微生物の系統樹，どう描くの？

飯野　隆夫*・伊藤　　隆

今日でも多くの微生物研究者がそれぞれの目的で自然環境から興味ある微生物を分離してくるが，最初にその系統学的位置を決めてから本来の研究目的に使うことが多くなってきているのではないだろうか．また自ら遺伝子の塩基配列を決定しなくても，公共のデータベースからデータをダウンロードし，系統解析プログラムを用いて，細かい系統解析することも可能になった．このように遺伝子塩基配列やアミノ酸配列に基づいた系統解析は微生物研究においても基本的なツールになったと言っても過言ではない．しかし，その一方で，少しぐらい誤ったデータ入力や操作をしても一応は系統樹が得られることが多いので，それが本当に自分の解析目的に合っていたのか吟味する必要性があろう．

本稿では微生物を研究対象としている研究者が，その系統関係を調べるために特定の遺伝子の塩基配列を決定し関連する微生物株のデータについても準備したものと仮定した上で，それ以降のアライメントや系統解析法についてその基礎的な原理とプログラムの基本的操作法について概説する．

アライメント

アライメントは比較する塩基（アミノ酸）配列の各座位の比較対象を決めるための位置合わせ，あるいはその結果として複数の塩基配列が整列配列されたものを示す．祖先を共有する二つの遺伝子配列同士を比較する場合，一方の塩基配列の座位に挿入や欠損が起きた場合にはどちらかの配列にギャップを入れて，他の座位が比較できるように位置合わせをする必要がある．系統解析に供するマルチプルアライメントの作成には一般にClustal X[1]などのプログラムが使用されるが，Clustal Xでは初めに比較する塩基配列二つずつのペアワイズアライメントが総当たりで行われ，配列間のスコアが計算される．続いて全配列ペアのスコアをもとに近隣結合法（後述）によってガイドツリーと呼ばれるデンドログラムが作成され，これに従ってもっとも近縁な配列からアライメントが行われ，逐次的にアライメントが並べられてマルチプルアライメントが作成される．系統解析を目的とする時は以下の点に注意してアライメントを行う必要がある．

1）データベースには同一の遺伝子が複数登録されているケースがあるので，その場合はなるべく信頼度が高い塩基配列を使用する．またイントロンなどを含んだものではあらかじめその配列を削除しておく．

2）プログラムでアライメントしたものは原則目を通してギャップの位置の確認を行い，必要に応じてマニュアルで修正を行う（アライメントの編集にはSeaViewなどが一般的に用いられる）．この場合，相同性の高いOTU（operational taxonomic unit；操作上の分類単位）同士が並ぶように順番を並び替えすることでギャップの位置をより明確に判断することができる．

3）マルチプルアライメントでギャップが集中している箇所は二次構造が異なっている可能性もある．高次構造の違いは他のサイトとは異なる要因で生じる可能性があるため目的によってはこのような箇所は削除する必要性がある．

分子系統解析法

近隣結合法（Neighbor-joining method；NJ法）[1,2]
複数のOTUの近隣を段階的に見いだし，最終的に無根系統樹を得る方法である．本法は各OTU間の距離行列を計算し，これを星状系統樹に配置した後，二つのOTUを結合したときにもっとも星状系統樹の枝長の総和を小さくするのに都合のよい系統樹を選択する．以下この操作を段階的に繰り返すことによって最終的な系統樹を得るものである．この方法は段階的クラスター法に属し，他の方法よりもはるかに短い計算時間で系統樹を作成できる利点がある．進化速度を一定と仮定しないため，進化速度が異なる系統であっても，比較的正しく系統樹を作成することが可能であるが，サイトごとに異なる進化速度は結果に反映されない．

最大節約法（Maximum parsimony method；MP法）[3]
塩基配列上に有意な塩基置換の起こった座位を抽出し，これらがもっとも少ない置換回数で説明できる系統関係を選択する方法である．プリン塩基同士あるいはピリミ

