

「分子系統樹の統計推測」 （下平英寿）

アルゴリズムの説明

1 プログラム B1

1.1 概要

DNA 配列データから生物種間の距離行列を計算する．DNA 配列データの各行は，生物種の DNA 文字列 (A,G,C,T) の対応する遺伝子部分を横に並べたものである．DNA 配列データから２行ずつとりだして比較すれば，その２種の生物がどれくらい昔に共通祖先から枝分かれしたかを推定できる．これが「距離」であり，すべての生物種をペアワイズに比較して得られる行列が距離行列である．

長い時間をかけて DNA は少しずつ変化する．これを確率過程（マルコフ過程）で数学的にモデル化する．このモデルには様々なものが提案されているが，本演習では「HKY モデル」と呼ばれるものを採用して枝分かれの時間を推定する．

- 参考資料 1: Joseph Felsenstein, “Inferring Phylogenies,” Sinauer Associates, 2003 の pp.196-211 (Cahpter 13, Models of DNA evolution).
- 参考資料 2: 下平 “Transition probability of the HKY model,” hkymodel20051004.nb (Mathematica ノートブックファイル), hkymodel20051004.pdf (PDF).

1.2 入力

- 生物種の数: S (整数)
- DNA 配列の長さ: N (整数)
- DNA 配列データ: $[X_{sn}]$, $s = 1, \ldots, S$, $n = 1, \ldots, N$ （サイズ $S \times N$ の文字配列で要素は {A,G,C,T} のいずれか）

1.3 出力

- 生物種の数: S （整数）
- 距離行列: $[D_{ij}]$, $i = 1, \ldots, S$, $j = 1, \ldots, S$ （サイズ $S \times S$ の実数行列．対称 $D_{ij} = D_{ji}$ で対角要素はゼロ $D_{ii} = 0$）

1.4 アルゴリズム

1.4.1 ステップ 1：塩基頻度の計算

- $[X_{sn}]$ に含まれる各塩基 A,G,C,T の個数を計算し，C_i, $i = \mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}$ とする．
- 塩基頻度を $\pi_i = C_i/(S \times N)$, $i = \mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}$ で計算する．

1.4.2 ステップ 2：塩基遷移頻度の計算

すべての生物種のペア (s_1, s_2), $1 \leq s_1 < s_2 \leq S$ に対してステップ 2，3 を実行して D_{s_1,s_2} を計算する．簡単のため，$s_1 = 1, s_2 = 2$ と仮定して説明する．

- $i, j = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ のすべての組み合わせについて，$n = 1, \ldots, N$ のうち $X_{1n} = i$, $X_{2n} = j$ となる回数 C_{ij} を数える．(塩基遷移頻度)
- $(C_{ij} + C_{ji})/2$ を計算して，C_{ij} と C_{ji} に代入する．(対称化)

1.4.3 ステップ 3：最尤法による距離の推定

- 実数 $t > 0$ を時間，実数 $r > 1$ を「α/β 比」（後述）とする．2 個の実数パラメタ t, r の関数である対数尤度関数 $\ell(t, r; [C_{ij}], [\pi_i])$ を最大化するようにパラメタ t を定める．このときの t を $\hat{t}$ と書く．本演習では $r = 4$ と固定しておく．
- $D_{s_1,s_2} = D_{s_2,s_1} = t$ と代入する．

1.5 詳細

1.5.1 HKY モデルの対数尤度関数

- $i, j \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ とする．時刻 0 に状態 i にいるという条件のもとで時刻 t に状態 j にいる確率を $P_{ji}(t)$ と書く．これを並べた 4×4 の行列 $[P_{ji}(t)]$ を遷移確率行列と呼ぶ．
- 以下を定義しておく．「α/β 比」は $r = \alpha/\beta$ である．

$$\pi_{\mathrm{R}} = \pi_{\mathrm{A}} + \pi_{\mathrm{G}}, \quad \pi_{\mathrm{Y}} = \pi_{\mathrm{C}} + \pi_{\mathrm{T}}$$

