

置を決める段階になって初めてこの仮定をしていることで、一貫性がない。これより良い方法は、問題となっている種全部よりも前に分岐したと明確にわかっている種（アウトグループ）を規準種として追加して系統樹を作成することである。このようにすれば、この規準種と問題となっている種全部をつなぐ枝に根があることになる。また、別の方法として、遺伝子重複を起こしている遺伝子について系統樹では、遺伝子重複が起こった時点が根の位置となることがある (§4・1, §4・3 参照)。

最後に、“改 Farris 法は系統樹作成法の考えの一つの流れのなかに位置する”と書いたが、これについて少し述べてみたい。この考えとは、一つには枝を一つずつ結合していくことであり、あと一つは節約的 (parsimonious) なことである。ただ、もう気がついたかもしれないが、式 (16)~式 (18) に関する議論からわかるように、全所的な節約ではなく局所的な節約である。したがって、改 Farris 法には、全所的な節約を考慮するように、さらに改良できる余地がある。

なお、筆者が FORTRAN で書いた改 Farris 法のコンピュータプログラムがあるので、希望者は筆者まで連絡されたい。

23・5 近隣結合法*

無根系統樹において、一つの結節だけでつながった二つの OTU (操作的分類単位) のことを“近隣”とよぶ。図 23・13 で、OTU1 と OTU2 は一つの結節 A だけでつな

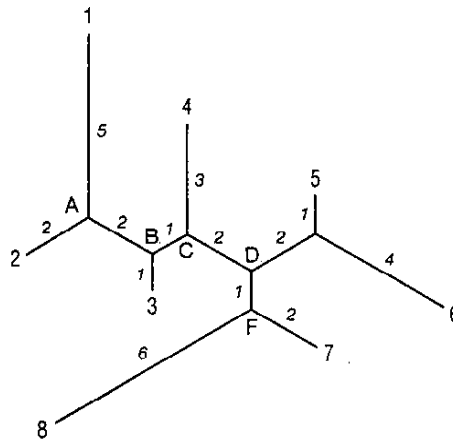


図 23・13 8 個の OTU から成る無根系統樹の例¹⁾

* 執筆担当：斎藤成也 (§23・5)

1) N. Saitou, M. Nei, *Mol. Biol. Evol.*, 4, 406 (1987).

新生化学実験講座 16 「分子進化実験法」, pp. 400-410
(五條堀・口野・西村・大沢共編) 東京化学同人, 1993年

がっているので近隣といえる。図 23・13 では、そのほかに 2 対の近隣 (OTU5 と OTU 6, OTU7 と OTU8) があるが、ほかはどれも近隣ではない。しかし、OTU1 と OTU2 を合体すると、この合体 OTU(1-2) は OTU3 と近隣になる。この近隣という概念を用いると、どのような樹形も近隣をつぎつぎに合体することによって規定することができる。図 23・13 の系統樹の樹形は、つぎのような近隣の対で規定することができる。[1, 2], [5, 6], [7, 8], [1-2, 3], [1-2-3, 4]。ここでは、もうひとつ [5-6, 7-8] という近隣も存在するが、これは近隣 [1-2-3, 4] と相補的であり、樹形の規定に必要ではない。いずれにせよ、これらの近隣をつぎつぎにみつけてゆけば、一つの樹形が得られることになる。一般に、 n 個の OTU から成る二分岐系統樹は、 $n-2$ 個の近隣でその樹形を規定することができる。

Saitou と Nei¹⁾ は、距離行列データから簡単なアルゴリズムを用いて段階的に近隣をみだし、最終的に 1 個の無根系統樹を得る方法を考案し、近隣結合法 (neighbor-joining method) と名づけた。この方法を用いると、OTU の数が 100 を超えても、短い計算時間で系統樹を得ることができる。しかも進化速度一定を仮定していないので、系統によって進化速度に違いがあっても、比較的正確に真の系統樹を復元することができる。この方法は、系統樹の枝の長さの総和を最小にする樹形を選ぶので、最小進化の原理を用いているが、そのアルゴリズムは、同じく最小進化の原理を形質状態データに用いる最大節約法 (§ 23・3 参照) とはかなり異なっている。

