受付 ID	15a-24
分野	生物

Non-Homogeneous 置換モデルに基づくγプロテオバクテリア大系統解析
Large−scale phylogenetic analysis of γ–proteobacteria with the Non−Homogeneous substitution model

石川　奏太
東京大学理学系研究科

１．研究目的

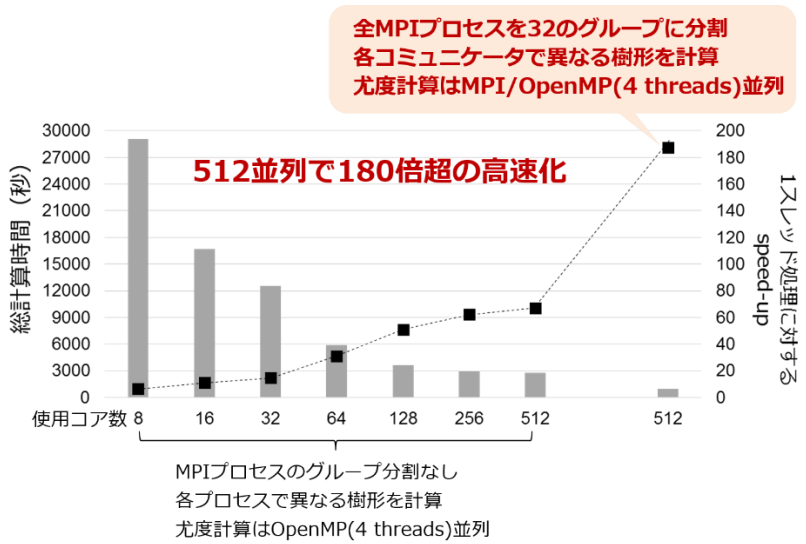
γプロテオバクテリアは真正細菌（バクテリア）の中でも特に巨大な分類群として知られ、海洋や深海底の極限環境、動物の腸内（例：大腸菌）など様々な環境から発見される。また、本分類群に属する個々の系統では異なる形質が独自の進化プロセスを経て獲得されてきた。例えば、生活様式についてみても、光合成を行うものや有機物・無機物を分解し化学合成を行うものなど多種多様である．γプロテオバクテリアでは、近年シークエンス技術の発展に伴い多くの種にてゲノム配列情報が解読されるともに、大規模遺伝子配列データに基づく分子系統解析により進化系統樹を推測し、本生物群における複雑な生物進化の歴史を解明する試みが活発に行われている。しかしながら、過去研究において行われた分子系統解析では、ゲノムレベルの遺伝子配列情報を用いたにも関わらず本生物群内の詳細な系統関係を頑健に推測することが出来なかった。

γプロテオバクテリア由来遺伝子配列データに基づく分子系統解析において頑健な進化系統樹を推測出来なかった原因として、「系統間における配列進化プロセスの不均一性」という問題が示唆される。従来の解析手法では、配列データに含まれる進化プロセスの不均一性を適切に評価できないため、新たなアプローチによる再解析が必要である。そこで、本計画では上記問題に対し有効性が実証されている「Non-Homogeneous 置換モデル（以下 NH モデル）」に基づく分子系統解析により、信頼性の高いγプロテオバクテリア系統樹を推測する。

ただし、Non-Homogeneous 置換モデルを用いた分子系統解析では系統樹上の個々の枝に対し異なるモデルパラメータを推定するため、全ての枝に同一のパラメータ値を適用する従来手法に比べ系統樹の尤度計算に要する演算量・計算時間が飛躍的に増加する。そこで本研究では、NH モデルを実装したプログラム(NHML: Galtier and Gouy, *Mol. Biol. Evol.*, 15(7), 871–879, 1998)に基づく分子系統樹推測をスーパーコンピュータ上で高速に行えるようにすることが第一の目的である。また並列版 NHML に基づき、γプロテオバクテリア由来の大規模遺伝子配列データの再解析を行うことで。本生物群の内部系統関係をより頑健に再推測することが最終的な目的である。

