# 保全生態学特論

## ― 系統推定法入門 ―

噺役


三中 信宏

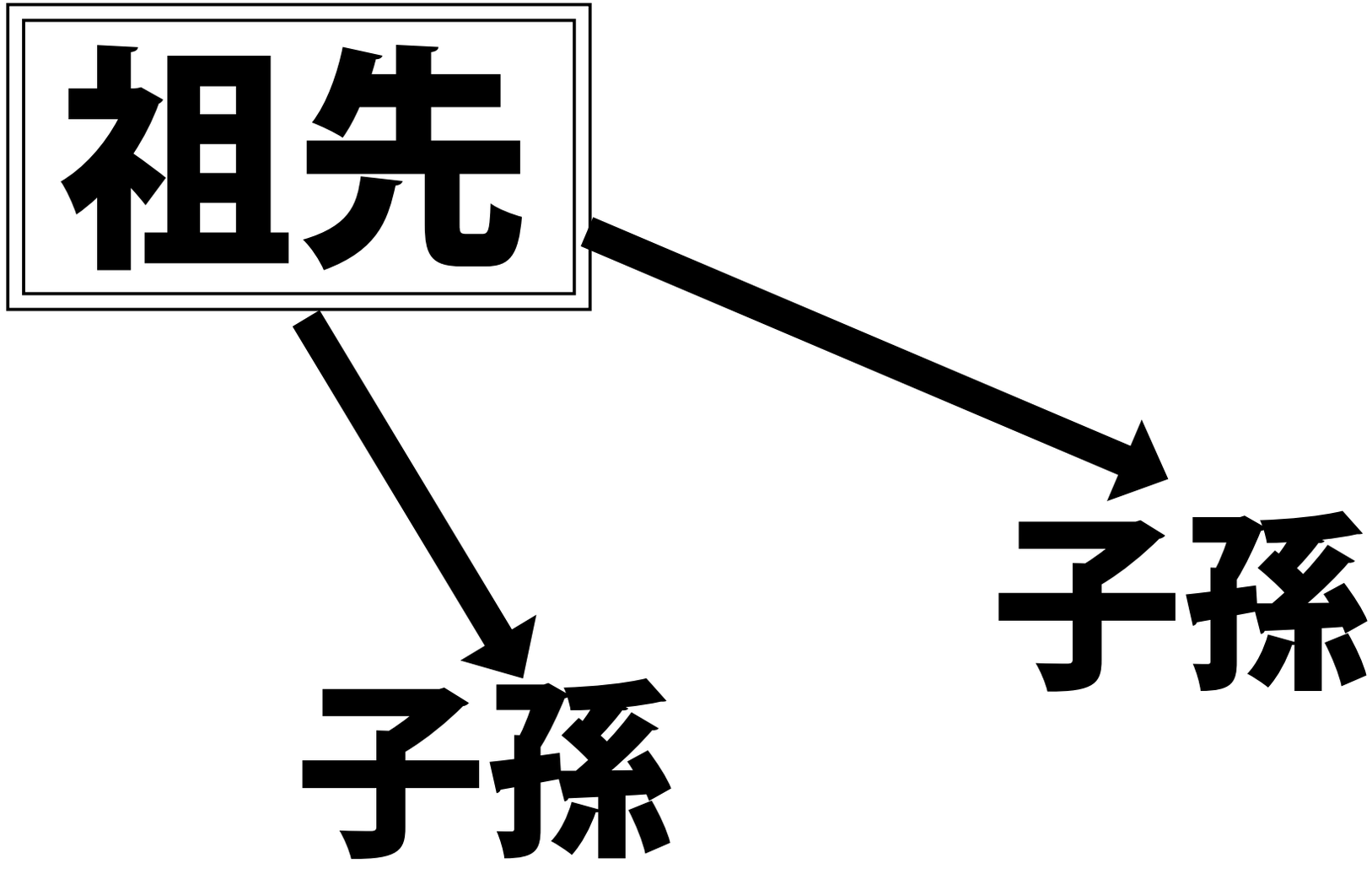
**MINAKA Nobuhiro**

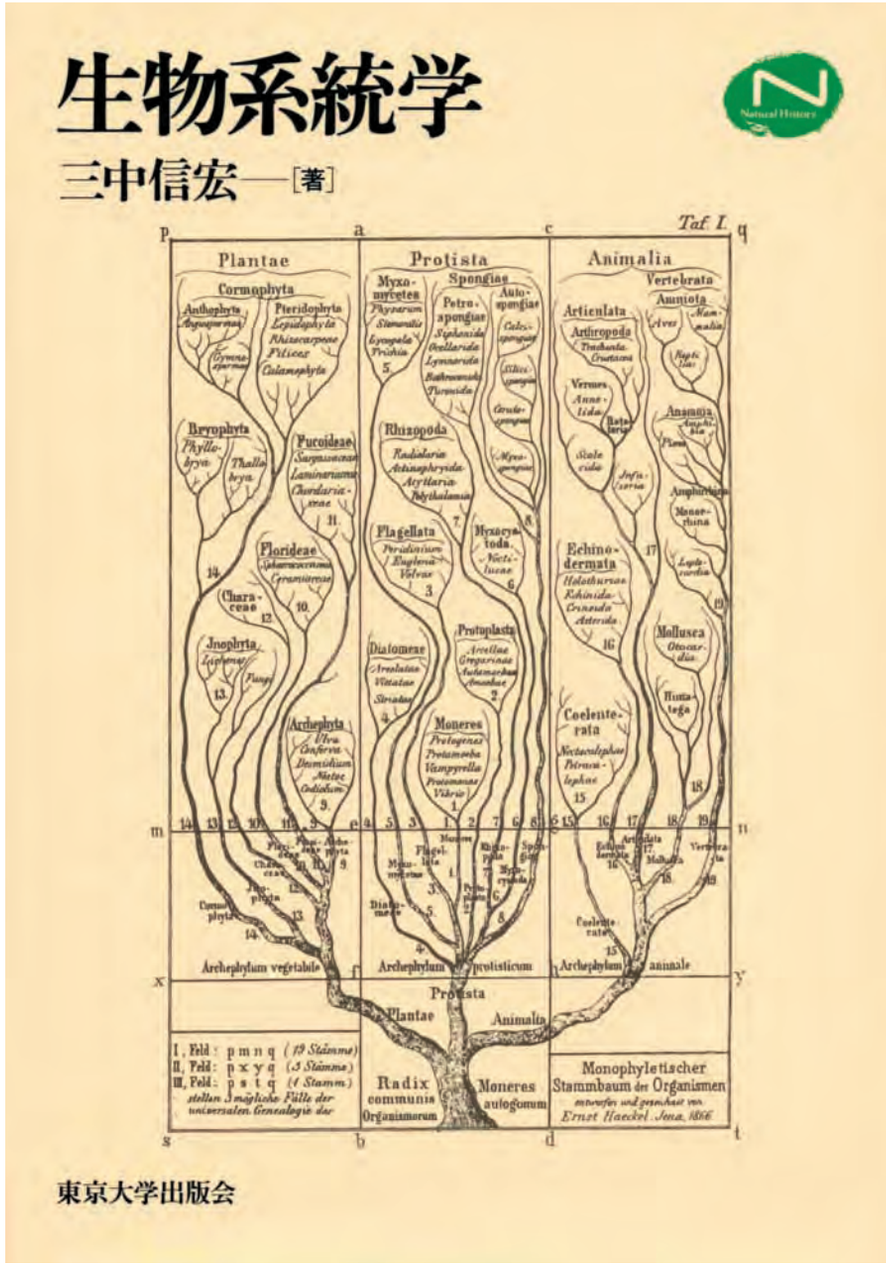
国立研究開発法人 農研機構 農業環境変動研究センター ユニット長［生物統計学］
東京大学大学院 農学生命科学研究科 生物・環境工学専攻 教授［生態系計測学］
東京農業大学大学院 農学研究科 客員教授［応用昆虫学］

**mail: minaka@affrc.go.jp**
**website: http://leeswijzer.org/**
**twitter: @leeswijzer**

# 系統

総合研究大学院大学講義〈先導科学考究〉
三中信宏：「進化系統学の先端で甦る歴史研究」
（22 May 2007，総研大，葉山）


祖先
子孫
子孫

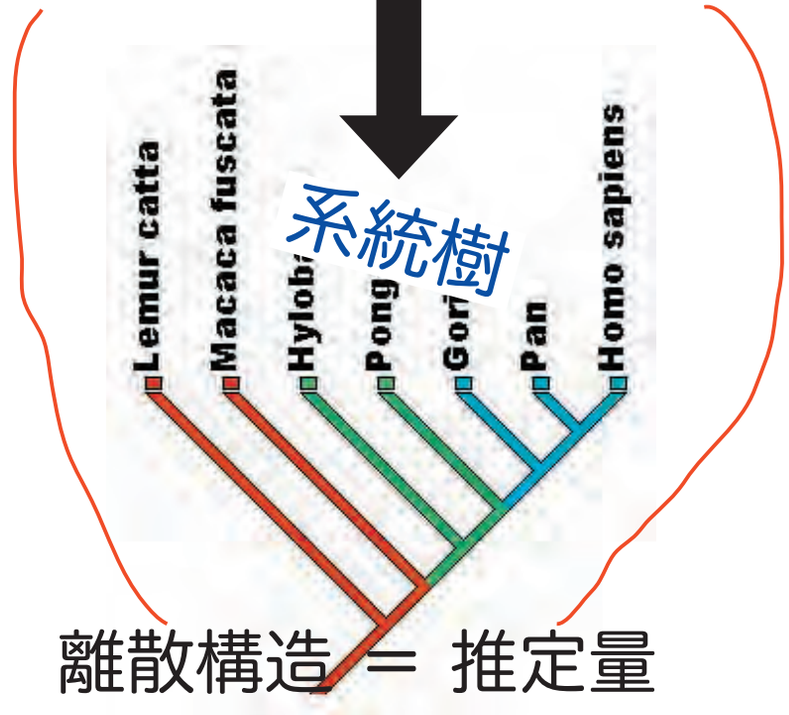


数学的問題

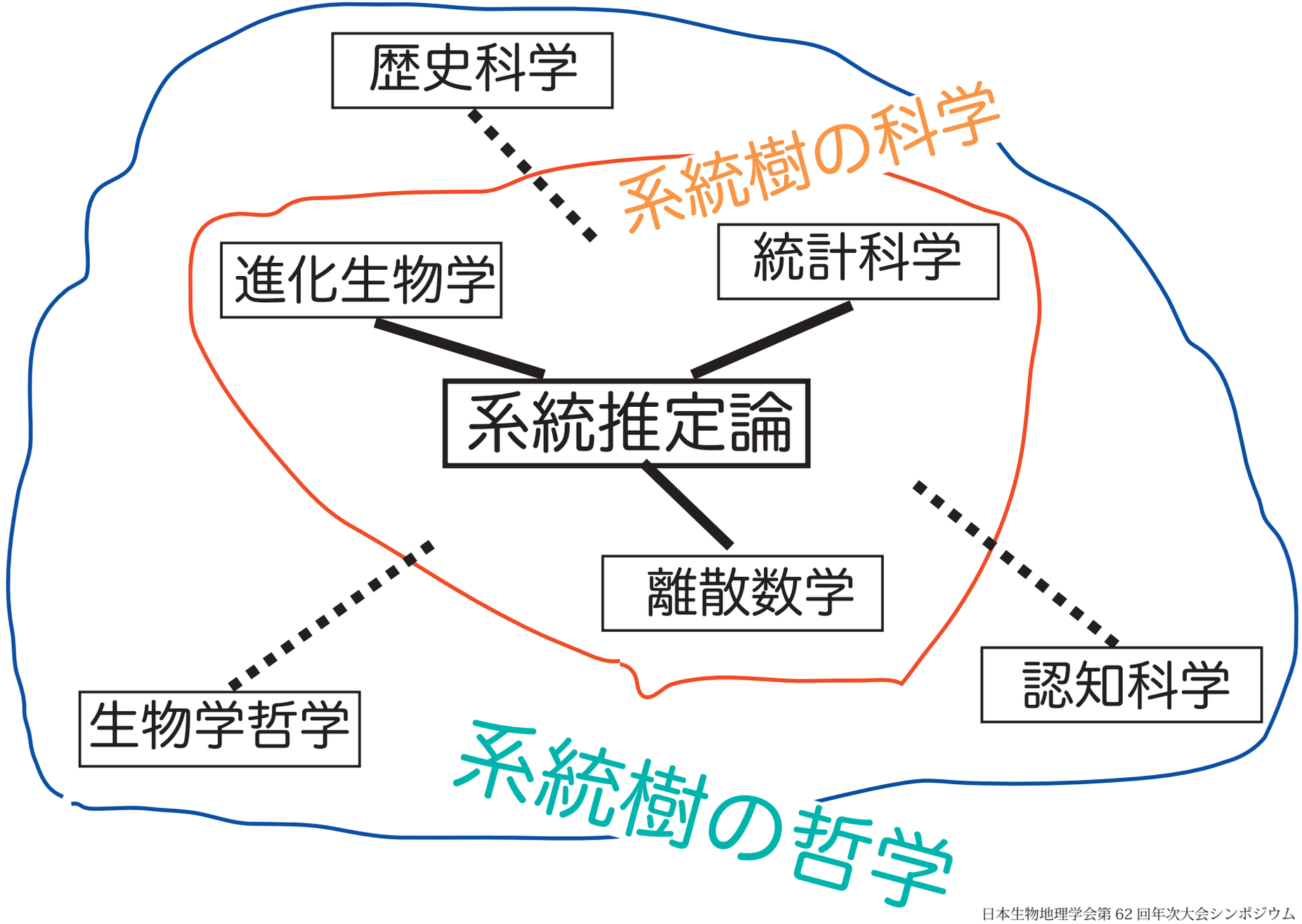
スタイナー問題
グラフ理論，etc.

統計学的問題

誤差，信頼性，
モデル選択 etc.




離散構造 = 推定量

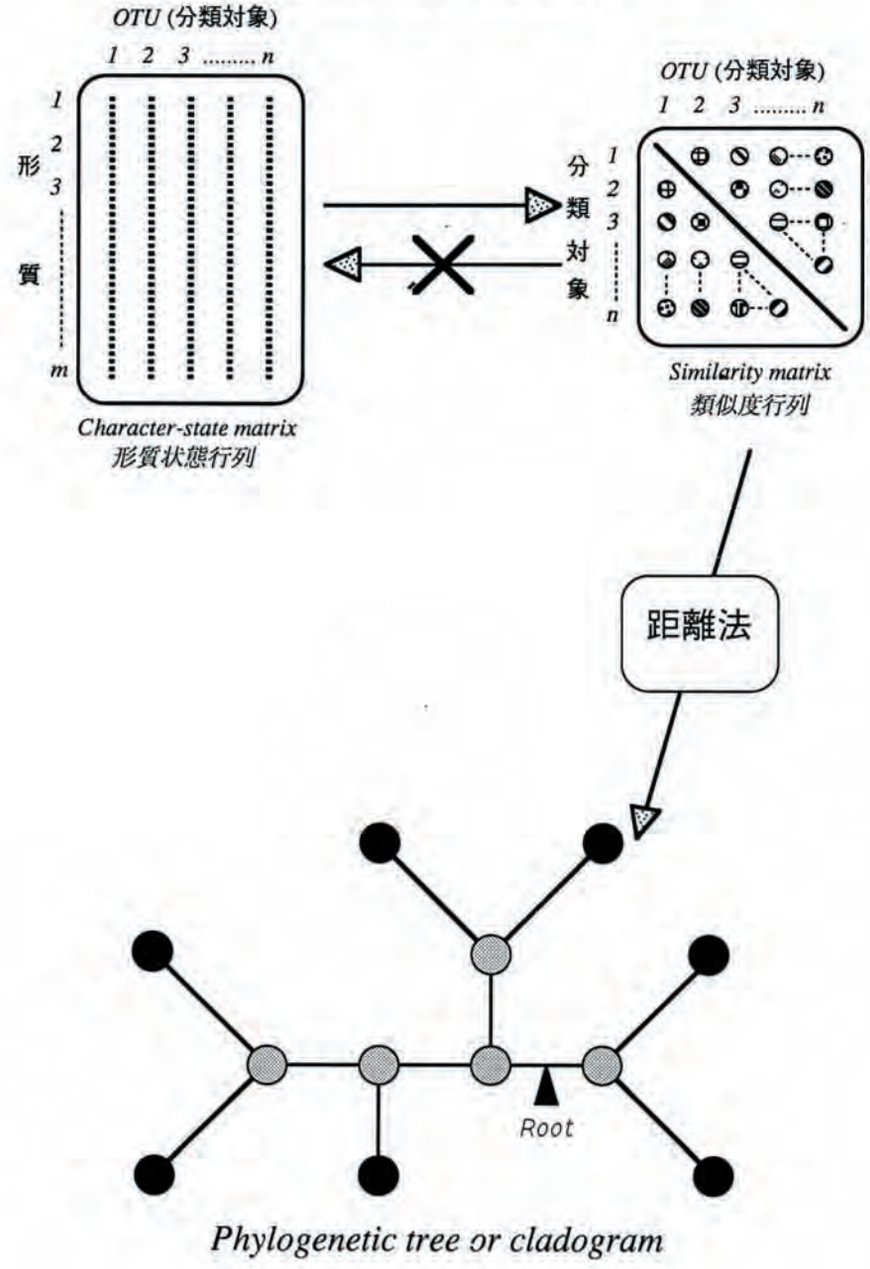



日本生物地理学会第 62 回年次大会シンポジウム
〈進化と系譜：ツリー，ネットワーク，視覚言語リテラシー〉
（8 April 2007，立教大学）

# 発展し続ける系統解析ソフトウェア

Here are 348 phylogeny packages and 49 free servers, all that I know about. It is an attempt to be completely comprehensive. I have not made any attempt to exclude programs that do not meet some standard of quality or importance. Updates to these pages are made roughly weekly. Here is a "waiting list" of new programs waiting to have their full entries constructed. Many of the programs in these pages are available on the web, and some of the older ones are also available from ftp server machines.

