2-2-b

# 遺伝子系統樹の構築法

遺伝子系統樹は，遺伝子自身の進化，遺伝子進化と表現型進化の関係を知るうえで必須の情報である．一方，遺伝子機能解析においても重要な情報を与えてくれる．遺伝子は遺伝子重複によって遺伝子族(gene family)を形成している場合が多く，近縁な遺伝子は似た機能をもっている場合が多い．したがって，どの遺伝子同士が近縁なのかを調べることにより，新規遺伝子の機能や遺伝子機能の冗長性を予測することができる．また，異種間で遺伝子機能を比較する場合，オーソログ*1同士を比較しているかを確認するためにも，遺伝子系統樹は必須である．遺伝子の系統関係の推定は，遺伝子機能の推定と同じように個々の遺伝子によって対処方法が異なるのが現状であるが，本稿では，筆者らの研究室で通常用いている方法を紹介する*2．

## 準　備

計算は原則としてUNIX環境で行う．Macintosh上で行うには，MachTen*3あるいはMac OS X*4を用いると，Macintosh用のアプリケーションが使える環境のままでUNIX環境が得られる．問題の規模により多量の演算処理が必要なので，できるだけ高速なコンピュータを用いる．自前でコンピュータを用意できない場合は，大学の計算機センターに問い合わせてアカウントを取得する．以下に概略を述べる*5．

①アラインメントに使うClustalW[1]と系統樹構築に使うMOLPHY[2]のソースコードは，それぞれftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/およびftp://ftp.ism.ac.jp/pub/ISMLIB/MOLPHY/から入手しコンパイルする*5．MOLPHYはUNIXシステムの上で使うのが最も便利である．

②PAUP*4.0[3]は2001年早々にSinauerから出版されることになっているので洋書として購入する．ワークステーション版はU.S.＄150，Macintosh版はU.S.＄100である．

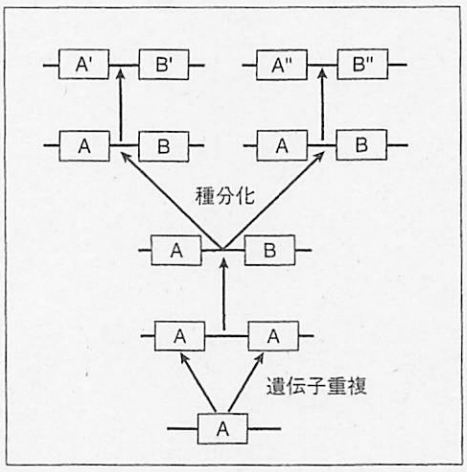


図1　オーソログとパラログ
A', B', A", B"は，祖先遺伝子であるAから遺伝子重複と種分化によって生じた．A'とA", B'とB"はそれぞれ互いにオーソログ，AとB，A'とB'，A"とB"はそれぞれ互いにパラログである．

＊1　オーソログ
2つの遺伝子が重複によってできたとき，それらはパラロガス(paralogous)であるといい，種分化にともなってできたときオーソロガス(orthologous)だという(図1)．

＊2　毎年春に行われる基礎生物学研究所バイオサイエンストレーニングコースで公開実習を行う予定である(http://www.nibb.ac.jp/参照)．

＊3　MachTenはTenon Intersystem社の商品で，http://www.tenon.com/からオンラインで購入するとU.S.＄249/licence＋送料U.S.＄50で1週間ほどで手に入る．

＊4　Mac OS 9(OS 9以前)上で計算処理を行うと，その間はコンピュータを実質的にほかのことには使えない．一方，Mac OS Xでは，一応他のことをしても大丈夫である．ただし，計算速度はMac OS 9で行う場合よりも若干遅くなる．Mac OS Xは現時点ではまだ発売されていないので，評価はMac OS X Public Betaに基づく．