２．研究成果の内容

本プロジェクトでは NHML を用いた分子系統解析のうち、①分子系統樹 1 本ごとの尤度計算＝モデルパラメータおよび枝長の最適化と、②系統樹空間の発見的探索による最尤系統樹選択という粒度の異なる 2 つのアルゴリズムを対象に並列化を行った。①においては遺伝子配列アライメント座位ごとの尤度計算は独立に行えること、また尤度計算におけるモデルパラメータおよび枝長の最適化は解析的な手法を用いることでパラメータ単位で独立に行えることのそれぞれに着目し、系統樹 1 本の尤度計算を OpenMP（＝座位ごとの計算のスレッド並列化）および MPI（＝異なるパラメータの並列的最適化）によりハイブリッド並列化した。さらに②において、最尤系統樹の発見的探索のためにはトポロジーの異なる複数の提案樹形について尤度計算を行い、それぞれを比較する必要があるが、これらの尤度計算は独立して行えることに着目したさらなる並列化も導入した。具体的には、MPI_COMM_WORLD の分割により複数の MPI プロセスからなるサブグループを作成し、異なる提案樹形の尤度計算をこれらのサブグループに分散させ、サブグループ内でも①のハイブリッドな並列計算を行う、という三層からなる Multi-grained な並列スキームを実装した。その結果、最大ノード数(32 ノード、512 コア)を使用時に、全 MPI プロセスを 1 グループ 4 プロセスの計 32 グループに分割した並列スキームにおいて 180 倍超の高速化を達成した(図)。



３．学際共同利用として実施した意義

本プロジェクトにおいて実装した「粒度の異なる三層のアルゴリズムのハイブリッド並列化」の性能を適切に評価するためには、多数の計算ノード。コアを使用できる計算環境は必須であった。COMA システムを利用することで我々はこの問題を解決でき、最大の並列効率を得るためのリソース配分について詳細な検証を行えた。またそ

の結果に基づき、GPU などぼアクセラレータ含むより多くの計算リソースを効率的に使用したさらなる展望(後述)をみることが出来た。

４．今後の展望

本プロジェクトは今後も継続する予定であり、既に平成 28 年度 HA-PACS システムにおける学際共同利用プロジェクトとして採択されている。平成 28 年度では、特にアクセラレータを利用したヘテロジニアスな並列化に着目し、Multi-grained 並列化に GPU 計算技術を新たに導入することで、より大規模な遺伝子配列データや、パラメータ数の多いより複雑な NH モデルに基づく分子系統解析にも耐えうる超並列アプリケーションの開発を目指す。平成 27 年度までの成果では、塩基配列データと最も単純な NH モデルによる分子系統解析でしかγプロテオバクテリアの内部系統関係を推測することが出来なかった。GPU 並列計算を実装した本プログラムを適用することにより。より充実した巨大アライメントとアミノ酸配列データに基づくより頑健な分子系統解析を行うことで、γプロテオバクテリア内部系統関係の最終的な評価を行うことが目標である。

５．成果発表

(1) 学術論文　上記成果に基づく論文を準備中

(2) 学会発表　(○は発表者)

○Sohta Ishikawa, Yuji Inagaki, Tetsuo Hashimoto, A multi-grained MPI/OpenMP parallelization of the maximum-likelihood phylogenetic inference with the non-homogeneous model, *Mathematical and Computational Evolutionary Biology 2016*, Montpellier, France, June, 2016, Poster

○石川奏太, Non-Homogeneous 置換モデルに基づく分子系統解析の大規模並列化, 生命情報科学若手の会第 7 回研究会, 慶應義塾大学鶴岡タウンキャンパス・鶴岡・山形, 2015 年 10 月, 口頭

〇Sohta Ishikawa. Computational challenge for the acceleration of the large-scale phylogenetic analyses based on the non-homogeneous models of evolution. *International HPC Summer School 2015*, Toronto, Canada, June 2015, poster

(3) その他

使用計算機	使用計算機に○	配分リソース※
HA-PACS		
HA-PACS/TCA		
COMA	○	2,100 時間

※配分リソースについては 32node 換算時間をご記入ください。