

配列から遺伝子の進化を探る

斎藤成也

はじめに

遺伝子の進化を扱う「分子進化学」では、塩基配列とアミノ酸配列データが解析の大部分を占める。塩基配列の場合には、DNAあるいはRNAの4文字から構成される配列であるのに対して、蛋白質のアミノ酸配列は通常20種類のアミノ酸の文字列である。しかしこれら2つとも文字列であることは共通しているため、本稿ではどちらにもあてはまる「配列」という言葉を用いることにする。

生物進化は時間軸に沿った現象なので、化石と現生生物を比較するように、理想的には過去の情報と現在の情報を直接比較すべきである。分子進化の研究でも、PCR法の発明によって、生物遺体からDNAを抽出して塩基配列を決定する古代DNAの研究¹⁾や、残存している蛋白質を抗体で検出するという研究も存在する。関与する進化時間は短いものの、HIVやインフルエンザウイルスなど、分子進化速度のきわめて速い場合にも、異なった時期に得られた配列を比較することができるので、広い意味では古代DNAに入るだろう。しかし大部分の配列データは、現生生物からとられたものである。

塩基配列やアミノ酸配列のデータから系統関係を復元するには、これらの配列をまず多重整列(multiple alignment)することが必要である。多重整列の方法については、本シリーズ第3回『配列に共通のパターンを探す』(7月号)を参照されたい。さて、とにかく多重整列はうまくいったと仮定しよう。すると次のステップは系統樹作成法の種類によって異なる。本稿で取り上げる近隣結合法のような距離行列法(distance matrix method)を用いる場合は、進化距離を計算し、距離行列を得る必要がある。距離行列法は、文字どおり進化距離などの「距離」を、比較するすべてのOTU(operational taxonomic unit; 操作上の分類単位)のペアで推定した「距離行列」

を用いるものである。距離行列法には、近隣結合法のほかにも、UPGMA、最小進化法など多数の方法がある。距離行列法とならぶもう1つの系統樹作成法のグループは、形質状態法(character state method)である。形質状態とは、文字どおり塩基やアミノ酸などの形質の状態(DNAの塩基であれば、A、C、G、Tのどれか)のことで、それらの多重整列結果がデータとなる。形質状態法には、最大節約法や最尤法がある。これらの系統樹作成法について詳しく知りたい方は、文献2と3をご覧ください。また、やさしい説明がApplied Biosystemsのホームページ(<http://www.appliedbiosystems.co.jp/>)に「遺伝子系統樹入門講座」として掲載されている。

数学的にはデータに対して最も妥当性の高い系統樹が得られたとしても、調べたその遺伝子本来の進化がそのとおりであったかどうかはわからない。これは、とくに組換えや遺伝子変換によってもはや系統樹という進化モデルが成り立たない場合には、最も深刻な問題点である⁴⁾。しかし、これについては誌面の関係上、割愛する。

I. 進化距離とは?

進化距離には、塩基配列間の進化距離、蛋白質間の免疫学的距離、集団間の遺伝距離などさまざまな種類があるが、本稿では塩基配列やアミノ酸配列についての進化距離についてだけ簡単に述べる。配列間の進化距離を推定するには、整列されている2本の配列間にギャップがある場合には、それらを取り除く。ときどき、これらのギャップの1つ1つを、塩基の違いと混ぜて進化距離を計算することがあるようだが、ギャップの生じる進化速度は塩基置換やアミノ酸置換の速度よりもかなり遅いことが知られている^{5,6)}。したがって、進化速度の大きく異なる両者を混ぜるのはよい方法ではない。

Saito Naruya, 国立遺伝学研究所進化遺伝研究部門 E-mail: nsaitou@genes.nig.ac.jp <http://sayer.lab.nig.ac.jp/~saitou/>
In search of gene evolution from sequences

