受付 ID	16a46
分野	生物

大規模遺伝子配列データに基づく分子系統解析の GPU 並列化

Parallelization of the large-scale phylogenetic inferences on the multiple graphic processors

石川　奏太
東京大学理学系研究科

１．研究目的

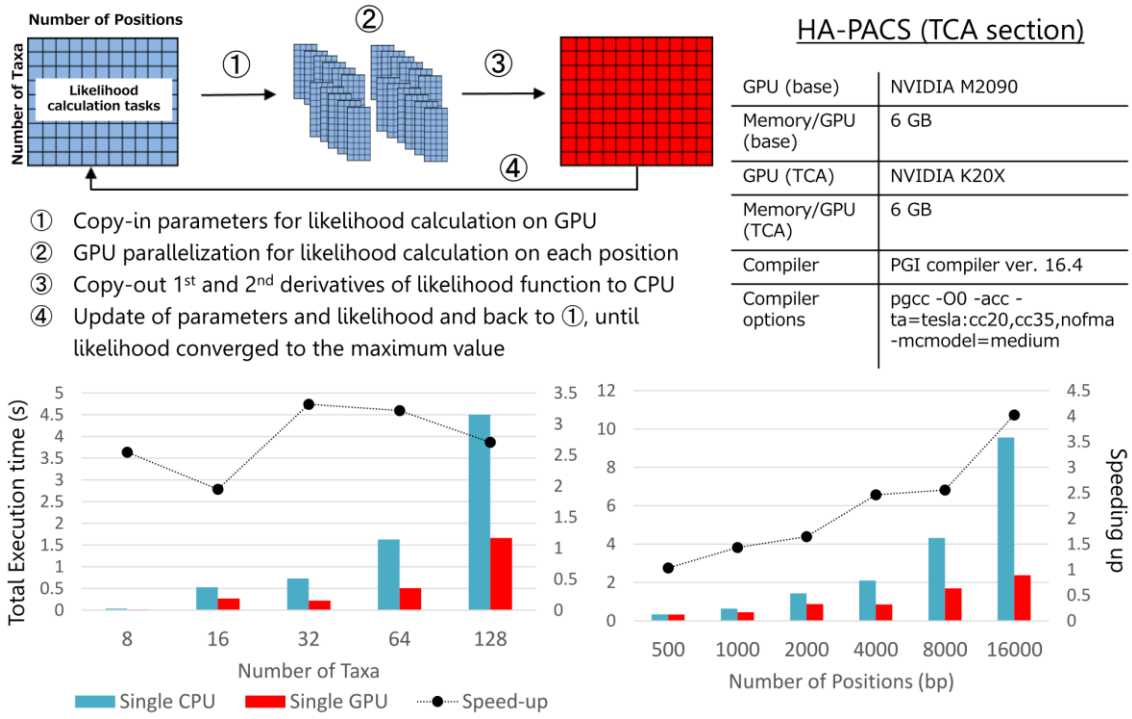
近年、シークエンス技術の急速な発展により、ゲノムデータに代表される膨大な量の遺伝子配列情報が急速に取得・蓄積され、生命の大系統や初期進化の解明に必要なデータ基盤が整いつつある。生命の進化研究には遺伝子配列データに基づき生物の系統関係を推測する「分子系統解析」が広く用いられている。また、生命の大系統の推測には多様な系統から得られた遺伝子配列データに基づく網羅的な分子系統解析が必須であるが、そのような解析では「系統間における遺伝子配列進化プロセスの不均一性」を考慮しなければならない。そこで本研究チームでは系統間で異なる進化プロセスをそれぞれ独立したパラメータとして推測する置換モデル（Non-Homogeneous モデル）に注目し、同モデルの頑健性について実データおよびシミュレーションによる評価を行ってきた。

一方、近年における大規模分子系統解析では数百以上の系統および遺伝子数からなる巨大アライメントが頻繁に用いられている。同規模のデータを Non-Homogeneous モデルに適用した場合、推測すべきパラメータ数が飛躍的に上昇し、実験室レベルの計算機では系統樹推測に数十日もの時間を要する問題が生じる。そのため、大規模遺伝子配列データと Non-Homogeneous モデルに基づく高速な分子系統解析を可能にするプログラムの開発は進化生物学における重要な計算科学的課題である。本プロジェクトでは平成 27 年度より継続して Non-Homogeneous モデルを実装した分子系統解析プログラムの汎用 CPU を用いた MPI/OpenMP 並列化に取り組んできた。しかし、現状飛躍的に増大する遺伝子配列データに基づく解析規模のスケーリングに対応するためには、アクセラレータを用いた更に高性能な並列計算の実装が不可欠であると考えている。そこで本年度においては超高速計算システム分野と連携し、同分野との専門的な議論に基づき HA-PACS システム上での分子系統解析プログラムの GPU 並列化を目指した。