*著者紹介　独立行政法人理化学研究所 バイオリソースセンター微生物材料開発室（研究員）　E-mail: iino@jcm.riken.jp

ジン塩基同士などの置換の起こりやすさを反映させることは可能である．一方，系統学的に深い位置で分岐したもの同士を含む場合や特定の系統で進化速度が速くなっている場合には同じ座位で起こる多重置換を無視することができず，偏った推測をしてしまう可能性も指摘されている．複数得られた系統関係の中からもっとも置換が少ない関係を選び出す目的には有用な方法である．

最尤法（Maximum likelihood method；ML法）[4,5] 塩基置換における何らかのモデルを仮定し（たとえば塩基置換におけるJukes-Cantorモデル，Kimura 2パラメータモデルなど），そのモデルに基づき対象とする塩基配列が時間経過に伴って先祖配列から子孫配列へ塩基置換する確率を計算し，もっとも尤度（確率）の高い樹形を導く方法である．最尤法は網羅的探索法に属する．そのため探索のアルゴリズムやデータによってはきわめて計算時間がかかることがある．

ベイズ法[6,7] マルコム連鎖モンテカルロ法に基づいて大量の系統樹を作成し，単系統群の出現頻度（事後確率）を求める方法である．すなわち初期系統樹を攪乱し，得られた系統樹が初期系統樹の置換モデル，樹形，枝の長さから導かれる確率によって受理・棄却されるかを決定し，受理される場合は新たな系統樹に対してこの操作を繰り返す（棄却された場合はもとの系統樹に対して再度この操作を行う）．このマルコフ連鎖を繰り返し定常状態に達した時に単系統群の出現頻度が事後確率に相当する．ベイズ法は比較的新しい系統解析法で今後に注目されているが，一方で計算時間がかかることや事後確率が過大評価されるなどの問題がある．

系統樹の統計学的評価

系統樹の統計的有意性を検定するにはブートストラップ確率がよく用いられる（ベイズ法では事後確率が用いられる）．これはN個の座位からなる塩基配列を比較した場合，N回の復元抽出によりN個の座位からなる仮想塩基配列データを作成する．この操作を100～1000回繰り返し，それぞれの仮想データから系統樹を再構築する．これによって特定の系統関係（系統枝）が再現できる確率をブートストラップ確率という．一般的にはブートストラップ値が95%以上であれば，その系統枝の系統関係は統計的に有意と見なされる．

プログラムの操作法

以下に系統樹作成でよく用いられるプログラムの操作について概説する（表1）．筆者らの使用するMacOS 10.6.8での使用（PAUPのみWindows 7）を前提としているので，他のOSではプログラムのインストール，操作性等に細かい違いが生じる可能性がある．

Clustal X（アライメント）

1. Clustal Xを起動する．
2. FileのLoad sequencesから塩基配列のデータセット（図1）を選択する．データを追加する時は，FileのAppend Sequenceからデータを追加する．塩基配列を削除する時は，削除する塩基配列を選択し，EditのCut Sequenceをクリックする．ファイルや保存場所に日本語が含まれるとファイルを読み込めないので注意する．
3. AlignmentのDo complete alignmentを選択する．データの保存先とファイル名を決定した後，OKをクリックするとアライメントが実施される（「.aln」と「.dnd」の拡張子のついたファイルが保存される．例．XXX.aln，XXX.dnd）．
4. 目的に応じ，FilesのSave sequences asを選択し，適当な形式で保存する．PAUPとMrBayesを使用

表1．系統樹作成のためのプログラム

使用目的	プログラム	入力形式	OS	プログラムのWebサイト
アライメント作成	Clustal X*	FASTA形式等	Mac/Win/Linux	http://www.clustal.org/
近隣結合法	同上	CLUSTAL形式	同上	同上
最大節約法	PAUP	NEXUS形式	Mac/Win/Unix/DOS	http://paup.csit.fsu.edu/
最尤法	Morphy*	GDE形式を改変	SUN-OS/HP-UX	http://www.ism.ac.jp/ismlib/softother.e.html
ベイズ法	MrBayse*	NEXUS形式	Mac/Win	http://mrbayes.sourceforge.net/
アライメント編集	SeaView*		Mac/Win/Linux	http://pbil.univ-lyon1.fr/software/seaview.html
系統樹の描画	NJplot*		Mac/Win/Linux等	http://pbil.univ-lyon1.fr/software/njplot.html
	TreeView*		Mac/Win/Linux/Unix	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

