配列からの遺伝子系統樹解析を役立てるには？

斎藤成也

昨年，このシリーズで，「配列から遺伝子の進化を探る」と題して，多重整列された塩基配列を出発点として，遺伝子系統樹解析について簡単に紹介した．用いるソフトウェアとしては，国立遺伝学研究所のDDBJで公開しているCLUSTALWを中心としたが，昨年末に，このCLUSTALWサーバーの機能を拡張して[1]，多重整列の結果を，論文などで馴染み深い表記法(ドットオプション)に変更できるようにしたほか，塩基置換数(進化距離)の推定法として，Tamura，Tajima-Nei，Gojobori-Ishii 6-parameter，Tamura-Neiの4種類の距離も計算できるようになった．

その後，コンピュータに描かせた系統樹がどの程度信頼性の高いものなのか，2つのグループを同じ群といっていいのか・違った群というべきなのかの判断のポイントを知りたい，という質問が読者からあったので，編集部から追加の解説をしてほしいという依頼を受けた．そこで今回は，ブートストラップ法や，異なる遺伝子を使って解析した際に異なる樹形(トポロジー)が得られた場合の解釈などについて説明する．

最初から系統樹だと決めてかかってはいけない

コンピュータソフトを使う・使わないに限らず，自分のデータを既存のモデルにあてはめて解析し，それでこと足れりとするのは危険である．この意味で，うまく多重整列できたら，ただちに系統樹作成ソフトを使って系統樹を作成するという盲目的なやり方に問題が生じる場合がある．同一種内の複数の配列を比較する場合や近縁種を比較する場合，組換えが生じて，系統樹構造を乱す可能性がある．直列重複遺伝子を比較する場合には，それらの間の遺伝子変換が生じると，やはり系統樹では表わすことのできない構造が生じうる．

このような場合には，われわれがABO式血液型遺伝子[2]やRh式血液型遺伝子[3]の解析で用いた，“系統ネットワーク”を作成して，系統樹構造が得られるかどうかをチェックするとよいだろう．系統ネットワークは系統樹の拡張と考えることができる．系統樹では表わすことのできない，網状の関係も表示できるし，データによっては系統樹が得られる場合もあるからだ．

1．ブートストラップ法を用いた系統樹の統計検定

さて，自分のもっている配列データについては，系統樹になるとしよう．そこで，さまざまなソフトウェアを用いて遺伝子の系統樹を作成することになる．結果として得られた樹形のそれぞれの枝が，どのくらい信頼性が高いのかは，いろいろな統計検定法があるが，最も広く使われているのが“ブートストラップ法”なので，この方法について簡単に説明する．この方法[4]は，分散を算出するのが困難な現象に対して，統計学者のEfron[5]が開発したものである．その後，Felsenstein[6]が系統樹の統計検定に最初に応用した．

ブートストラップとは靴紐のことである．その昔，「ほらふき男爵の冒険」という物語があったが，そのなかで，自分で自分の体を引っ張って空中に浮かぶという方法が書かれている．同様に，靴紐を引っ張って自分の体を持ち上げるというのがブートストラップ法の名前の由来である．もっとも，別にほらをふいているわけではなく，きちんと前提が満たされていれば，この方法はコンピュータを用いて分散を正しく推定することのできる，すばらしい方法である．ただ，前提が満たされない場合にはとんでもない結果を生むことになるので，盲目的な利用は危険である．もっとも，これは統計検定すべてにあてはまることである．

ブートストラップ法を用いる場合には，「重複を許した標本抽出」をまず行なう．多重整列された塩基配列データの場合，1つ1つの標本(サンプル)は1つの塩基サイトにあたる．たとえば，500塩基の配列があったとしよう．1番目から500番目のサイトが存在するが，コンピュータで一様疑似乱数を発生させて，1～500の数の1つを選ぶ．それが234であれば，234番目の塩基サイトを選んで，それが新しい1番目のサイトとなる．同じことを500回くり返すのである．この場合，常に500個のサイトのどれかを選ぶので，サイトが重複する場合がある．逆に，500個の標本のなかに，一度も選ばれなかったサイトも出てくるだろう．こうして，同じ500個のサイトといっても，1番から500番まで残らず順に並んでいるもともとの配列とは少し異なる配列が得られることになる．なお，塩基サイトの順番は，大多数の系統樹作

Saitou Naruya，国立遺伝学研究所集団遺伝研究部門　E-mail：nsaitou@genes.nig.ac.jp　http://sayer.lab.nig.ac.jp/~saitou/
How to improve gene phylogeny analysis from sequence data

成法では問題にしないので，サイトをソートしたりする必要はない．

これで1つのブートストラップ標本が得られたわけである．これには500個の疑似乱数を発生させているが，コンピュータを用いれば，あっという間である．このプロセスをたくさん，通常は1,000回以上くり返して，ブートストラップ標本を生成する．多重整列された同一の塩基配列データから得られているものの，1つ1つの新しい配列は少しずつもとの配列とは異なっている．このばらつきが分散を与えるのである．ブートストラップ法の理論では，もし本来の塩基配列データが，仮想的な無限に伸びる母集団配列からの任意抽出標本としての性格をもっていれば，そこから得られたブーツストラップ標本のばらつきは，十分くり返せば母集団の分散に近づくことが知られている．

これら1つ1つのブートストラップ標本(それらがすべて多重整列された塩基配列となっている)をデータとして，いろいろな系統樹作成法を用いて，それぞれのデータから系統樹を作成する．このとき，ある特定の配列グループをまとめる(クラスターにする)枝が，十分高い割合のブートストラップ標本で観察されたら，その枝の信頼性が高いと判定するのである．

