

スペクトラルグラフ理論に基づく新規系統解析手法の開発

Developing the novel phylogenetic method based on the spectral graph theory

京都大学化学研究所 松井求

研究成果概要

AlphaFold の誕生以来、タンパク質立体構造の高精度な推定がもたらす恩恵はさまざまな分野に波及している。そして、AlphaFold3 や ESMFold はその精度もさることながら、リガンドとの結合が高精度に推定可能である、多重配列アライメントに非依存であるといった優れた特徴を持ち、応用研究が加速度的に進められている。分子系統学の分野では近年、これらのソフトウェアを駆使してタンパク質の立体構造情報を系統樹推定法に取り込もうという取り組みが活発に行われている。しかしながら、タンパク質立体構造の進化過程に関する理解がまだ不十分なことから、進化モデルをどのように構築すべきか？ 立体構造の情報を既存の系統樹推定のスキームにどのように接続すべきか？ など基礎的な知見の蓄積すら未だ十分ではない。

本研究では、昨年度に引き続き、これまでに開発してきた Graph Splitting (GS) 法、Edge Perturbation (EP) 法、および PA+NJ+EP (PANJEP) 法をタンパク質立体構造も利用できるようにアルゴリズムを改善し、実データに適用することで有効な手法であるか検証した。アルゴリズムの開発と検証は、京都大学化学研究所スーパーコンピュータシステムにおいて、AlphaFold DB に登録されている膨大な立体構造データセットに上記の新手法を適用することで行った。まず、(1) 立体構造の類似性、(2) 立体構造の縮約配列 (FoldSeek) の類似性、(3) モチーフの共有度、(4) アミノ酸配列、の四つの情報のうち、どの情報が系統樹推定に最も有用であるか比較・評価したところ、一般的には(1) (2) および(4) のそれぞれから構築した距離行列を組み合わせた方法が最も精度の良い系統樹を推定できた。ただし、全体的な立体構造が保存されていない、あるいはドメインシャッフリングが頻繁に起こっている、といったようなタンパク質群において同様の検証を行ったところ、(3) の情報を加えた方が系統樹の推定精度が向上することが分かった。タンパク質立体構造予測の手法開発は日進月歩で進んでおり、その進捗を積極的に分子系統学に取り込んでいくことは、機能未知遺伝子の機能予測やその進化解析など幅広い波及効果も期待できることも相まって、重要課題となっている。本成果は、その基盤整備につながるものと考えている。

発表論文(謝辞なし)

Morigasaki S, Matsui M, Ohtsu I, Doi Y, Kawano Y, Nakai R, Iwasaki W, Hayashi H, Takaya N. **Temporal and fertilizer-dependent dynamics of soil bacterial communities in buckwheat fields under long-term management**, *Scientific Reports*, 14(1):9896, 2024