＊5　入力するコマンドなど，詳細をhttp://www.nibb.ac.jp/~mhasebe/に記してあるので，参考にしていただきたい．

## 方法 (図2)

### ▶ 1. 配列収集

1) 関連する配列を集める．遺伝子系統樹を書く場合，まず，その遺伝子が含まれる遺伝子族のすべての配列を集めるべきである．アノテーションに依存せずに遺伝子族のメンバーをすべてみつけるためには，データベースに対してホモロジー検索を行う．ここでは，NAC遺伝子族[4]を例にあげる．

↓

2) NCBIのウェブサイトでCUC2(AB002560)[4]のアミノ酸配列をもとに，データセットnr(非重複タンパク質データベース)に対してblastp searchを行う．*6

↓

3) CUC2に対してヒットする配列は100個みつかった．この結果は，ブラウザからHTMLソースとして保存する．このファイルには関連配列へのリンクが含まれているので，関連配列を調べるのに便利である．

↓

＊6　この検索ではESTは対象にならないので，ESTを解析に含める場合は別途検索する必要がある．ESTは不確定な部分があり，系統解析に使えない場合が多い．

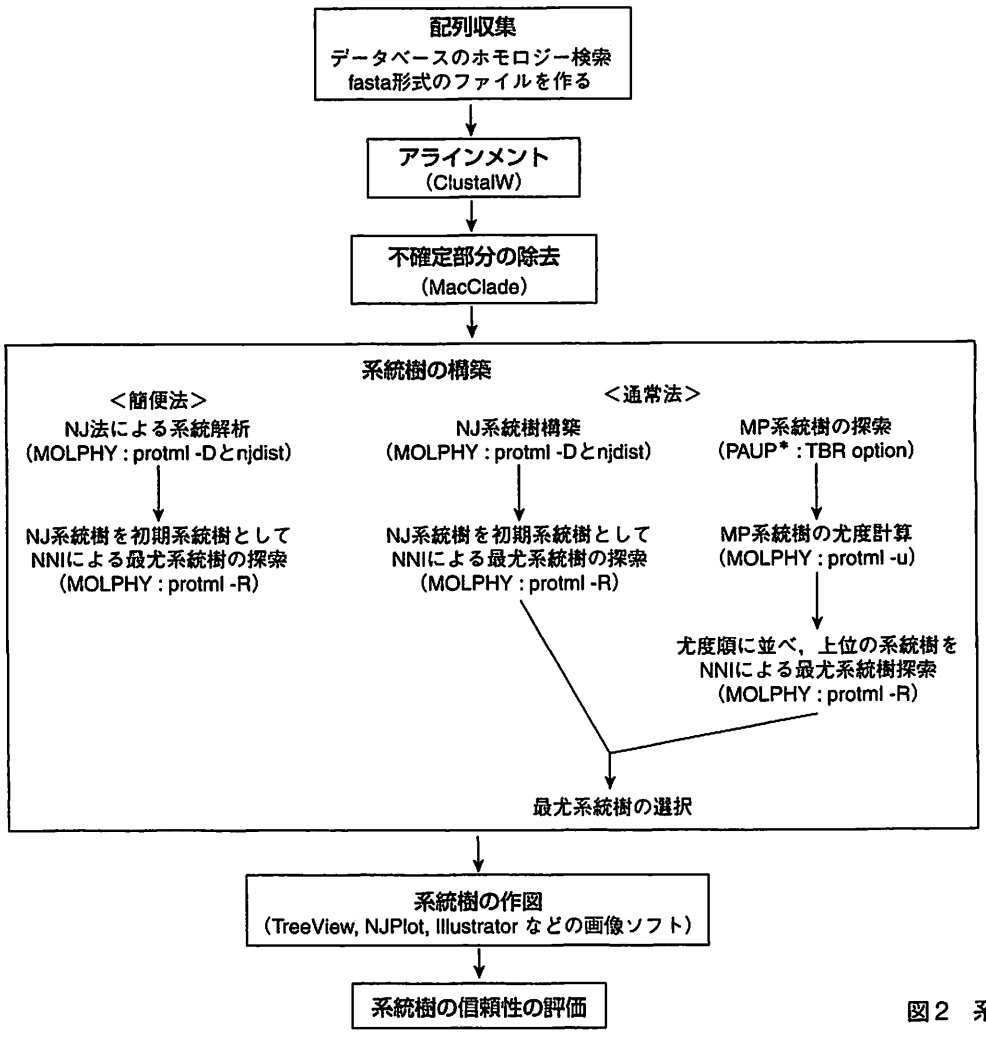


図2　系統樹構築の手順

4) 得られた配列をfasta形式にまとめる．すなわち，不等号(>)で始まる行に遺伝子の名前を書き，次の行から実際のアミノ酸配列を並べる．遺伝子の名前は英字で始めて英数字で10文字以内とし，“．(ピリオド)，(カンマ)ー(ハイフン)”などの記号は使わないようにする．

↓

5) 1つのアミノ酸配列が終わったら，次の行に再び不等号に続いて遺伝子名を入れる．これを解析対象の全配列について行う．このようにして，関連配列すべてを1つのfasta形式のファイルに集める．CUC2の場合には，次のようになる(実際はもっと長い)．

```
>CUC2
MDIPYYHYDHGGDSQYLPPGFRFHPTDEELI
THYLLRKVLDGCFSSRAIAEVDLNKCEPWQL
>BAB105131
MAPVSLPPGFRFHPTDEELITYYLKRKINGQ
DRKYPNGSRTNRATKGGYWKATGKDRRVSWR
```

