

系統樹作成アルゴリズムの性能評価実験

Experimental Performance Evaluations of Algorithms of Constructing a Phylogenetic Tree

<http://www.tani.cs.chs.nihon-u.ac.jp/g-2005/hisaka/>

谷 研究室 西村 久香

Hisaka NISHIMURA

概要

3つの手法で作成された塩基配列の距離表と2種類の距離法をそれぞれ組み合わせて用い、進化系統樹を推定し比較実験する。

1 はじめに

系統樹という、生物が祖先からその子孫へ突然変異により枝分かれしながら進化していく様子を木で表したものが有名であるが、その基本となるのはDNAの塩基配列やアミノ酸配列等のデータを用い、遺伝子の進化過程を木構造で表したものである。系統樹を作成する方法や、配列間の距離を求める方法がいくつか存在し、その組み合わせにより生成される系統樹に違いがでる場合がある。

そこで今回は、距離表作成法である『塩基配列の変換数を用いた距離』と2種類の『Kolmogorov 記述量を用いた距離』、一方は2001年にM. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang[1]らにより定義されたもので、もう一方は2003年にMing Li, Xin Chen, Xin Li, Bin Ma, Paul Vitanyi[2]で定義されたもの(前者を配列間の距離1とし、後者を配列間の距離2とする)、系統解析の方法の一つである距離法の『UP-GMA』と『近隣結合法』をそれぞれ組み合わせ、系統樹を作成し、比較実験を行う。

2 生物学からの準備

遺伝子

遺伝の法則を発見したメンデルが、実験結果を説明するために遺伝のもとになる粒子状の単位として仮定した要素のこと。遺伝形質を支配し、親から子へと伝えられる因子。

メンデルの時代には、遺伝子の実体は何であるか全く分かっていなかったが、今日では遺伝子の実体はDNA(デオキシリボ核酸)であることが解明された。

DNAの構造

塩基と糖(五炭糖)とリン酸からなるヌクレオチドがDNA構成単位で、DNAの糖はデオキシリボース、塩基はアデニン(A)・グアニン(G)・シトシン(C)・チミン

(T)の4種類である。

塩基配列

ここでは、DNAの塩基配列を指すので、A, G, C, Tの列のこと。

例

AAGTCCTG
GTCA
TAAAAAAGGGGCCCC
...等

系統樹

系統樹とは、生物間の進化関係を示す二次元グラフであり、配列の場合は、別個の生物から得られる特定の遺伝子間の進化を表す。

構成

系統樹は、分類単位を表す枝先(葉)と分類単位間を表す節(node)と枝からなっている。

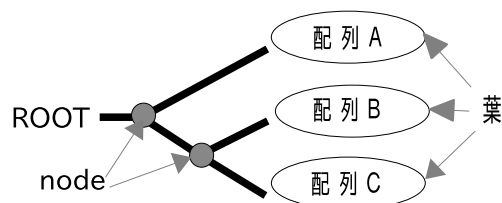


図 1: 系統樹

分類単位

系統樹上での系統学的に異なる単位として定義される。別個の生物から得られる配列のこと。

有根系統樹と無根系統樹

系統樹には有根と無根があり、作成方法により決まる。有根の代表的な距離法はUPGMAで無根の代表的な距

離法は近隣結合法である．有根と無根の違いは，以下の通り．

有根の場合

根の基部にある節がその他の全配列の共通祖先（最も古い祖先）であることを示している．図 1 が有根系統樹．

無根の場合

最も古い祖先の位置を明らかにしていない．

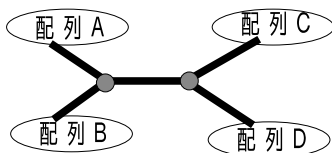


図 2: 無根系統樹

3 距離について

系統樹作成の際必要となる配列間の距離を例を使って示す．以下のような 4 つの配列が与えられたとする．

配列

配列 A : ACGCGTTGGGCGATGGCAAC

配列 B : ACGCGTTGGGCGACGGTAAT

配列 C : ACGCATTGAATGATGATAAT

配列 D : ACACATTGAGTGATAATAAT

配列 A と配列 B の間の距離をもとめる．

$$\begin{aligned} D_1(A, B) &= 1 - \frac{bzip(A) - bzip(A|B)}{bzip(AB)} \\ &= 1 - \frac{47 - 6}{53} \\ &= 0.226415 \end{aligned}$$

