Part 2 赤池統計学の展開

生物多様性の理解をめざして

長谷川政美
復旦大学生命科学学院教授

近年、大量のDNA塩基配列データの蓄積に伴って、分子系統学が盛んになってきた。分子系統解析におけるモデルの重要性と、そこで赤池情報量規準が果たす役割について紹介する。

DNA配列から系統樹をつくる

地球上には数千万種ともいわれる多様な生物が生息している。このような生物の多様性を理解するためには、進化的な視点が不可欠である。地球上のあらゆる生物は、1つの共通祖先から種分化を繰り返しながら進化してきたものであり、生物多様性の起源は、系統樹という形ではじめてとらえられる。

生物進化の歴史は、現在生きている生物のゲノムのなかに刻まれており、ゲノムDNAの配列を解析することによって系統樹を推定することができる。これが分子系統学である。以前は形態の比較による系統学が主流であったが、形態レベルでは似た環境で似たような形質が独立に進化するといった収斂（しゅうれん）進化がたびたび起こり（図1）、形態の比較だけでは間違った系統樹が得られる危険性が高いことがしだいに明らかになってきた。

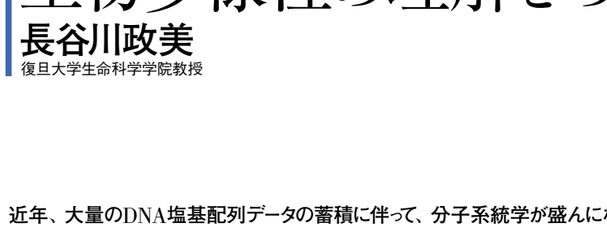
進化におけるDNA塩基やタンパク質アミノ酸の置換は、確率過程とみなすことができる。分子レベルでの変異の出発点は、まず個体の生殖細胞中のDNA上で突然変異が生じることであり、これはその名前が示すように確率的な現象である。

しかしながら、突然変異は個体レベルの現象であり、一方、進化とは生物種が集団として世代を超えて変化していくことである。個体レベルで起こった突然変異が子孫に受け継がれ、その遺伝子をもった子孫が増えることによって、突然変異遺伝子が集団全体に広がることが必要である。これを突然変異遺伝子の集団への固定というが、ここでも偶然的な要素が重要であることが明らかになってきた。

分子レベルでの進化的な変化の多くは、自然選択に必ずしも有利なものではなく、良くも悪くもない中立的な変異のなかで運のよいものがたまたま選ばれるというものである。これが木村資生（もとお）の分子進化の中立説（1968年）である。したがって、そのような進化の結果として生成された現生生物の分子配列データから進化の歴史を再構築することは、統計的推測の問題になる。

複数の生物種から得られた塩基配列（あるいはアミノ酸配列）を縦に並べると、挿入や欠失があって対応する座位（塩基配列などの位置）がうまくそろわないことがある。そのような場合に挿入や欠失に手を加えて配列をそろえることを「アラインメント」という。アラインメント上で同じ座位の塩基が生物種によって異なる場合には、共通祖先から進化する間に蓄積した変異とみなすことができる。こうして得られたデータセットを、「データ行列」（図2）という。

ハリネズミ



ハリテンレック

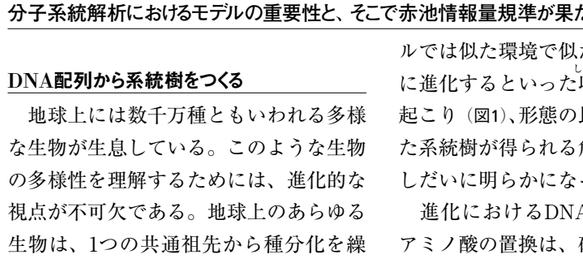


図1　収斂進化の例。ハリネズミとマダガスカル固有のハリテンレック。

最節約法から最尤法へ：分子系統樹推定

このデータ行列から系統樹を推定するわけであるが、統計的推測の最も自然な枠組みは、最尤法である。現在広く用いられるようになってきた最尤法による分子系統樹推定法は、フェルゼンシュタイン（Joseph Felsenstein）によってはじめて定式化された（1981）ものである。最尤法では、まず進化過程で起こる塩基（あるいはアミノ酸）置換の法則性を確率モデルとしてとらえて、そのようなモデルにもとづいた進化の結果として当該のデータ行列が実現する確率を計算する。これが尤度である。それを可能な系統樹のトポロジー（枝分かれの順番）について計算し、尤度が最大になるトポロジーを真のトポロジーの最有力候補として選ぶのである。