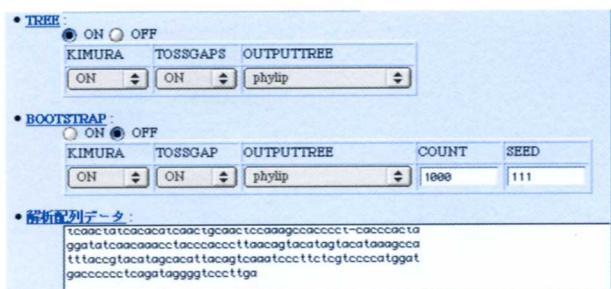


図1 DDBJのWWW ClustalW システムのページの一部

この考え方のもとに、DDBJのホームページ(<http://www.ddbj.nig.ac.jp>)で提供しているClustalW⁷⁾では、ギャップをすべて取り除く(TOSSGAP=on)オプションをデフォルトにしている(図1)。

通常の進化距離とは、これらギャップを除いた2本の配列が共通祖先から進化してきた間に蓄積した塩基置換数(あるいはアミノ酸置換数)を部位あたりで表現したものである。部位あたりにするのは、進化距離を推定するための理論的なモデルとして、通常はどの塩基部位、アミノ酸部位も同じ速度で変化すると仮定しているの、比較した部位数で割って基準化している。ClustalWでは、進化距離を配列の異なる割合(p)で表わす場合(KIMURA=off)と、木村の式(塩基配列の場合には文献8, アミノ酸配列の場合には文献9)を用いて置換数を推定する場合(KIMURA=on)があるが、DDBJのWWWで提供しているClustalWでは、後者をデフォルトとしている(図1)。

塩基配列の場合、転位型の塩基の異なる割合を P 、転換型の塩基の異なる割合を Q とすると、Kimuraの2変数法⁸⁾では、進化距離 d は、 $d = -(1/2)\log[1-2P-Q] - (1/4)\log[1-2Q]$ で推定される。ほかにも多数のモデルが提唱されているが、ClustalWで用いられているのはこの2つだけである。そこで、DDBJでは他の推定法、たとえばJukes-Cantorの1変数法、Tamura-Neiの方法なども計算できるような変更を行ない、近くWWW版で公開する予定である。

一方、アミノ酸配列の場合には、アミノ酸の異なる割合を p とすると、アミノ酸の置換数(進化距離)のKimura⁹⁾による推定値 d は、 $d = -\log[1-p-0.5p^2]$ で与えられる。ここで \log はどちらも自然対数である。

数学でいう「行列」(matrix)とは、一般には n 行と m 列の数字(行列の要素)の集まりだが、系統樹の作成に用

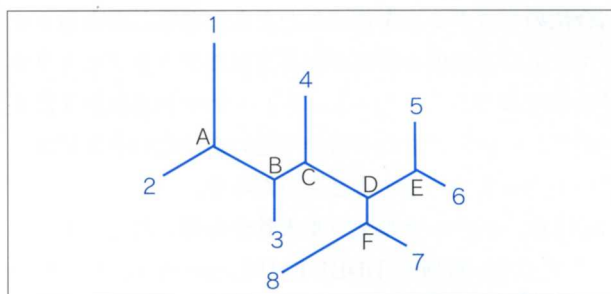


図2 無根系統樹の例¹⁰⁾

いる「距離行列」は、進化距離(evolutionary distance)が要素である特殊な行列である。配列 i と配列 j の間の進化距離を $D(i, j)$ と表わすことにしよう。自分自身との距離はゼロなので、行列の対角成分である $D(i, i)$ はすべてゼロである。また、 $D(i, j) = D(j, i)$ なので、距離行列は対称行列である。このような性質があるために、進化距離行列は、行列の下または上三角部分だけを表示する場合が多い。

無根系統樹において、1つの結節だけでつながった2つのOTUのことを「近隣」(neighbors)とよぶ。図2で、OTU1とOTU2は、1つの結節Aだけでつながっているので、近隣である。これらを合体すると、この合体OTU(1-2)はOTU3と近隣になる。このように、近隣をつぎつぎに合体することによって樹形を規定することができる。一般に、 n 個のOTUからなる二分岐系統樹は、 $n-2$ 個の近隣でその樹形を規定することができる。

近隣結合法(Neighbor-Joining method)¹⁰⁾は、段階的にこれらの近隣を見いだしていく方法であるが、詳しいアルゴリズムについては、文献11, 12を参照されたい。

