研　究　報　告　書

「基礎医学と社会医学をつなぐ離散幾何学的モデリング」

研究タイプ：通常型
研究期間：　2016 年 10 月～2020 年 3 月
研　究　者：　早水　桃子

1．　研究のねらい

対象物同士の距離（非類似度）が与えられ，近さや遠さの全体像をグラフ（ネットワーク）で記述するという問題設定は幅広い分野に現れるが，特に医学や生物学分野では頻出する重要なものである．典型的な例はゲノム配列の違いに基づく生物同士の距離から進化系統樹を構築するというもので，系統樹で実現できる距離（木距離）については理論から応用まで多くの既存研究があるが，系統樹以外のグラフと距離の関係性について数学的に解明されていることは少なく，木距離から系統樹を構築する Saitou & Nei (1987) の Neighbor-Joining 法のように普及している計算手法もない．

本研究では，多様な生物学的現象のモデル化やデータ解析を可能にする方法論の創出によって社会的にも重要な基礎医学の課題解決に貢献することを目指し，有限距離空間からグラフ構造を抽出するデータ解析全般を「離散幾何学的モデリング」という枠組みで捉え，その枠組みの構築に関わる概念や方法を数理科学・生命科学の協働を通じて整備する．具体的には，細胞の分化と病原体の進化という二つの現象に関するデータ解析上の諸問題を取り上げる．

- **課題 A　細胞分化の木モデルを構築するための基礎と応用**
  個々の細胞における多数の遺伝子の発現量を網羅的に測定するシングルセル技術は近年の生物学の技術革新の一つである．この技術がもたらす細胞の詳細なデータは新発見の宝庫と期待されているが，その解析方法に関しては多くの問題が残されている．本研究では，シングルセルの遺伝子発現データから最小全域木問題を解くアルゴリズムを活用して細胞分化の木構造を解明するデータ解析手法を確立させ，細胞生物学者のニーズを満たす実用的なソフトウェアを開発する．
- **課題 B　細菌やウイルスなどの進化をモデル化するための基礎**
  木距離から系統樹を構築する Saitou & Nei (1987) の Neighbor-Joining 法は距離から系統樹を構築するツールとして普及しているが，例えば微生物の進化は遺伝子の水平伝播などが原因で複雑なネットワーク構造になり得るので，従来の系統樹モデルでは記述しきれない．また系統樹で十分に記述可能な現象を考察する場合でも，ノイズを含む現実のデータを木構造で完璧に記述することは不可能で，複雑なネットワークを分析して情報を抽出する方法が必要となる．本研究では，系統樹の拡張版といえる新しい進化のモデルを提案するとともに，系統ネットワークの数理的性質を解明することで革新的な系統解析の方法論を創出する。

2．　研究成果

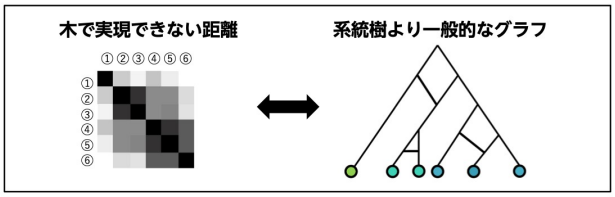
（1）概要

本研究では①色々なグラフと距離空間の関係性の解明，②新しいデータ解析を可能にする高速なアルゴリズムの設計，③実用的なデータ解析ソフトウェアの開発という幅広い成果を上げた．