２．研究成果の内容

本年度における研究では、既存のプログラムである「NHML (Galtier, N. & Gouy, M. Mol. Biol. Evol., 15.7, 871-879, 1998.)」を対象に、本プログラムによる系統樹の尤度計算

アルゴリズムに GPU 並列化を適用した。並列化にあたり、筑波大学ハイパフォーマンス・コンピューティング・システム研究室（高性能計算分野）との共同研究のもと、同研究室にて開発中である GPU クラスタ用の並列プログラミング言語 XcalableACC (M. Nakao, et al., Workshop on accelerator programming using directives, 2014.)の実アプリケーション実装を行った。同大学計算科学研究センターの提供する GPU スーパークラスタ「HA-PACS」にて性能評価では、128 種 16,000 塩基座位までの大規模データサイズに対し良好な速度向上を得た（図）。



NONHOMO プログラムにおける系統樹の尤度計算アルゴリズムの GPU 並列化および HA-PACS システムにおける性能評価

３．学際共同利用として実施した意義

本研究は学際開拓プロジェクトとして申請し、高性能計算分野との共同体制のもと新規言語に基づく分子系統解析プログラムの GPU 並列化を行った。分子系統解析では、系統間の進化プロセスを均一とする単純な置換モデルを実装したプログラムの GPU 並列化については幾つかの先行研究があるが、Non-Homogeneous モデルを実装したプログラムの並列化については本プロジェクト以外には例が無い。特に生命の大系統や初期進化を頑健に推測するためには、Non-Homogeneous モデルを含めた複数のモデルに基づき系統樹の評価を行うことが必要不可欠である。本プロジェクトの達成により、Non-Homogeneous モデルに基づく分子系統解析の計算科学的課題が解決され、ゲノム情報ビッグバン時代における生命の進化研究全体に大きく寄与できると期待される。また、本研究で適用した

XcarableACC 言語は GPU 並列コードの生産性と性能の両立を目指した言語であり、本プロジェクトにて並列化したプログラムを公開することで、今後より複雑なモデルを搭載した他の分子系統解析プログラムの GPU 並列化も容易に行うことが可能となり、「大規模分子系統解析の高速並列化」という研究分野の発展に寄与できるものと考える。学際開拓型プロジェクトとしては、生物分野から高性能計算分野へは XcarableACC の適用および実践的な性能評価に用いるための実アプリケーションの提供を行うことで、双方の研究プロジェクトのさらなる進展を促進できた。

４．今後の展望

本プロジェクトは平成 29 年度も継続申請プロジェクトとして遂行する。平成 29 年度においても引き続き高性能計算分野との共同研究のもと、XcarableACC 言語の更なる拡張機能を利用したさらに効率的な GPU 並列化を実現する。特に、数十万座位からなるゲノム規模の遺伝子配列データに基づく分子系統解析では全座位における尤度計算に要求されるメモリ量が GPU の搭載メモリ（HA-PACS では GPU ごと 6GB）を大幅に凌駕する問題が生じる。そこで、部分的な座位データをまず GPU メモリに転送し、これらの座位における尤度計算を行う間に次の部分的な座位データを GPU メモリに新たに転送するという、「計算とデータ転送のオーバーラップ化」の機能を実装する。さらに、上記の尤度計算を単一 GPU で行わせるだけでなく、複数 GPU に配列データを分割し、異なる座位の尤度計算を複数の GPU（最大 32 ノード×4 台）で同時並列的に行わせることで更なる高速化を目指す。

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

Sohta A. Ishikawa, A multi-grained MPI/OpenMP parallelization of the maximum-likelihood phylogenetic inference with the non-homogeneous model, Mathematical and Computational Evolutionary Biology 2016, June 12~16, 2016, Hameau du l'etoile, Montpellier, France

石川奏太, 田渕晶大, 朴泰祐, 稲垣祐司, 佐藤三久, 橋本哲男, 大規模遺伝子配列データに基づく分子系統解析の GPU 並列化, 第 8 回「学際計算科学による新たな知の発見・統合・創出」シンポジウム–発展する計算科学と次世代の計算機–, ポスター, 2016 年 10 月 17 日～2016 年 10 月 18 日, 筑波大学, 茨城県つくば市

(3) その他

使用計算機	使用計算機に○	配分リソース※

HA-PACS	○	1350
HA-PACS/TCA	○	400
COMA		
※配分リソースについては 32node 換算時間をご記入ください。		