# *Tree topology reconstruction 2*
# *Distance method*



**sequences**

sites

| sequences | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **1** | T | T | A | T | T | A | A |
| **2** | A | A | T | T | T | A | A |
| **3** | A | A | A | A | A | T | A |
| **4** | A | A | A | A | A | A | T |

**distances**

| sequences | **1** | **2** | **3** |
|---|---|---|---|
| **2** | 3 | | |
| **3** | 5 | 4 | |
| **4** | 5 | 4 | 2 |

sequences

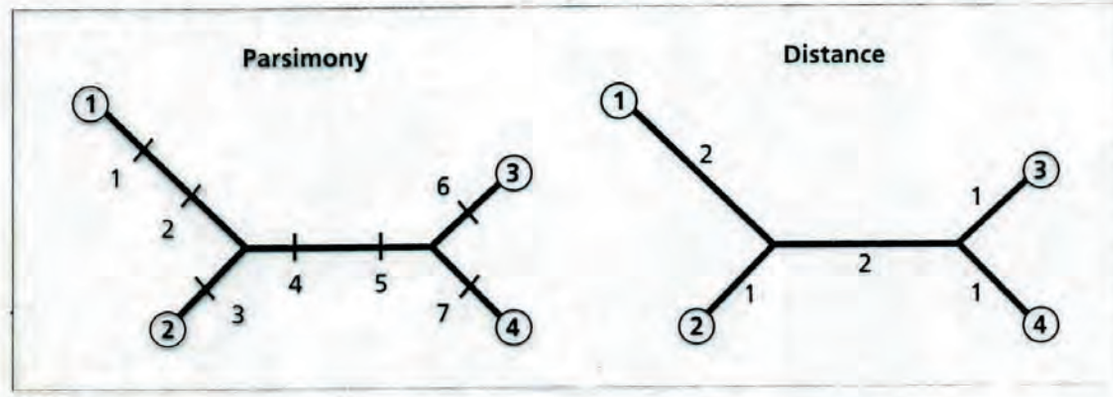



**Fig. 6.1** A parsimony tree and a distance tree for the same sequence data. Note that both trees have the same topology and branch lengths, but that the parsimony tree identifies which site contributes to the length of each branch.

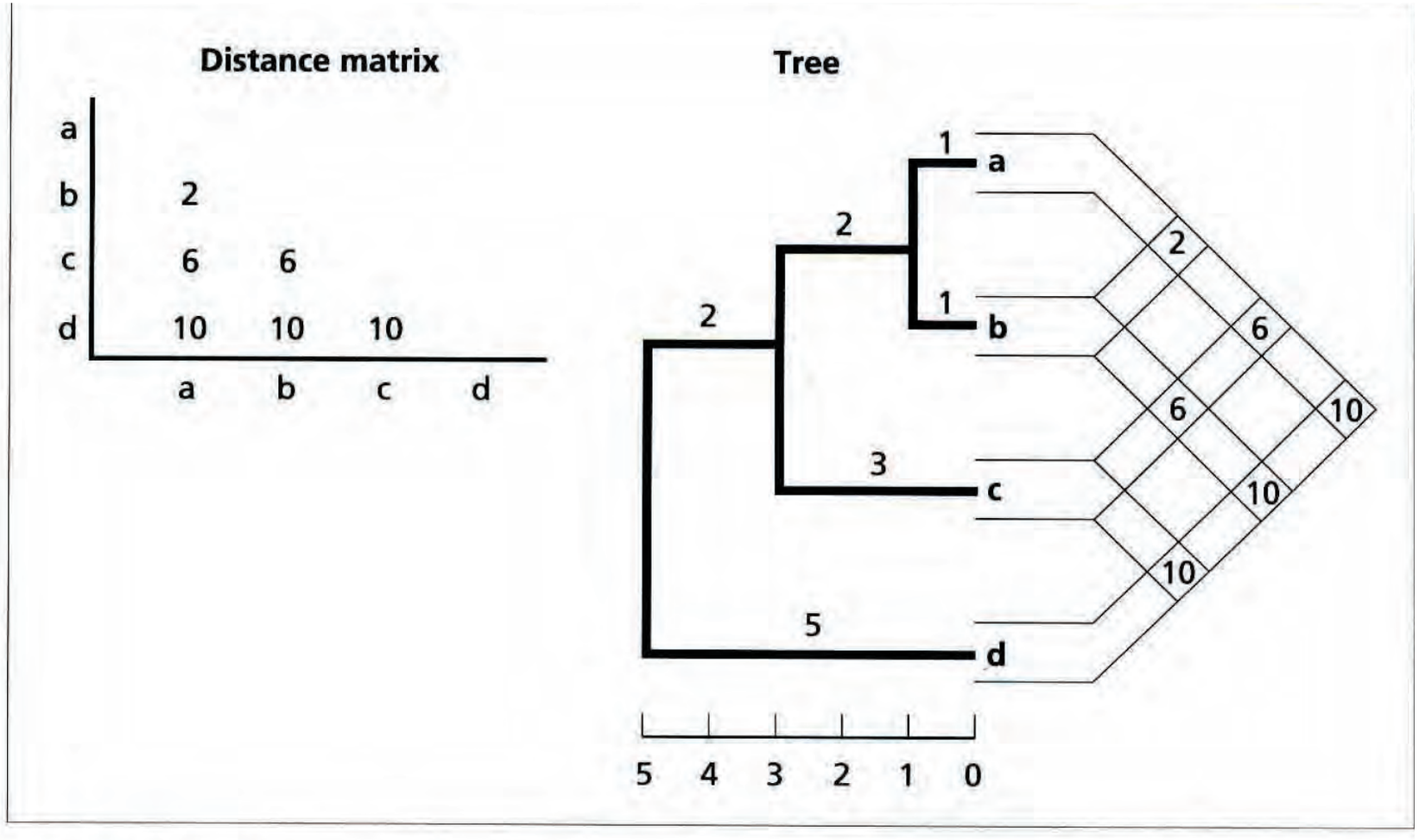



**Fig. 2.17** An ultrametric distance matrix between four sequences a–d and the corresponding ultrametric tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

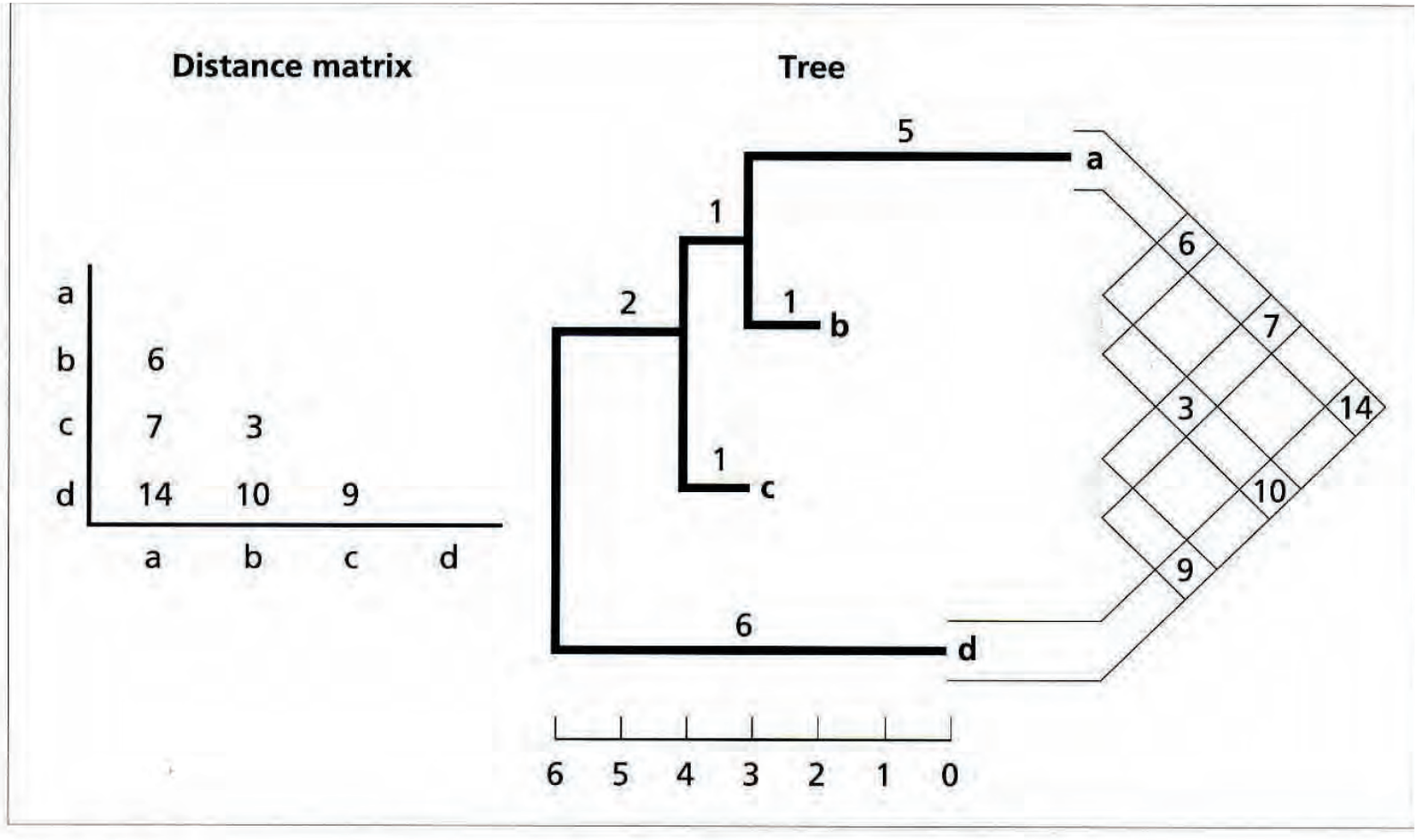
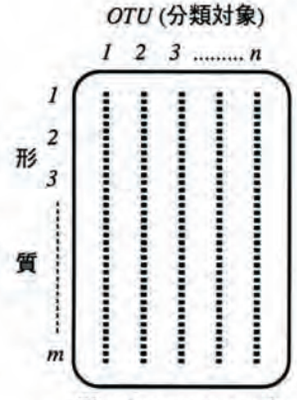
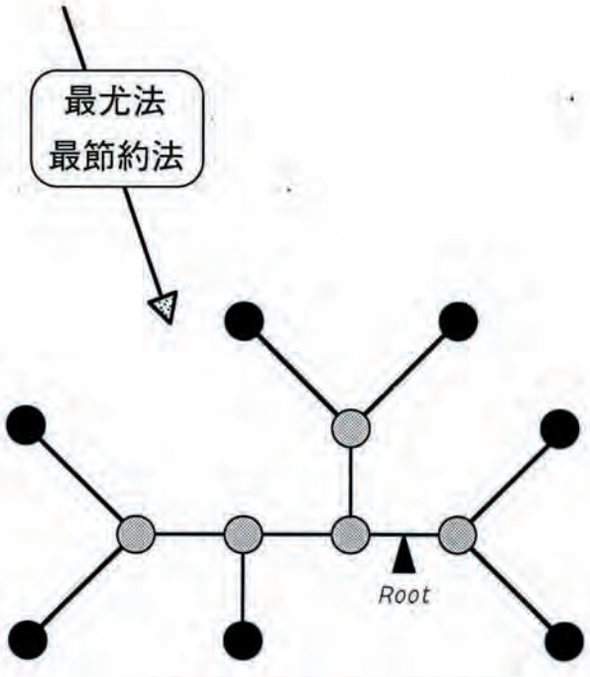



**Fig. 2.18** An additive distance matrix between four sequences and the corresponding additive tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