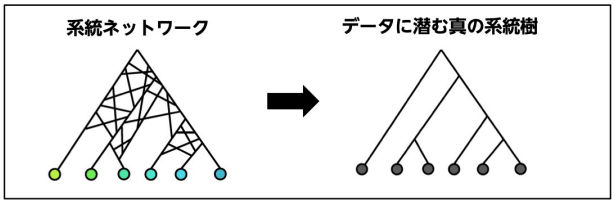
研究の概要

数理的な基礎研究から，諸分野との連携・融合により医学・生物学にインパクトを与える応用研究まで行う

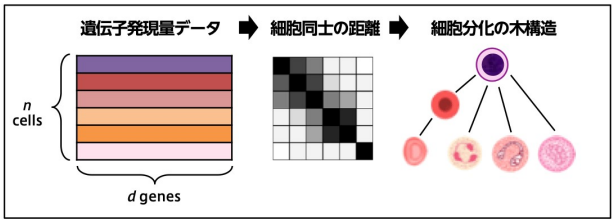
① 色々なグラフと距離空間の関係を調べる基礎研究



② 新しいデータ解析を可能にする高速なアルゴリズムの設計



③ 実用的なデータ解析ソフトウェアの開発，医学・生物学を含む諸分野との協働



課題 B は①と②に該当する．一つ目は系統樹の拡張版を創出するために木距離を一般化するという発想に基づくもので，系統樹／木距離の概念を一般化した系統カクタス／カクタス距離を提案し，カクタス距離が木距離と同様に多くの良い性質を持つことを示した（講演 5，論文 2）．またカクタス距離から系統カクタスを計算するための，Neighbor-Joining 法と同程度に効率的な方法も与えた（論文 2）．二つ目は理論系統学で話題の tree-based network（TBN）と全域系統樹に関するもので，全域系統樹の数え上げ・列挙・最適化といった一連の問題を定式化し，それらを統一的な視点で解く枠組みとなる TBN の構造定理を証明し，その系として各問題を解く線形時間／線形時間遅延アルゴリズムを与えた．これにより TBN に関する既知の結果をまとめて一般化し，全域系統樹の数え上げと列挙の計算量という複数の未解決問題を同時に解決し，さらに統計学的な応用も初めて開拓した（プレプリント 1，講演 2・3，受賞 1）．さらにランキング上位の全域系統樹を上から順に抽出する線形時間遅延アルゴリズムというより実用的な方法に発展させた（プレプリント 2）．

課題 A はどの側面もあるが特に③に該当する．近年の細胞生物学分野では，個々の細胞における大量の遺伝子の発現量を網羅的に計測したデータ（高スループットのシングルセル RNA-seq データ）を利用して細胞分化などの全体像を推定するためのデータ解析技術の開発・確立が大きな課題になっており，細胞同士の遺伝子発現パターンの非類似度から細胞分化の全体像（木構造）を推定する問題に関しては離散数学の古典的な知見の一つである最小

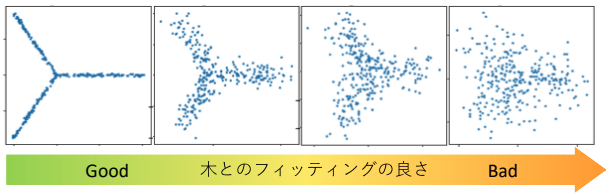
全域木（MST）を求めるアルゴリズムが経験則として定着し始めている．確かに高次元ユークリッド空間の点群データから距離行列を計算してMSTを求めればデータの可視化がしやすくなるので，MST を求めるアルゴリズムは便利なツールだが，それで意味のある木構造が得られるかはケースバイケースで，そもそもデータと木モデルのフィッティングの良さをはかる定量的尺度がないので，もとのデータの情報をどれほど良く反映した木構造が得られたのかを事後評価できないという大きな問題点がある．本研究では，データと木モデルが完璧にフィットするときはその木は MST に他ならないことを示した（論文 1）．また，データと木モデルのフィッティングの良さを定量的に評価する手法を提案し，“データの主成分”といえる木構造を抽出する方法も与え，シングルセルの遺伝子発現データから細胞分化の木構造を定量的に推定するデータ解析ソフトウェア Treefit を開発した（ソフトウェア 1）．