23・5・1 アルゴリズム

近隣結合法では、図 23・14 (a) にあるような星状の系統樹から出発する。最初は OTU 間の系統関係の情報がなにもないからである。この樹形を仮定して、すべての枝の長さの和 S_0 を距離行列データから計算してみよう。ここで、

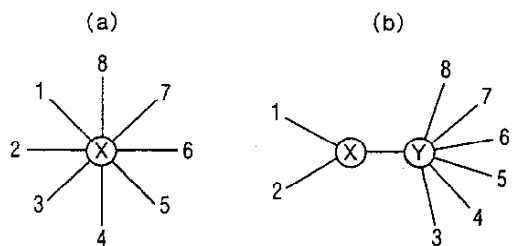


図 23・14 星状系統樹の例 (a) と OTU1 と OTU2 が近隣と仮定された無根系統樹 (b)¹⁾

1) N. Saitou, M. Nei, *Mol. Biol. Evol.*, 4, 406 (1987).

$$S_0 = \sum_{i=1} A_{iX} \quad (26)$$

A_{iX} は、図 23・14(a) の星状系統樹において、OTU i と結節 X のあいだの枝の長さである。一方、 D_{ij} を OTU i と OTU j のあいだの OTU 間距離とすると、

$$D_{ij} = A_{iX} + A_{jX} \quad (27)$$

という関係が仮定される。ところで、 n 個の OTU に対しては、 $n(n-1)/2$ 個の OTU 間距離が存在するが、それらを全部加えた量を Q としよう。すなわち、

$$Q = \sum_{i < j} D_{ij} \quad (28)$$

このとき、式 (26) と式 (27) の関係が仮定されるため、

$$Q = \sum_{i < j} (A_{iX} + A_{jX}) = (n-1)S_0 \quad (29)$$

となるので、

$$S_0 = \frac{Q}{n-1} \quad (30)$$

となる。

つぎに、図 23・14(b) の樹形を考えてみよう。これは、特定の一对の OTU (1 と 2) を取出し、それらが近隣となる仮定したものである。その全枝長の和 S_{12} は、

$$S_{12} = (B_{1X} + B_{2X}) + B_{XY} + \sum_{i=3} B_{iY} \quad (31)$$

で与えられる。 B_{ij} は、OTU (または結節) i と j 間の距離である。ここで、図 23・14(b) の樹形を仮定しているので、以下の関係が成立する。

$$D_{12} = B_{1X} + B_{2X} \quad (32a)$$

$$D_{1i} = B_{1X} + B_{XY} + B_{iY} \quad (i \geq 3) \quad (32b)$$

$$D_{2i} = B_{2X} + B_{XY} + B_{iY} \quad (i \geq 3) \quad (32c)$$

$$D_{ij} = B_{jX} + B_{jY} \quad (i, j \geq 3) \quad (32d)$$

式 (32d) の関係は式 (27) と同等なので、式 (30) を適用して、

$$\sum_{i=3} B_{iY} = \frac{\sum_{3 \leq i < j} D_{ij}}{n-3} \quad (33)$$

となる。式 (32a) と式 (33) を式 (31) に代入して、

$$S_{12} = D_{12} + B_{XY} + \frac{\sum_{3 \leq i < j} D_{ij}}{n-3} \quad (34)$$

B_{XY} を OTU 間距離で表すため、ここで再び Q を考えてみよう。式 (32a)~式 (32b) の関係が成り立つので、

$$\begin{aligned}
 Q &= (n-1)(B_{1x} + B_{2x}) + 2(n-2)B_{xy} + (n-1) \sum_{i=3} B_{ix} \\
 &= (n-1)D_{12} + 2(n-2)B_{xy} + \frac{(n-1) \sum_{3 \leq i < j} D_{ij}}{n-3} \quad (35)
 \end{aligned}$$

よって,

$$B_{xy} = \frac{Q - (n-1)D_{12} - \frac{(n-1) \sum_{3 \leq i < j} D_{ij}}{n-3}}{2(n-2)} \quad (36)$$

式(36)を式(34)に代入すると,

$$S_{12} = \frac{(n-3)D_{12} + Q + \sum_{3 \leq i < j} D_{ij}}{2(n-2)} \quad (37)$$

