分子系統と分子進化

予測発見戦略研究センター・ゲノム解析グループ
モデリング研究系・グラフ構造モデリンググループ

准教授　　足立　淳

生物の進化の道筋を遺伝情報から解明する分子系統樹の推定法として、最尤法はその有効性を高く評価されてきた。しかしゲノム時代が到来し、大量の遺伝子を同時に扱うことが求められるようになり新たな局面を迎えている。最尤法のこれまでのやり方では少数の遺伝子を解析するのには有効であったが、大量の遺伝子を対象にする場合には様々な障害が浮かび上がってくる。ゲノム上の個々の遺伝子は互いに独立に進化しており、塩基やアミノ酸の置換傾向やその速度は異なることが多い。ゲノム時代の分子系統学では、大量の遺伝情報を同時に扱うための新たな戦略が求められている。

1　ゲノム時代の最尤系統樹推定とそのソフトウェア

複数の遺伝子を対象とした最尤法のソフトウェアは、まだ決定的なものはなく、解析の状況に応じて使い分けられているのが現状である。例えば、系統樹間の統計的な評価や複数遺伝子の静的な統合解析では筆者らが開発したMOLPHY、座位間の置換速度の不均一性を取り入れた解析ではPAML、大規模な種数の系統樹探索にはMrBayes等が使われる。また、GUIを取り入れた使いやすさから PAUP*も良く利用されている。また新興勢力として、より高速なRAxMLやGARLIも台頭してきている。

これらの状況を打破すべく、新しいMOLPHYの開発を行っている。特徴としては、第１にGUI等を取り入れた使いやすさであり、第２に最新の進化モデルの実装であり、第３として最尤系統樹の効率的な探索法の開発である。MOLPHYは .NET framework上で開発し、最尤推定などの数値計算モジュールはC++、データの入力や結果の出力、樹形の操作などのインターフェイス部分はC#でコーディングを行っている。これにより、Linuxではコマンドライン上で各種のプログラムを組み合わせることにより効率よく解析でき、またWindows上ではインタラクティブに樹形を操作しながら快適に解析ができる。また、Windows をクライアントとして Linux の計算サーバーにジョブを分散処理できるようにもなる。

処理速度の向上の面では、並列化に対応することも重要である。一般的なPCのCPUがマルチコア化されて以来、一台のPCに搭載されるコア数は純増する傾向にある。系統樹推定法は互いに独立した数値計算部分が多く並列処理に向きマルチコアとの親和性が高いので、積極的にマルチコアへの対応を行っている。また、系統樹の尤度の計算には倍精度浮動小数点の行列演算が多用されるので、GPGPUなどの積極的な利用もコストパフォーマンスの向上に有効である。

2　進化モデルの改良と最尤系統樹探索

進化モデルの向上を目指す点では、座位間の進化の不均一性を考慮したモデルとして、離散ガンマモデルを実装した。置換モデルでは、これまで塩基置換モデルとアミノ酸置換モデルが主に使われてきたが、これらは互いに独立ではないため解析の統合が難しかった。そこで両者の利点を組み合わせたコドンモデルの開発を行った。アミノ酸置換とアミノ酸を変えない塩基の置換（同義置換）を組み合わせたコドン置換モデルを複数実装し、どのコドンモデルが有効かを検証した。さらに、系統によって進化速度や置換様式が変化することが実在することが分かってきたので、それらをコバリオンモデルとしてモデルの構築とプログラムへ実装を検討した。

系統樹の探索法としてこれまで実用化されてきた方法は、大きく３つのグループに分けられる。第１のグループは、一つの根から全ての種が同時に分かれたと仮定したスター型を初期木とし、

その中から最近隣のペアを見つけて独立を繰り返し、最終的に二分木を得る方法である。第２は、３種からなる木に、残りの種を順次最適となる位置に分岐させて最終的な二分木を得る方法である。第３は、何らかの簡便法により二分木を推定し、それを初期木としてトポロジカルに近傍な系統樹を走査して、より最適な二分木を発見していくというヒューリスティックな方法である。これは現実的な方法ではあるが最適解を保障することは困難となる。上記の３方法は、どれも初期木の仮定やその探索方法から局所最適解に陥ることがあり、新たな探索方法が求められていた。

そこで第１グループの欠点を改良した、新しい方法（star decomposition）の開発を行ってきた。この方法の利点は、スター型の初期木の二分化において、最近隣の２種を見つけるのではなく、全種をある２グループ（部分系統樹）に分けた時に最適となる２グループを見つける点にある。種分化において短時間に続けて分岐をした種群があった場合、２種の最近隣を見つける方法では局所解落ち込む可能性があるが、新しい２グループに分ける方法では、尤度比の改善が高い２グループから分けて行くので、尤度比の改善が低い分岐を決定するのが後回しになり、結果として局所解に落ち込む可能性が低くなる傾向がある。

2　ゲノム系統学

近年のゲノムプロジェクトの急速な進行とともに、全ゲノム規模のデータを用いた系統樹推定が行われるようになった。配列データ量の増加は系統解析に有用ではあるが、もし系統樹推定の際に仮定する進化モデルに偏りがあった場合、誤った結論を導いてしまうことがある。遺伝子配列データが膨大であっても全配列に対して同一の進化モデルを仮定してしまうと誤った結論を導くことがあり、それを避けるためには進化速度・進化パターンが遺伝子ごとに異なることを仮定したモデルを適用すべきである。

ゲノム上の個々の遺伝子は互いに独立に進化しており、塩基やアミノ酸の置換傾向やその速度は異なることが多い。遺伝子ごとにモデルやパラメータの独立を仮定しなければならず、推定しなければならないパラメータの総数は膨大となる。そのため実用的な解析を行うためには、進化の傾向が近い遺伝子同士を分類し統合することが求められる。また、パラメータを共有できる遺伝子群の数を、計算可能な数に収める必要がある。しかし、それができたとしても遺伝子群の数が多くなるほど最尤系統樹の探索もより困難なものとなる。

3　ゲノム構造の進化

現在のゲノム系統学は、共通祖先のある遺伝子から由来した相同な遺伝子のみを比較するに留まっている。それはゲノム情報の一部に過ぎない遺伝子間の点突然変異の違いを比較していることになる。この情報からは、進化の歴史的な流れを表す系統樹を推定するには十分であるが、進化そのものを解明するためには不十分である。系統樹推定は進化の解明の第一歩に過ぎず、進化のメカニズムを知るためにはゲノム構造の変化を追わなければならない。

現実の進化は染色体の構成を変化させるほど、ゲノム構造に様々な変異を起こし、新たな遺伝子を生み出したり、既にある遺伝子の機能を変えたりもする。そういった進化の原動力になっている変異の歴史を追うためには、共通祖先のゲノム構造を推定し復元する必要がある。もし復元が可能となれば、どのような突然変異がどんな機能の獲得や喪失に関わったかを説明できる情報を手にすることになる。

現生生物のゲノム構造を比較し、共通祖先のゲノム構造を推定し復元することができれば、過去に起きた突然変異の歴史を遡ることができる。そうすれば、ある遺伝子やその機能の獲得が、どのような突然変異に由来したのかが推定でき、どのような突然変異が進化の原動力となってきたかが分かるであろう。