# *Tree topology reconstruction 1*
# *Character-state method*




*Character-state matrix*
*形質状態行列*




*Phylogenetic tree or cladogram*

# 最節約基準

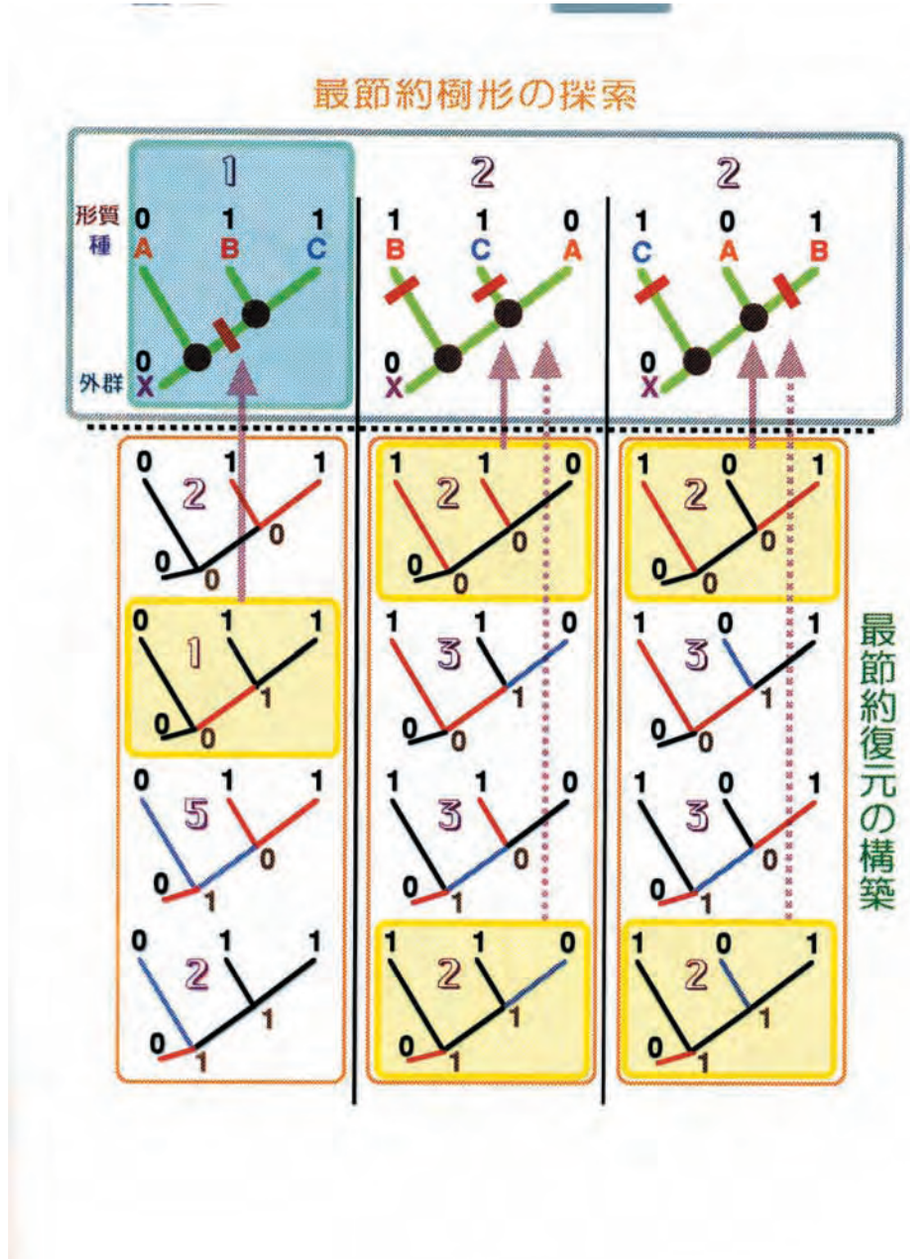
# 尤度基準

# ベイズ基準

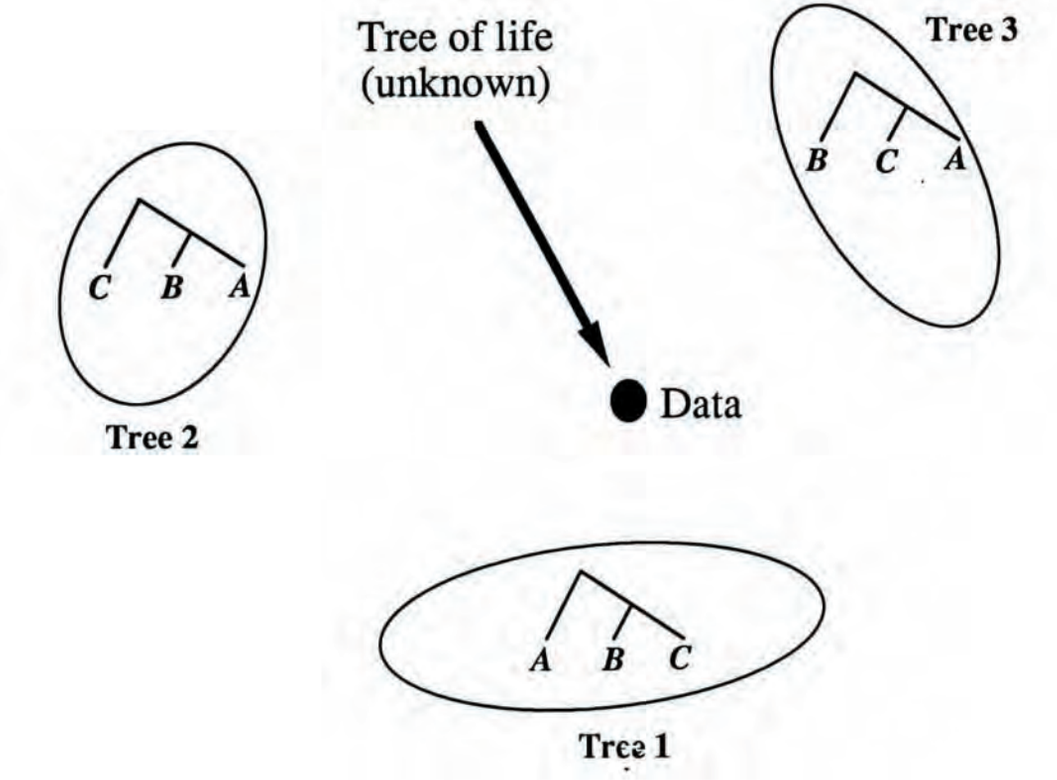
● ● ● ● ● ●

# 最節約法による
# 系統関係の推定

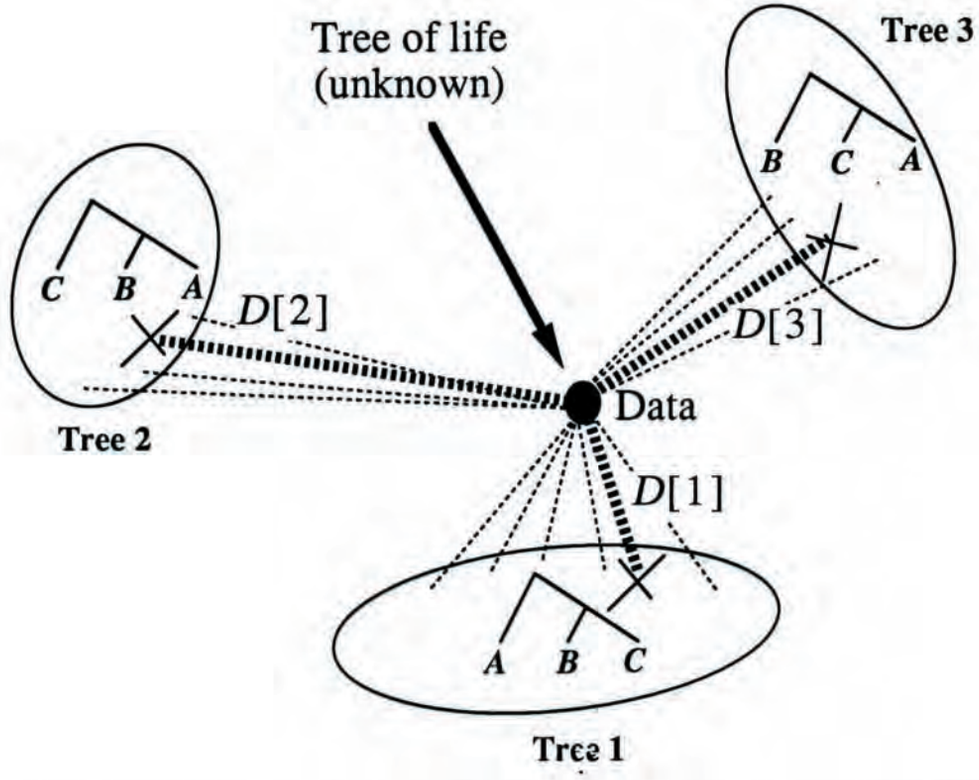
第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法




第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法


Tree of life
(unknown)
Tree 3
B C A
C B A
Tree 2
Data
A B C
Tree 1

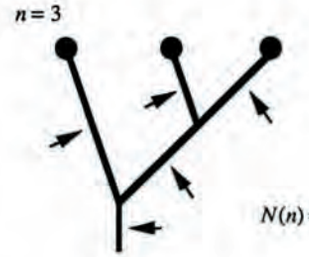
## *Maximum parsimony*




第１最節約原理＝ＭＰＲ問題（min(j)$D$[i;j]）
第２最節約原理＝Steiner問題（min(i)$D$[i]）

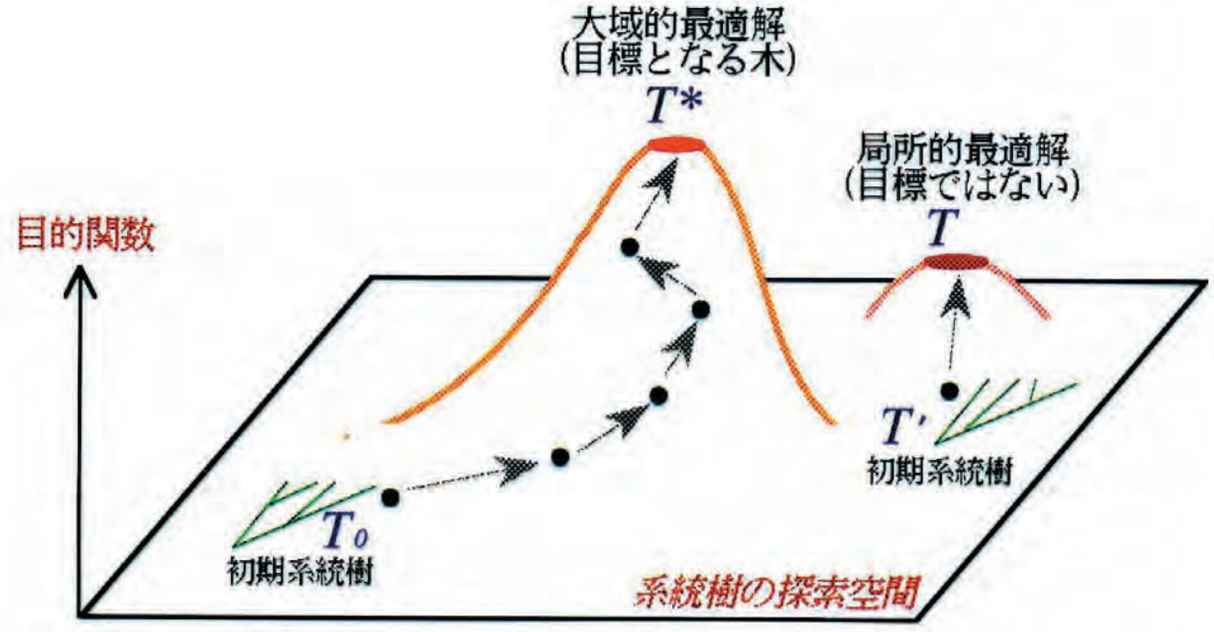
「距離和 $D$→Min」

# Number of bifurcating trees




$$N(n) = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \prod_{k=2}^{n} (2k-3)$$

| $n$ | $N(n)$ |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 3 |
| 4 | 15 |
| 5 | 105 |
| 6 | 945 |
| 7 | 10,395 |
| 8 | 135,135 |
| 9 | 2,027,025 |
| 10 | 34,459,425 |
| 11 | 654,729,075 |
| 12 | 13,749,310,575 |
| 13 | 316,234,143,225 |
| 14 | 7,905,853,580,625 |
| 15 | 213,458,046,676,875 |
| 16 | 6,190,283,353,629,375 |
| 17 | 191,898,783,962,510,625 |
| 18 | 6,332,659,870,762,850,625 |
| 19 | 221,643,095,476,699,771,875 |
| 20 | 8,200,794,532,637,891,559,375 |
| 21 | 319,830,986,772,877,770,815,625 |
| 22 | 13,113,070,457,687,988,603,440,625 |

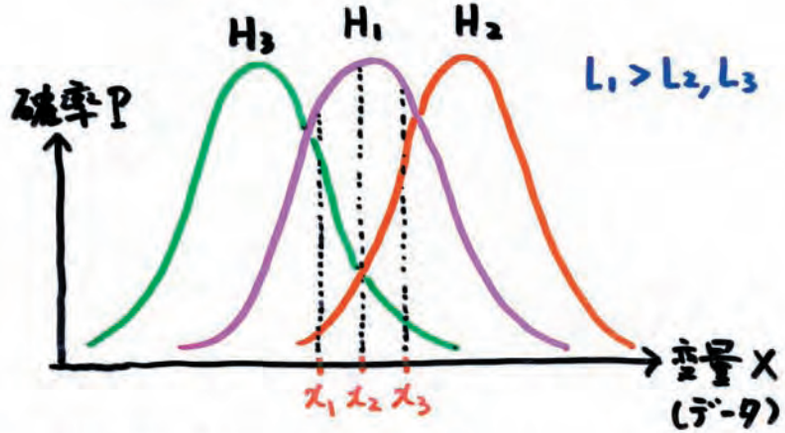

大域的最適解
(目標となる木)
T*
局所的最適解
(目標ではない)
T
目的関数
T'
初期系統樹
T0
初期系統樹
系統樹の探索空間

# 最尤法

# *Maximum Likelihood method*

# 尤度（Likelihood）

データが仮説に与える支持の強さ．




仮説 $H_1$, $H_2$, $H_3$ } 確率分布の平均値（パラメーター）が異なる．

尤度 $L = P(x_1) \cdot P(x_2) \cdot P(x_3)$
（確率積）

↓

最尤推定：尤度Lが最大となる仮説を受容する．

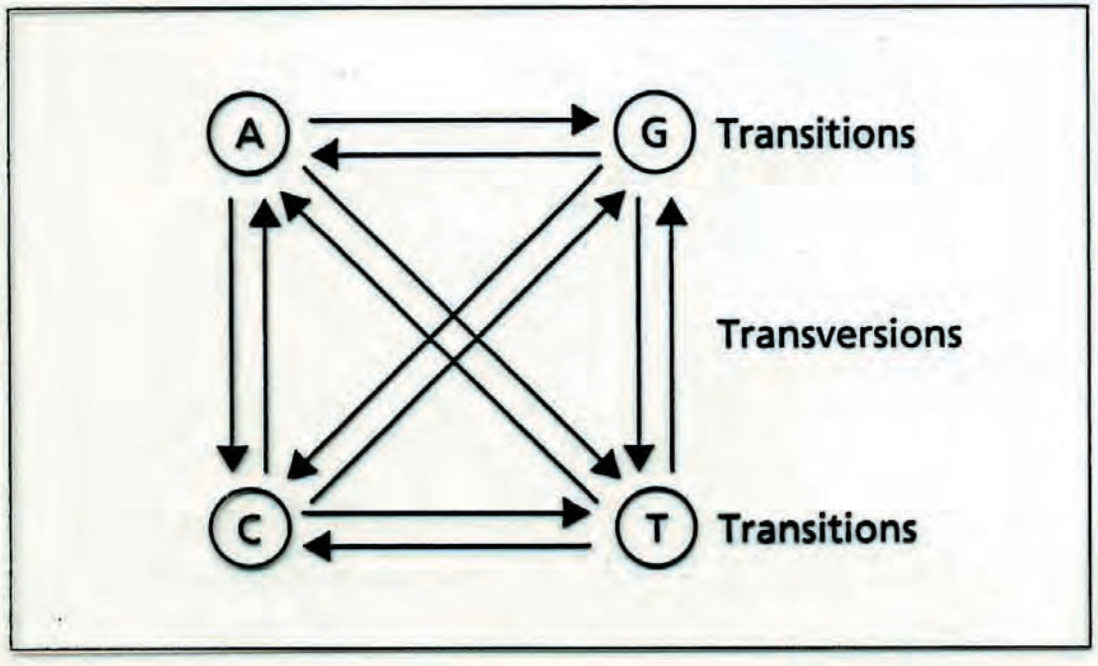



**Fig. 5.10** The possible substitutions among the four nucleotides.

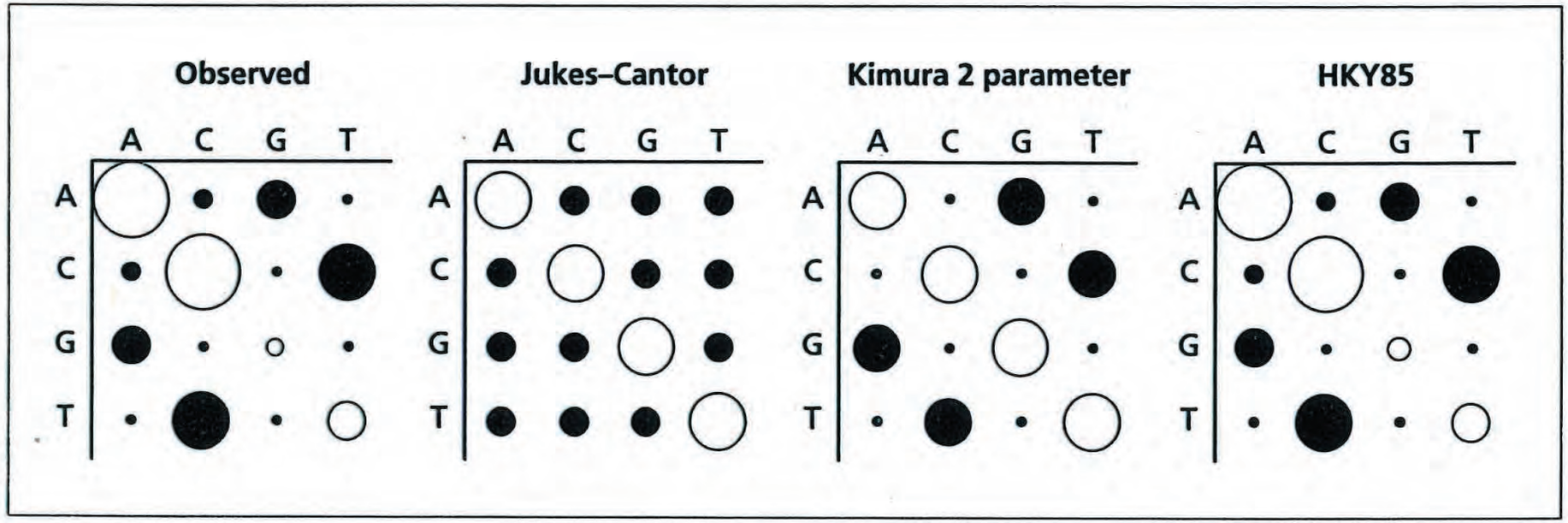



**Fig. 5.15** Observed and expected numbers of nucleotide pairs between human and chimpanzee mtDNA sequences for three different models. As the models add parameters they more closely approximate the observed pattern. Data from Tamura (1994).