ここで, $R_i = \sum_{j=1} D_{ij}$ とすると,

$$\sum_{3 \leq i < j} D_{ij} = Q - (R_1 + R_2 - D_{12}) \quad (38)$$

という関係が成り立つので, 式(37)に代入して,

$$S_{12} = \frac{D_{12}}{2} + \frac{2Q - R_1 - R_2}{2(n-2)} \quad (39)$$

したがって, 一般に, OTU i とOTU j を近隣に仮定すると, そのときの枝長の総和 S_{ij} は,

$$S_{ij} = \frac{D_{ij}}{2} + \frac{2Q - R_i - R_j}{2(n-2)} \quad (40)$$

という一般式で表すことができる。このように簡単な式なので, S_{ij} の計算は, OTUの数が比較的少なければ, コンピュータープログラムの助けを借りなくても, 電卓を用いて容易に行うことができる。

さて, OTU i とOTU j が近隣であるとき, はたして S_{ij} は S_0 よりも小さくなっているであろうか。ここで, 話を簡単にするため, 図23・14(b)の樹形が正しいとする。このとき, 式(30)と式(35)より,

$$\begin{aligned}
 S_0 &= \frac{Q}{n-1} \\
 &= D_{12} + \frac{2(n-2)B_{xy}}{n-1} + \frac{\sum_{3 \leq i < j} D_{ij}}{n-3} \quad (41)
 \end{aligned}$$

よって, 式(34)と式(41)から,

$$S_0 - S_{12} = \frac{(n-3)B_{xy}}{n-1} \quad (42)$$

図23・14(b)の樹形が正しいと仮定しているので, $B_{xy} > 0$ である。したがって, S_{12}

は S_0 よりも小さいことが証明された。

実際には、どの OTU の対が近隣であるかわからないので、 $n(n-1)/2$ 個のすべての OTU の対について、図 23・14 (b) のような樹形を考え、それに対応する枝長の総和 S_{ij} を計算する。そうして、最小の S_{ij} を与える OTU の対を近隣とみなすのである (最小進化の原理)。もちろん、実際には最小の S_{ij} を示す OTU の対が近隣ではない可能性がある。しかし、用いられた距離が相加的である (系統樹において、任意の 2 個の OTU を結ぶすべての枝の長さを加えると必ず OTU 間距離に等しい場合) ときには、この方法は必ず近隣を見いだすことがわかっている^{1),2)}。また、距離が必ずしも相加的ではなくても、正しい系統樹では S_{ij} の期待値が最小になることが、最近 4 個の OTU から成る系統樹について示されている³⁾。

いずれにせよ、最小の S_{ij} を与える OTU の対が見つかり、つぎにそれらを合体して一つの OTU にし、またうえの操作を繰返すのである。OTU $_i$ と OTU $_j$ が合体した合体 OTU($i-j$) と他の OTU $_k$ との進化距離 $D_{(ij)k}$ はつぎのように単純平均を用いて計算する。