## ▷ 2. アラインメント

ClustalWなどの多重アラインメントを行うソフトウエアを利用して，アミノ酸配列の対応する位置をあわせる．核酸レベルでアラインメントしてから翻訳してもよいが，アミノ酸に翻訳してからアラインメントしたほうが簡単である．MacTenを使った操作法は，ウェブサイト(＊5)を参照していただきたい．国立遺伝学研究所でもClustalWのサービスを行っている*7のでこれを利用してもよい．しかし，手許のコンピュータで行ったほうが結果が早く得られる．

＊7 http://www.ddbj.nig.ac.jp/htmls/E-mail/clustalw-e.html

## ▷ 3. 不確定部分の除去

アラインメントをした時に対応関係がつかなかったり，確定できない部分が出てくる．この領域は，系統解析で誤った解析の原因となる可能性があるので除去し，曖昧さがなくアラインメントできた部分のみを用いる．すなわち，原則としてギャップのある座位は利用しない．また，あまりにも欠失の多い配列は解析から除く．

曖昧なアラインメントの例

```
gene1  LDEELAFWQIINS
gene2  LEEDLAFWQVINS
gene3  LE-ELAFWQIIKS
gene4  LEDEL--------
```

この例では，gene3のギャップ(−)を2〜4番目のうちどこに入れるのが適切か判定できないため，2〜4番目の3アミノ酸はどれも使わないのが妥当である．またgene4は除くほうがよい．NAC遺伝子族を用いた例を，ウェブサイト(＊5)に示してある．このとき，MacClade[5]を用いて，解析に使用する部分と使わない部分をマークしたファイルを作ると便利である．しかし，紙にアラインメントを印刷したうえでどの部分が使えるかを検討すれば，MacCladeを使わなくても可能である．

アラインメントおよびそのうちのどの領域を用いたかの情報は，系統解析を検証するのに必要な情報である．しかし，論文には紙幅の都合上載せられないことが多い．EBIでは，アラインメントのデータベースを作っているので，アミノ酸配列の場合は登録するとよい*8．また，核酸配列のアラインメントはNCBIに

＊8 http://www.ebi.ac.uk:80/embl/Submission/alignment.html

登録することもできる*9.