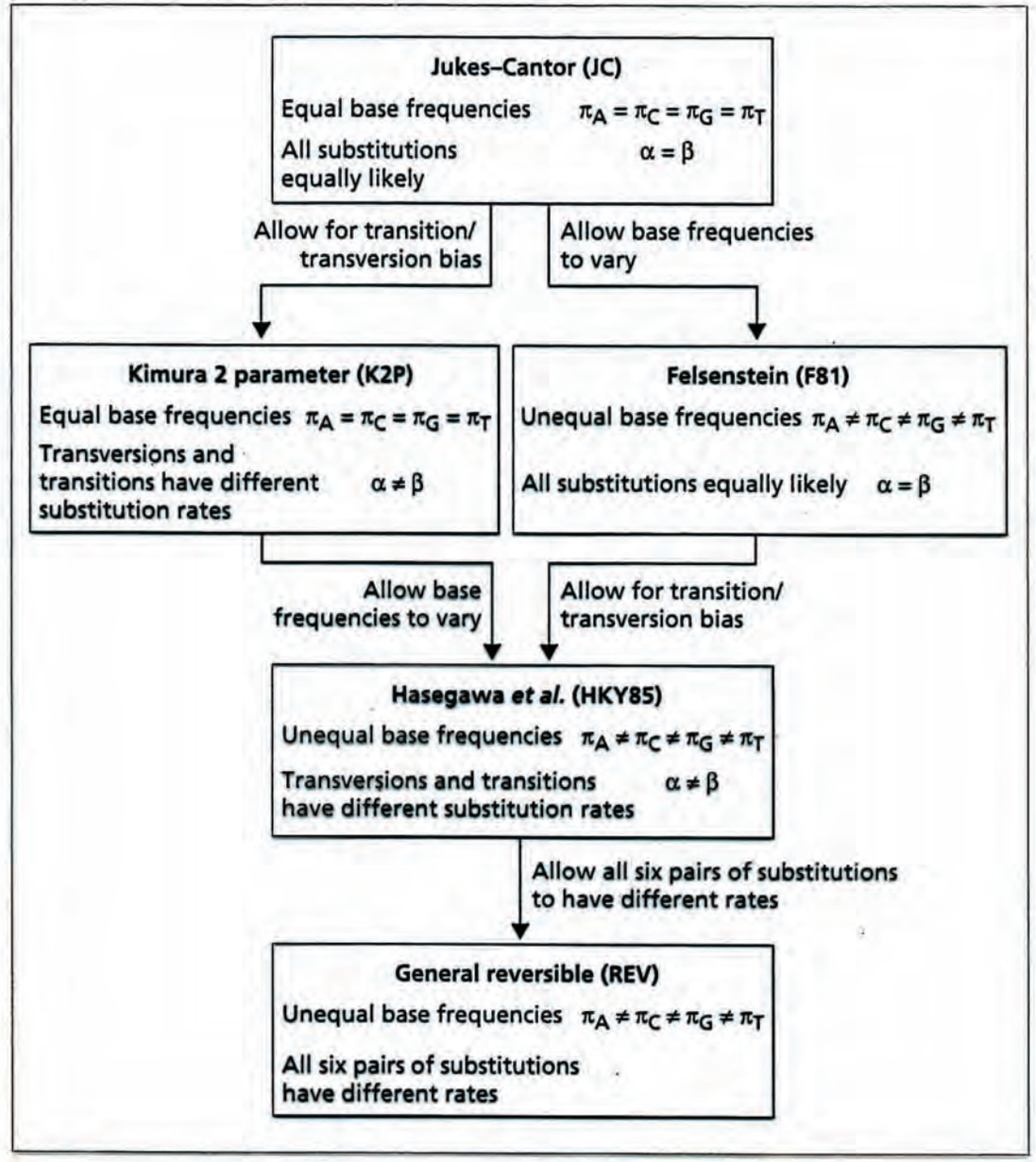
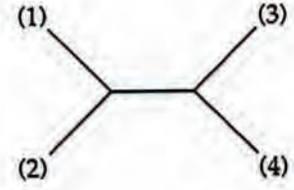



**Fig. 5.14** Interrelationships among five models for estimating the number of nucleotide substitutions among a pair of DNA sequences. The JC, K2P, F81 and HKY85 models can all be generated by constraining various parameters of the REV model.

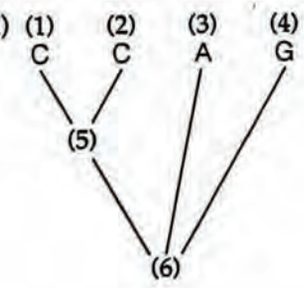
(A)

| | 1 | | | | | | | j | | | | | | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | C | ... | G | G | A | C | A | C | G | T | T | T | A | ... C |
| (2) | C | ... | A | G | A | C | A | C | C | T | C | T | A | ... C |
| (3) | C | ... | G | G | A | T | A | A | G | T | T | A | A | ... C |
| (4) | C | ... | G | G | A | T | A | G | C | C | T | A | G | ... C |

(B)




(C)




(D)

$$L_{(j)} = \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & A & & \\ & & A & \end{array}\right) + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & C & & \\ & & A & \end{array}\right)$$

$$+ \quad \ldots \quad + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & G & & \\ & & C & \end{array}\right)$$

$$+ \quad \ldots \quad + \text{Prob}\left(\begin{array}{cccc} C & C & A & G \\ & T & & \\ & & T & \end{array}\right)$$

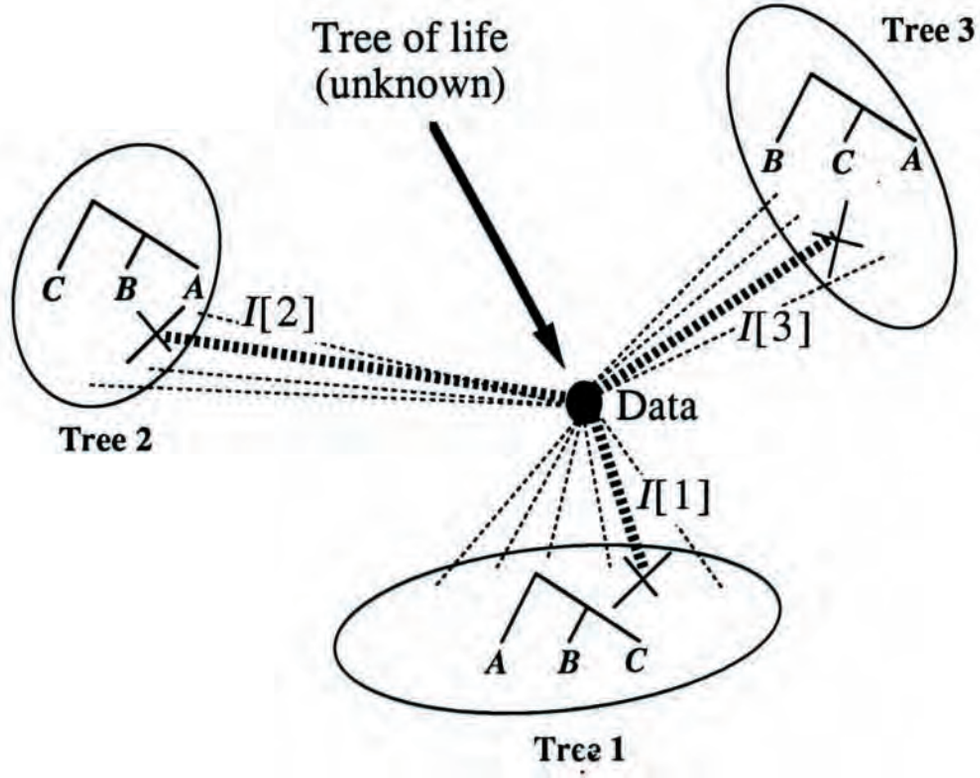
(E)

$$L = L_{(1)} \cdot L_{(2)} \cdot \ldots \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

(F)

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ldots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}$$

## *Maximum likelihood*



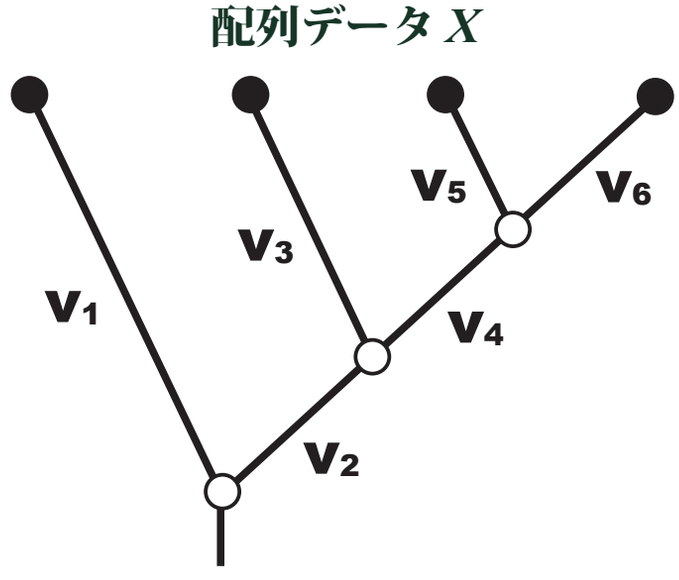

第 1 最尤原理＝min(j)$I$[i;j]
第 2 最尤原理＝min(i)$I$[i]

「Kullback-Leibler情報量 $I$ →Min」

# ベイズ事後分布を求める

## モデルが複雑になると計算量が膨大になる

［例］分子系統樹のベイズ推定




パラメータ群

- 樹形：$\tau$ → 推定対象
- ~~枝長：$v$~~
- ~~置換遷移確率：$\theta$~~
- ~~サイト間変異：$\alpha$~~ → 対象外（撹乱母数）

樹形の事後分布

$$f(\tau_i|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\tau_i)f(\tau_i)}{\sum_{j=1}^{N} f(\boldsymbol{X}|\tau_j)f(\tau_j)}$$

樹形の事前分布

組合せ論的爆発

樹形の尤度

$$f(\boldsymbol{X}|\tau_i) = \int_{\boldsymbol{v}_i}\int_{\boldsymbol{\theta}}\int_{\alpha} f(\boldsymbol{X}|\tau_i, \boldsymbol{v}_i, \boldsymbol{\theta}, \alpha) f(\boldsymbol{v}_i) f(\boldsymbol{\theta}) f(\alpha) d\boldsymbol{v}_i d\boldsymbol{\theta} d\alpha$$

撹乱母数は期待値計算で消去

多重積分の重荷

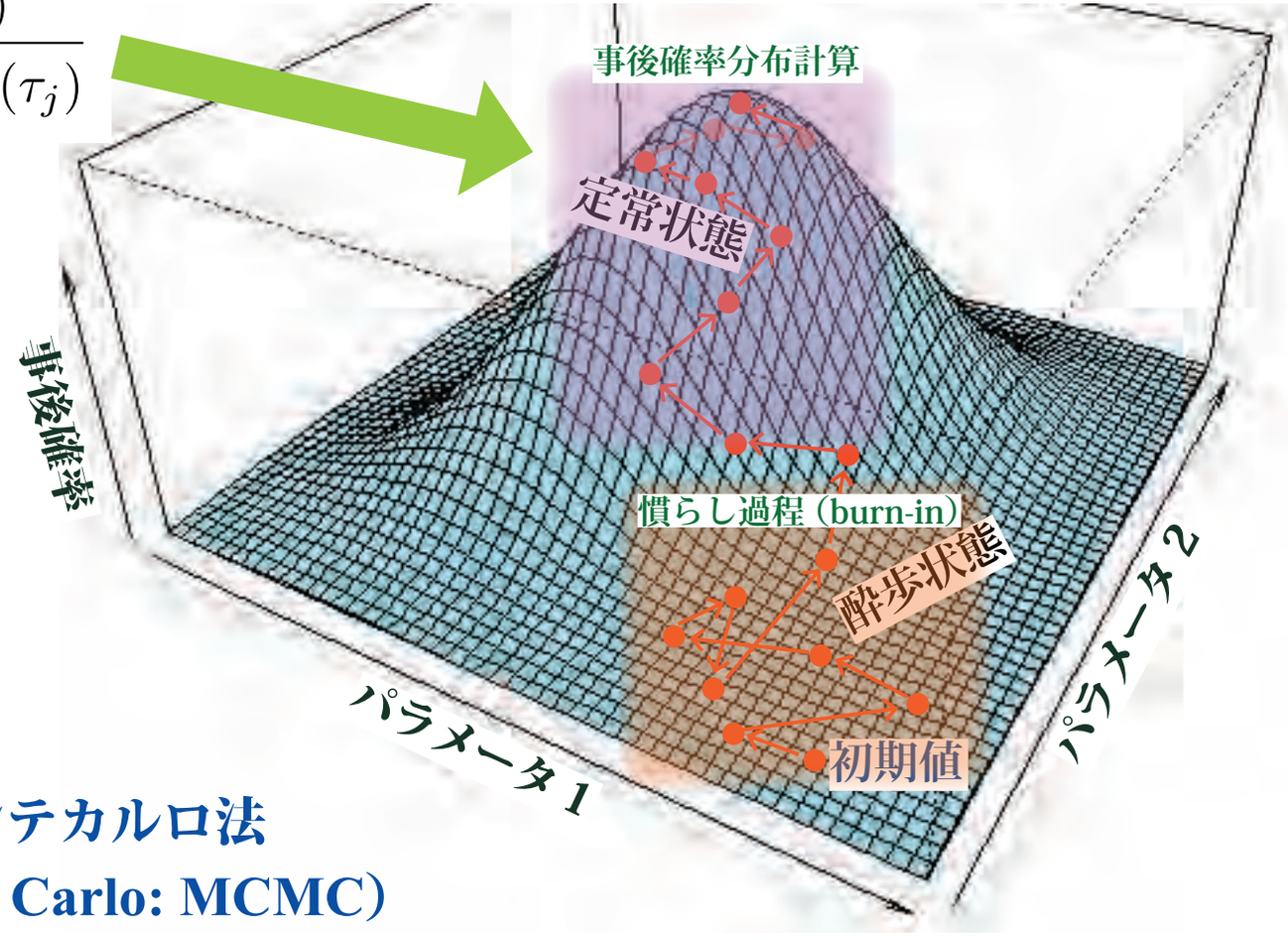
# ベイズ事後分布を求める

## マルコフ連鎖モンテカルロ法（MCMC）という最終兵器

**事後確率分布**

$$f(\tau_i|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\tau_i)f(\tau_i)}{\sum_{j=1}^{N} f(\boldsymbol{X}|\tau_j)f(\tau_j)}$$

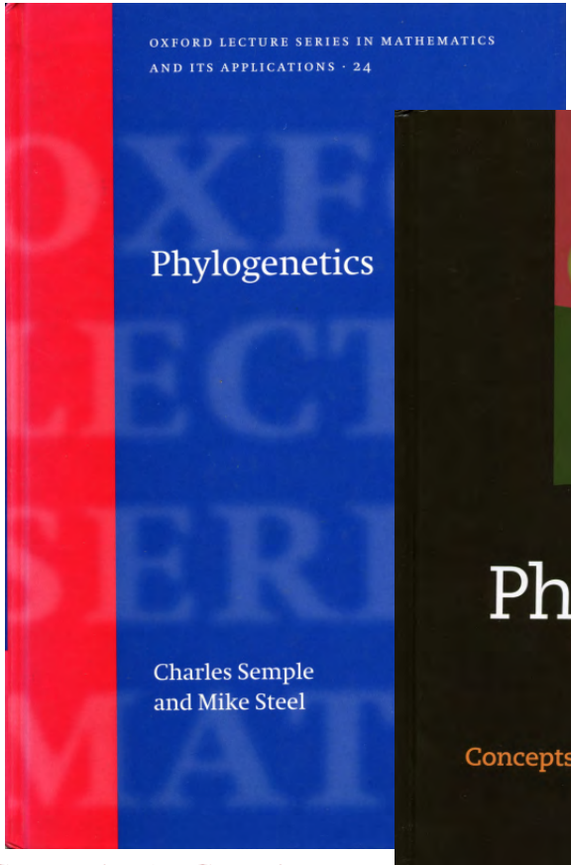
**パラメータ初期値を無作為に与え，マルコフ過程を用いて事後確率分布を系統的にサンプリングし，定常状態になるまで探索させる．**