$$D_{(ij)k} = \frac{D_{ik} + D_{jk}}{2} \quad (43)$$

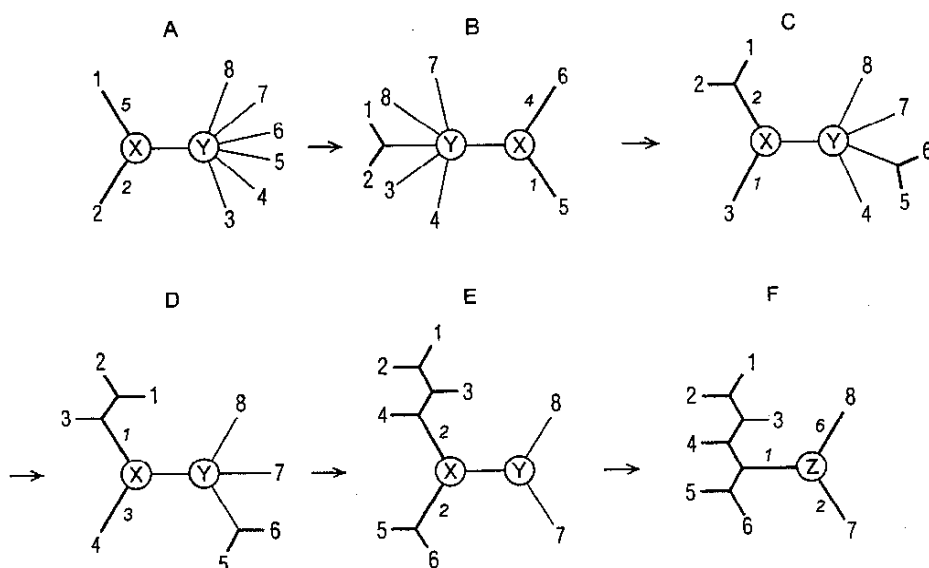


図 23・15 近隣結合法により段階的に系統樹が作成されてゆくようす¹⁾

- 1) N. Saitou, M. Nei, *Mol. Biol. Evol.*, 4, 406 (1987).
- 2) J. A. Studier, K. J. Keppler, *Mol. Biol. Evol.*, 5, 729 (1988).
- 3) A. Rzhetsky, M. Nei, *J. Mol. Evol.*, 35, 367 (1992).

これらの進化距離を含めて、新しい $(n-1)$ 個の OTU 間の距離行列ができる。この行列についてうえと同じ計算を行えばつぎの近隣を決めることができる。この繰返し操作によりつぎつぎに近隣が見つかるので、それによって最終的な系統樹を作成することができる。図 23・15 は、そのようにして、表 23・3 で与えられている距離行列 (図 23・13 の系統樹から相加性を仮定して作成した) をもとに、図 23・13 の系統樹が復元されてゆくようすを示したものである。

表 23・3 図 23・12 の系統樹をもとにした距離行列 (下三角行列), S_{ij} の値 (上三角行列), および R_i の値

OTU	R_i	OTU							
		1	2	3	4	5	6	7	8
1	85	—	36.67	38.33	39.00	40.33	40.33	40.17	40.17
2	67	7	—	38.33	39.00	40.33	40.33	40.17	40.17
3	53	8	5	—	38.67	40.00	40.00	39.83	39.83
4	63	11	8	5	—	39.67	39.67	39.50	39.50
5	59	13	10	7	8	—	37.00	38.83	38.83
6	77	16	13	10	11	5	—	38.83	38.83
7	61	13	10	7	8	6	9	—	37.67
8	85	17	14	11	12	10	13	8	—

近隣結合法で枝長を決めるには、Fitch-Margoliash 法¹⁾を用いる。この場合、図 23・14(b) のように OTU1 と OTU2 が最初に結合すると仮定すれば、枝 B_{1x} と枝 B_{2x} はつぎのようにかくことができる。

$$B_{1x} = \frac{D_{12}}{2} + \frac{R_1 - R_2}{2(n-2)} \quad (44a)$$

$$B_{2x} = \frac{D_{12}}{2} - \frac{R_1 - R_2}{2(n-2)} \quad (44b)$$

このようにして与えられる B_{1x} と B_{2x} は、図 23・14(b) が正しい系統樹であるときには最小二乗推定値である²⁾。また、進化距離がすべて相加的である場合には、この方法は、最終的に得られる二分岐系統樹についてもすべての枝について正しい枝長を与える。

23・5・2 計算例

表 23・3, 表 23・4 と図 23・15 を用いて、近隣結合法の具体的な計算例を示そう。

1) W.M. Fitch, E. Margoliash, *Science*, 155, 279 (1967).

2) N. Saitou, M. Nei, *Mol. Biol. Evol.*, 4, 406 (1987).

表 23・3 の下三角行列の距離 D_{ij} の総和をとると、 $Q=275$ となる。したがって、式 (30) から $n=8$ なので、 $S_0=39.29$ が得られる。つぎに、 R_i を計算するが、その結果

表 23・4 OTU1 と OTU2 を合体させたあとの距離行列 (下三角行列)、 S_{ij} の値 (上三角行列)、および R_i の値

OTU	R_i	OTU						
		1-2	3	4	5	6	7	8
1-2	69	—	31.50	32.30	33.90	33.90	33.70	33.70
3	46.5	6.5	—	32.30	33.90	33.90	33.70	33.70
4	53.5	9.5	5	—	33.70	33.70	33.50	33.50
5	47.5	11.5	7	8	—	31.30	33.10	33.10
6	62.5	14.5	10	11	5	—	33.10	33.10
7	49.5	11.5	7	8	6	9	—	31.90
8	69.5	15.5	11	12	10	13	8	—

