)$ for all u in V . We call $(\mathbf{Rec}(T), \leq)$ a *REC-poset*.

On the other hand, from a phylogenetic point of view, Minaka [4, 5] has introduced the two partial orderings on $\mathbf{Rmp}(T)$ to investigate the relationships among MPRs. One is the usual ordering and the other is a partial ordering that depends on a state of a specified root of a given el-tree.

We now give a mathematically explicit formulation for those partial orderings. Let T be an el-tree. The usual ordering $\lambda \leq \mu$ on $\mathbf{Rmp}(T)$ is defined by $\lambda(u) \leq \mu(u)$ for all u in V . Let T be a rooted el-tree (T_s, r) . A binary relation $a \leq_{\sigma(r)} b$ on Ω is defined by $\sigma(r) \leq a \leq b$ or $\sigma(r) \geq a \geq b$. Then a binary relation $\lambda \leq_{\sigma(r)} \mu$ on $\mathbf{Rmp}(T)$ is defined by $\lambda(u) \leq_{\sigma(r)} \mu(u)$ for all u in V . It is easily shown that those relations are partial orderings. $(\mathbf{Rmp}(T), \leq)$ is called a *usual MPR-poset* which is really an induced subposet of $(\mathbf{Rec}(T), \leq)$, and $(\mathbf{Rmp}(T), \leq_{\sigma(r)})$ is called a $\sigma(r)$ -*version MPR-poset*. Note that the usual MPR-poset is uniquely defined for an el-tree, but the $\sigma(r)$ -version MPR-poset, depending on the character-state of a specified root, is defined for each rooted el-tree.

We here illustrate some $\sigma(r)$ -version MPR-posets in Fig.3 for the el-tree T in Fig.1. From $\sigma(l) = \sigma(n) = 1 \leq \lambda(u)$ ($\lambda \in \mathbf{Rmp}(T)$, $u \in V$) in Table 1 and the definition of $\sigma(r)$ -version MPR-poset, we see that each $\sigma(r)$ -version MPR-poset of (c) and (d) in Fig.3 is order-isomorphic to the usual MPR-poset.

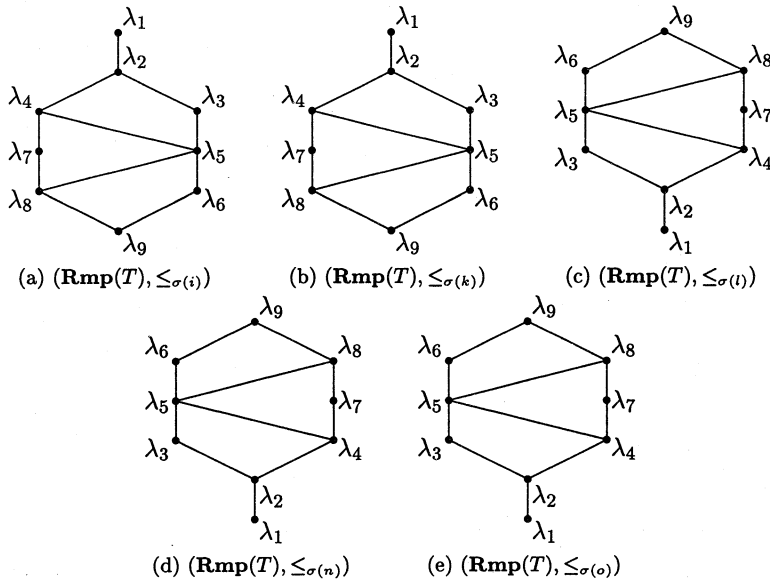


Fig. 3: Examples of $\sigma(r)$ -version MPR-posets

It is easily shown that the REC-poset $(\mathbf{Rec}(T), \leq)$ is a complete distributive lattice. At the start of investigating the completeness of MPR-posets and the distributivity, that is, whether “an MPR-poset is a complete sublattice of the lattice $(\mathbf{Rec}(T), \leq)$ or not”, in the main section, we first restate the the following which is Proposition 5 in [9].

Proposition D *Let T be an el-tree. Let λ_{\max} (λ_{\min}) denote a reconstruction λ on T such that $\lambda(u) = \max S_u$ ($\min S_u$) for any internal node u . Then the reconstruction λ_{\max} (λ_{\min}) on T is*

the greatest (least) element of the MPR-poset $(\mathbf{Rmp}(T), \leq)$.

It is well-known theorem in lattice theory that a lower-complete poset with the greatest element is a complete lattice. Hence, from Proposition D we see that the lower-completeness (upper-completeness) of $(\mathbf{Rmp}(T), \leq)$ is sufficient to show the following, the first main theorem in this paper.

Theorem 1. *Let T be an el-tree. Then the usual MPR-poset $(\mathbf{Rmp}(T), \leq)$ is a complete distributive lattice.*

We next investigate lattice-theoretic properties of $\sigma(r)$ -version MPR-posets. First of all, recall that the usual MPR-poset is uniquely defined for an el-tree, but the $\sigma(r)$ -version MPR-poset, depending on the character-state of a specified root, is defined for each rooted el-tree. The same framework as usual MPR-posets applies to $\sigma(r)$ -version MPR-posets. The $\sigma(r)$ -version ordering $\lambda \leq_{\sigma(r)} \mu$ on $\mathbf{Rec}(T)$ is defined by $\lambda(u) \leq_{\sigma(r)} \mu(u)$ for all u in V . We call $(\mathbf{Rec}(T), \leq_{\sigma(r)})$ a $\sigma(r)$ -version REC-poset. Then we easily see that there exists the infimum of any nonempty subset of $\mathbf{Rec}(T)$ on a $\sigma(r)$ -version REC-poset, that is, a $\sigma(r)$ -version REC-poset is a lower-complete semilattice.