$$\beta = \frac{1}{2r(\pi_{\mathrm{A}}\pi_{\mathrm{G}} + \pi_{\mathrm{C}}\pi_{\mathrm{T}}) + 2\pi_{\mathrm{R}}\pi_{\mathrm{Y}}}, \quad \alpha = r\beta$$

$$\alpha_{\mathrm{R}} = (\alpha - \beta)\pi_{\mathrm{R}}, \quad \alpha_{\mathrm{Y}} = (\alpha - \beta)\pi_{\mathrm{Y}}$$

- HKY モデルの遷移確率

$$P_{\mathrm{CA}}(t) = \pi_{\mathrm{C}}(1 - e^{-\beta t})$$

$$P_{\mathrm{GA}}(t) = \pi_{\mathrm{G}}(1 - e^{-\beta t}) + \frac{\pi_{\mathrm{G}}}{\pi_{\mathrm{R}}} e^{-\beta t}(1 - e^{-\alpha_{\mathrm{R}} t})$$

４種類の塩基 A,G,C,T は，似た性質をもつ二つのグループに分かれている．塩基 A と塩基 G はグループ R に属する．塩基 C と塩基 T はグループ Y に属する．グループ内の遷移確率は高く，グループ間の遷移確率は低い．一般に i と j が異なるグループに属すれば，

$$P_{ji}(t) = \pi_j(1 - e^{-\beta t})$$

である．一般に i と j が同じグループ k に属すれば ($i \neq j$ として)，

$$P_{ji}(t) = \pi_j(1 - e^{-\beta t}) + \frac{\pi_j}{\pi_k} e^{-\beta t}(1 - e^{-\alpha_k t})$$

である．遷移確率行列の対角成分 $i = j$ は

$$P_{ii}(t) = 1 - \sum_{j \neq i} P_{ji}(t)$$

より定まる．

- HKY モデルは「可逆」なマルコフ過程であり，時間軸を反転しても同じモデルになっている．遷移確率で言えば，$i, j \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$ に対して

  $$P_{ji}(t)\pi_i = P_{ij}(t)\pi_j$$

  となる．

- 対数尤度関数は

  $$\ell(t, r; [C_{ij}], [\pi_i]) = \sum_{i \in \{\mathrm{A,G,C,T}\}} \sum_{j \in \{\mathrm{A,G,C,T}\}} C_{ji} \log(P_{ji}(t)\pi_i)$$

2 プログラム B2

2.1 概要

生物種間の距離行列から進化系統樹を推定するアルゴリズムを実装する．このようなアルゴリズムは統計学一般では「階層型クラスタリング」とよばれ，また進化系統学では「距離行列法」と総称される．その中でも特に進化系統樹の推定精度が高いとされる近隣結合 (Neighbour-joining; 略して NJ) 法をつかう．

- 参考資料: Joseph Felsenstein, “Inferring Phylogenies,” Sinauer Associates, 2003 の pp.161-171 (Cahpter 11, Distance matrix methods).