（2）詳細

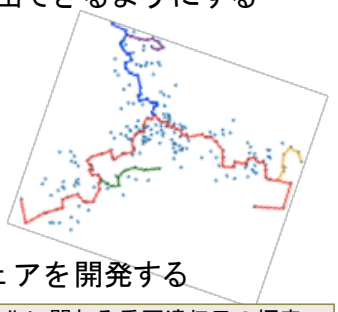
課題 A 「細胞分化の木モデルを構築するための基礎と応用」

研究を開始した当初は，木距離（tree metric）を特徴づける四点条件の類似物である「四点目条件」を活用することで「MST 距離空間（最小全域木で実現できる有限距離空間）」を厳密に特徴づけたり，「与えられた有限距離空間の MST-likeness」を定量化したりするアプローチを考えていた．この方向性による研究成果として論文 1 を発表したが，本さきがけ研究を進める中で細胞生物学，病理学，バイオインフォマティクス，コンピューターサイエンス，物理学，数学といった多様な専門分野の研究者や企業のエンジニアと協働したことにより，別のアプローチのほうが適切であると考えるようになった．具体的には MST-likeness の尺度はノイズに対して頑健であることが必要であると再認識したため，もとのデータに摂動を加えたときの最小全域木構造の安定性によってデータの MST-likeness を評価するという方針に至った．この方針転換によって，異なる木グラフ同士の非類似度をどのように定量化すれば良いかという新しい課題が生まれたが，木グラフのラプラシアンの固有ベクトルの集合同士の類似性をグラスマン距離を用いて評価するという方法で木グラフ同士の近さや遠さを定量的に評価できるようにするという解決策を与え，実際のデータ解析における有用性を実験的に示し，R および Python の両方でオープンソースのデータ解析ソフトウェア Treefit を開発して GitHub で公開した（ソフトウェア 1）．Treefit についてはホームページで生物学者向けの実践的なチュートリアルも公開し，また生物学分野の国際シンポジウムで招待講演を行うなど（講演 1），研究成果の普及促進にも努めた．

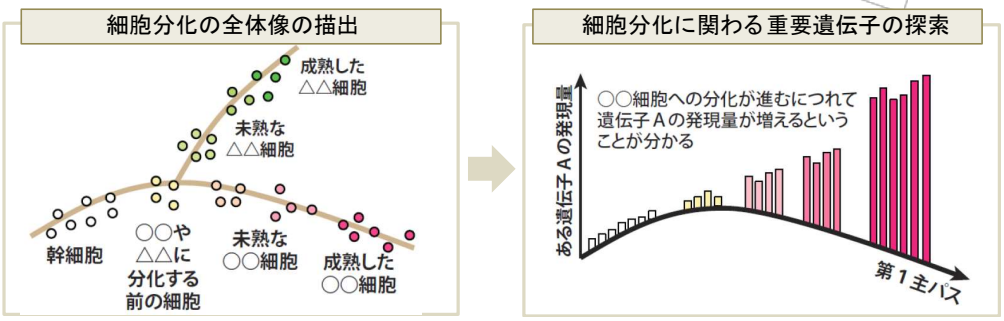
❶ 距離空間と木モデルのフィッティングを定量的に評価できるようにする



❷ 距離空間から主要な木構造を抽出できるようにする



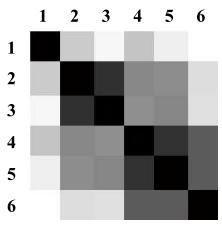
❸ 細胞の遺伝子発現データ解析に使えるソフトウェアを開発する



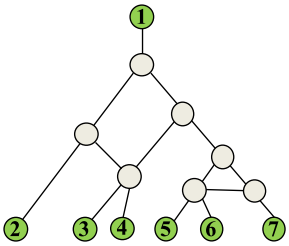
課題 B 「細菌やウイルスなどの進化をモデル化するための基礎」

進化は系統樹で記述されることが多いが，実際には植物の異種交雑や，微生物間の遺伝子の水平伝播といった現象があるため，全ての生物の進化を系統樹で記述することはできず，「系統樹の拡張版」といえる進化のモデルが必要とされている．本研究の1つ目の成果では，

