Minimum Number of Leaf-Covering Subtrees Covering Phylogenetic Networks
（系統ネットワークを被覆する最小個数の部分系統樹）

数理情報学専攻 48226236　吉田 勇輝
指導教員　岩田 覚 教授

1　はじめに

(根付き) 系統ネットワークとは，系統樹に reticulation という構造が追加されたものであり，異種交雑や遺伝子の水平伝播の反映等の応用が広い一方で，生物学的分類において本質的な木構造を喪失している．そこで，葉集合を保存する全域木を持つ tree-based な系統ネットワーク [4] が重要なクラスとして近年様々に研究されている．研究の一つが，tree-based からの離れ具合 (逸脱度) について定義及び考察するものである．

本論文では，逸脱度を表す指標の一つである「系統ネットワークを被覆する部分系統樹の最小個数」について，計算手法の構築及び min-max 定理の証明を行う．

2　準備

非空かつ有限な集合 X に対し，DAG $N=(V,E)$ が**系統 X-ネットワーク**であるとは，N が唯一の根 ρ と葉集合 X を持ち，各葉の入次数は 1 であり，根でも葉でもない頂点は次のいずれかを満たすものと定義する．
(a) 入次数が 1, 出次数が 2 以上 (tree-vertex)
(b) 入次数が 2 以上, 出次数が 1 (reticulation)
以降，N は系統 X-ネットワークを指すとし，$n=|V|, m=|E|$ とする．根が ρ，葉集合が X であるような N の有向部分木 T を**部分系統樹**といい，T として全域木が取れるような N を **tree-based** という．N の部分系統樹全体の集合を $\mathcal{T}(N)$ で表す．

N の逸脱度として，図 1 に示す $\mu(N)$, $\eta(N)$, $\kappa(N)$ が考えられる [3]．N を tree-based とする葉の追加個数の最小値 $\mu(N)$ や，部分系統樹で覆えない N の頂点数の最小値 $\eta(N)$ は，多項式時間アルゴリズム [1, 3] 等の既存研究があるが，葉集合や頂点集合を保存しない構造に基づく．一方，それらを保存する構造に基づいた指標である，N の頂点を覆う部分系統樹の最小個数 $\kappa(N)$ については，計算手法等は知られていなかった．

本論文では，$\eta(N)$ の計算手法 [1] を基に，$\kappa(N)$ と被覆部分系統樹 $T_1,\dots,T_{\kappa(N)}$ を計算する $O(mn)$ 時間アルゴリズムを構築する．また，緩和問題を用いて，N の頂点集合による $\kappa(N)$ の min-max 定理を証明する．

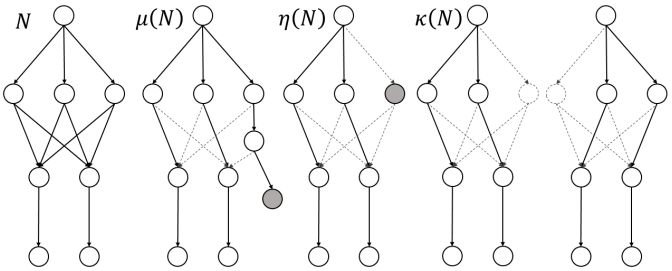

図 1. 逸脱度 $\mu(N),\eta(N),\kappa(N)$ の定義

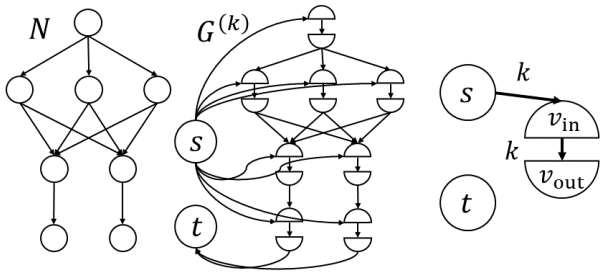

図 2. 補助ネットワーク $G^{(k)}$ の定義

3　提案手法

ネットワーク $G^{(k)}, G^{(k)}_{\mathrm{lb}}, H^{(k)}$ $(k\in\mathbb{N})$ を定義する．

$G^{(k)}$:N の各頂点 v を $v_{\mathrm{in}}, v_{\mathrm{out}}$ に分け，頂点 s,t と図 2 に示すような辺を追加．各辺の容量は $u(e)=k$.