以前から広く使われていた分子系統樹推定法に、「最節約法」がある。これはデータ行列を説明するために置換数がなるべく少なくてすむようなトポロジーを選ぶという方法である。この方法はわかりやすく、最尤法にくらべると計算も簡単なので、現在でも広く使われているが、いろいろな欠点もある。

とくに、進化速度が系統によって異なる場合には、間違った系統樹が選ばれる危険性が高い。最尤法では通常、進化速度は系統によって違っていても構わないとして（分子時計を仮定しない）解析が行われる。最節約法でも分子時計は仮定されないが、必要最小限の置換しか考えないので、枝が長くなると短い枝に比べて相対的に多重置換（同じ座位に繰り返し置換が起こること）の効果が過小に評価されることになる。そのために最節約法では、長い枝同士が間違って組んでしまう傾向が強いのである。

一方、最尤法でも置換モデルが単純だと多重置換が過小に評価されて、同じように長い枝同士が組んでしまう傾向があるが、モデルの改善によりそのような推定の偏りを回避できる。

1980年代初頭に最尤法による系統樹推定法が定式化されたが、長い間この方法は実際の分子系統樹解析にはあまり使われてこなかった。その理由としては、計算に膨大な時間がかかるため、一般の研究者が使うことのできたコンピューターでは、実際のデータをなかなか解析できなかったということがある。私は幸い、統数研で大型計算機をふんだんに使うことのできる環境にいたために、生物学の実際問題にはじめて最尤法を適用することができた。岸野洋久さんと共同で開発した最尤系統樹の検定法は、現在では広く使われている。

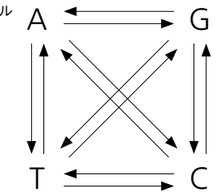
最節約法を使っていた研究者からは、次のような批判があった。「最尤法で使われている置換モデルはおよそ現実からかけ離れた単純なものだから、結果は信頼できない。それに対し、最節約法はいかなるモデルも仮定しない方法だから、その点で最尤法よりもすぐれている」。たしかにこの批判の前半は当たっている面があるが、後半は間違いである。

いかなる推定法も何らかのモデルの上に成り立っている。モデルを明示的に仮定しているか、暗示的かの違いだけである。最節約法は明示的にはモデルを仮定しないが、何らかの仮定の上で成り立っているはずであり、その仮定が最尤法のようにははっきりとしていないだけである。仮定がはっきりしている場合、それが間違っていることが明らかになれば改めていく余地があるわけで、科学的なデータ解析法としてはそちらの方がすぐれているといえるだろう。