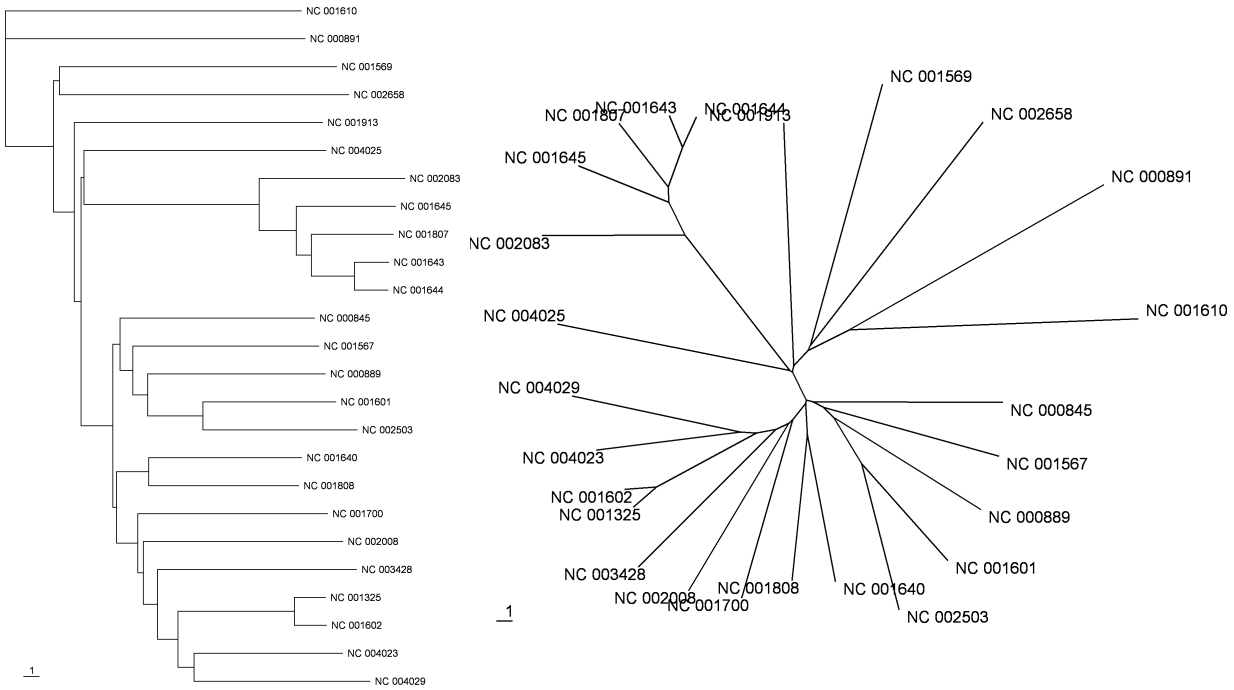
2.2 入力

- 生物種の数: S （整数）
- 距離行列: $[D_{ij}]$, $i = 1, \ldots, S$, $j = 1, \ldots, S$ （サイズ $S \times S$ の実数行列．対称 $D_{ij} = D_{ji}$ で対角要素はゼロ $D_{ii} = 0$）
- 生物種の名前: S 個の文字列だが，ここでの説明では $1, \ldots, S$ と番号で表現する．

2.3 出力

- 無根系統樹: T．つまり，入力された生物種の無根系統樹．各枝には枝長（枝分かれから次の枝分かれまでの時間），末端ノードには生物種の名前が入った木のこと．


```
(((((((((NC_001601:8.594,NC_002503:9.981):3.587,NC_000889:11.501):0.961,NC_001567:12.039):0.823,NC_000845:12.586):0.465,((NC_001640:9.901,
NC_001808:9.701):2.090,(((((NC_001325:2.001,NC_001602:2.063):7.544,(NC_004023:9.607,NC_004029:11.383):1.041):1.317,NC_003428:12.877):0.921,
NC_002008:12.899):0.361,NC_001700:12.278):1.384):0.233):2.062,(((((NC_001643:2.214,NC_001644:2.167):2.784,NC_001807:5.297):0.976,
NC_001645:6.422):2.410,NC_002083:9.421):11.340,NC_004025:15.670):0.190):0.434,NC_001913:16.065):1.354,(NC_001569:17.930,NC_002658:18.717):0.395):3.112,
NC_001610:19.162,NC_000891:19.399);
```

