受付 ID	17a52
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、生命の進化機構解明に必要なデータ基盤が整いつつある。研究代表者はこれまで大規模遺伝子配列データに基づく進化的解析に携わる中で、特にこれらのデータ蓄積が近年著しい生物としてヒトや家畜に甚大な被害をもたらす寄生虫や病原性細菌、ウィルスなどに注目するようになった。これらの病原体の遺伝子配列解析が積極的に行われている背景には、①ある病原体の流行に際し個々の株が遺伝子レベルでどのように進化してきたのか、②遺伝子レベルの進化がどのような表現型の出現を招き、薬剤耐性など疫学上重要な性質がどのような経緯で獲得されてきたのか、という点を明らかにすることが公衆衛生学において最も重要な課題のひとつと考えられていることが挙げられる。これに対し、近年では “Phylodynamics” と呼ばれる手法による研究が盛んに行われている。Phylodynamics では遺伝子配列データおよびそれに基づき推測された分子系統樹と、個々の配列に紐付けられるメタデータ（サンプリングされた日時や場所、薬剤耐性に関する変異の有無などの表現型）という性質の異なる複合的なデータに基づき①②を説明する進化シナリオを明らかにすることを目的とする。しかしながら、特にウィルス由来データに基づく Phylodynamics 解析では、対象となる株数は 10,000 以上にのぼることが予想され、既存のプログラムおよび実験室レベルの計算機では解析に数日～数週間の時間を要する問題があった。また、そのような巨大系統樹上において推測された進化シナリオを分かりやすく可視化できる解析ツールはなかった。そこで、本研究ではこれらの方法論的・計算科学的問題を GPU 並列化など高性能計算技術の導入を含めたアプローチによって解決し、大規模遺伝子配列およびメタデータに基づく Phylodynamics 解析に有用なバイオインフォマティクスツールの開発を目的とした。

２．研究成果の内容

Phylodynamics では一般的に確率モデルと最尤法あるいはベイズ法に基づき分子系統

樹上での祖先形質復元を行うことで進化シナリオを推測する。そこで、本研究では高速・高精度な祖先形質復元法を新たに開発・実装し、大規模シミュレーションおよび実データ解析に基づく評価を行った。本手法では系統樹上の各ノード（＝各祖先）における形質状態について、それぞれの事後確率を計算し、確率の小さいものを除外する。また、除外の際はスコア関数の導入により残された祖先状態セットの尤もらしさを評価することで、系統樹全体で尤もらしい形質進化シナリオの推測が可能である。以上の計算は node-by-node の greedy algorithm によって行われるが、本アルゴリズムは系統樹の規模、すなわち種数（株数）N に対し N = 10,000 に近い巨大データセットの解析においても O(NlogN)に近いパフォーマンスを達成した（図 1）。

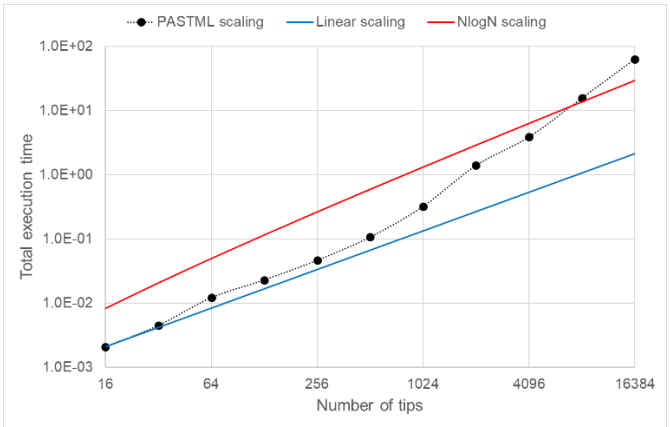

図 1：シミュレーションデータに基づく新たな祖先復元法の評価

以上の成果に基づき、マルチプラットフォームで大規模 Phylodymamics を行えるプログラム「PASTML」を開発した。また、復元された祖先形質に基づき各ノードで生じた主な形質進化イベントを集約し、巨大系統樹上でも容易に解釈可能な形質進化シナリオの可視化機能も新たに実装した。さらに、ヒト免疫不全ウイルスサブタイプ C（HIV-1C）3,036 株の Phylodynamics 解析に本プログラムを適用することで、HIV-1C の地理的拡散とそれに伴う薬剤耐性形質の獲得経緯を明らかにした（図 2）。

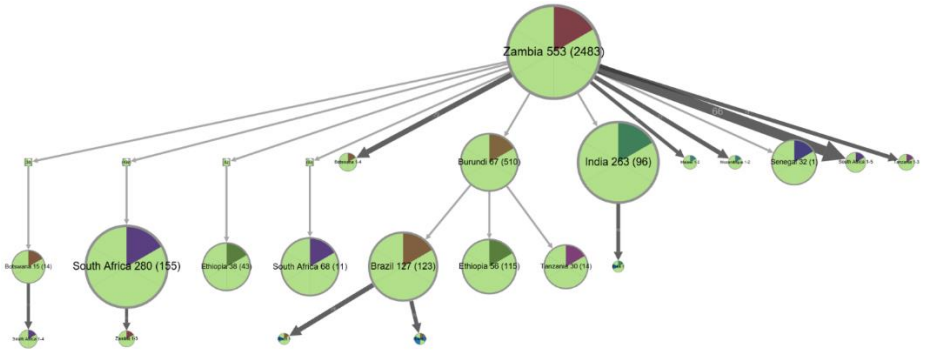

図 2：PASTML による HIV-1C 3,036 株の進化疫学解析

ブラジルおよびインドへの HIV-1C 株の流入後、特定の薬剤耐性変異（RT:D67N, RT:K70R, RT:M184V）の獲得と拡散が起こったことが示された。

３．今後の展望

本研究にて開発した PASTML を、HIV 以外の様々なウィルス由来大規模データセット（例：インフルエンザ、ジカウィルス）に基づく Phylodynamics 解析に実践し、これらの病原体の地理的拡散やそれに伴う病原形質の獲得経緯を明らかにすることで、公衆衛生のための網羅的知見の蓄積に貢献する。また、病原体のみならず、ヒト腸内や各種土壌・水質から得られた微生物群集メタゲノミクスデータに基づく大規模 Phylodynamics 解析にも応用することで、これらの特定環境に定着する微生物群が「共同体として」もつゲノム機能がどのように進化してきたかを明らかにする。また、PASTML プログラムは実行時間・スケーリングともに良好な性能を示した（図 1）が、将来的に数万種（株）以上からなる超巨大データセットの解析を想定し、マルチ GPU デバイス上での並列化など更なる高性能計算技術の導入にも引き続き取り組む予定である。

４．成果発表

(1) 学術論文

上記成果に基づく論文を投稿準備中。

(2) 学会発表

第 9 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－, 巨大ウィルス系統樹における祖先配列復元アルゴリズムの開発, ポスター, 2017 年 10 月 10 日~2017 年 10 月 11 日, つくば国際会議場, 茨城県

Mathematical and Computational Evolutionary Biology 2017, A probabilistic model-based prediction of virus character evolution on large phylogenies, Poster, June 12-16, 2017, Porquerolles Island, France

(3) その他

使用計算機	使用計算機に○	配分リソース※	
		当初配分	追加配分
HA-PACS/TCA	○	12800	
COMA			
Oakforest-PACS			
※配分リソースについてはノード時間積をご記入ください。			