最尤法は当初、系統学の研究者からはあまり評価されなかったが、近年のコンピューター性能の飛躍的な進歩と実用的

確率過程
- 突然変異
- 変異遺伝子の集団への固定

塩基置換の確率モデル



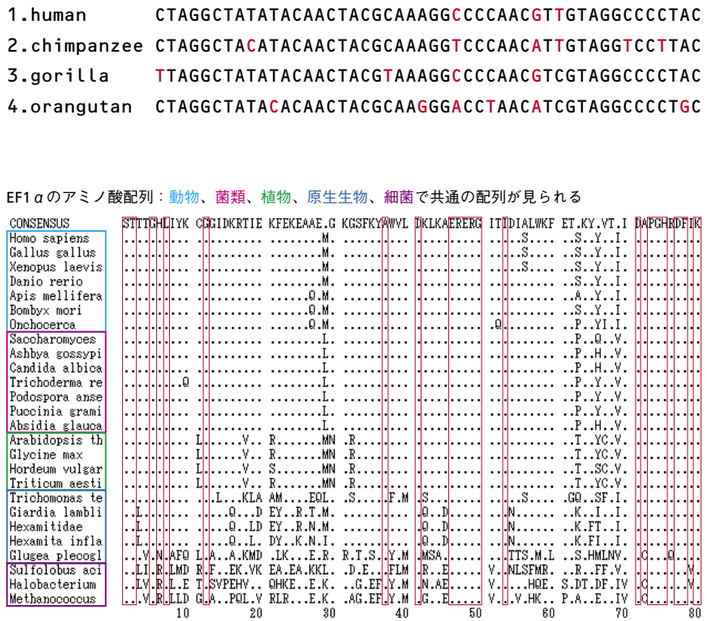
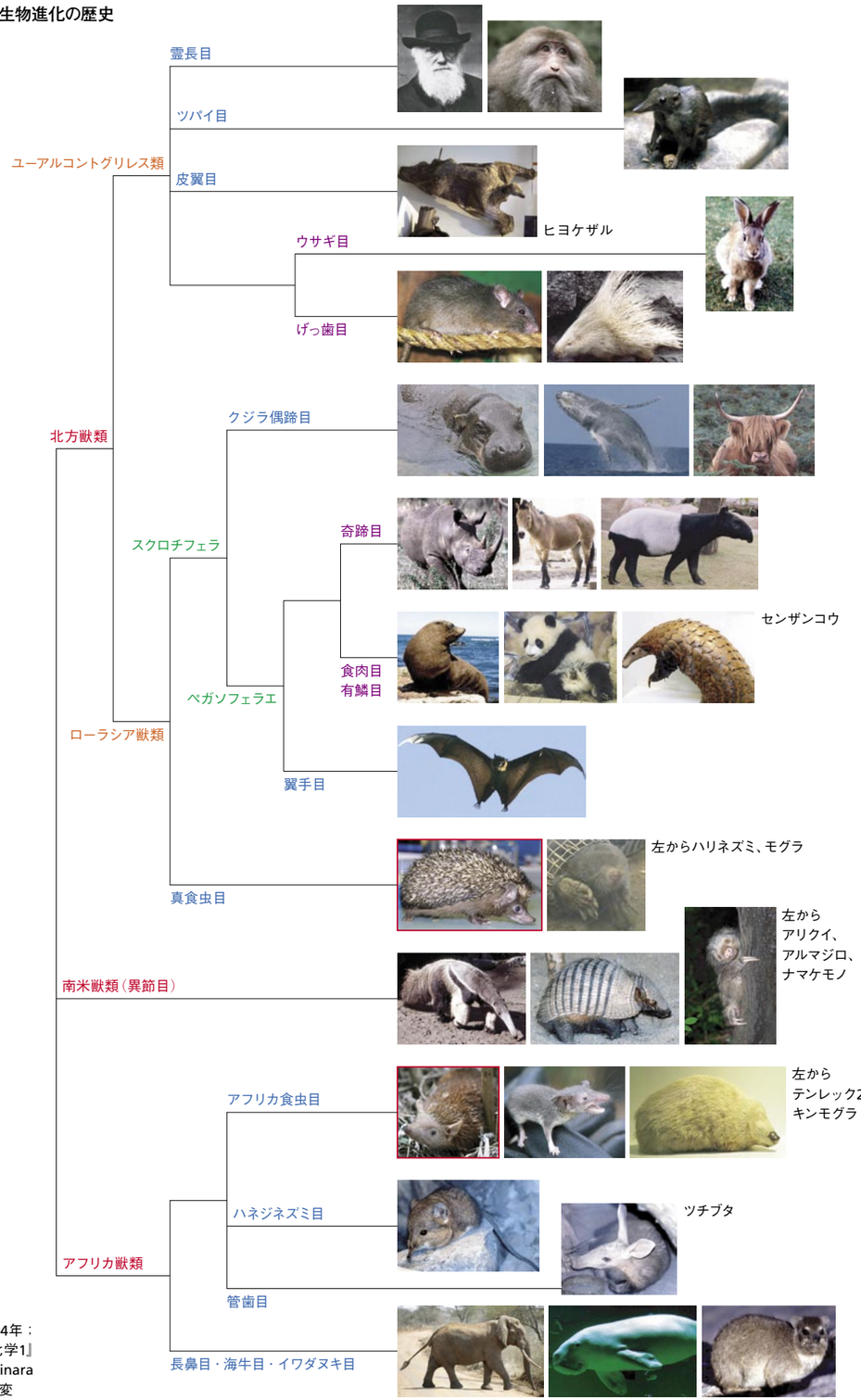


図2　**アラインメントされたデータ行列の例**
動物から細菌に至るまでの多様な生物の間でEF1αのアミノ酸配列に共通性が見られるということは、地球上のあらゆる生物が1つの共通祖先から進化してきたことを示している。したがって、あらゆる生物は1本の巨大な系統樹のどこかに位置づけられるはずである。

図3　分子系統でたどる生物進化の歴史



出典：長谷川政美、2004年：岩波書店『シリーズ進化学1』pp-51-91、図14を、Nishinara *et al.*（2006）をもとに改変

なソフトウエアの開発もあって、分子系統学の分野でしだいに認められるようになってきた。その結果、計算時間の制限から非現実的な極めて簡単なモデルしか使えなかったのが、現実に即したモデルに基づいた解析を可能にしつつある。そこで重要になってきたのが、解析に際してどのような置換モデルを採用するのがよいか、というモデル選択の問題である。