は表 23・3 の左端カラムに示してある。これらの情報から、すべての OTU の対についてそれに対応する枝長の総和 S_{ij} を計算し、その結果が表 23・3 の上三角行列に示してある。たとえば、OTU1 と OTU2 に対する枝長の総和 S_{12} は、式 (39) を用いて、 $S_{12}=7/2+(2 \times 275-85-67)/(2 \times 6)=36.67$ が得られる。実際、すべての S_{ij} のなかでいちばん小さいのはこの S_{12} である。したがって、OTU1 と OTU2 を近隣とする。図 23・13 からわかるように、これらの OTU は真の近隣の一つである。また、OTU1 と OTU2 の結節 X からの枝長は、式 (44a) と式 (44b) を用いて、それぞれ 5 と 2 になる。この枝長の推定値も真の値と同一である (図 23・13 参照)。ここで注意しておきたいが、 S_{ij} が S_0 より小さいからといって、OTU i と OTU j が近隣であるとは限らないことである。たとえば、 $S_{13}=38.33$ は $S_0(=39.29)$ より小さいが、OTU1 と OTU3 は近隣ではない (図 23・13 参照)。一方、 S_{15} や S_{28} のように S_0 より大きな S_{ij} もある。

つぎに、OTU1 と OTU 2 を合体し、式 (43) を用いてこの合体 OTU(1-2) と他の OTU との距離を計算する。これらの新しい距離と、他の OTU 間の距離をいっしょにして、新しい $n=7$ の距離行列が得られる (表 23・4 の下三角行列)。この新しい距離行列について、うえと同じ計算を繰返すと、つぎに近隣として OTU5 と OTU6 が得られる (表 23・4 上三角行列の S_{ij} の値を参照)。それらと結節 X とのあいだの枝長として、それぞれ 6 と 2 が得られる (図 23・15 参照)。これらも真の値と同一である (図 23・13 参照)。このような計算を繰返すと、図 23・15 に示したように、つぎつぎと近隣が決まってゆき、第 6 ステップで、最終的に図 23・13 と同一の樹形と枝長をもつ系統

樹が得られる。一般に、 n 個の OTU を比較すると、最終の系統樹を得るのに $n-2$ 個のステップが必要である。

うえの例において、真の樹形と真の枝長が完全に復元されたのは、用いられた距離が相加的であったからである。実際のデータは相加的でないのが普通なので、得られた系統樹が必ずしも正しいとは限らない。しかし計算機シミュレーションによると、近隣結合法は他の方法に比べて真の樹形を復元する効率の高いことが知られている^{1)~3)}。

23・5・3 得られた系統樹の統計的検定法

近隣結合法によって得られた系統樹の信頼性を調べるには、大きくわけて、二つの方法がある。第1の方法は、得られた系統樹の内部にあって系統樹の樹形を決めている枝の長さの標準誤差を計算し、それらが有意かどうかを調べる方法である。Li⁴⁾ および Tajima⁵⁾ は、OTU が5個以下である場合についてこの標準誤差の計算方法を発表している。図23・16はミトコンドリア DNA の塩基配列データから近隣結合法で系統樹を作成して、標準誤差を計算したものである⁵⁾。[ヒト, チンパンジー] と [ゴリラ, オランウータン] の近隣を結ぶ枝の長さが 0.0038 ± 0.0016 なので、この枝は統計的に有意である。この値は、塩基置換数の推定に1-パラメーター法を用いた場合だが、トランジ