Cactus metric

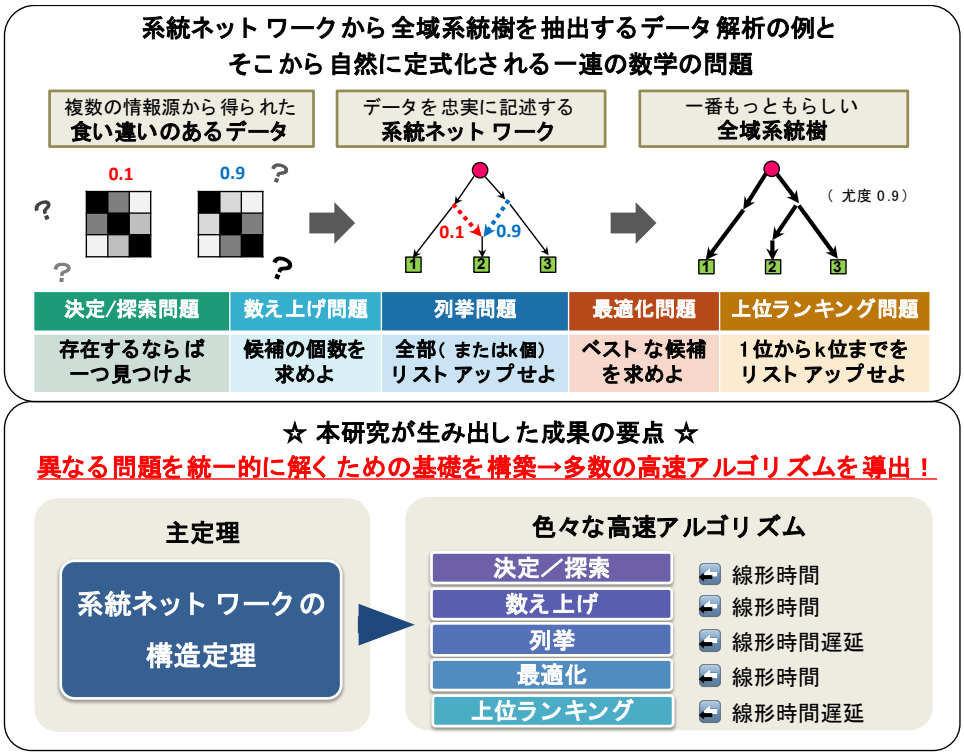


系統カクタス



木距離を拡張したカクタス距離（cactus metric）を提案して，カクタス距離を実現する最適な系統カクタスが一意に定まることや，カクタス距離から系統カクタスを計算する問題が，木距離から系統樹を計算する問題と同様に $O(n^3)$時間で解けることなどを示し（ただし n は距離空間の点の個数），木距離の良い性質の多くが系統カクタスに継承されていることを証明した．なお，どのようなときに距離の最適な実現は一意に定まるのかという一般的な問題についても考察し，部分的な場合の特徴づけを与えた（準備中論文1）．

もう一つの研究成果は，tree-based network という系統ネットワークに関するものである．現実的には系統樹モデルを使いたい場合も多数あるが，現実のデータにはノイズや不確実性がつきもので，単純な木構造では表現しきれないという難しさがある．系統ネットワークから真の系統樹を探る方法を創ることは進化生物学的に重要であるとともに，多様な理論的問題が自然に現れるため数学的にも興味深いテーマである．本研究では，全域系統樹にまつわる個々の問題の解き方を考えるのではなく，系統ネットワークの構造（どのような部分構造に一意的に分解されるのか）を解明し，様々な問題を統一的に解くための枠組みとなる構造定理を与え，その定理の系として多くの問題に対する高速なアルゴリズムを導出して，これまでにはできなかった様々な系統学的なデータ解析を可能にした．



3.　今後の展開

- 課題 A：実験研究者との連携をさらに強化し，本研究で開発したソフトウェアの応用を推進して生物学的に新しい知見の獲得を進めたい．最小全域木を活用したデータ解析の枠組みに関する理論的な考察を深めるとともに，将来的にはもっと複雑なグラフを用いたデータ解析ができるようにしたい．
- 課題 B：位相的データ解析（TDA）や統計学分野の研究者と融合して理論を深める．海外の系統学研究者コミュニティとも交流を続けるとともに，例えば感染症医学分野の研究者とも協力して，医学的な応用分野にも貢献したい．

4.　自己評価

全体を通して，「社会的課題の解決に向けた数学と諸分野の協働」という領域の趣旨に合致する当初の研究計画を完遂しただけでなく，さらにプラスアルファの良い成果も上げることができ

た. 課題 A は数理科学の研究者・生命科学の研究者・統計科学の研究者・オープンソースソフトウェア開発者の良い協働事例になったと考えており, 今後もこのような異分野・異業種連携の取り組みを続けたい. 課題 B は当初の目標をクリアしただけでなく, 当初の計画を上回る数学的成果(TBN の構造定理)を上げることができたと考えている.