モデル選択と赤池情報量規準

分子系統解析にあたって仮定する置換モデルは、なるべく現実の進化過程に合うものが望ましい。しかし、限られたデータを解析するのに、むやみに複雑なモデルを使うのは問題である。情報の少ないデータに対して複雑なモデルに含まれる多くのパラメーターを適合させようとすると、過適合（over-fitting）が起こる。赤池さんは情報理論的な考察から、

AIC＝−2（最大対数尤度）＋2（モデルのパラメーター数）

で定義される赤池情報量規準（AIC：Akaike Information Criterion）が最小になるようなモデルが、当該のデータを表現するのに最もふさわしいモデルであることを示した。

モデルが複雑になればデータとの当てはまりが良くなるので、最大対数尤度が大きくなってマイナス符号のついた第1項は小さくなるが、逆に第2項は大きくなる。つまり、第2項はモデルを複雑にしたことに対するペナルティを表している。モデルを複雑にしてパラメーターを増やしたことに見合うだけのデータとの当てはまりの改善が見られなければ、なるべく簡単なモデルにとどめておくべきということである。こうして、モデルを改善していく際の客観的な規準が得られたことになる。

赤池さんがAICに関する論文を最初に書いたのは1973年だったので、分子系統学でモデル選択が問題になりはじめた1990年ごろには、AICは統計学の世界ではすでに確立した方法になっていた。そのころ、赤池さんは統数研の所長をしており、その下で研究していた私は、AICを分子系統学の世界に導入すべき立場にあった。

分子系統学のモデル選択にAICを最初に導入したのはわれわれであったが、この分野で広く使われるようになったのは、1998年にポサダ（David Posada）とクランダール（Keith A. Crandall）が「MODELTEST」というプログラムを公開してからのことである。これにはさまざまな塩基置換モデルが実装されていて、ユーザーはAICを使ってそのなかから自分の扱っているデータに最も適合したモデルを選択し、それを用いてさまざまなプログラムで分子系統解析ができるようになった。

いまMODELTESTは多くの研究者に使われるようになり、これを使っていないと論文の査読者から忠告を受けるほどである。それに伴って、モデル選択とAICの重要性が広く認識されるようになってきたことは喜ばしいことであるが、新たな問題も浮かびあがっている。それは、MODELTESTに実装されているモデルがいずれも塩基置換のモデルであり、タンパク質遺伝子の進化を近似するには現実から離れ過ぎているということである。

たとえば、タンパク質をコードしている遺伝子は3連塩基コドンを単位として構成されており、コドン内のそれぞれの塩基の置換は決して独立ではない。ところが、MODELTESTに実装されているモデルは、いずれも独立性を仮定している。アミノ酸に対応したコドンは61種あるので、本来は61×61の遷移行列を扱うコドン置換モデルを用いることが望ましい。非現実的なモデルのセットのなかから最良のものを選び出しても、あまり意味はない。今後、モデルに取り入れていかなければならないことは多い。

分子系統解析は現在、多くの生物群について行われている。とくに研究が進んでいる真獣類（有胎盤哺乳類）の系統進化で、最近明らかになってきたことを図3に示す。分子系統解析により、図1で示した非常に良く似た動物が、まったく異なった由来をもった収斂進化の結果であることがわかる。

より現実に近づくために

上で述べたMODELTESTは、多くのモデルをAICで比較したうえで、データに最も適合したモデルを用いて系統樹推定ができる。だから、ユーザーの多くはこれでよいのだという自己満足に陥る傾向がある。しかしここで問題なのは、用意されているモデルはいずれも現実から離れた未熟なものだということである。

モデルはあくまでも現実の過程を近似するものにすぎない。だから、限られたデータをうまく近似したモデルであっても、データ量が増えてくると現実とのずれがしだいに目立つようになってくる。

したがって、常に最新の知見を取り入れてモデルをより現実に即したものに改善する努力を続けていくことが必要である。その際の道標として、AICの重要性は今後も変わることはないだろう。AICの長所は自由なモデル構築と、それらのモデルを客観的な規準で比較することを可能にしたことにあるのだから。

長谷川政美（はせがわ・まさみ）
2007年3月に総合研究大学院大学先導科学研究科生命体科学専攻／統計数理研究所教授を定年退職し、現職。さまざまな生物の系統進化の研究を行っていますが、とくに哺乳類の進化やマダガスカルの自然史に興味をもっています（写真はチベット調査中の著者）。