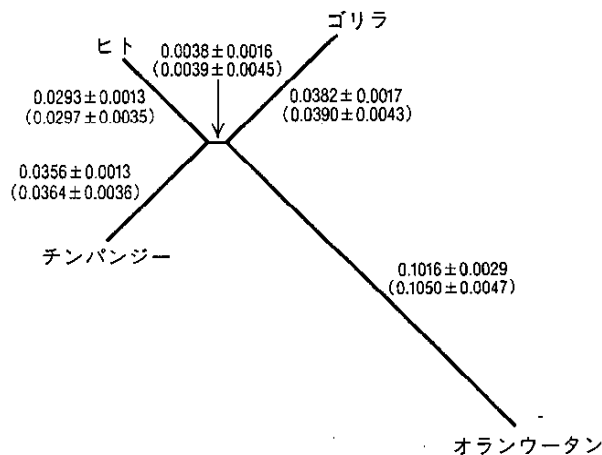


図 23・16 近隣結合法で得られた系統樹の各枝に標準誤差を与えた例⁵⁾

- 1) N. Saitou, M. Nei, *Mol. Biol. Evol.*, 4, 406 (1987).
- 2) J. Sourdis, M. Nei, *Mol. Biol. Evol.*, 5, 298 (1988).
- 3) N. Saitou, T. Imanishi, *Mol. Biol. Evol.*, 9, 514 (1989).
- 4) W.-H. Li, *Mol. Biol. Evol.*, 6, 424 (1989).
- 5) F. Tajima, *Mol. Biol. Evol.*, 9, 168 (1992).

ションとトランスポージョンの違いを考慮した2-パラメーター法を用いると 0.0039 ± 0.0045 となり、標準誤差が大きくなるために、統計的に有意ではなくなる（塩基置換数の推定については第21章を参照）。

もっとも、OTUが5個以上の場合には、これらの方法に従って標準誤差を計算するのはかなり面倒になってくる。したがって、多数のOTUから成る系統樹の場合、興味のある枝に関係する4個のOTUのみを考えて、その枝の長さの標準誤差を計算する方法が考えられる。たとえば、図23・13で枝CDを考えると、OTU3, 4, 5, 7のみを考慮すればよい。一方、RzhetskyとNei¹⁾は、最終的な系統樹の枝の長さを最小二乗法で推定し、各枝に標準誤差を与える方法を開発している。

第2の方法は、ブーツストラップを用いるものである。ブーツストラップ法については、§24・1の解説を参照されたい。ブーツストラップ系統樹は、もとの系統樹に比べて、一部の分岐パターンだけ異なっていることが多い。したがって、この方法では樹形の部分的な分岐パターンが正しいかどうかを問題にすることが多い。たとえば、図23・17²⁾は脊椎動物の系統樹を臍臓由来のタンパク質のアミノ酸配列をもとに近隣結合法で作成したものであるが、一部の枝に数字がかいてある。これは、2000個のブーツストラップ系統樹をつくり、そのなかの何パーセントが図23・17の系統樹の樹形を復元するかを調べたものである。95%以上の頻度でもとの系統樹の分岐パターンが得

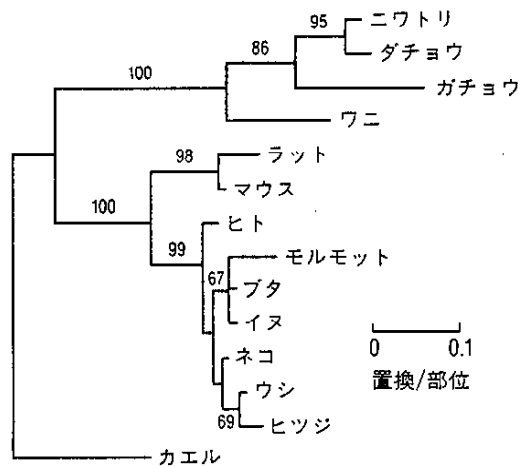


図 23・17 近隣結合法にブーツストラップ法を応用した例²⁾

- 1) A. Rzhetsky, M. Nei, *Mol. Biol. Evol.*, 9, 945 (1992).
- 2) S. B. Hedges, L. R. Maxson, *Mol. Biol. Evol.*, 8, 888 (1991).

られている場合、それはかなり信頼できると考えられる。逆に、それよりずっと小さい数字の場合（図 23・17 でいうと、哺乳類のなかの系統関係で 67 や 69 というパーセントしかない枝）、このデータからははっきりしたことはいえないことになる。

近隣結合法は最終的に一つの系統樹が得られる。しかし、近隣結合法で得られた樹形は、たいていの場合、最小進化法によって得られるものと同一であることがわかっている¹⁾。また、同一のデータを用いれば、どの系統樹作成法を用いても似かよった系統樹が得られる場合が多い（文献²⁾を参照）。したがって、最小進化法、最大節約法、あるいは最尤法を用いる場合でも、近隣結合法による系統樹をまず求めておいて、その樹形の周辺を調べるという方法をとることが合理的であろう。