## ▶ 4. 系統樹の構築

系統樹構築法にはいくつかの方法があるが，ここでは最尤法[6]を用いる．最尤法は，最も尤度*10の高い系統樹を選択する方法であり，最も良い推定法であると考えられる[6]．この方法では考えられるすべての樹形の尤度を比較し，そのなかから最も尤度の高いものを選ぶ(完全探索 exhaustive search) のが理想的である．しかし，配列数に対して考えられる樹形の数は指数的に増えるので*11，一部の樹形の尤度を評価することでよりよい樹形を探索する(発見的探索, heuristic search) ことになる．

### 1) 簡便法

予備的に短時間で系統関係を推定したい場合に用いる．正しい系統関係が得られない場合もあるので，最終的には次に述べる通常法で解析しなおしたほうがよい．

1) MOLPHYのprotml -Dとnjdistを用いて，近隣結合法(NJ法)*12により系統樹を構築する．
↓
2) MOLPHYのprotml -Rを用いて，1)で得たNJ系統樹を初期系統樹とし，NNI*13により最尤系統樹を探索する．このとき，局所的ブーツストラップ確率*14も一緒に計算する．

### 2) 通常法[7]

1) 簡便法により最尤系統樹を探索する．
↓
2) PAUP*[3]のTBR*13によって最大節約系統樹*15を探索する．
↓
3) MOLPHYのprotml -uにより2)で得られた系統樹の尤度を計算する．
↓
4) 尤度の順に最大節約系統樹を並べる．最大節約系統樹が100本以下の場合はすべての系統樹をそれぞれ初期系統樹として，MOLPHYのprotml -Rを用いて，NNIによりさらに最尤系統樹を探索する．配列によっては数百以上の最大節約系統樹が得られるときがあるので，その場合は，コンピュータの能力に応じて尤度が高いものから順に選んで探索する．このとき，局所的ブーツストラップ確率も一緒に計算する．
↓
5) 1)と4)で得られた系統樹から，尤度の最も高い系統樹(最尤系統樹)を選ぶ．NAC遺伝子族を用いた例をウェブサイト(*5)に示してある．

## ▶ 5. 系統樹の作図

系統樹構築プログラムの出力では，通常，系統樹は括弧書きで表現される．これを線画に変換して系統樹の図を作る．このとき，通常は縦軸は任意で，横軸が進化距離(座位あたりの置換数)に比例するようにする．系統樹の作図にはTreeView*16やNJPlot*17といったプログラムを使う．また，MOLPHYでは線画のepsファイルも出力される．NAC遺伝子系統樹をウェブサイト(*5)に

*9 http://www.ncbi.nlm.nih.gov/Web/Newsltr/Winter99/winter99.htm#Alignment

*10 尤度 (likelihood)
ある確率モデルのもとで，用いているデータが得られる確率．ここでは，ある系統樹のもとで系統解析に用いた配列データが得られる確率のことをいう．

*11 n個 (n＞2) の配列の二分岐だけからなる無根系統樹の樹形の数は，1・3・5・…・(2n－5) である．これはn＝13のとき137億に達する．

*12 近隣結合法 (neighbor joining method, NJ法)
配列の数が多いときに計算時間が少なくてすむ方法である．その反面，誤った系統推定となる確率が相対的に大きい．

*13 NNI (Nearest Neighbor Interchange), TBR (Tree Bisection and Reconnection)
発見的探索法では，初期系統樹を作りその初期系統樹の枝を位置交換することによって，より尤度の高い樹形を探す．NNIは最近隣枝を交換する方法，TBRは系統樹の切断と再接続をする方法である．一般にTBRのほうが広い範囲を探索することになり局所的最適解にトラップされにくいので，TBRを行うことが望ましい．ただし，計算時間がかかる．局所的最適解とは，すべての樹形のなかで最も尤度の高い樹形ではないが，1回の枝交換によってできる樹形のなかで最も尤度の高い樹形である．

