PhyCAP
Phylogenomic-noise Cleaning Approach by PCA

■ はじめに

分子系統解析は DNA やタンパク質の配列データを比較し，その間で生じた塩基やアミノ酸の置換を解析することで，それらの配列の進化過程を推定する．これにより得られる遺伝子やゲノム領域の系統関係は生物進化を理解するうえで欠かすことのできない重要な情報である．

分子系統解析法の多くでは，塩基やアミノ酸の置換パターンを数理モデル化し，その置換モデルに基づいて推定を行う．分子系統解析の精度が低下してしまう主要な原因には，確率誤差（stochastic error）と系統誤差（systematic error）がある．確率誤差とは，配列の長さが有限であるため，偶然によって結果がばらつくことで，系統誤差とは，置換モデルに適合しないデータが含まれることによって，推定を誤ることである．

従来，分子系統解析では少数の遺伝子配列から推定を行っていたため，確率誤差が重要な問題であった．しかし近年，シークエンシング技術の飛躍的発展によりゲノム規模の配列データが比較的容易に取得できるようになり，多数の遺伝子やゲノム領域を用いた解析が可能となった．これにより確率誤差の問題は激減した一方，大規模データ中に置換モデルに適合しない領域が一定以上含まれる問題が生じた．その結果，分子系統解析における主要な課題は確率誤差から系統誤差に移り変わった．

■ 活動内容

1．既存の方法

系統誤差の問題を解決するため，これまで置換モデルの改良や置換モデルに適合しないデータの除外など，様々なアプローチが考案されてきた．特に不適合なデータを除外するアプローチは，必要量以上のデータを確保しやすいゲノム規模解析において有効性が高い．しかしながら，これまでの方法では，①効果がある場合が限られている，②配列データ以外に必要な情報がある，③計算時間が長大であるなど，多くの課題が残されている．そこでこれらの問題を解消する新規の方法理論を研究し，PhyCAP（Phylogenomic Cleaning Approach based on Principal component analysis）を開発した．

2．PhyCAP

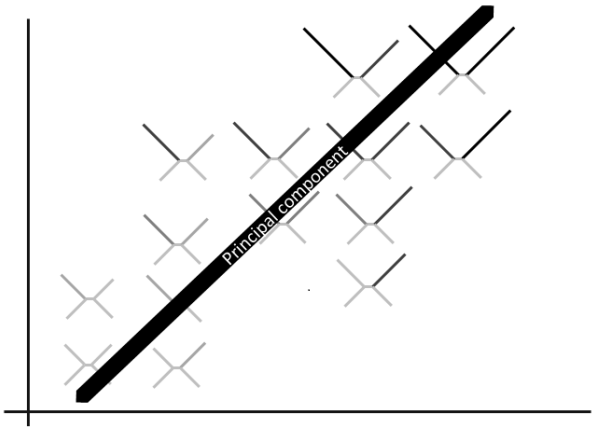
PhyCAP では，系統誤差の要因を一種類に限定せず，系統樹推定の最終結果である枝長から系統誤差の影響を大きく受けた領域を発見する．このため，特定の要因に制限されず広範な要因に対して有効であるとともに，特定の要因の影響を評価するための情報も必要としない．具体的には，遺伝子系統樹の枝長に基づいて主成分分析（Principal component analysis）を行うことで，枝長の違いに基づいて遺伝子を並び替え（図），確率誤差が増大しすぎない範囲で，系統解析に不適切な領域を除外する．

2．性能評価

コンピューターシミュレーションによって複数の条件下で生成されたデータと，ハチ，鳥類，脊椎動物の配列データを用いて PhyCAP の効果を既存の他の方法と比較した結果，ほとんどの場合において PhyCAP は他の方法より高い効率で系統解析の精度を改善した．また計算時間についても，PhyCAP は他の方法の1%程度しか必要としなかった．

■ 関連情報等（特許関係、施設）

なし



図．遺伝子系統樹の枝長に対する主成分分析．枝の色は置換パターンを表す．

代表発表者　**岩本　栄介（いわもと　えいすけ）**
所　　属　**生体システムビッグデータ解析オープン・イノベーションラボラトリ，産業技術総合研究所**
問合せ先　**〒169-8555　新宿区大久保 3-4-1　早稲田大学　西早稲田キャンパス 63 号館 5-20**
TEL：050-6867-1806
cbbdoil.eisuke-iwamoto@aist.go.jp

■キーワード：（1）ゲノム規模系統解析
（2）システマティック エラー
■共同研究者：田村　浩一郎
所属　首都大学東京 理学研究科 生命科学専攻