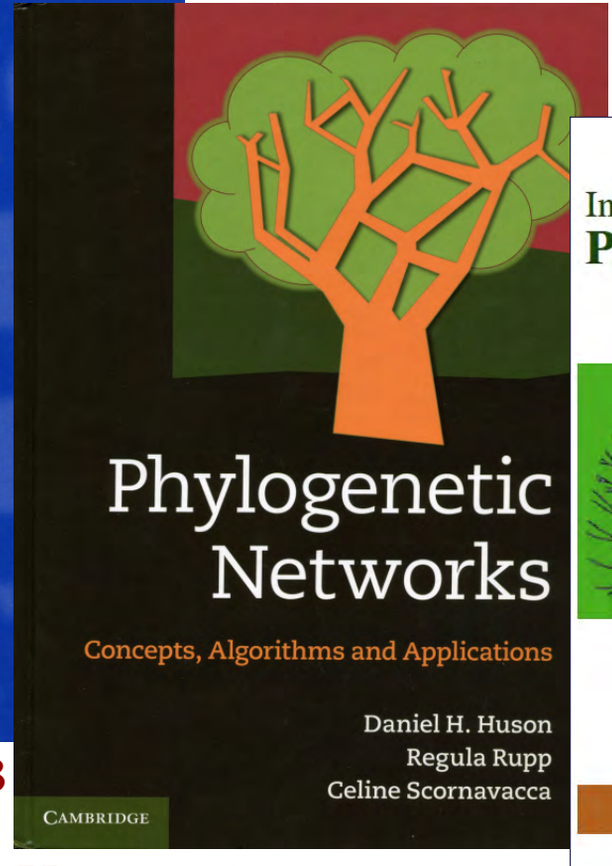

**マルコフ連鎖モンテカルロ法**
**（Markov chain Monte Carlo: MCMC）**

# チェイン・ツリー・ネットワーク

# 数理系統学：系統ダイアグラムの体系

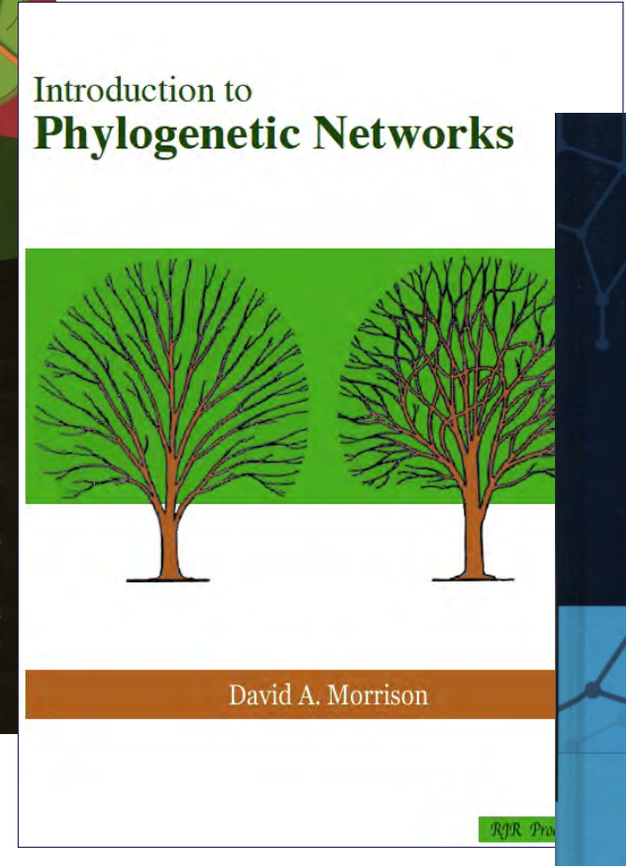



Semple & Steel 2003

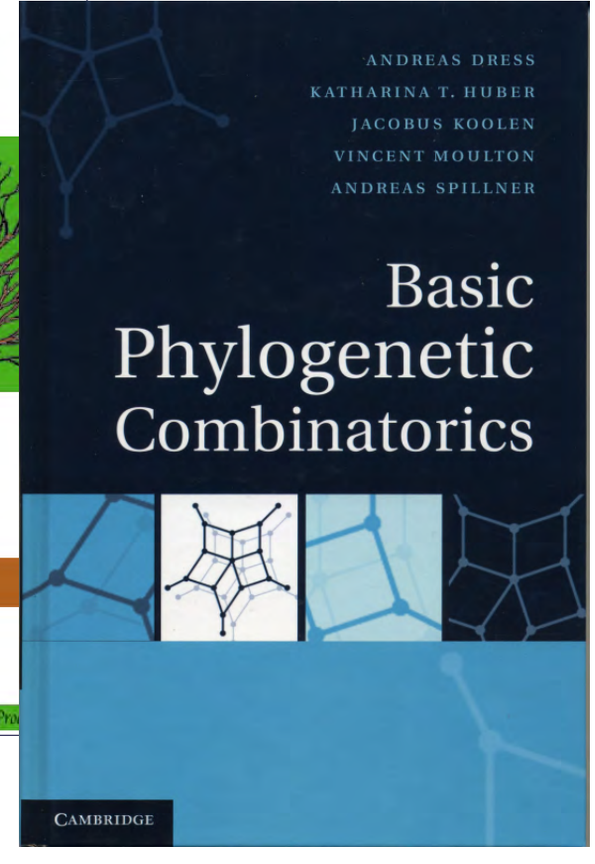



Huson *et al.* 2010

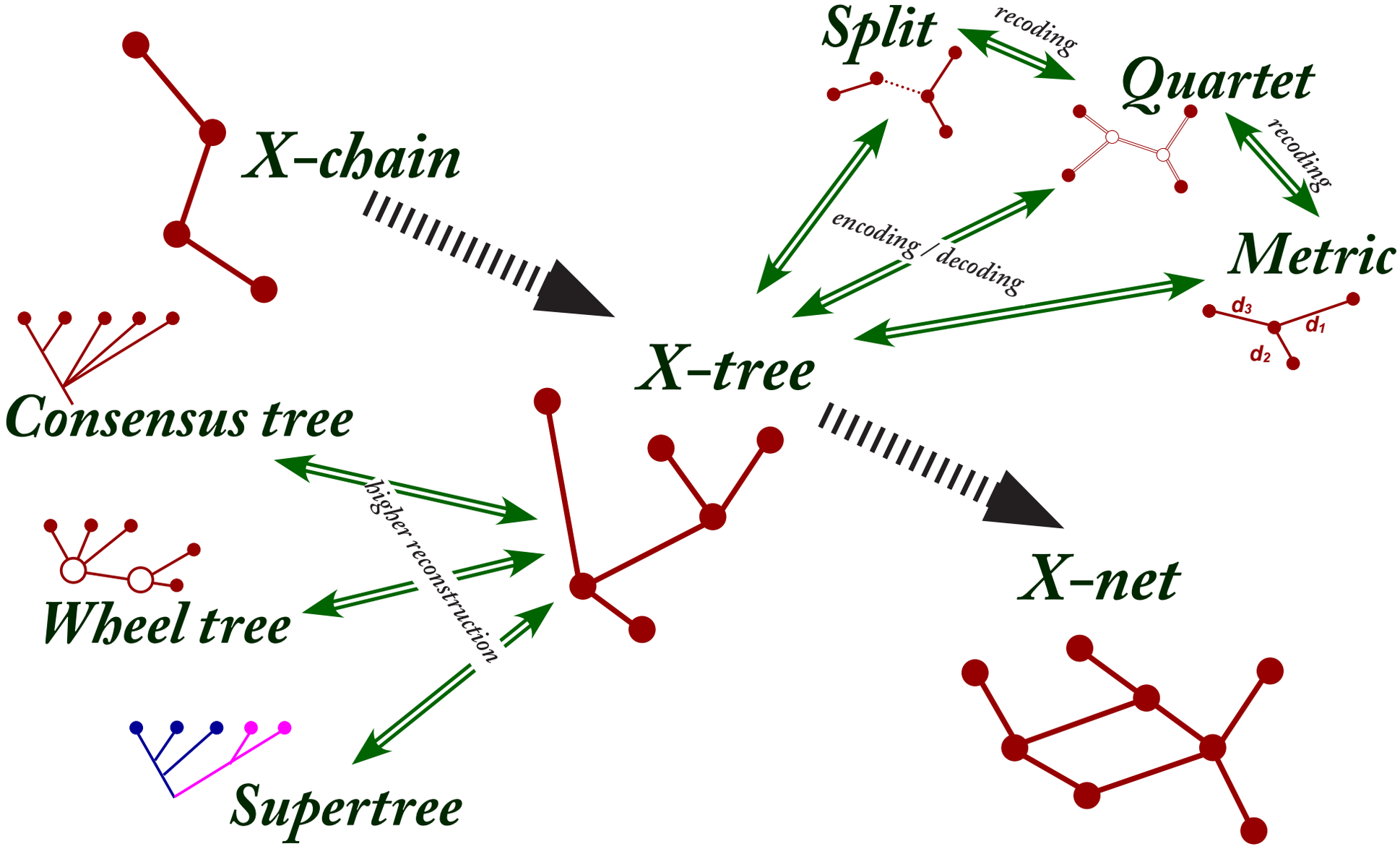
Morrison 2011

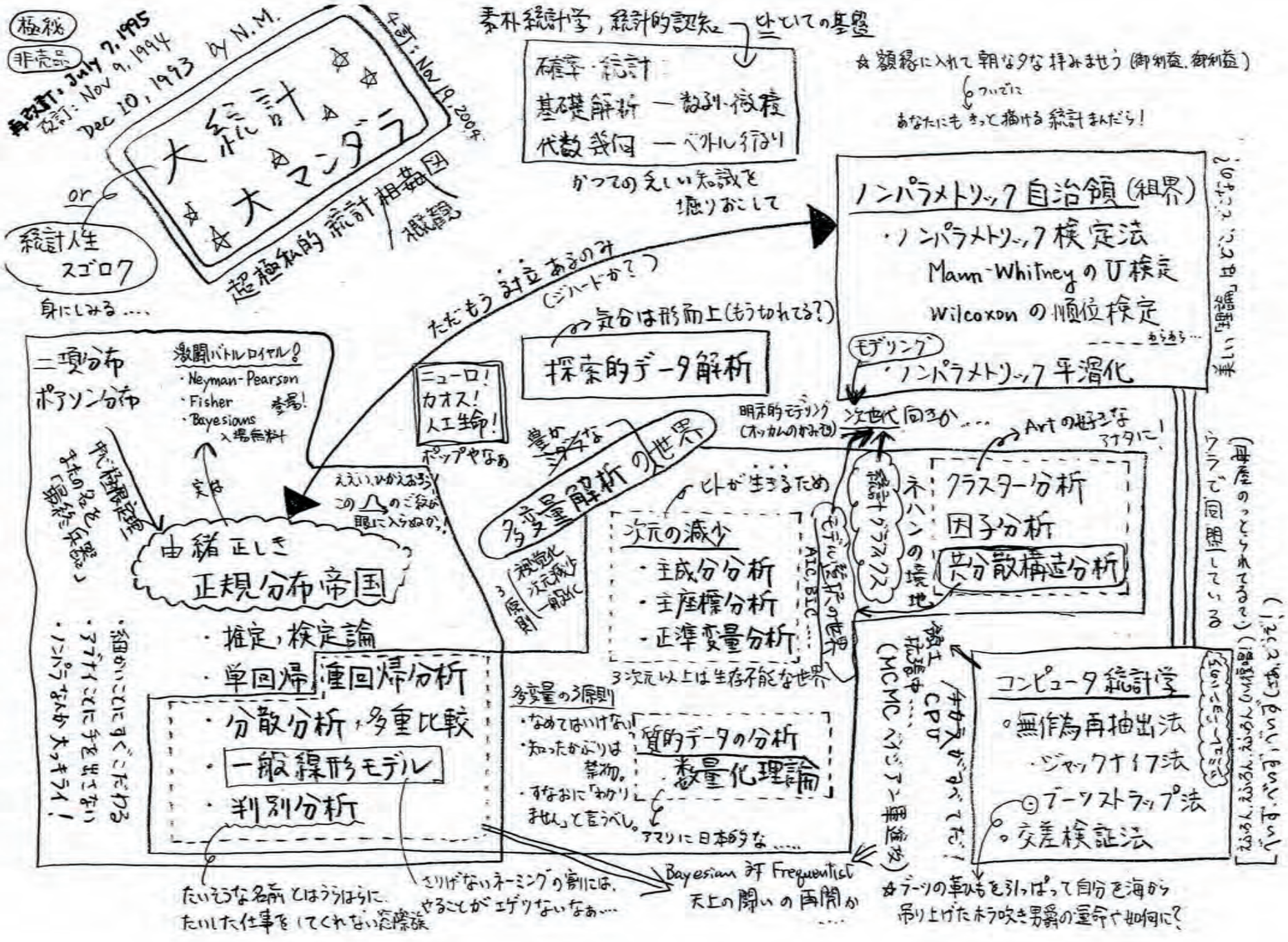


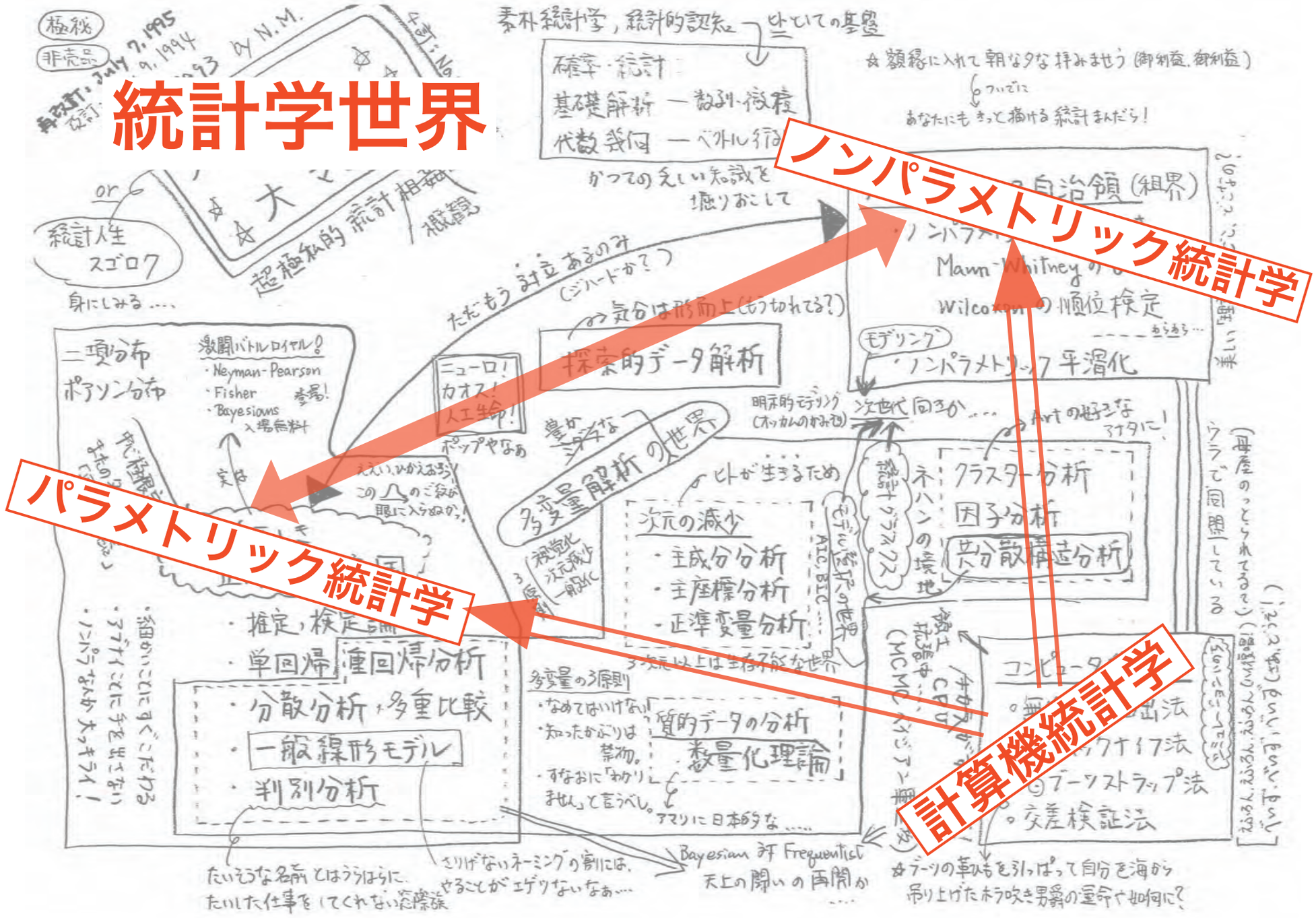

Dress *et al.* 2012



# 数理系統学：系統ダイアグラムの体系

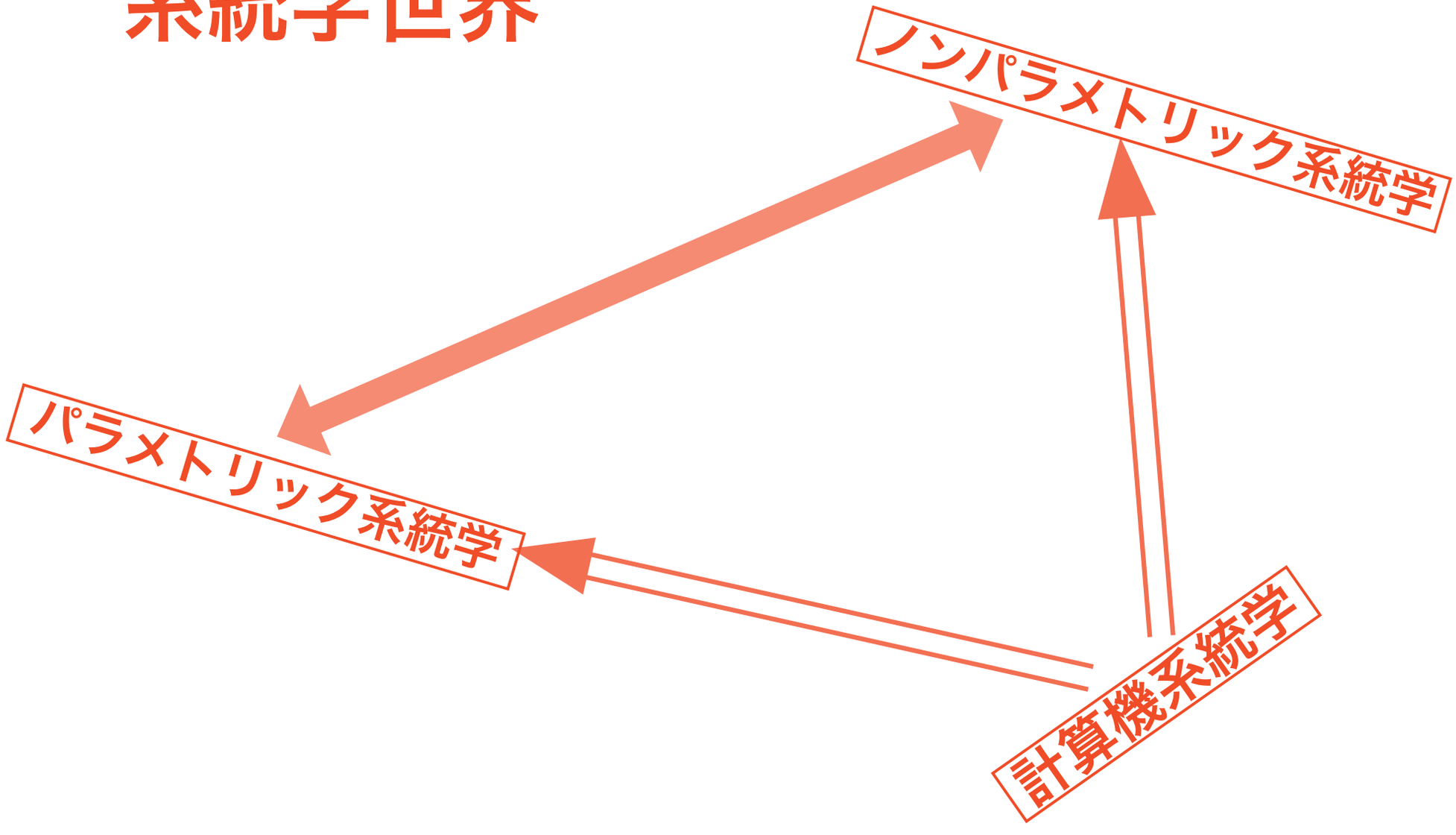



極秘
非売品
再改訂: July 7, 1995
改訂: Nov 9, 1994
Dec 10, 1993 by N.M.
大統計 大マンダラ
超極私的統計相関図
ver 1.0 Nov 19, 2004
or
統計人生スゴロク
身にしみる....
素朴統計学，統計的認知 にいたっての基盤
確率・統計
基礎解析 ― 数列・微積
代数幾何 ― ベクトル行列
かつての えい知識を 憶りおこして
☆額縁に入れて朝な夕な拝みませう (御利益、御利益)
ついでに
あなたにもきって描ける統計まんだら!
ノンパラメトリック自治領(組界)
・ノンパラメトリック検定法
Mann-Whitney の U 検定
Wilcoxon の順位検定
からきら…
モデリング
・ノンパラメトリック平滑化
ただもう対立あるのみ (ジハードかて?)
気合は形而上(もうかれてる?)
探索的データ解析
ニューロ! カオス! 人工生命!
ポップやなあ
二項分布
ポアソン分布
激闘バトルロイヤル!
・Neyman-Pearson
・Fisher 参戦!
・Bayesians 入場無料
実は
由緒正しき 正規分布帝国
・推定，検定論
・単回帰，重回帰分析
・分散分析，多重比較
・一般線形モデル
・判別分析
細かいことにすぐこだわる
・アプローチごとに手を出さない