The following is the second main theorem in this paper, which is proved by using Theorem A.

Theorem 2. *Let T be a rooted el-tree (T_s, r) . Then the $\sigma(r)$ -version MPR-poset is a lower-complete semilattice.*

We see from Theorem 2 that there exists the least element $\inf_{\sigma(r)}(\mathbf{Rmp}(T))$ in any $\sigma(r)$ -version MPR-poset. Let's here show a more concrete characterization for the least element.

Proposition 1. *Let T be a rooted el-tree (T_s, r) . Let's define a reconstruction λ on T as follows; for each u in V_H ,*

$$\lambda(u) = \begin{cases} \min(S_u) & (\sigma(r) \leq \min S_u) \\ \sigma(r) & (\min S_u < \sigma(r) < \max S_u) \\ \max(S_u) & (\sigma(r) \geq \max S_u). \end{cases}$$

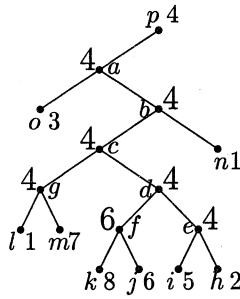
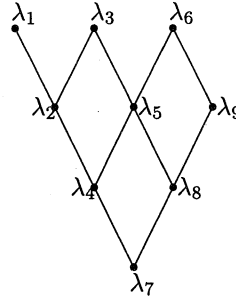
Then the reconstruction λ is the least element of the $\sigma(r)$ -version MPR-poset.

The reconstruction λ defined in Proposition 1 is particularly written as $\lambda_{\min}^{<\sigma(r)>}$. We here show some examples for the least element $\lambda_{\min}^{<\sigma(r)>}$. Let the el-tree T in Fig.1 be rooted at p . Then from the MPR-sets in Fig.Fig:MPRsets and Proposition 1, we obtain the least element $\lambda_{\min}^{<\sigma(p)>}$ is shown in Fig.4, with the rooted el-tree $T = (T_a, p)$. We also see that $\lambda_{\min}^{<\sigma(p)>}$ is really equal to λ_7 in $\mathbf{Rmp}(T)$ shown in Table 1, i.e, the least element of the $\sigma(r)$ -version MPR-poset $(\mathbf{Rmp}(T), \leq_{\sigma(p)})$ shown in Fig.5.

We see from Fig.5 that there is not necessarily the greatest element in a $\sigma(r)$ -version MPR-poset, that is, a $\sigma(r)$ -version MPR-poset is not necessarily a complete distributive lattice, and so we here examine the lattice-theoretic properties on intervals of any $\sigma(r)$ -version MPR-poset.

For any λ, μ which are comparable elements in a REC-poset, it is easily shown that the interval subposet $([\lambda, \mu], \leq_{\sigma(r)})$ of the REC-poset has the greatest element μ and the infimum of any nonempty subset of $[\lambda, \mu]$. Furthermore, when A is of any two elements in $[\lambda, \mu]$, $\inf_{\sigma(r)} A$ is to be the *meet*(\wedge) of the two elements, which is really the minimum of the two elements on $\leq_{\sigma(r)}$, and $\sup_{\sigma(r)} A$ is the *join*(\vee) of the two elements, which is really the maximum of the two elements on $\leq_{\sigma(r)}$. Then we see easily that the distributive laws on the lattice-theoretic operations hold in $([\lambda, \mu], \leq_{\sigma(r)})$. Thus we have that any interval poset $([\lambda, \mu], \leq_{\sigma(r)})$ in $(\mathbf{Rec}(T), \leq_{\sigma(r)})$ is a complete distributive lattice.

By the similar way stated above, we obtain the following which is the third main theorem in this paper.

Fig. 4: $\lambda_{\min}^{<\sigma(r)>}$ on (T_a, p) Fig. 5: $(\mathbf{Rmp}(T), \leq_{\sigma(p)})$

Theorem 3. *Let T be a rooted el-tree (T_s, r) . Then any interval poset $([\lambda, \mu], \leq_{\sigma(r)})$ in $(\mathbf{Rmp}(T), \leq_{\sigma(r)})$ is a complete distributive lattice.*

We finally give some remarks. It is easily shown that the results in this paper are naturally generalized to the multi-character case for an el-tree. One also see easily that an MPR-lattice is not always a complemented lattice. In a later paper, we investigate in detail the order-theoretic structures of a $\sigma(r)$ -version MPR-poset. We particularly give some characterizations of maximal elements in that poset and then a necessary and sufficient condition for that poset to have the greatest element, i.e., to be a complete distributive lattice.

References

- [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] H. Narushima and M. Hanazawa, A more efficient algorithm for MPR problems in phylogeny, *Discrete Applied Mathematics* 80 (1997) 231-238.
- [7] 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 Combinatorics and Applications* 28-30, June, 1996, Tianjin, P.R.China, pp. 306-313.
- [8] 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.
- [9] H. Narushima and N. Misheva, On characteristics of ancestral character-state reconstructions under the accelerated transformation optimization, preprint.
- [10] H. Narushima, On extremal properties of ACCTRAN reconstructions in phylogeny, preprint.
- [11] D. L. Swofford and W. P. Maddison, Reconstructing ancestral character states under Wagner parsimony, *Mathematical Biosciences* 87 (1987) 199-229.
- [12] 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.