II. DDBJ WWWのClustalWで近隣結合系統樹を作成する

DDBJ WWWのClustalWでは、多重配列から系統樹作成まで一連の計算をすることができる。図1でTREEオプションがonになっていれば、指定された方法で進化距離を計算したあとに自動的に近隣結合系統樹が生成される。ただし、生成されるのはデフォルトのOUTPUTTREEオプション(phytip)の場合、Newickフォーマット形式である。それを別のソフトウェアのインプットファイルとして与えて初めて系統樹が図示される。OUTPUTTREEオプションでclustalを選ぶと、筆者が近隣

結合法のプログラムを作成したときに考案した出力フォーマットで系統樹の関係が記述されたアウトプットファイルが生成される。Newick フォーマットはわかりにくいので、系統樹の構造を自分で追いかけてみたいときには、この出力結果を利用するとよいだろう。

図3は、ネアンデルタール人骨から得られたミトコンドリア DNA 配列¹³⁾ (DDBJ/EMBL/GenBank アクセス番号: AF011222) をクエリとして DDBJ WWW の BLAST で相同性検索を行ない、得られた相同配列 100 本を DDBJ WWW の ClustalW で多重整列し、近隣結合系統樹を DendroMaker¹⁴⁾ で表示したものである。根は「最長経路の中点」に設定するオプションを用いた。また、クラスター整理(中央寄せ)オプションを用いてみた。DendroMaker はこのようにいろいろなオプションがあり、使いやすいソフトである。ただ、現在のところブートストラップ確率を表示することはできず、query.ph ファイルしか用いることができない。

そこで、図4は、DDBJ WWW の Clustal W 画面で BOOTSTRAP オプション(図1参照)を ON にして、ブートストラップ確率を含むアウトプットファイル query.phb を生成し、それを TreeView (<http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>) で描いたものである。遺伝子系統樹では、各枝の長さの情報が重要なので、筆者は常に phylogram オプションを用いている。しかし、図3と比べると、同一の樹型であるにもかかわらず、かなり印象が異なっている。また TreeView では 3 分岐パターンしか描くことができず、DendroMaker に比べると不自然な形の系統樹となる。ブートストラップ確率をみると、ここでは 1,000 個のブートストラップ系統樹の結果であるが、ネアンデルタールの 3 個の塩基配列のクラスターは、確率 100% で現代人のクラスターと結合していることがわかる。一部の枝にブートストラップ確率の値が示されていないのは、ClustalW または TreeView のバグのようである。なお、同一の塩基配列(たとえば図4の query 配列と AF011222) がクラスターした場合にはこのクラスターは 1 つの配列と考えるべきなので、これらに対して query.phb で出力されていたブートストラップ確率は意味がないため、すべて削除してある。