*，フリーウェア

する際はNEXUS形式（拡張子は「.nxs」），Morphyを使用する際はGDE形式とPHYLIP形式で保存する（「.gde」，「.phy」）．MrBayesではハイフン「-」を認識できないため，ファイル名にハイフン「-」を使用しない．

以下の系統解析プログラムではClustal Xでアライメントしたファイルを入力ファイルとして使用することを前提とする．また入力ファイルのアライメントはギャップ，不確定塩基のある座位を取り除き，各OTUの塩基数を揃える必要がある．

Clustal X（近隣結合法）およびNJplot（系統樹の描画）

以下に一般的な流れを示す．必要に応じてオプションを選択すること．

1. Clustal Xを起動し，FileのLoad sequencesから入力ファイル（CLUSTAL形式）をロードする．
2. TreesのExclude Positions with Gapsをオンにする．
3. TreesのCorrect for Multiple Substitutionsをオンにする．
4. TreesのOutput Format Optionsを選択する．
5. Output FilesをPhylip format treeにチェックを入れる．Bootstrap labels onをBranchに設定し，OKを押す．
6. TreesのBootstrap N-J Treeを選択する．
7. Random number generator seedとNumber of bootstrap trialsの値を設定する．指定がなければ初期値（111と1000）で良い．データの保存先とファイル名を決定した後，OKを押すと，進化距離計算（ブートストラップ計算）と系統樹作成が行なわれる．
8. 解析後，ファイル名にphbの拡張子が付いたファイルが作成される．
9. 出力されたファイルをNJplot[8]で開く（図2A）．
10. Operationのチェック項目を選択し，系統樹を編集する．
11. Displayのチェック項目を選択し，ブートストラップ値を表示する．

Clustal Xのオプション：

- Draw Tree：統計的な進化距離計算なしに系統樹を作成する．解析は早く終わるので，系統樹の正確さの確認に使用すると有効である．「.ph」の拡張子がついたファイルが作成される．
- Bootstrap N-J Tree：統計的な進化距離計算と系統樹作成を行う．「.phb」の拡張子がついたファイルが作成される．
- Exclude Positions with Gaps：系統解析の際に，配列中のギャップを含んだ座位を取り除く．
- Correct for Multiple Substitutions：系統解析の際に，配列間の距離を補正する．
- Output Format Options：出力するファイルの種類や保存形式を指定する．系統樹の描写にNJplotを使用する際は初期設定（Output Files：Phylip format tree，Bootstrap labels on：Branch）でよい．TreeViewを使用する際はOutput Files：Nexus format tree，Bootstrap labels on：Nodeを選択する．
- Clustering Algorithm：系統解析の方法（UPGMA法とNJ法）を選択する．近隣結合法では初期設定（NJ）を選択する．
- Random number generator seed [1-1000]：乱数の種を設定する．何らかの整数を与えることで，ブートストラップ計算を実施した際にその整数に基づいてランダムにアライメントを作成する．通常は初期値（111）でよい．
- Number of bootstrap trials [1-10000]：ブートストラッ