・ノンパラなんか オッキライ!
ええい、ひかえおろう この八のご紋が 眼に入らぬか!
豊かなデータ
多変量解析の世界
明示的モデリング (オッカムのかみそり)
次世代 向きか…
Art の好きな アナタに!
ヒトが生きるため
次元の縮少
・主成分分析
・主座標分析
・正準変量分析
3次元以上は生存不能な世界
3原則
視覚化
次元縮少
一般化
認知グラフィクス
ネハンの境地
クラスター分析
因子分析
共分散構造分析
ゲノム 探索の 世界
AIC, BIC
多変量の3原則
・なめてないけない
・知ったかぶりは 禁物。
・すなおに「わかり ません」と言うべし。
質的データの分析
数量化理論
アマリに日本的な…..
コンピュータ統計学
・無作為再抽出法
・ジャックナイフ法
・ブートストラップ法
・交差検証法
(「ええっ?」ばかり 言いやがって…)
(母屋のっとられてるぞ…) (間抜けの ヘタクソ 演奏)
ウラで「回想」している
CPU 酷使 (MCMC ベイジアン 推論など)
☆ブーツの革ひもを引っぱって自分を海から 吊り上げたホラ吹き男爵の運命や如何に?
Bayesian 対 Frequentist 天上の闘いの再開か
たいそうな名前とはうらはらに、たいした仕事をしてくれない変態族
さりげないネーミングの割には、やることがエゲツないなあ…

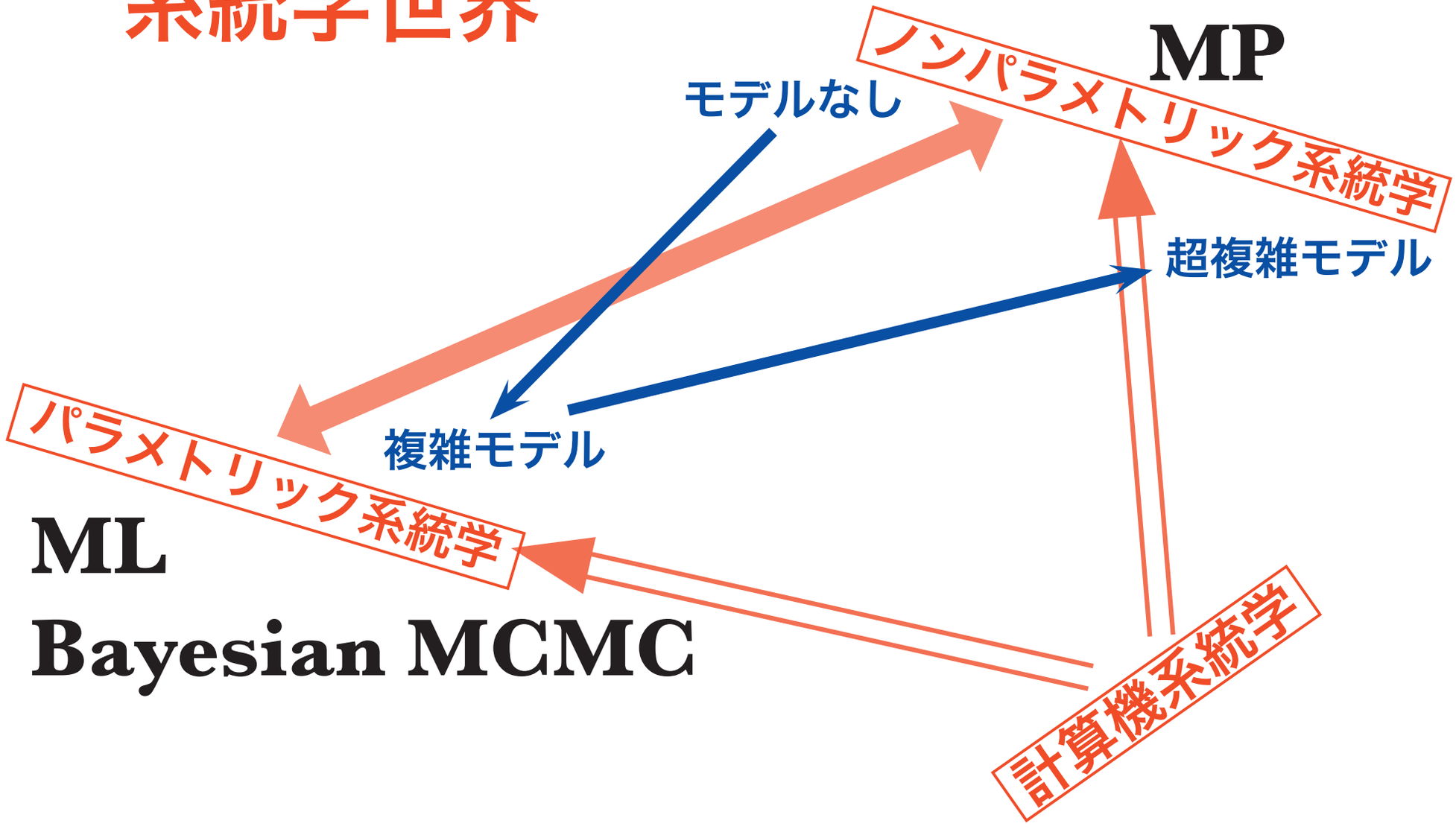



第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法

# 系統学世界



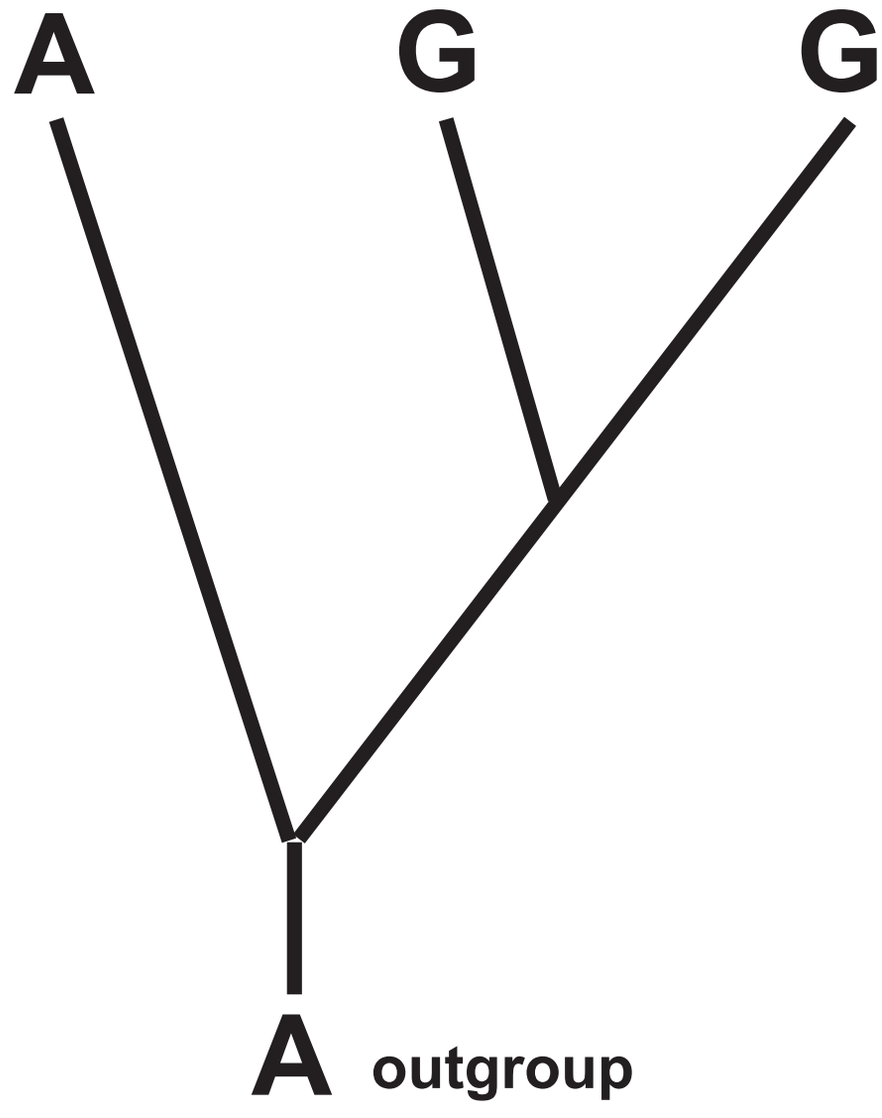
# 系統学世界



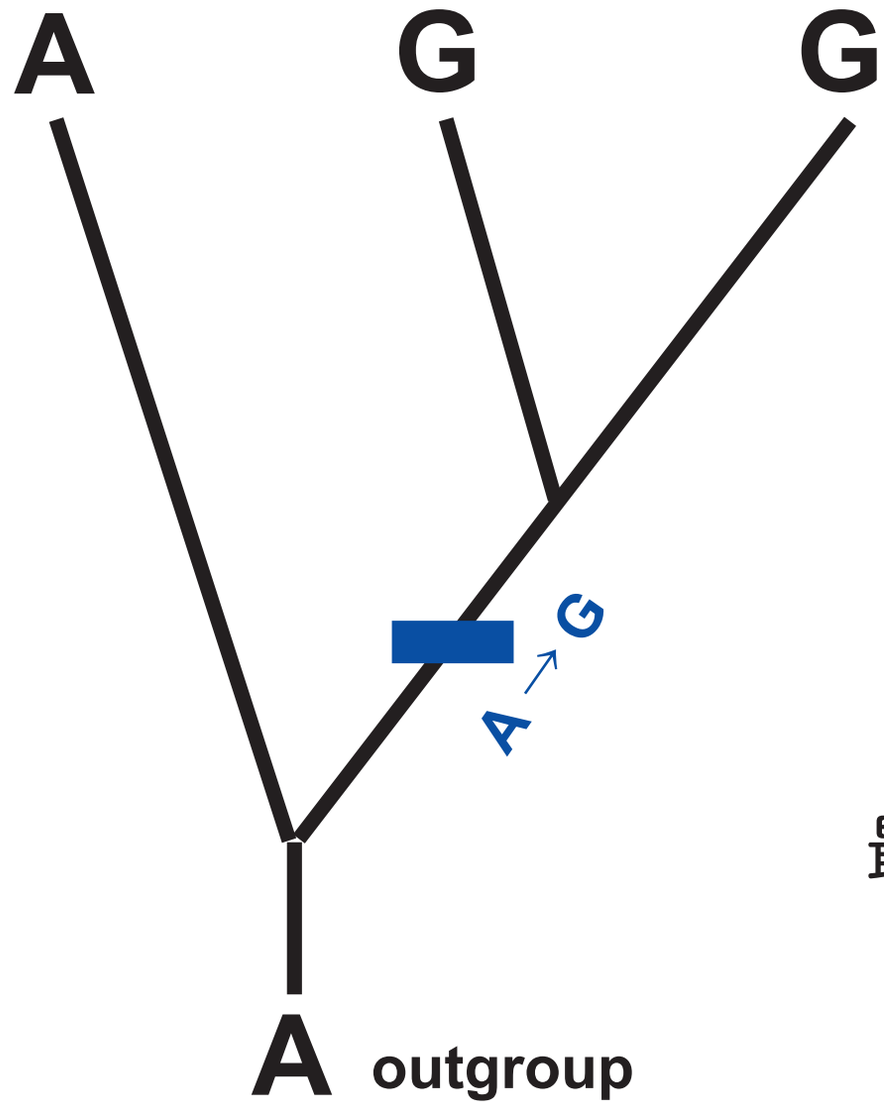

第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法

T ：系統樹

$\chi$ ：形質



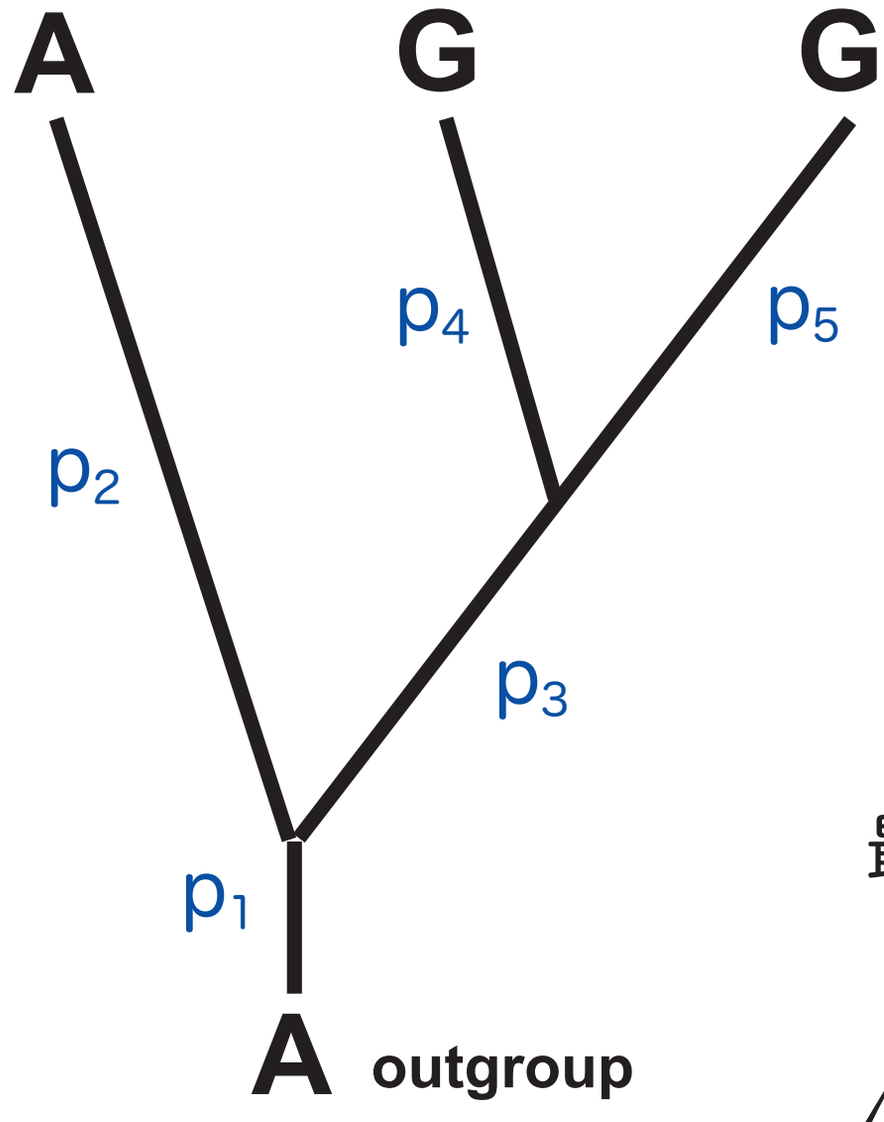
MP




最節約スコア

$\ell(x, T)$

# ML




最尤スコア

$$\mathcal{L}(\chi \mid T, p)$$

# "No Common Mechanism"

1)「根」の形質状態は均一分布

2) 形質状態変化確率 p は枝ごとに変化し，他のすべての形質状態に等確率で変化する.