$G^{(k)}_{\mathrm{lb}}$:$G^{(k)}$ に下限制約 $l(v_{\mathrm{in}},v_{\mathrm{out}})=1$ $(\forall v)$ を追加.

$H^{(k)}$: 頂点 σ,τ を追加し，$G^{(k)}_{\mathrm{lb}}$ の下限制約を変換 [2].

これらの補助ネットワークと，N を被覆する部分系統樹の個数 k に関して，次の定理が成り立つ．

定理 次の 3 条件は同値.
1. $\exists T_1,\dots,T_k\in\mathcal{T}(N), \bigcup_i V(T_i)=V$.
2. $G^{(k)}_{\mathrm{lb}}$ 上の流量 $k|X|$ の st-フローが存在.
3. $H^{(k)}$ 上の流量 n の $\sigma\tau$-フローが存在.

証明は省略する．これにより以下の等式が成り立つ．

$$\kappa(N)=\min\left\{k\in\mathbb{Z}\;\middle|\;\exists f \text{ in } H^{(k)}, \mathrm{val}(f)=|V|\right\}. \quad (1)$$

$\kappa(N)$ を求める提案手法として，逐次的に最大流を求めるシンプルかつ効率的なアルゴリズムを構築する．

$\kappa(N)$ と $H^{(\kappa(N))}$ 上 $\sigma\tau$-フロー f の計算手法

1. $k \leftarrow 1$ 及び $f(e) \leftarrow 0$ と初期化
2. $\mathrm{val}(f) < n$ の間，残余ネットワーク $H_f^{(k)}$ 上に増大路があれば f を増大，なければ k に 1 加算
3. $\mathrm{val}(f) = n$ となったら $k\,(=\kappa(N)), f$ を出力

ステップ 2 はちょうど $n + \kappa(N) - 1\,(\leq 2n-1)$ 回行われることから，計算量は $O(mn)$ である．

次に，f を $G_{\mathrm{lb}}^{(\kappa(N))}$ 上の st-フロー f' に変換し，$G^{(1)}$ 上の $\kappa(N)$ 個の st-フロー $f'_1, \ldots, f'_{\kappa(N)}$ に分解する．各 f'_i から得た部分系統樹 $T_i\,(1 \leq i \leq \kappa(N))$ は，f' が満たす下限制約に対応して $\bigcup_i V(T_i) = V$ を満たす．

フローの分解では，各辺の流量の下限及び上限が $0, 1$ のいずれかである最大流問題を $\kappa(N) - 1$ 回解くことになる．本研究では，下限制約を負のコスト -1 へと変換した最小費用流問題に増大路アルゴリズム [6] を適用し，$G^{(k)}$ が DAG であることとコストの小ささを用いて高速化することで，$\kappa(N)$ の大きさに関わらず合計 $O(mn)$ 時間でフローを分解する手法を提案した．

4 min-max 定理

$\kappa(N)$ は，$\sum_{T: v \in V(T)} a_T \geq 1\ (\forall v \in V)$ の制約下で $\{a_T\} \in \{0,1\}^{\mathcal{T}(N)}$ の和を最小化する問題の最適値である．a_T の定義域を非負実数とした緩和 LP 問題の最適値を $\kappa_{\mathrm{r}}(N)$ とすると，次式が成立することを示せる．

$$\kappa_{\mathrm{r}}(N) = \min\left\{ a \in \mathbb{Q} \;\middle|\; \exists f \text{ in } G_{\mathrm{lb}}^{(a)} \right\}. \tag{2}$$

（$G_{\mathrm{lb}}^{(a)}$ は $a \in \mathbb{Z}$ と同様の定義．）式 (2) より $\kappa(N) = \lceil \kappa_{\mathrm{r}}(N) \rceil$ を示せる．また，$G_{\mathrm{lb}}^{(a)}$ に容量 n の辺 (t, s) を追加したネットワークに循環フローが存在することは，

$$\forall U \subseteq V(G_{\mathrm{lb}}^{(a)}),\ \sum_{e \in \delta^-(U)} l(e) \leq \sum_{e \in \delta^+(U)} u(e) \tag{3}$$

が成立することと同値である [5]．これらから次を得る．

$$\kappa_{\mathrm{r}}(N) = \max_{U \subseteq V(G_{\mathrm{lb}}^{(a)})} \frac{\sum_{e \in \delta^-(U)} l(e)}{|\delta^+(U)|}. \tag{4}$$