図1．FASTA形式で作成した塩基配列のデータセット

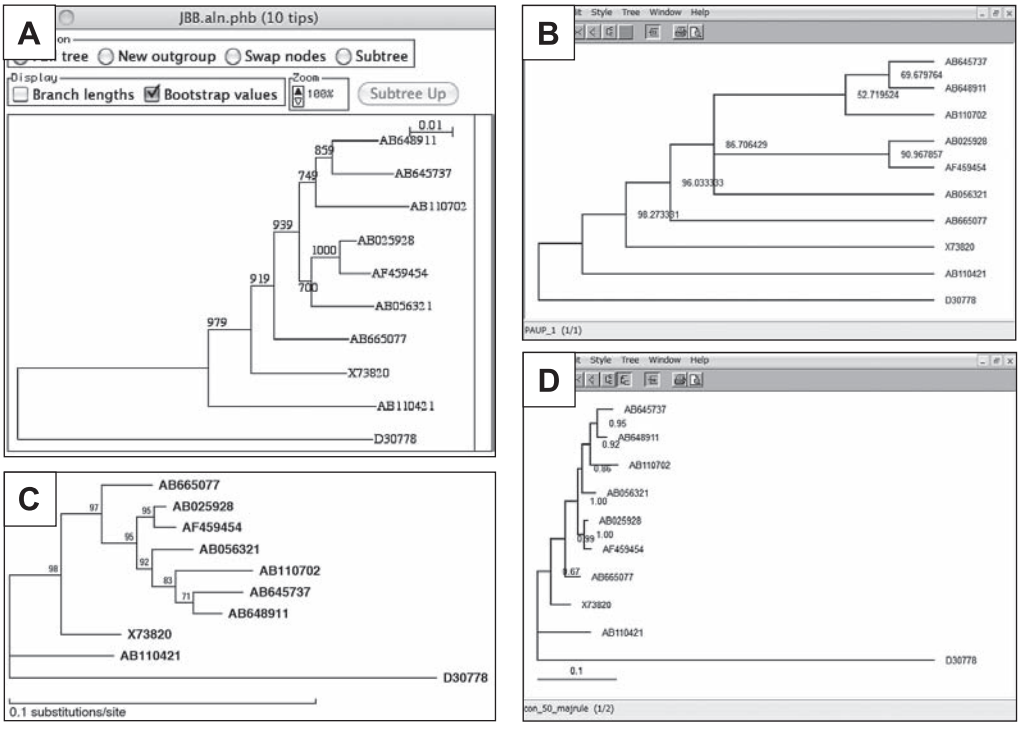

図2. 各種方法で解析した酢酸菌の16S rRNA遺伝子系統樹の描写．A，近隣結合法；B，最大節約法；C，最尤法；D，ベイズ法．AB665077, *Acetobacter acet* JCM 7641^{T}; AB025928, *Asaia bogorensis* 71^{T}; AB056321, *Kozakia baliensis* Yo-3^{T}; AB110421, *Saccharibacter floricola* S-877^{T}; AB110702, *Acidomonas methanolica* NRIC 0498^{T}; AB645737, *Komagataeibacter xylinus* JCM 17840^{T}; AB648911, *Gluconacetobacter liquefaciens* JCM 17840^{T}; AF459454, *Swaminathania salitolerans* PA51^{T}; X73820, *Gluconobacter oxydans* DSM 3503^{T}; D30778, *Rhodospirillum rubrum* ATCC 11170^{T}（アウトグループ）.

プ計算の回数を設定する．1000回行うことが一般的である．

・Save Phylip tree as：出力したファイルの保存場所と名称を指定する．

NJplotのオプション：

・Full tree：系統樹の全体を表示する．

・New outgroup：アウトグループを選択する．“#”をクリックすると，その枝を系統樹の外側に表示する．

・Swap nodes：分岐を並び替えます．“#”をクリックすることで枝の上下が入れ替わる．

・Subtree：系統樹の一部を表示する．“#”をクリックすると，その分岐より下位だけが表示される．subtree upをクリックすることで一つ上位の分岐が表示される．また，Full treeをチェックすることで全体の表示に戻る．

