← 結界Ｔ山門

2012年度　玉川大学「分子系統進化学」実習1

●はじめに：系統樹の「ことば」を読み解く●

オブジェクトの多様性を「系統樹」という図形言語を用いて記述することが，一般的な意味での「系統学（phylogenetics）」の出発点である．ポイントは，グラフとして描かれる系統樹がどのような情報を伝達しているかを理解した上で，系統樹を誤読（まちがった解釈）したりあるいは過剰解釈（伝えられていない架空の情報の深読み）をしないように気をつけることである．系統樹リテラシーの出発点は系統樹を正しく「読む」ことにほかならない．

今回の実習では，系統樹に含まれる「単系統群（monophyletic group）」の情報がどのように図式表現されているのかを理解するために，生物系統学で広く用いられている「Newick形式」（→解説記事：「系統推定の基本用語」／「The Newick tree format」）のコード化の方法を学び，「ことば」としての系統樹に関する基本概念を学習する．

Windows PC による実習に用いるソフトウェアは下記の２つで，いずれもフリーソフトウェアとして公開されている：

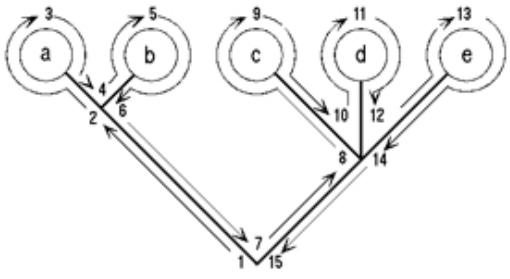
- PhyloWidget：「Web version」または「Standalone version」を用いる
- TreeView

なお，PhyloWidget は Java 言語で書かれたソフトウェアなので，Standalone version を用いる際は，事前に Java 計算環境を構築しておく必要がある．Java ホームページを参照し，自分のPCに Java がインストールされていない場合は，このサイトから適切なバージョンの Java をダウンロードして各自インストールされたい．Web version はインターネット接続されていればそのまま使える．

各ソフトウェアのインストールを確認した上で，次に進む．

●系統樹と Newick format（1）：単系統群の階層構造●

ことばとしての系統樹にはさまざまな情報を盛り込むことができる．この点で系統樹は図形言語としての自由度が高いといえるが，裏を返せばユーザー（読み手）のリテラシーが問われるということでもある．以下では，Newick format を用いて系統樹が伝達できる情報をひとつずつ付け加えることにより，ことばとしての性質を理解しよう．



まずはじめに，Newick format による系統樹のコード化について概説する．左図に示された系統樹は「根（root）」をもつ「有根系統樹（rooted tree）」の例である．グラフとして画像的に示されたこの系統樹のもつ情報のひとつとして，「単系統群」の構造に関する情報がある．単系統群とは「ある共通祖先に由来するすべての子孫から成る群」と定義さ

れる．そして，ある単系統群に含まれる子孫は互いに「姉妹群（sister group）」の関係にあると呼ばれる．

この図に含まれる単系統群は：

a+b
c+d+e
a+b+c+d+e

の三つだけである（それぞれの単系統群の共通祖先がどれかを確認されたい）．

いま，「根」を出発点として，この系統樹を時計回りに“一筆書き”しながらたどる作業を行なう．その際，きわめて単純な「コード化」の規則を下記のように与えよう．

- 共通祖先を往路で通過するときは「**(**」を，同じ共通祖先を復路で通過するときは「**)**」を付加する
- 末端の子孫をまわりこむときはその「**ラベル（名前）**」を付加する
- 共通祖先で折り返すときは「**,**」を付加する
- ひとまわりして根に戻ってきたならば最後に「**;**」を付加する

この四つの規則を上の系統樹に図中の番号にしたがって逐次的に適用すると，次のようになる：

```
1  (
2  ((
3  ((a
4  ((a,
5  ((a,b
6  ((a,b)
7  ((a,b),
8  ((a,b),(
9  ((a,b),(c
10 ((a,b),(c,
11 ((a,b),(c,d
12 ((a,b),(c,d,
13 ((a,b),(c,d,e
14 ((a,b),(c,d,e)
15 ((a,b),(c,d,e))
16 ((a,b),(c,d,e));
```

つまり，図の系統樹は `((a,b),(c,d,e));` という文字列によって一意的に記述されることになる．この文字列を Newick format による系統樹のコード化と呼ぶ．

次に，エディターソフト（たとえば「メモ帳」や「ノートパッド」）を起動し，下記のスクリプト：

```
#NEXUS
```

```
Begin trees;
tree test = ((a,b),(c,d,e));
End;
```

をテキストファイル（ファイル名：test.tre）として保存する．この形式のファイルを「NEXUS format」によるツリーファイルと呼ぶ．スクリプトの構造を補足説明すると：

```
#NEXUS                              [NEXUS format 宣言]
Begin trees;                        [ツリーブロック開始]
tree test = ((a,b),(c,d,e));        [Newick format によるツリー記述（ツリー名：test）]
End;                                [ツリーブロック終了]
```

となる．この NEXUS format でつくられたツリーファイルは，さまざまな系統推定ソフトウェアで共通に利用できる．たとえば，すでにインストールした〈TreeView〉を起動してこのツリーファイルを開くと，コード化された文字列をグラフとしての系統樹に図像化できることが確認できる．

さて，Newick format でコード化された系統樹をよく見ると，上で指摘した三つの単系統群 (a,b)，(c,d,e)，(a,b,c,d,e) がすべて含まれていることがわかる．つまり，Newick format の括弧 (と) は単系統群の分割を指定する書式であることがわかる．そして，コンマ , は末端の子孫の順列を指定する書式である．このように，系統樹に含まれる単系統群の階層構造は Newick format によって一意的な文字列として正確に表現されていることが理解できるだろう．

上では，Newick format による系統樹の単系統群構造のコード化について説明した．続いて，系統樹のもつそれ以外の情報について考えてみることにしよう．

Last Modified: 3 October 2012 by MINAKA Nobuhiro