3) 枝ごとの形質変化は互いに独立である.

4) 形質状態変化確率 p は $0 \leqq p \leqq (r - 1)/r$ を満たす.
ここに，r は形質状態の個数を表す.

# "No Common Mechanism"

定理

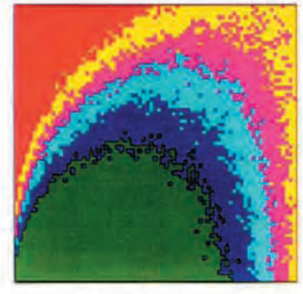
$$\max_p [p(x \mid T, p)] = r^{-\ell(x, T)}$$

**ML = MP**

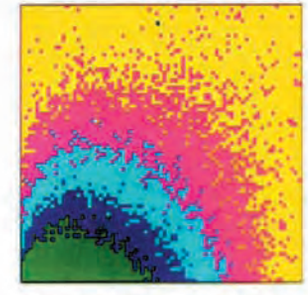
(C. Tuffley and M. Steel 1997)

# Systematic Biology

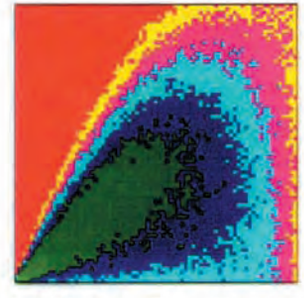
PUBLISHED BY THE
Society of Systematic Biologists



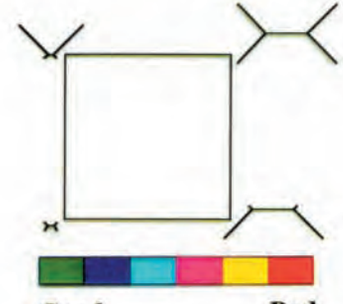
Parsimony



Lake's Invariants



UPGMA (Similarity)




Success

VOLUME 42 SEPTEMBER 1993 NUMBER 3

Syst. Biol. 42(3):247–264, 1993

# SUCCESS OF PHYLOGENETIC METHODS IN THE FOUR-TAXON CASE

JOHN P. HUELSENBECK AND DAVID M. HILLIS

*Department of Zoology, University of Texas, Austin, Texas 78712, USA*

*Abstract.*—The success of 16 methods of phylogenetic inference was examined using consistency and simulation analysis. Success—the frequency with which a tree-making method correctly identified the true phylogeny—was examined for an unrooted four-taxon tree. In this study, tree-making methods were examined under a large number of branch-length conditions and under three models of sequence evolution. The results are plotted to facilitate comparisons among the methods. The consistency analysis indicated which methods converge on the correct tree given infinite sample size. General parsimony, transversion parsimony, and weighted parsimony are inconsistent over portions of the graph space examined, although the area of inconsistency varied. Lake's method of invariants consistently estimated phylogeny over all of the graph space when the model of sequence evolution matched the assumptions of the invariants method. However, when one of the assumptions of the invariants method was violated, Lake's method of invariants became inconsistent over a large portion of the graph space. In general, the distance methods (neighbor joining, weighted least squares, and unweighted least squares) consistently estimated phylogeny over all of the graph space examined when the assumptions of the distance correction matched the model of evolution used to generate the model trees. When the assumptions of the distance methods were violated, the methods became inconsistent over portions of the graph space. UPGMA was inconsistent over a large area of the graph space, no matter which distance was used. The simulation analysis showed how tree-making methods perform given limited numbers of character data. In some instances, the simulation results differed quantitatively from the consistency analysis. The consistency analysis indicated that Lake's method of invariants was consistent over all of the graph space under some conditions, whereas the simulation analysis showed that Lake's method of invariants performs poorly over most of the graph space for up to 500 variable characters. Parsimony, neighbor-joining, and the least-squares methods performed well under conditions of limited amount of character change and branch-length variation. By weighting the more slowly evolving characters or using distances that correct for multiple substitution events, the area in which tree-making methods are misleading can be reduced. Good performance at high rates of change was obtained only by giving increased weight to slowly evolving characters (e.g., transversion parsimony, weighted parsimony). UPGMA performed well only when branch lengths were close in length. [Phylogeny estimation; simulation; parsimony; Lake's invariants; UPGMA; neighbor joining; weighted least squares; unweighted least squares; tree space.]

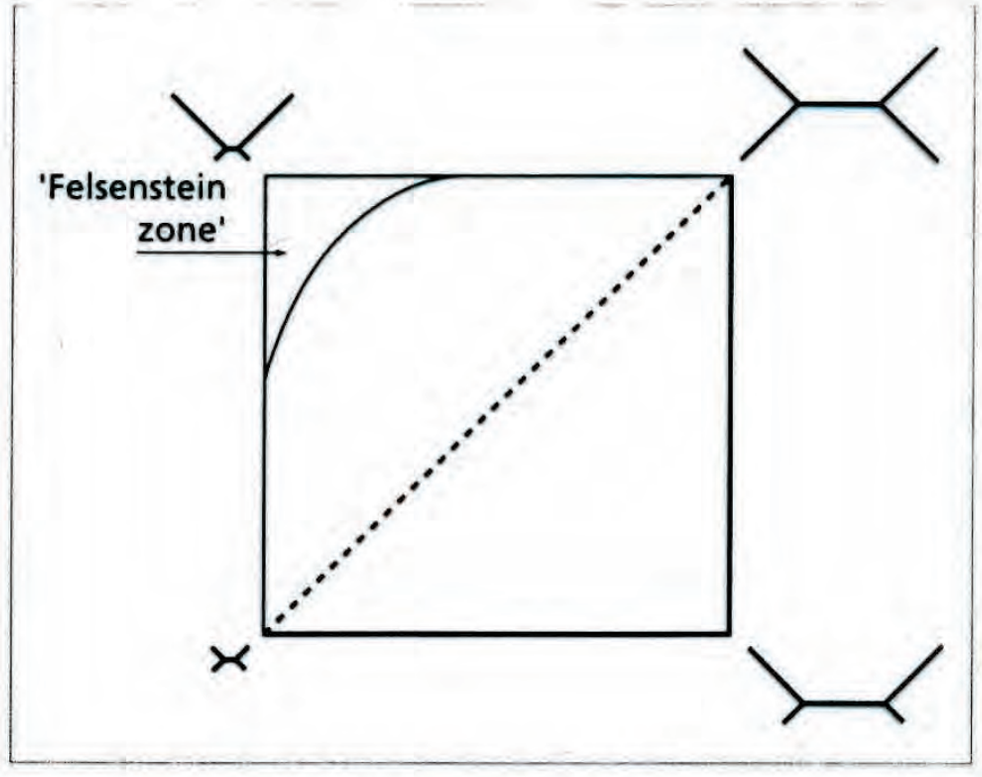
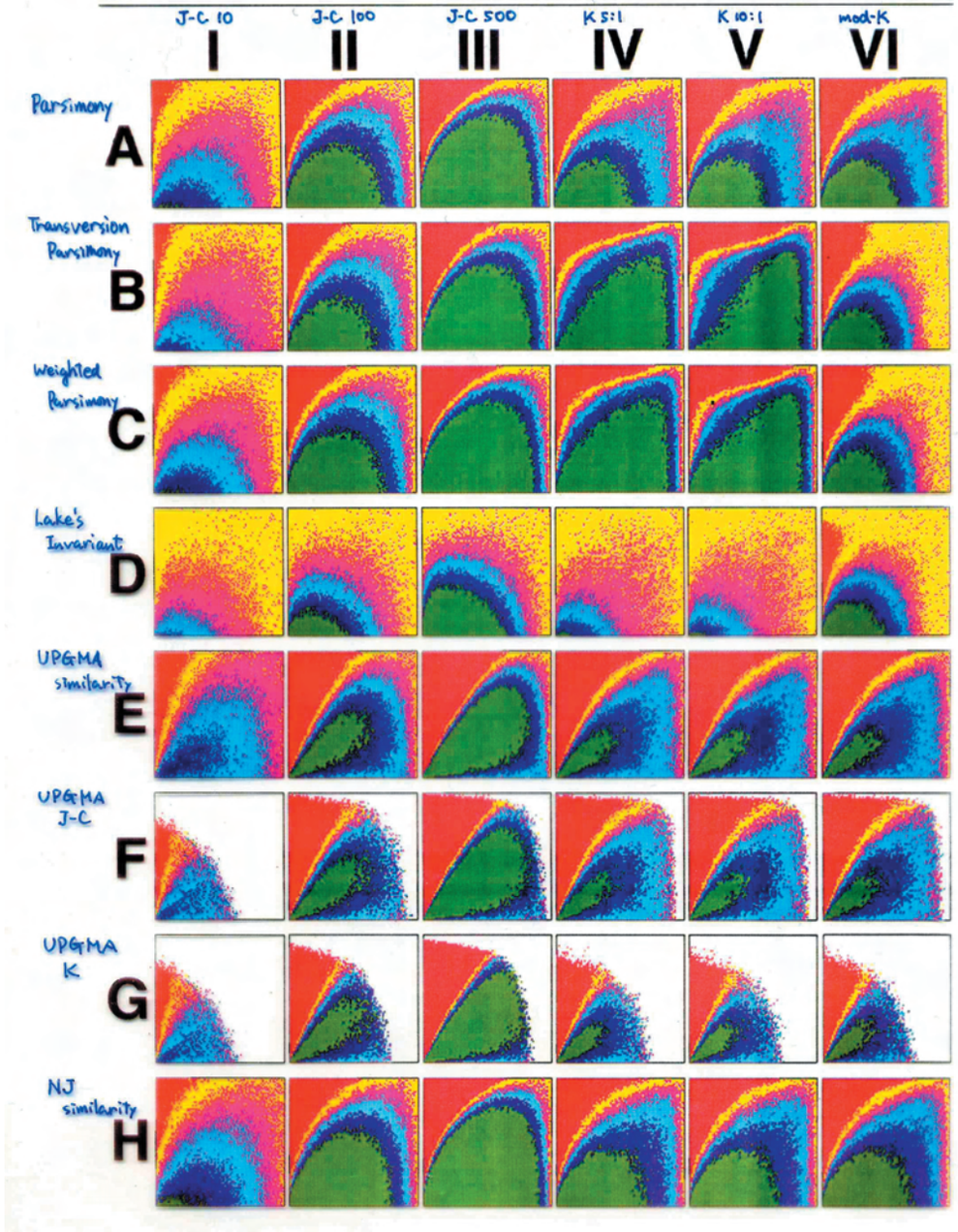
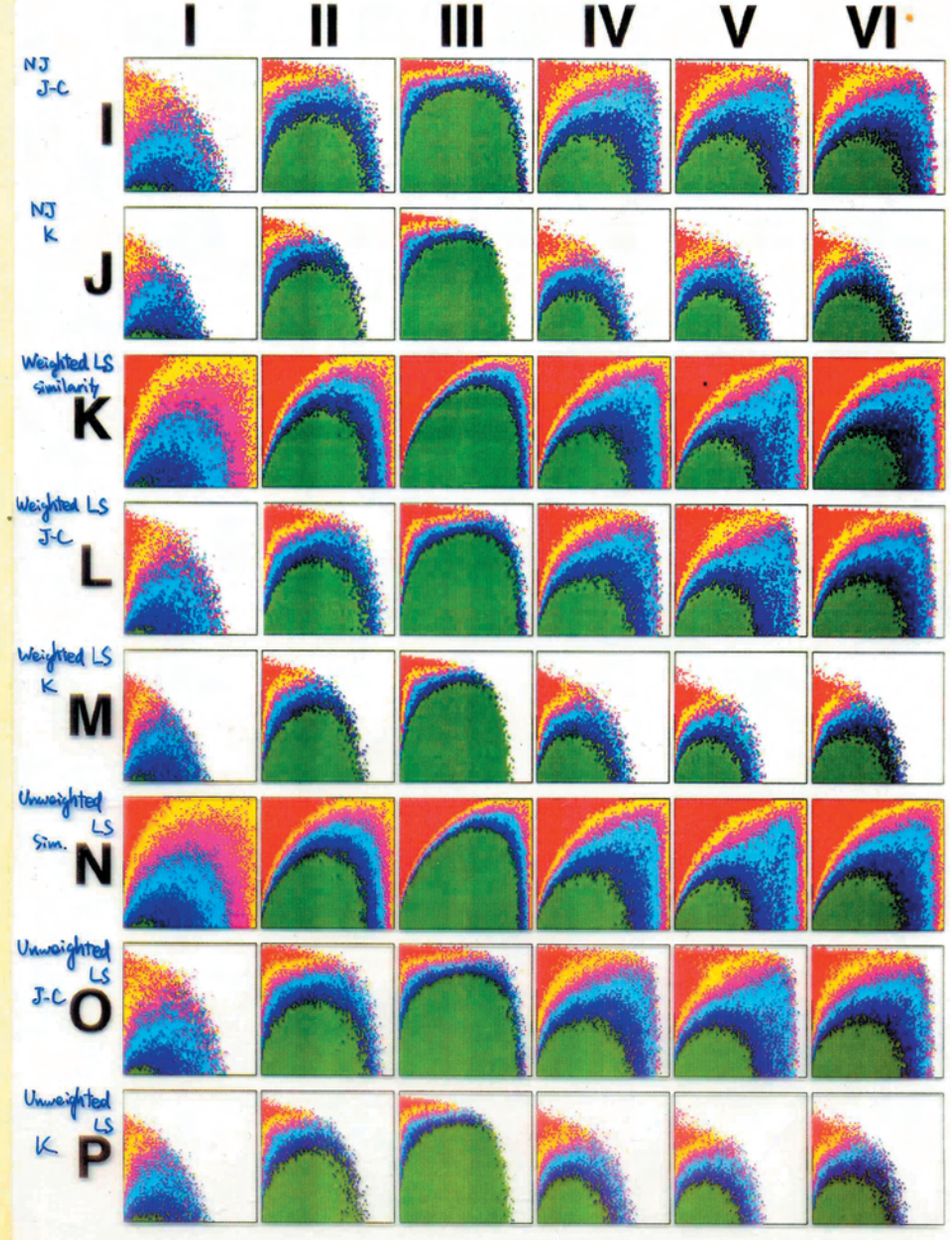



**Fig. 6.29** The tree space explored by the computer simulations analysed in Fig. 6.30. At each corner of the square the corresponding tree is drawn. Along the dotted line the rate of substitution is the same on each edge in the tree. The trees at top left and bottom right show two extremes in rate variation, differing in how much change occurs along the central edge. The region at the top left-hand corner has become known as the 'Felsenstein zone', and corresponds to parameters Felsenstein (1978) used to show that parsimony could be inconsistent. After Huelsenbeck and Hillis (1993: Fig. 4).


