統計数理研究所 75 周年記念事業（2019）

組合せ論的系統学における最近の話題

－グラフ理論が拓く系統解析の新展開－

Recent topics in combinatorial phylogenetics

—From graph theory to statistical analysis—

モデリング研究系　早水 桃子（Momoko Hayamizu）

要　　旨

生物の進化の道筋を解明する系統学的なデータ解析において，今日では系統樹を拡張した「系統ネットワーク」というグラフ構造が広く用いられるようになっているが，記述能力が高い汎用的なモデルが必ずしも優れたモデルであるとは限らないため，系統ネットワークのサブクラスの中で生物学的な妥当さと数学的な性質の良さを兼ね備えたものを見出すことは重要である．特に Francis and Steel (2015)が定義した「系統樹ベースのネットワーク」(tree-based phylogenetic network; TBN）は系統樹に辺を追加する単純な操作で得られる系統ネットワークのサブクラスで，TBN の数学的性質や計算複雑性に関する未解決問題は理論生物学分野のホットトピックスになっている．Hayamizu (2018)は，Francis and Steel が取り上げた決定／探索問題や数え上げ問題だけでなく，TBN に関する列挙問題や最適化問題にもスポットライトを当て，これらの問題を高速に解くアルゴリズムを統一的な視点で生み出す「TBN の構造定理」を証明し，多様な統計学的な応用を可能にした．本稿では，その研究成果の一端を紹介する．

キーワード：系統樹推定，系統ネットワーク，細分系統樹，離散アルゴリズム

1. 研究の背景

生物の進化は古くから系統樹（phylogenetic tree）を用いて記述されてきたが，例えば植物，菌類，細菌類が進化する過程では異種交雑（hybridization）や遺伝子の水平伝播（horizontal gene transfer; HGT）といった木構造で記述しきれない現象が起きうるため，あらゆる種の進化を系統樹だけで描写することはできないといわれている．また，仮にそのような現象を考慮しなくてもよい種を対象にした系統解析を行う場合でも，現実のデータを扱い，その情報を忠実に描写したいなら，木構造よりも融通の利くグラフ構造が欲しいと考えるのは自然であろう．

このようなニーズに動機づけられ，**組合せ論的系統学**（*combinatorial phylogenetics*）という理論生物学の一領域では，系統樹を拡張した**系統ネットワーク**（*phylogenetic network*）とその様々なサブクラスに関する研究がこれまでに多数行われている（Huson *et al.* (2010); Steel (2016)）．その研究成果は既に実際のデータ解析に応用されており，例えば，SplitTree などのソフトウェアは系統ネットワークを使ってデータを可視化するツールとして広く使われている（Bryant and Moulton (2004)）．ただし，このトレンドは系統ネットワークが系統樹に取って代わることを意味しているのではなく，系統樹は依然として進化を記述するファンダメンタルなモデルであることを強調しておく．

2. 系統樹ベースのネットワーク（TBN）と細分系統樹

興味のある現存種の集合 X を葉とする根付き二分系統ネットワーク N が与えられているとき，これらの種が辿った進化の道筋を系統樹モデルで記述するとなれば，N の中に X を葉とする何らかの系統樹 T を見出したくなる．そこで Francis and Steel (2015)は，系統樹に余分な辺を加えてできる**系統樹ベースのネットワーク**（*tree-based phylogenetic network*; *TBN*）を定義した．TBN は**細分系統樹**（*subdivision tree*）という全域木を持つ系統ネットワークと定義することもできるため，TBN を論じるうえで細分系統樹という概念は本質的な役割を果たす．

3. TBN の構造定理が導く一連のアルゴリズムと統計学的な意義

Hayamizu (2018)は，根付き二分系統樹 N の細分系統樹の集まり $\{T_1, \ldots, T_{\alpha(N)}\}$ を特徴づける構造定理を示し，次の一連の問題を解く高速なアルゴリズムを統一的な視点で記述した．

(1)**決定／探索問題：** 系統ネットワーク N が与えられたとき，N が TBN か（すなわち細分系統樹が存在するか）否かを決定し，存在するならば一つ見つける問題．Francis and Steel (2015)はこの問題を解く線形時間アルゴリズムを与えたが，次のように数え上げ問題に拡張すると，多項式時間では解けないかもしれないと予想していた．

(2)**数え上げ問題:** 系統ネットワーク N が与えられたとき，N の細分系統樹の個数 $\alpha(N) \in \mathbb{Z}_{\geq 0}$ を求める問題．Hayamizu (2018)は，これを解く線形時間アルゴリズムを与え，N が TBN のとき，$\alpha(N)$ は N の複雑さを評価する尺度になるため，モデル選択の文脈に関連する．

(3)**列挙問題：** 系統ネットワーク N が与えられたとき，N の全ての細分系統樹 $T_1, \ldots, T_{\alpha(N)}$ を列挙する問題．入力 N のサイズに関する多項式時間でこの問題を解くアルゴリズムが存在しないことはすぐに分かる（列挙したい解の個数 $\alpha(N)$ 自体が N のサイズに関する指数関数で表される場合があるため）．しかし，Hayamizu (2018)は，これを高速に解く線形時間遅延アルゴリズム（列挙アルゴリズムの中で最も効率的なクラスに属すもの）を与えた．全ての解ではなく指定の個数 $k \in \mathbb{N}$ の解のみを列挙するには，$O(k|V(N)|)$ 時間で十分である．これにより細分系統樹の一様サンプリングなどの応用が可能になる．

(4)**最適化問題:** 系統ネットワーク N と辺の重みづけ関数 $w \geq 0$ が与えられたとき，ある目的関数の値 $f(T)$ を最大化（または最小化）する細分系統樹 T を求める問題．力まかせ探索では指数時間を要するが，Hayamizu (2018)の構造定理は最適解を線形時間で求めるアルゴリズムを導く．この最適化問題は，N の各辺が存在するか否かの不確かさに応じた確率 w が与えられ，尤度または対数尤度 $f(T)$ を最大化するベストな細分系統樹 T を求めるという最尤推定の文脈で現れる問題である．

参 考 文 献

Bryant, D. and Moulton, V. (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks, *Molecular Biology and Evolution*, **21**, 255–265.

Francis, A. and Steel, M. (2015). Which phylogenetic networks are merely trees with additional arcs?, *Systematic Biology*, **64**, 768–777.

Hayamizu, M. (2018). A structural theorem for tree-based phylogenetic networks, preprint available at `arXiv:1811.05849[math.CO]`.

Huson, D. H. and Rupp, R. and Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*, Cambridge University Press.

Steel, M. (2016). *Phylogeny: Discrete and random processes in evolution*, SIAM.