*14 ブーツストラップ確率 (bootstrap probability)，ブーツストラップ法 (bootstrap method)
ブーツストラップ確率は，ブーツストラップ法により得られた系統樹の枝の統計的確からしさを評価する値．ブーツストラップ法は，得られている配列が真の分布であると仮定して，そこから解析に用いた座位と同数の座位を重複を許して無作為に抽出したときにそれぞれの枝が得られる確率を求める方法．MOLPHYでは局所的ブーツストラップ確率[6]が計算される．

示してある．さらに，アドビイラストレーターなどのドロー系グラフィック処理アプリケーションを用いて，各枝のブーツストラップ確率，どの生物から単離された遺伝子であるか，どの範囲がどのような特徴をもっているかなどの情報を書き加える．

## ▶ 6．系統樹の信頼性の評価

以上の計算で得られた系統樹は，必ずしも正しい系統関係を推定しているとは限らない．そこで，系統樹の樹形の信頼性を統計的手法によって検定する必要がある．系統解析を乱す要因として，以下のようなものが知られている．

①統計的誤り
②系統間の進化速度の違い
③座位間の進化速度の違い
④適応的な進化

統計誤差は解析に用いることのできる配列長が限られていることによるものである．一般に真の確率分布を知ることはできないのでブーツストラップ法*14によって評価する．ブーツストラップ確率が低い枝は信頼性が低いといえる．逆にブーツストラップ確率が高くてもそれだけでは信頼できるとは限らない．適応的な進化や，系統間あるいは座位間の進化速度の違いなどの系統的誤差の影響を受けている可能性があるからである．MOLPHYでは局所的ブーツストラップ確率が計算される．通常のブーツストラップ確率よりも高めの値が出るので，70%以下の枝はあまり信用できないと考えるべきである．

＊15 最大節約系統樹 (most parsimonious tree)
最小の置換回数で現在の配列が説明できる系統樹．置換率が小さい場合は最尤系統樹と樹形が一致するが，置換率が大きいと必ずしも合わない．

＊16 http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

＊17 http://pbil.univ-lyon1.fr/software/njplot.html

## コメント

①非常に近縁な遺伝子のみを解析するためには，塩基配列データを用いたほうが精度よく解析できる場合がある．アミノ酸で書いた系統樹では，枝の長さが非常に短くて解けない場合には，その周辺の遺伝子について塩基配列レベルの解析を試みるとよい．

②J. Felsensteinのウェブサイト(http://evolution.genetics.washington.edu/phylip/software.html)は系統解析のためのソフトウエアを網羅的に紹介しており，どのようなソフトウエアがあるか，どのように入手すればよいかがわかるようになっている．

### ◆参考文献


1) Thompson, J.D., Higgins, D.G. & Gibson, T.J. : CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, Nucl. Acids Res. 22, 4673-4680 (1994)
2) Adachi, J. & Hasegawa, M: MOLPHY Version 2.3: Programs for Phylogenetics, ver. 2.3. Institure of Statistical Mathematics, Tokyo (1996)
3) Swofford, D.L.: PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts (2001)
4) Aida, M., Ishida, T., Fukaki, H. et al.: Genes involved in organ separation in *Arabidopsis*: an analysis of the *cup-shaped cotyledon* mutant, Plant Cell 9, 841-857 (1997)
5) Maddison, W.P. & Maddison, D.R.: MacClade: analysis of phylogeny and character evolution, ver. 4. Sinauer Associates, Inc. MA, USA (2000)
6) 長谷川政美，岸野洋久:「分子系統学」，岩波書店 (1996)
7) Sakakibara, K., Nishiyama, T., Kato, M. et al.: Isolation of homeodomain-leucine zipper genes from the moss *Physcomitrella patens* and the evolution of homeodomain-leucine zipper genes in land plants, Mol. Biol. Evol. 18, in press (2001)



西山智明・長谷部光泰† 岡崎国立共同研究機構 基礎生物学研究所 種分化機構第2研究部門
† E-mail：mhasebe@nibb.ac.jp