このように、よく用いられているソフトでもそれぞれ問題点があるので、それらについて注意を払いつつ使用するべきである。

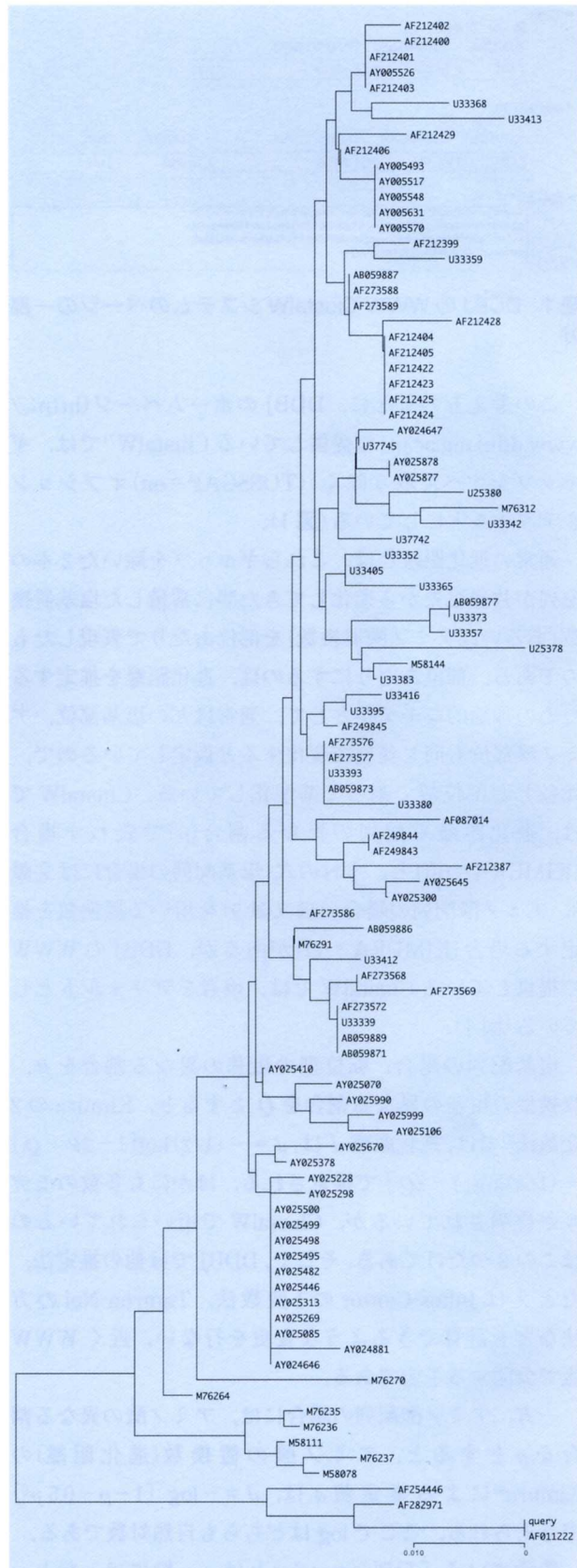


図3 DendroMaker で描いた近隣結合系統樹

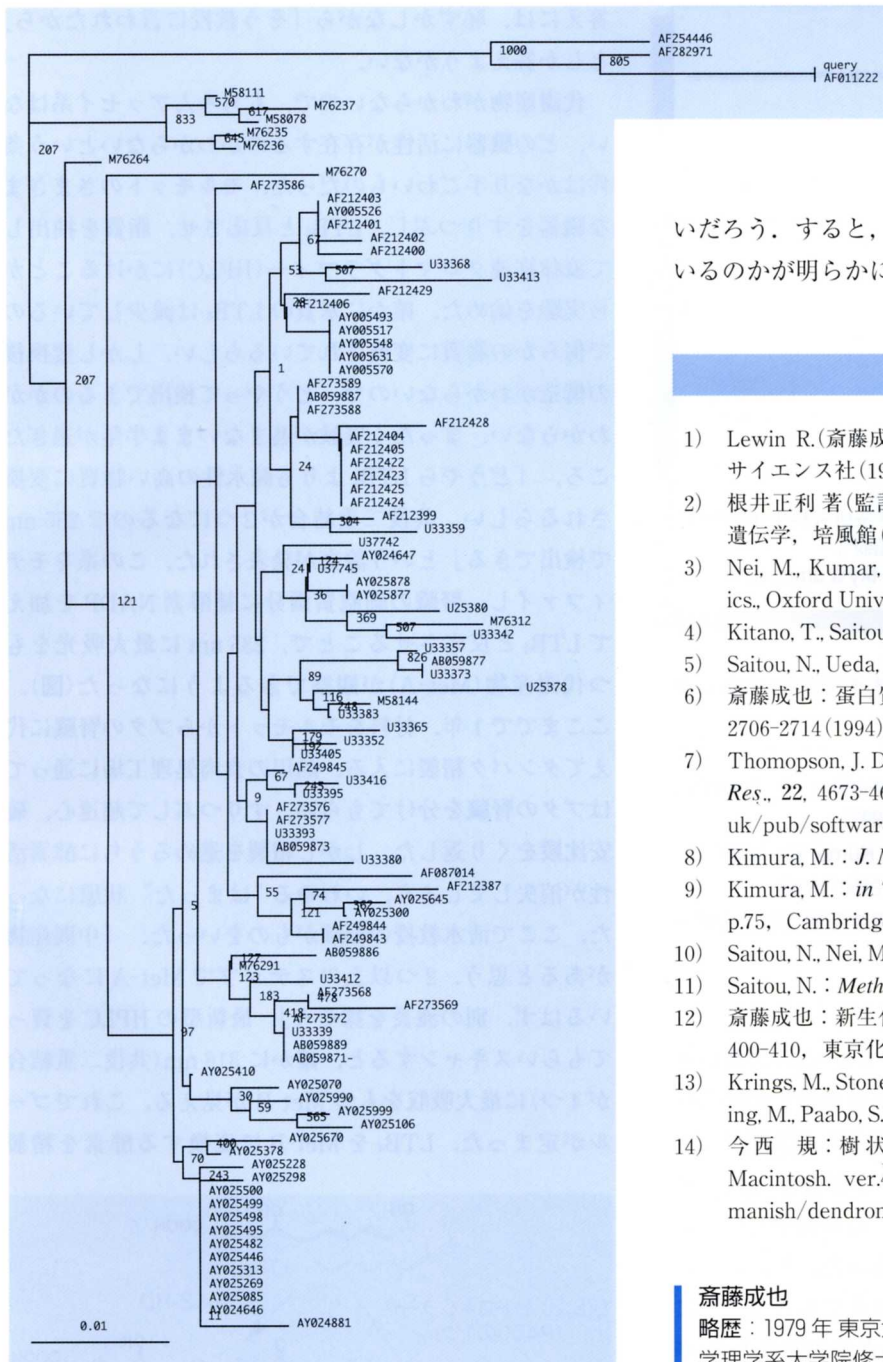


図4 図3と同一の多重整列データを用いて、TreeViewで描いた近隣結合系統樹
ブツストラップ系統樹で実現した枝の回数も示されている。

■ おわりに

なんでもそうだと思うが、「習うより慣れろ」である。とにかく自分でコンピュータを使って、解析をしてみることである。最初は簡単なデータ、それこそ手計算でチェックできるような数配列のデータを料理してみるといい

いだらう。すると、プログラムが実際には何を計算しているのかが明らかになるからである。

文献

- 1) Lewin R. (斎藤成也 監訳) : DNA から見た生物進化, 日経サイエンス社 (1998)
- 2) 根井正利 著 (監訳), 五條堀孝・斎藤成也 (訳) : 分子進化遺伝学, 培風館 (1990)
- 3) Nei, M., Kumar, S. : Molecular evolution and phylogenetics, Oxford University Press (2000)
- 4) Kitano, T., Saitou, N. : *J. Mol. Evol.*, 49, 615-626 (1999)
- 5) Saitou, N., Ueda, S. : *Mol. Biol. Evol.*, 11, 504-512 (1994)
- 6) 斎藤成也 : 蛋白質 核酸 酵素 (増刊), エボルーション, 39, 2706-2714 (1994)
- 7) Thomopson, J. D., Higgins, D. G., Gibson T. J. : *Nucl. Acids Res.*, 22, 4673-4680 (1994) (ソフトウェアは <ftp://ftp.ebi.ac.uk/pub/software/> からダウンロードできる)