$$D_1(AB) = 0.226415$$

$$D_1(AC) = 0.245283$$

$$D_1(AD) = 0.192308$$

$$D_1(BC) = 0.211538$$

$$D_1(BD) = 0.192308$$

$$D_1(CD) = 0.176471$$

	A	B	C	D
A	—	0.226415	0.245283	0.192308
B	—	—	0.211538	0.192308
C	—	—	—	0.176471
D	—	—	—	—

距離表 1

Kolmogorov 記述量を用いた距離 2

配列間の距離 2

$$K(y) > K(x)$$

のとき

$$d_2(x, y) = \frac{K(xy) - K(x)}{K(y)}$$

で表される．

前節と同様計算ができないので，bzip を用いて計算可能な距離 D_2 を以下に示す．

$$d_2(x, y) = \frac{K(xy) - K(x)}{K(y)}$$

$$D_2(x, y) = \frac{bzip(xy) - bzip(x)}{bzip(y)}$$

配列 A と配列 B の間の距離をもとめる．

$bzip(A) = 47$, $bzip(B) = 47$, $bzip(AB) = 53$ より

$$\begin{aligned} D_2(A, B) &= \frac{bzip(AB) - bzip(A)}{bzip(B)} \\ &= \frac{53 - 47}{47} \\ &= 0.127659 \end{aligned}$$

Kolmogorov 記述量を用いた距離 1

配列間の距離 1

$$d_1(x, y) = 1 - \frac{K(x) - K(x|y)}{K(xy)}$$

で表される．

$K(x|y)$: y を用いたときの x の情報量

$K(x) = K(x|\epsilon)$: x の情報量 (ϵ は空の文字列)

$K(xy)$: x と y を結合した情報量

しかし，これでは計算ができないので，bzip を用いた計算可能な距離 D_1 を以下に示す．

$$d_1(x, y) = 1 - \frac{K(x) - K(x|y)}{K(xy)}$$

$$D_1(x, y) = 1 - \frac{bzip(x) - bzip(x|y)}{bzip(xy)}$$

ここで $K(x|y) \approx K(xy) - K(y)$

(y を用いた時の x の情報量) と (x と y を結合した情報量) - (y の情報量) は同値．なので

$$bzip(x|y) = bzip(xy) - bzip(y)$$

$$D_{2(AB)} = 0.127659$$

$$D_{2(AC)} = 0.148936$$

$$D_{2(AD)} = 0.106383$$

$$D_{2(BC)} = 0.127659$$

$$D_{2(BD)} = 0.106383$$

$$D_{2(CD)} = 0.106383$$

	A	B	C	D
A	—	0.127659	0.148936	0.106383
B	—	—	0.127659	0.106383
C	—	—	—	0.106383
D	—	—	—	—

距離表 2

塩基配列の変換数を用いた距離

配列間の距離 3

一方の配列をもう一方の配列に変換するのに必要なステップ数

配列 A と配列 B の間の距離をもとめる。

配列 A : ACGCGTTGGGCGA **T** GG **C** AA **C**

配列 B : ACGCGTTGGGCGA **C** GG **T** AA **T**

網掛け部分が配列 A と配列 B で異なる部分，すなわち変換する部分である。3箇所あるので，変換するのに必要なステップ数は3であり，これが配列 AB 間の距離である。他の配列についても同様に求める。

配列 i, j 間の距離を n_{ij} とすると，

$$n_{AB} = 3$$

$$n_{AC} = 7$$

$$n_{AD} = 8$$

$$n_{BC} = 6$$

$$n_{BD} = 7$$

$$n_{CD} = 3$$

(注： n_{ii} (変換する必要がないから 0)， $n_{ij}(=n_{ji})$ は表記しない。)

	A	B	C	D
A	—	3	7	8
B	—	—	6	7
C	—	—	—	3
D	—	—	—	—

距離表 3

4 UPGMA

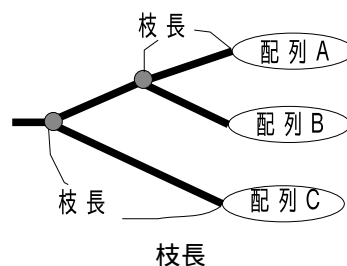
UPGMA 法 (The Unweighted Pair Group Method with Arithmetic Mean) は単純な系統作成法である。配列の対を作ったりクラスタリングを行うためのこの方法の変法も多数存在している。UPGMA 法は，最も近縁の配列間の枝長を計算することから始まり，この平均化された距離を用いて次のクラスタを作り，すべての配列が木に取り込まれるまで，この手順が続けられる。最後に，その系統樹の根の位置が推定される。