・Branch lengths：各枝の進化距離が表示される．

・Bootstrap values：ブートストラップ値が表示される．通常は値の百分率をブートストラップ確立として記述する．

・Zoom：系統樹の拡大や縮小を行う．

PAUP（最大節約法）

1. PAUPを起動する．FileのOpenから解析するNEXUS形式の入力ファイル選択する．File Open ModeでExecuteが選択されていることを確認し，Open/Executeで入力ファイルを読み込む．
2. コマンドラインにBandBもしくはhsearchを入力することで，Branch and Bound解析もしくはHeuristic解析が行われる．解析終了後，ステータスウィンドウを閉じる．
3. コマンドラインにshowtree xを入力すると樹形を表示できる（xには樹形番号を入力する．作成された樹形の数はNumber of trees retainedに表示されている）．
4. コマンドラインにbootstrap nrep = xxxx search = heuristic brlens = yesを入力し，Excuteをクリックするとブートストラップ解析が行われる．xxxxにはブートストラップ計算の回数を設定する．1000回行うことが一般的で，この場合，1000を入力する．
5. ブートストラップ計算が終了した後，ステータスウィンドウを閉じると，ブートストラップ値が表

```
 10  1392
AB665077
agcgaacgctggcggcatgcttaacacatgcaagtcgcacgaaggcttcggccttagtgg
cggacgggtgagtaacgcgtaggaatctatccatgggtgggggataactccgggaaactg
gagctaataccgcatgatacctgagggtcaaagggcaatcgcctgtggaggagcctgcgt
ttgattagcttgttggtggggtaaaggcctaccaaggcgatgatcaatagctggtctgag
```

図3. Morphyで解析するためのデータファイルの作成

示される.

6. コマンドラインにsavetrees file = XXX.tre savebootp = nodelabels from = 1 to = 1を入力し, Excuteをクリックすることで系統樹が保存される（XXXにはファイル名を入力する）.
7. 出力されたXXX.treをTreeViewで開く．Tree/Show internal edge labels を選択すると，ブートストラップ確立が表示される（図2B）.

Morphy（最尤法）

1. テキストエディットなどを用いて, ファイル上のデータ定義をMorphyで認識できる形式に修正する.
   1a. PHYLIP形式の入力ファイルの1行目の情報（データ数 塩基数）をコピーし，GDE形式の入力ファイルの先頭に追加する.
   1b. GDE形式の入力ファイルの“#”を削除する（図3）.
2. GDE形式をbinのフォルダ内に移す.
3. ターミナルを起動する.
4. ターミナル上で，PATH = $HOME/bin:$PATHを入力する.
5. 続けて，cd binを入力し，binにアクセスする.
6. nucml -D -topt XXX.gde > distanceを入力する（XXXはファイル名）.
7. njdist -t njtree distanceを入力すると，系統樹が形成される.
8. binに作成されたnjtree.tplを開き，系統樹情報をGDE形式の入力ファイルの末端に追加する.
9. ターミナル上で，nucml -topt -R -u XXX.gdeを入力するとブートストラップ計算が行われる（XXXはファイル名）.
10. binに作成されたnucml.epsをAcrobatもしくはPhotoshopで開くとブートストラップ確立つきの系統樹が描写される（図2C）.

MrBayse（ベイズ法）

1. テキストエディットなどを用いて，NEXUS形式の入力ファイル上のデータ定義をMrBayesで認識できる形式に修正する（図4）.

```
#NEXUS
BEGIN DATA;
dimensions ntax=10 nchar=1392;
format missing=?
symbols="ABCDEFGHIKLMNOPQRSTUVWXYZ"
interleave datatype=DNA gap= -;

matrix
AB665077  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGAAGGCTTCG
AB025928  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGGACCCTTCG
AB056321  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGGACCTTTCG
AB110421  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGAACCTTTCG
AB110702  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGAGGGTTTCG
AB645737  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGAACCTTTCG
AB648911  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGAACCTTTCG
AF459454  AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGGACCTTTCG
X73820    AGCGAACGCTGGCGGCATGCTTAACACATGCAAGTCGCACGAAGGTTTCG
D30778    GACGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGCATCCTTCG
```