- 8) Kimura, M. : *J. Mol. Evol.*, 16, 111-120 (1980)
- 9) Kimura, M. : *in* The neutral theory of molecular evolution, p.75, Cambridge University Press (1983)
- 10) Saitou, N., Nei, M. : *Mol. Biol. Evol.*, 4, 406-425 (1987)
- 11) Saitou, N. : *Method. Enzymol.*, 266, 427-449 (1996)
- 12) 斎藤成也 : 新進化学実験講座第16巻 分子進化実験法, pp. 400-410, 東京化学同人 (1993)
- 13) Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., Paabo, S. : *Cell*, 90, 19-30 (1997)
- 14) 今西 規 : 樹状図作成プログラム DendroMaker for Macintosh, ver.4.1 (1998) (<http://www.cib.nig.ac.jp/dda/timanish/dendromaker/home-j.html>)

斎藤成也

略歴 : 1979年 東京大学理学部生物学科卒業。1981年 東京大学理学系大学院修士課程人類学専攻修了。1986年 テキサス大学ヒューストン校生物医科学大学院博士課程修了(Ph.D.)。東京大学理学部助手を経て、1991年より国立遺伝学研究所進化遺伝研究部門助教授。総合研究大学院大学生命科学研究科遺伝学専攻教授を併任。研究テーマ : 類人猿とヒトの比較ゲノム解析, 人類集団の遺伝的近縁性, 遺伝子の進化一般, 分子進化解析の手法開発など。関心事・抱負 : 進化ゲノム学(Evolutionary Genomics)の確立, 人間性を規定する遺伝子変化の発見。

▶ 次号「アミノ酸配列に既知のパターンを探す」 池尾一穂