具体例をあげてみよう．300塩基の長さからなる，5本の塩基配列(A～E)を用いたとしよう．1,000個のブートストラップ標本配列を生成し，それぞれについて無根系統樹を近隣結合法で生成した．その結果，991標本では，配列Aと配列Bがクラスターする系統樹が得られた．もとの配列データを用いて作成した系統樹でも，同じクラスターが得られた．一方，配列Aと配列Cがクラスターする系統樹も6標本で，配列Aと配列Dがクラスターする系統樹が残りの3標本で認められた．しかし，それらのクラスターはもとの配列データを用いて作成した系統樹では現われていないので，無視する．次に，もとの配列データを用いて作成した系統樹では，配列Dと配列Eがクラスターしていたとしよう．ところが，ブートストラップ標本配列から得られた系統樹を分類すると，522標本だけがこのクラスターを示し，残りの478標本ではそれ以外のクラスターであった．このような結果が得られた場合，5個の配列のうち，配列Aと配列Bはブートストラップ確率99％という高い信頼度でクラスターしているが，配列Dと配列Eのクラスターは，もとの配列データを使った系統樹で認められるものの，ブートストラップ確率は52％であり，信頼性は低いことになる．

ただし，ブートストラップ確率が高いからといって，手放しで信用してはいけない．使われた進化距離の算出モデルや系統樹作成法に問題があれば，ブートストラップ確率には偏りが生じるからである．実際に，脊椎動物の系統関係をミトコンドリアDNAの完全配列で調べると，外群種の取り方によっては，どの系統樹作成法を用いてもすべての枝で100％のブートストラップ確率となるのに，その系統樹は従来の常識とはまったく異なる(魚類と四足類に二分されてしまう)ということが知られている．これは，魚類と四足類ではアミノ酸配列の変化パターンが違うからではないかと考えられている[7]．

逆に，ブートストラップ確率が低いからといって，がっかりすることはない．50個の配列を比較して系統解析をしたところ，ある2個の配列がクラスターする傾向にあったが，そのブートストラップ確率は70％だったとしよう．通常の5％有意レベルでは，これはまったく低い値である．しかし，50個の配列から任意に2個の配列を選び出す確率(1/2205)からすると，これはきわめて高い値であり，何らかの意味でこれら2個の配列は共通性があると考えるべきである．

■2．複数の遺伝子を用いて種系統樹を推定する

遺伝子系統樹が，種の系統関係を推定する目的で使われる場合には，いろいろな遺伝子を用いることができる．種分化に伴って，ゲノム全体が同じように分岐してゆくからである．ただし，これは進化的に遠い関係にある生物ばかりを比較する場合であり，近縁種や，急速に分岐した生物群を比較する場合には，遺伝子によって異なる系統樹が生じることがある．これは，祖先種で存在していた，同一遺伝子座のなかの対立遺伝子の系統関係が子孫種に引き継がれる，いわゆる“lineage sorting”が生じることがあるからである．この現象が最初に注目された，ヒト，チンパンジー，ゴリラの系統関係を例にとってみよう．生物種の系統関係としては，ヒトとチンパンジーがクラスターする(姉妹群となる)にもかかわらず，チンパンジーとゴリラ，あるいはヒトとゴリラというクラスターが，遺伝子によっては見いだされることがあるということだ．ただ，このような場合でも，遺伝子間で進化速度に大きな違いがなければ，全遺伝子配列をくっつけて1つの長い配列として，それから単一の系統樹を作成するという方法がある．各遺伝子の系統樹の平

均像しかみることができないが，得られた系統樹は，種の系統関係を反映していることが期待される．

進化的に遠く離れた種を比較する場合には，塩基配列よりもアミノ酸配列を用いることが多いだろう．この場合，遺伝子によって進化速度が大きく異なる可能性がある．すると，全体の配列を合体して単一の系統樹を作成した場合，どうしても進化速度の高い遺伝子のパターンにひきずられる可能性がある．もちろん，進化速度が高いということは，それだけ系統樹の樹形を決定するためのシグナルが多いということを意味しているが，同時に，並行置換などのノイズも入ってくるだろう．それに対して，進化速度の低い遺伝子では，わずかなアミノ酸の置換でも，系統樹の枝を確定するには重要である場合もあるだろう．それなのに，進化速度の高い遺伝子と一緒にされると，このような少数派の影響力は弱まってしまうだろう．通常の統計検定でこのあたりの細かい対応を行なうのはむずかしい．結局，打開策としては，より多くの遺伝子を比較するということになるだろう．系統進化は過去の1回限りの現象であり，それを復元しようとしたら，あれこれ手を尽くすべきだろう．

文 献

1) 日本DNAデータバンク編：CLUSTALWサーバーの機能拡張．DDBJオフラインニュース，16，8(2001)
2) Saitou, N., Yamamoto, F.-I.：*Mol. Biol. Evol.*, 14, 399-411 (1997)
3) Kitano, T., Saitou, N.：*J. Mol. Evol.*, 49, 615-626 (1999)
4) Efron, R., Tibshirani, R. J.：An introduction to the bootstrap. Chapman & Hall, London (1993)
5) Efron, R.：*Ann. Statist.*, 7, 1-26 (1979)
6) Felsenstein, J.：*Evolution*, 39, 783-791 (1985)
7) Takezaki, N., Gojobori, T.：*Mol. Biol. Evol.*, 16, 590-601 (1999)

著者プロフィール
→2001年8月号(p. 1413)

▶次号 Q&A(その2)

MolScriptを使いこなすには？ 楠木正巳

アルゴリズムって何？ 隠れマルコフモデルって何？ 中井謙太