図4. NEXUS形式の入力ファイルの初期状態

1a. symbols =“ABCDEFGHIKLMNOPQRSTUVWXYZ”を削除する.
1b. interleave datatype = DNA gap = -;をinterleave = yes datatype = DNA gap = -;に変更する.
2. ファイルをmrbayesのフォルダ内に移す.
3. MrBayesを起動する.
4. MrBayes上でexecute XXX.nxsを入力し，入力ファイルを読み込む（XXXはファイル名）.
5. 続けてlset nst = 6 rates = invgammaを入力する.
6. mcmc ngen = 10000 samplefreq = 10を入力し，マルコフ連鎖を開始する．ngen = には世代数の指定，samplefreqには何世代ごとにデータをサンプリングするかを指定する.
7. 10,000世代の試行が終了すると，解析を続行するか尋ねられる．Average standard deviation of split frequenciesの値が0.01未満に達するまで解析を繰り返す．解析を続行する場合，Continue with analysis? (yes/no):の問いにyesと入力する．次にAdditional number of generations:と問われるので，追加の世代数を入力する.
8. Frequenciesが0.01未満に達したら，Continue with analysis? (yes/no):の問いにnoと入力する.
9. sump burnin = xxxを入力し，結果の信頼性を確認する．xxx =（最終的な世代数）/ Samplefreq数 / 4 例）10,000/10/4 = 250
10. ParameterのPSRFが0.9~1.1程度の範囲であることを確認する
11. sumt burnin = xxxを入力し，樹形を出力する（xxxにはsumpの時と同じ値を入力する）.
12. 出力されたXXX.conをTreeViewで開くと系統樹が描写される．Tree/Show internal edge labelsを選択すると，事後確率が表示される（図2D）.

以上，系統解析に使われるプログラムの一部を紹介したが，解析例として酢酸菌菌株の16S rRNA塩基配列に基づく系統樹を見てみよう（図1参照，OTUはすべて塩基配列のアクセッション番号で表示）．図1の場合，AB025928，AB056321，AB110702，AF459454，AB645737，AB648911で構成されるクラスターの分岐にあるブートストラップ値はすべての解析法で94%以上の数値を得ており，信頼性あるクラスターと言える．クラスター内の詳細を見ると，AB025928とAF459454の分岐はすべての解析法で90%以上のブートストラップ値を得ており，信頼性があると言えよう．一方，AB110702，AB645737，AB648911はすべての方法で同様の樹形が得られているもののブートストラップ値は低い．AB056321に至っては，解析法によって系統位置が異なる．これらの正確な系統位置を知りたい時は再度詳細な解析が必要となる．前述のように，理想の系統樹を得るためには，常により信頼性の高い塩基配列情報を集めることが重要である．同種内でも1株の配列に依存せず，複数の菌株でアライメントをとり，より正確な塩基配列を取得することが望ましい．また，アウトグループの配置も重要であり，進化距離の遠近だけでなく，解析する塩基配列数が不用意に短くならないよう適切なアウトグループを選択することをお薦めする．アウトグループの選択には，広範囲な系統関係を理解することが理想であるが，それには経験が必要である．熟知するまでは，時間はかかっても塩基配列を精査して系統樹を作り直し，客観性をもって信頼性の高い樹形に近づけることである．原著や関連のサイトが数多くあるので，さらに細かい解析をしたい時はそちらも参照されたい．

文　献

1) Thompson, J. D. *et al.*: *Nucleic Acids Res.*, **24**, 4876 (1997).
2) Saitou, N. and Nei, M.: *Mol. Biol. Evol.*, **4,** 406 (1987).
3) Swofford, D. L.: PAUP*. Phylogenetic Analysis Using Parsimony (* and Other. Methods), Sinauer Associates (1998).
4) Felsenstein, J.: *J. Mol. Evol.*, **17**, 368 (1981).
5) Hasegawa, M. *et al.*: *J. Mol. Evol.*, **22**, 160 (1985).
6) Huelsenbeck, J. P. and Ronquist, F.: *Bioinformatics.*, **17**, 754 (2001).
7) Ronquist, F. and Huelsenbeck, J. P.: *Bioinformatics.*, **19**, 1572 (2003).
8) Perrière, G. and Gouy, M.: *Biochimie.*, **78**, 364 (1996).