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. Momoko Hayamizu and Kenji Fukumizu. On minimum spanning tree-like metric spaces. Discrete Applied Mathematics. 2017, 226, 51-57. DOI: 10.1016/j.dam.2017.04.001
2. Momoko Hayamizu, Katharina T. Huber, Vincent Moulton, Yukihiro Murakami. Recognizing and realizing cactus metrics. Information Processing Letters 2020, 157 (105916). DOI: 10.1016/j.ipl.2020.105916

(2)特許出願

研究期間累積件数:0 件(公開前の出願件名については件数のみ記載)

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

【プレプリント】

1. Momoko Hayamizu. A structure theorem for tree-based phylogenetic networks. arXiv:1811.05849 [math.CO], 2018 年 11 月.
2. Momoko Hayamizu and Kazuhisa Makino, Ranking top-k trees in tree-based phylogenetic networks, arXiv:1904.12432 [math.CO], 2019 年 4 月.

【準備中論文】

1. Momoko Hayamizu and Andrew Francis: Optimal realisations of cubic graph metrics.

【講演】(※代表的な講演のみ)

1. (招待, 国際)Momoko Hayamizu. How to reconstruct a cell differentiation trajectory using scRNA-seq snapshot data. The 1st International Symposium on Human InformatiX (X-Dimensional Human Informatics and Biology), ATR, 京都, 2020 年 2 月 27-28 日
2. (招待, 国際)Momoko Hayamizu, A structure theorem for tree-based phylogenetic networks and its algorithmic applications, Combinatorics Seminar, Institute of Mathematics, Academia Sinica (台湾), 2019 年 4 月 19 日
3. Tree-based network の構造定理と系統樹推定に関する諸問題への応用, 早水 桃子 日本応用数理学会 2019 年研究部会連合発表会 離散システム(1), 2019 年 3 月 4 日
(▸下記の受賞情報を参照)
4. Universal tree-based network とその最小サイズについて 早水桃子(実講演者), 鍛冶静雄 山口大学大学院創成科学研究科, 藤重 悟 京都大学数理解析研究所 日本数学会 2017 年度秋季総合分科会, 2017 年 9 月 11 日
(▸未解決問題ワークショップを通じた他のさきがけ研究者等との共同研究成果)
5. (国際) X-cactus trees and cactus tree metrics Momoko Hayamizu The 21st Annual New Zealand Phylogenomics Meeting (Waiheke 2017) - The Interface of Mathematics and

Biology　2017 年 2 月 14 日

【受賞】

1. 早水 桃子, 日本応用数理学会 2019 年研究部会連合発表会優秀講演賞,「Tree-based network の構造定理と系統樹推定に関する諸問題への応用」, 2019 年 6 月 28 日.

【ソフトウェア】

1. Momoko Hayamizu, Kouhei Sutou, Ryohei Suzuki, Hiromi Ishii. Treefit. 2020 年 2 月 7 日. https://hayamizu-lab.github.io/treefit/

【インタビュー取材】

1. 「数学の言葉で、世界を新たに描く」,『Someone』2018 年秋号, Vol.44 （研究者に会いに行こう！p.20-21）
   https://lne.st/download/40419/
2. 離散数学で解き明かす「細胞分化の木構造」,『統計数理研究所ニュース』No. 146, p.2-5,統数研プロジェクト紹介, 2019 年 11 月
   https://www.ism.ac.jp/gaiyo-news/News/No146.pdf

【寄稿】

1. 「進化の系統樹とデータ解析」,『数学セミナー』(リレー連載「数理のクロスロード」), 2020 年 1 月号・2 月号