クラスタリング

関連のある大きな対象群の中から，よく類似した対称のセットをグループ化する。本発表の場合は近縁の配列を結合して，1つの配列とみなす行為。

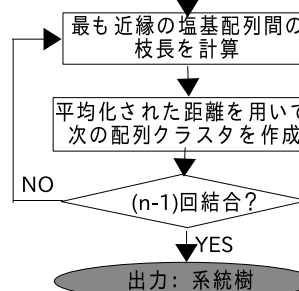
枝長

配列解析において，系統樹の特定の枝に沿って起こった配列の置換数。



UPGMA のアルゴリズム

入力：塩基配列，配列間の距離



UPGMA の流れ

STEP1

まず，各塩基配列間の距離が距離表として与えられる。その表から最小値を成す2配列を見つける。これらの配列が最も近縁の配列である。その最小値の平均が，最小値を持つ配列と共通祖先間の枝長となる。選んだ2つの配列は結合して1つの配列とみなす。

STEP2

表の更新をする。
結合しなかった配列から結合した 2 配列への距離の平均を最初に与えられた距離表から求める。その結果を新しい表に入れていく。

STEP3

塩基配列が結合することにより配列数が 1 つずつ減るので、

- ・ $n - 1$ 回繰り返したら → STEP4 へ
- ・ $n - 1$ 回未満なら → STEP1 へ

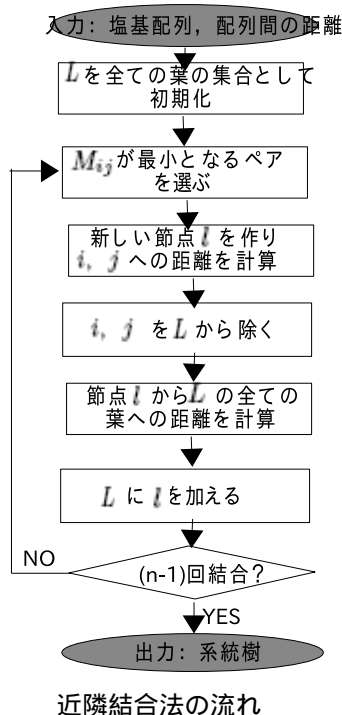
STEP4

近隣を正しく位置付け、できるだけ正確に元のデータを再現する枝長をもつ木を一つだけ同定できた。
系統樹を出力して終了。

5 近隣結合法

無根系統樹作成法の一つである。近隣結合法は個々の系統の進化速度が異なっているときに特に有効である。
(Saitou and Nei 1987)

近隣結合法のアルゴリズム

**STEP1**

- ・ L を全ての葉の集合として初期化
- ・ $r(i)$ をもとめる

$$r(i) = \sum_{k \in L, k \neq i}^{N-1} d(i, k)$$

STEP2

- ・ M_{ij} を計算

$$M(i, j) = d(i, j) - \frac{[r(i) + r(j)]}{N - 2}$$

- ・ M_{ij} が最小となるペア i, j を選ぶ

N : 葉の数 (塩基配列の数)

STEP3

- ・ 新しい節点 l を作る
- ・ 節点 l から i, j への枝長を計算

$$S(il) = \frac{d(ij)}{2} + \frac{[r(i) - r(j)]}{2(N - 2)}$$

$$S(jl) = d(ij) - S(il)$$

STEP4

- ・ L から i, j を除く
- ・ l から L に含まれる葉 q への距離 d を計算する

$$q \in L, q \neq i, j$$

$$d(ql) = \frac{d(iq) + d(jq) - d(ij)}{2}$$

- ・ l を一つの葉とみなす
- ・ l を L に追加

$$N = N - 1$$

STEP5

- ・ if ($N = 2$) → STEP6 へ
- ・ else → STEP1 へ

STEP6

L が 2 つだけになったら、その間に枝を張る。
系統樹を出力して終了。

参考文献

- [1]M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, An information-based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinformatics*, 17:2(2001), 149-154.
- [2]Ming Li, Xin Chen, Xin Li, Bin Ma, Paul Vitanyi. The similarity metric. In *Proceedings of the 14th annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pp. 863-872, Baltimore, Maryland , 2003.
- [3]David W.Mount (著) バイオインフォマティクスゲノム配列から機能解析へ, メディカル・サイエンス・

インターナショナル

- [4]Ming Li, Paul Vitanyi. An introduction to Kolmogorov complexity and its applications.2nd ed. *Texts and Monographs in Computer Science*. Springer-Verlag, 1997, XX, 637 p. 41 illus.
- [5] 渡辺 治, コンピュータ基礎理論ハンドブック : アルゴリズムと複雑さ, 丸善, 翻訳分担 : 3, 4 章, 1994.
- [6] 浅野 哲夫 (著), 小保方 幸次 (著) LEDA で始める C/C++ プログラミング入門からコンピュータ・ジオメトリまで, サイエンス社
- [7]<http://www.icp.ucl.ac.be/opperd/private/upgma.html>
- [8]<http://www.icp.ucl.ac.be/opperd/private/neighbor.html>