2.4 アルゴリズム

- (ステップ 1)　末端ノード (葉) のリスト L を作成．$|L| \geq 3$ とする．

$$L = \{1, 2, \ldots, S\}$$

- (ステップ 2)　次の量を $i \in S$ に対して計算．

$$u_i = \frac{1}{|L| - 2} \sum_{j \in L \setminus \{i\}} D_{ij}$$

- (ステップ 3)　$i, j \in S \quad (i \neq j)$ のうち，次の量を最小にするペアを探す．

$$D_{ij} - u_i - u_j$$

- (ステップ 4)　新たにノードを作成する．その名前を (ij) と書いておく．この新ノードから i, j までの距離（枝長）を次式で与える．

$$v_i = \frac{D_{ij} + u_i - u_j}{2}, \quad v_j = \frac{D_{ij} + u_j - u_i}{2}$$

- (ステップ 5)　新ノード (ij) から，i, j 以外のすべての末端ノード $k \in L \setminus \{i, j\}$ までの距離を次式で計算する．

$$D_{(ij)k} = \frac{D_{ik} + D_{jk} - D_{ij}}{2}$$

- (ステップ 6)　末端ノードのリスト L から i, j を取り除き，新ノード (ij) を追加する．

- (ステップ 7)　もし $|L| \geq 3$ ならばステップ 2 へ戻る．さもなければ，$L = \{\ell, m\}$ の 2 個を長さ $D_{\ell m}$ の枝でつないで木を完成する．

2.5 詳細

- もし距離行列が無根系統樹の各枝長を足し合わせて得られているならば，NJ 法はこの系統樹を正しく再構成する．
- ステップ 3 では，“neighbour” を選んでいる．どうしてこの方法で neighbour が得られるかは，ちょっと考えてみないと分からない．
- ステップ 5 が正しいことは，i, j, k の 3 点だけからなる木を考えるとすぐに理解できる．i, j, k を結びつけるノードは (ij) であるから，

  $$D_{ik} = D_{i(ij)} + D_{(ij)k}, \quad D_{jk} = D_{j(ij)} + D_{(ij)k}, \quad D_{ij} = D_{i(ij)} + D_{(ij)j}$$

  である．これよりステップ 5 の $D_{(ij)k}$ の式が得られる．

3 プログラム B3

3.1 概要

データにはランダムネスがあるから，推定された系統樹を１００%完全に信用してよいわけではない．ここではブートストラップ法をつかって，推定した系統樹を何パーセントくらい信用してよいかを見積もる．

- 参考資料: Joseph Felsenstein, “Inferring Phylogenies,” Sinauer Associates, 2003 の pp.335-342 (Cahpter 20, Bootstrap, jackknife, and permutation tests).

3.2 入力

- ブートストラップ法の反復数: B （整数）
- 上記に加えて，プログラム B1 への入力と同じもの

3.3 出力

- ブートストラップ法の反復数: B （整数）
- プログラム B2 の出力と同じものを B 個ならべたリスト

3.4 アルゴリズム

以下のステップ１，２，３を B 回反復する．

- (ステップ 1)　ブートストラップ標本 $[X^*_{sn}]$ を生成する．まず $\{1, 2, \ldots, N\}$ から各要素を等確率 $(1/N)$ でランダムに N 個選んで並べたものを $\{i_1, i_2, \ldots, i_N\}$ とする．その際同じ数が重複して選ばれることを許す．そして

$$X^*_{sn} = X_{si_n}, \quad n = 1, \ldots, N$$

  によって $[X^*_{sn}]$ を定義する．

- (ステップ 2)　$[X^*_{sn}]$ をプログラム B1 に入力して距離行列 $[D^*_{ij}]$ を計算する．

- (ステップ 3)　$[D^*_{ij}]$ をプログラム B2 に入力して系統樹 T^* を計算する．

- 上記を B 回繰り返して得られた B 個の系統樹を

$$T^*_1, T^*_2, \ldots, T^*_B$$

  とする．$[X_{sn}]$ から計算した系統樹 T と，ブートストラップ法で得られた B 個の系統樹を比較する．もし T が C 回含まれていれば，T のブートストラップ確率 (bootstrap probability; 略して BP) は C/B となる．

- S が大きい場合，この方法で計算した T の BP 値はとても小さくなることがある．T の BP 値よりも，T に含まれる各枝（クラスタ）の BP 値を計算することのほうが一般的である．