23・5・4 プログラムパッケージ

1) 筆者の作成したプログラムパッケージ：PASCAL でかいた日本語版 MS-DOS でうごくもの。すでにアラインメントされている塩基配列データから塩基置換数を推定し、近隣結合法で系統樹を作成する。筆者の連絡先はつぎのとおり。国立遺伝学研究所 進化遺伝研究部門。〒411 三島市谷田 1111。電子メール：nsaitou@genes.nig.ac.jp。

2) Whittam の作成した NJBOOT プログラムパッケージ：FORTRAN でかかれており、配列から出発して、ブーツストラップ法が使える。Dr. Thomas Whittam の連絡先はつぎのとおり。Institute of Molecular Evolutionary Genetics, Pennsylvania State University, State College, Pennsylvania, U. S. A. 電子メール：tsw1@psuvm.psu.edu。

3) Felsenstein の作成したプログラムパッケージ PHYLIP 中の NEIGHBOR というプログラムは、近隣結合法で系統樹を作成する。PHYLIP のプログラムはすべて PASCAL でかかれており、ソースコードを配布しているが、最近マッキントッシュ版も配布を始めた。Dr. Joseph Felsenstein の連絡先はつぎのとおり。Department of Biology, University of Washington, Seattle, U. S. A. 電子メール：joe@genetics.washington.edu。

4) Higgins の作成したプログラムパッケージ CLUSCAL は、配列データを多重アラインメントしたあと、塩基置換数を推定して距離行列を作成し、そこから平均距離法 (§ 23・2 参照) または近隣結合法で系統樹を作成する。Dr. Desmond Higgins の連

1) N. Saitou, T. Imanishi, *Mol. Biol. Evol.*, 6, 514 (1989).

2) N. Saitou, "Handbook of Statistics", ed. by R. Chakraborty, C.R. Rao, p.317, Elsevier Science Publishers B.V, Amsterdam (1991).

絡先はつぎのとおり. EMBL Data Library, Heidelberg, Germany. 電子メール: higgins@embl-Heidelberg.DE.

5) 伊奈康夫の作成したプログラムパッケージ ODEN 中の treenj というプログラムは, 近隣結合法で系統樹を作成する. このためには塩基配列データから塩基置換数を推定して距離行列を作成する必要があるが, これについても, 多種類の方法が使えるようになってきている. さらに, Hein の作成した多重アラインメント (多数配列の同時整列) のプログラムも含まれており, 使いやすくなっている. ODEN のプログラムは, ほとんど C でかかっている. ODEN は, 現在のところ, 国立遺伝学研究所の計算機にアクセスして使うしかないが, 近いうちにプログラムパッケージの配布を始めるそうである. 伊奈康夫氏の連絡先はつぎのとおり. 国立遺伝学研究所遺伝情報研究センター. 〒411 三島市谷田 1111. 電子メール: yina@dbj.nig.ac.jp.

23・6 最尤法*

最尤法はデータの生じてくる背景について仮説をたて, そのもとで構造を統計的に推定し, 一方で仮説の妥当性を検証していく方法である. 仮説を表す統計モデルは, データの生じる確率 (尤度) を未知パラメーターの関数として表現した尤度関数の形で具体化される. この尤度関数を最大にするように, データから未知パラメーターを推定する.

23・6・1 最小二乗法と最尤法

最尤法を具体的に論ずるために, x と y の 2 変量からなるデータ $(x_i, y_i), i=1, \dots, n$ に対して直線を当てはめる問題を考える. すなわちこの場合の統計モデルは, 従属変数 y が説明変数 x と平均的には直線関係にあるというもので,

$$x_i = a + bx_i + \epsilon_i, i = 1, \dots, n$$

と表される. ϵ_i は誤差項で, 測定誤差やその他さまざまな要因を総合したものである. 平均は 0 である.

ところでこれらが互いに独立で, さらに分散 σ^2 の正規分布に従っているようなときは, データ y_1, \dots, y_n が得られる尤度は

$$L(a, b, \sigma^2 | y_1, \dots, y_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - a - bx_i)^2 \right\}$$

と表される. 計算上の理由から, 通常は尤度を最大にする代わりにその対数をとった対数尤度を最大にする. したがっていまの場合は, 回帰係数 a, b の最尤推定量 (最尤法

* 執筆担当: 岸野洋久 (§23・6)