右辺の最適解 U^* に対し，$\{v \in V \setminus X \mid v_{\mathrm{out}} \in U^*\}\ (=: W^*)$ により定義される U^{**} (図 3) も最適解であることが示せる．$\Delta(W) := \{v \in V \mid \exists w \in W, (w, v) \in E\}$ として最適値と対応させ，次の min-max 定理を得る．

$\kappa_{\mathrm{r}}(N)$ 及び $\kappa(N)$ の min-max 定理

$$\kappa_{\mathrm{r}}(N) = \max_{W \subseteq V \setminus X} \frac{|W \setminus \Delta(W)|}{|\Delta(W) \setminus W|}, \tag{5}$$

$$\kappa(N) = \max_{W \subseteq V \setminus X} \left\lceil \frac{|W \setminus \Delta(W)|}{|\Delta(W) \setminus W|} \right\rceil. \tag{6}$$

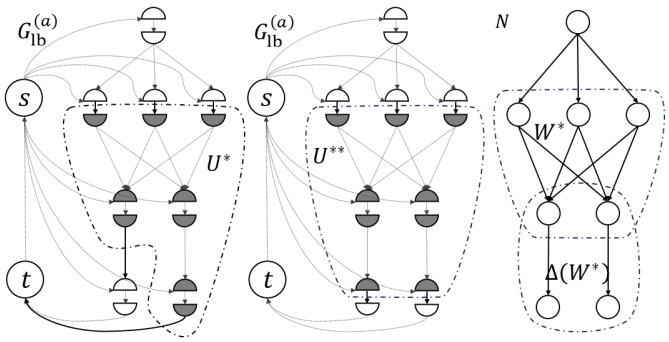

図 3. 式 (4) の最適解 U^*, U^{**} と対応する W^*

$\kappa_{\mathrm{r}}(N)$ の分母は n 以下であることを用いて，$\kappa_{\mathrm{r}}(N)$ 及び最適解 W^* の $O(mn \log n)$ 時間アルゴリズムが構築できる．また，$|V(T) \cap (W \setminus \Delta(W))| \leq |V(T) \cap (\Delta(W) \setminus W)|$ が示せて，定理に表れる関数が $\kappa(N), \kappa_{\mathrm{r}}(N)$ の下界であることの直感的な理解が得られる．さらに，$\kappa_{\mathrm{r}}(N) - 1$ は劣モジュラ関数の最小比問題の一種であることが示せる．一方で，既存の結果 [3] に基づき $\mu(N) = \max_{W \subseteq V \setminus X} |W \setminus \Delta(W)| - |\Delta(W) \setminus W|$ が示せて，$\kappa(N)$ と $\mu(N)$ の関係性を得ることが出来る．

5 むすび

本論文では，N の逸脱度 $\kappa(N)$ に対し，$O(mn)$ 時間アルゴリズム及び min-max 定理を導出した．今後の展望として，N の特殊な構造を用いたアルゴリズム及び特徴付けを改善することや，計算手法が近い $\kappa(N)$ と $\eta(N)$ の関係性を考察することが挙げられる．

参考文献

[1] N. Davidov, A. Hernandez, J. Jian, P. McKenna, K. A. Medlin, R. Mojumder, M. Owen, A. Quijano, A. Rodriguez, K. St. John, K. Thai, and M. Uraga. Maximum covering subtrees for phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 18, No. 6, pp. 2823–2827, 2021.

[2] L. R. Ford and D. R. Fulkerson. *Flows in Networks*. Princeton University Press, Princeton, NJ, 1962.

[3] A. Francis, C. Semple, and M. Steel. New characterisations of tree-based networks and proximity measures. *Advances in Applied Mathematics*, Vol. 93, pp. 93–107, 2018.

[4] A. Francis and M. Steel. Which phylogenetic networks are merely trees with additional arcs? *Systematic Biology*, Vol. 64, No. 5, pp. 768–777, 2015.

[5] A. J. Hoffman. Some recent applications of the theory of linear inequalities to extremal combinatorial analysis. *Combinatorial Analysis, AMS, Providence*, pp. 113–128, 1960.

[6] N. Tomizawa. On some techniques useful for solution of transportation network problems. *Networks*, Vol. 1, No. 2, pp. 173–194, 1971.