J-C 10
J-C 100
J-C 500
K 5:1
K 10:1
mod-K
I
II
III
IV
V
VI
Parsimony
A
Transversion Parsimony
B
Weighted Parsimony
C
Lake's Invariant
D
UPGMA similarity
E
UPGMA J-C
F
UPGMA K
G
NJ similarity
H




I
II
III
IV
V
VI
NJ J-C
I
NJ K
J
Weighted LS similarity
K
Weighted LS J-C
L
Weighted LS K
M
Unweighted LS Sim.
N
Unweighted LS J-C
O
Unweighted LS K
P

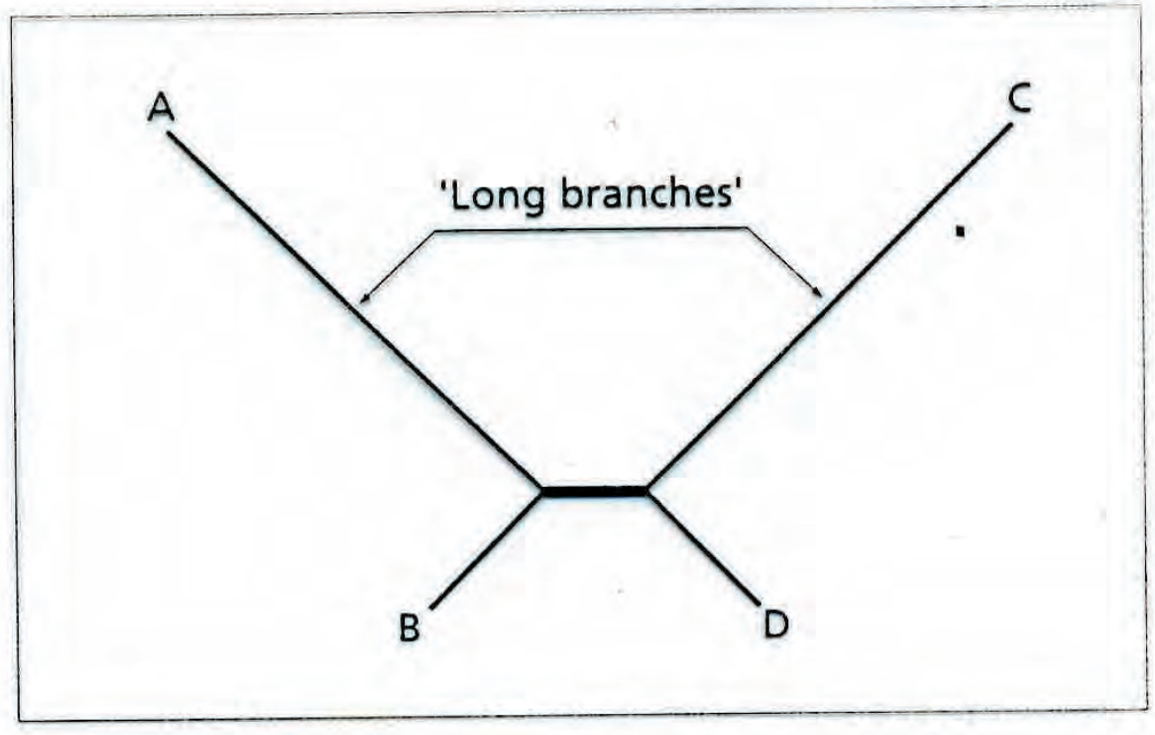


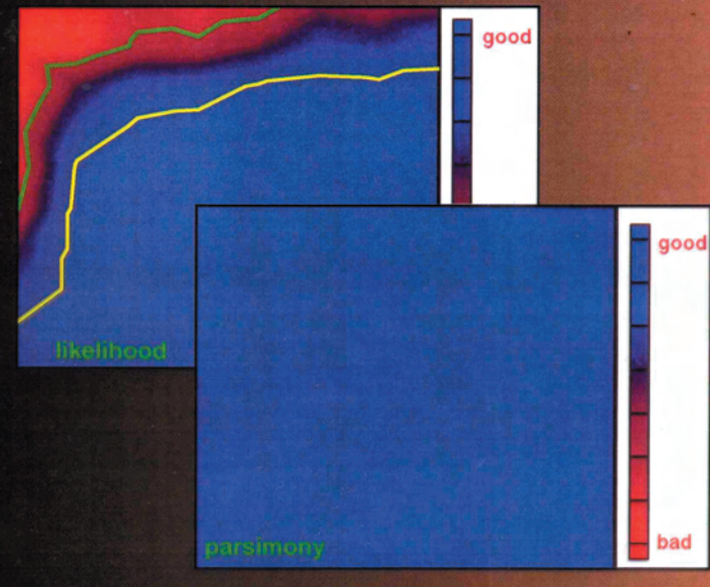

**Fig. 6.14** The problem of 'long branches attracting'. The edges leading to sequences A and C are long relative to the other branches in the tree, reflecting the relatively greater number of substitutions that have occurred along those two edges. If the difference in numbers of substitutions is sufficiently great, there may be more apparent support for the split {{A, C},{B, D}} than for the split {{A, B},{C, D}}.

ISSN 0748-3007
Volume 14, Number 3, September 1998

1999.11.20
農業環境技術研究所

# Cladistics

## The International Journal of the Willi Hennig Society




ACADEMIC PRESS

Cladistics 14, 209–220 (1998)
WWW http://www.apnet.com
Article No. cl980063

# Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone

Mark E. Siddall

*Laboratory of Phylohirudinology, Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.*

Received for publication 25 February 1998

**The accuracy of phylogenetic methods is reinvestigated for the four-taxon case with a two-edge rate and a three-edge rate. Unlike previous studies involving computer simulations, the two-edge rate relates to branches that are sister taxa in the model tree. As with previous studies, certain methods are found to behave inaccurately in a portion of the parameter space where the two-edge rate is proportionally large. This phenomenon, to which parsimony is immune, is termed "long-branch repulsion" and the region of poor performance is called the Farris Zone. Maximum likelihood methods are shown to be particularly prone to failure when closely related taxa have long branches. Long-branch repulsion is demonstrated with an empirical case involving Strepsiptera and Diptera.**

© 1998 The Willi Hennig Society

## INTRODUCTION

Molecular systematists may choose among many methods of phylogenetic inference. These include parsimony (Hennig, 1950; Kluge and Farris, 1969; Farris et al., 1970), phenetics (Sokal and Michener, 1958), neighbor-joining (Saitou and Nei, 1987), genetic distances (e.g. Fitch and Margoliash, 1967), or maximum likelihood (Felsenstein, 1973, 1981a). The ability of many of these methods to achieve their stated goals has already been abundantly countered, whereas others have not. For example, phenetic algorithms like UPGMA explicitly require that rates of change do not differ markedly across lineages on a tree if degree of similarity is to be a measure of phylogenetic relatedness. And so, in the absence of clock-like behavior, the data more often than not contravene a primary premise of UPGMA. Distance methods, while not necessarily requiring equal rates of change, still require metricity. In so far as data may violate the requisite triangle inequality or result in negative branch lengths, again the fundamental premise for these methods denies their legitimacy (Farris, 1981, 1985, 1986). Some methods, like neighbor-joining, although fast, always yield singular solutions to data that may properly have multiple optima. Phylogeneticists should be concerned that their data lack the power to resolve relationships and it is unfortunate that neighbor-joining methods are so prone to this sort of over-resolution (Farris et al., 1996).

Correspondence to: Mark Siddall. E-mail: msiddall@umich.edu; Fax: (313) 663 4080.

0748-3007/98/030209 + 12 $30.00/0
Copyright © 1998 by The Willi Hennig Society
All rights of reproduction in any form reserved

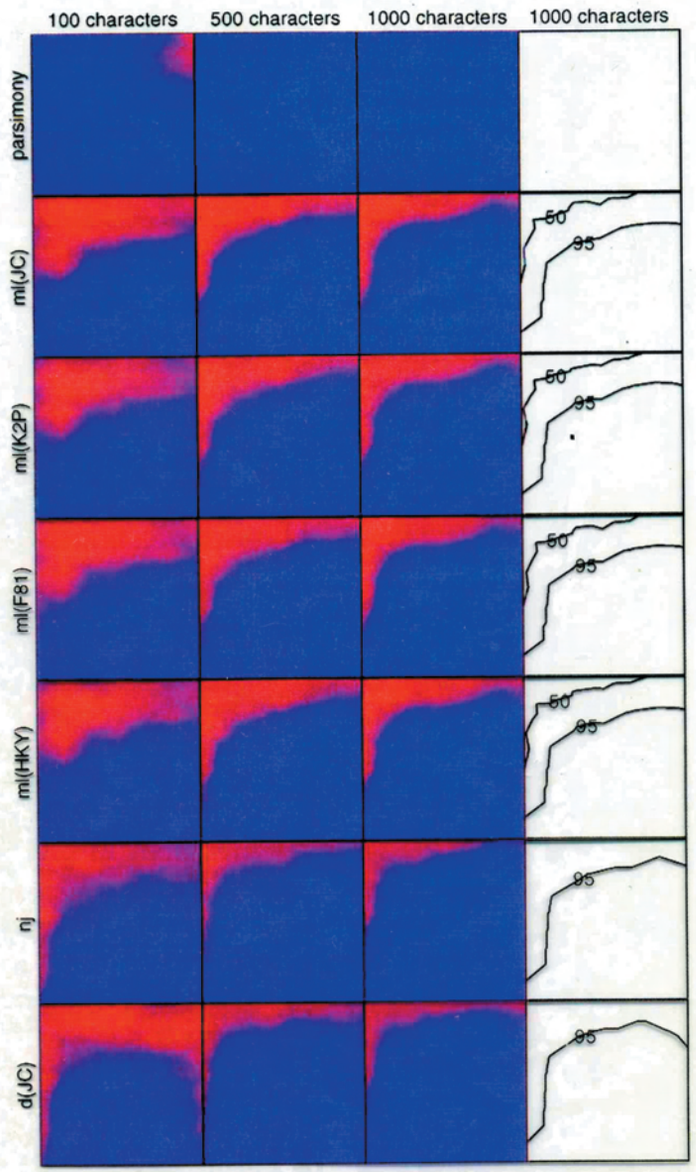



FIG. 4. The frequency of estimating the correct model tree for seven phylogenetic methods in the parameter space outlined in Fig. 3. Spline interpolated values of performance range from dark blue (100% recovery) to red (30% recovery) when 100, 500 or 1000 characters are free to vary. In addition, the 50% and 95% frequency contour lines are plotted for the case involving 1000 characters. These contours are absent for parsimony because, with 1000 characters free to vary, this method reconstructed the correct model tree more than 95% of the time across the whole parameter space. Similarly, the 50% contour lines are absent for nj and d(JC) plots. ml, maximum likelihood; nj, neighbor-joining; d, distance (a.k.a. "minimum evolution"); JC, Jukes–Cantor (1969) model; K2P, Kimura (1980) two-parameter model; F81, Felsenstein (1981a) model; HKY, Hasegawa–Kishino–Yano (1985) model.

# Felsenstein Zone

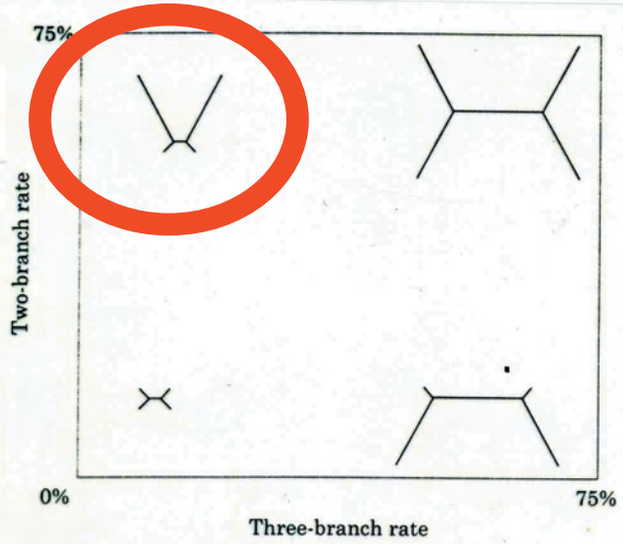



FIG. 1. Conditions previously employed in four-taxon data simulations in which the two-branch length on the ordinate relates to two unrelated terminal branches.

# Farris Zone

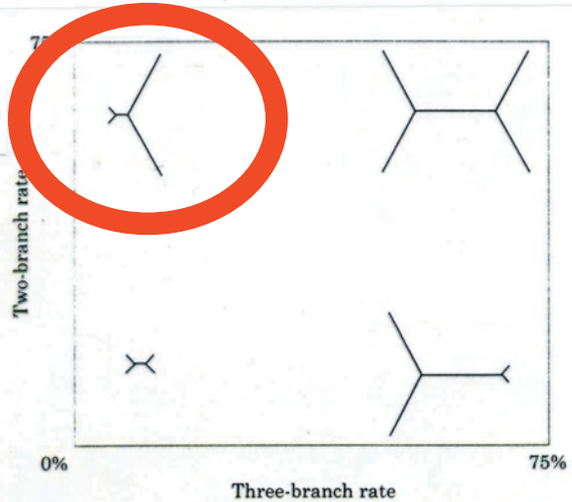
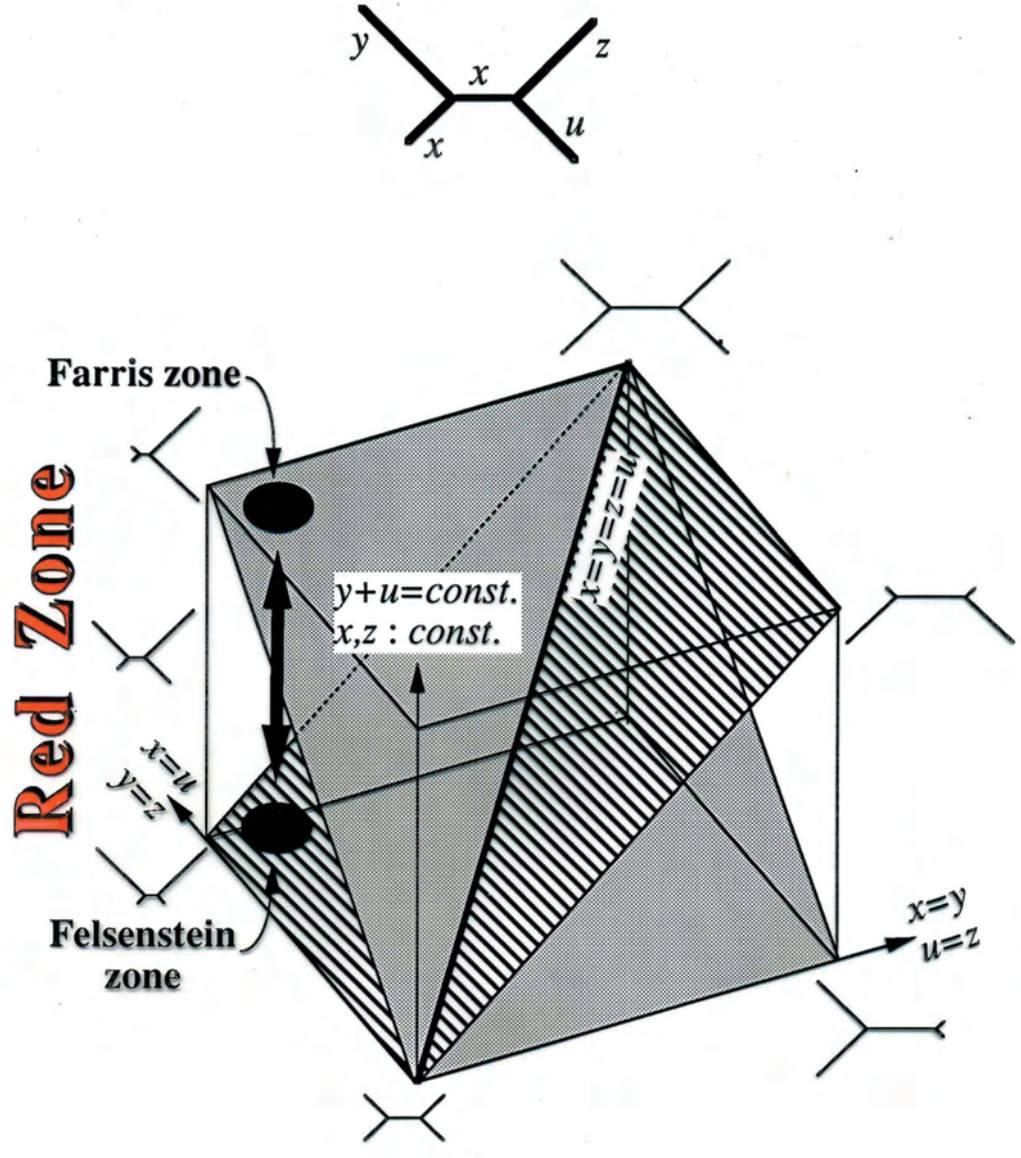



FIG. 3. Parameter space examined for the four-taxon model tree. The two-branch length is on the ordinate and the three-branch length is on the abscissa. Amounts of expected internodal change ranged from 0.75 to 75% in increments of 7.5%.

# 4-dimensional Tree-space



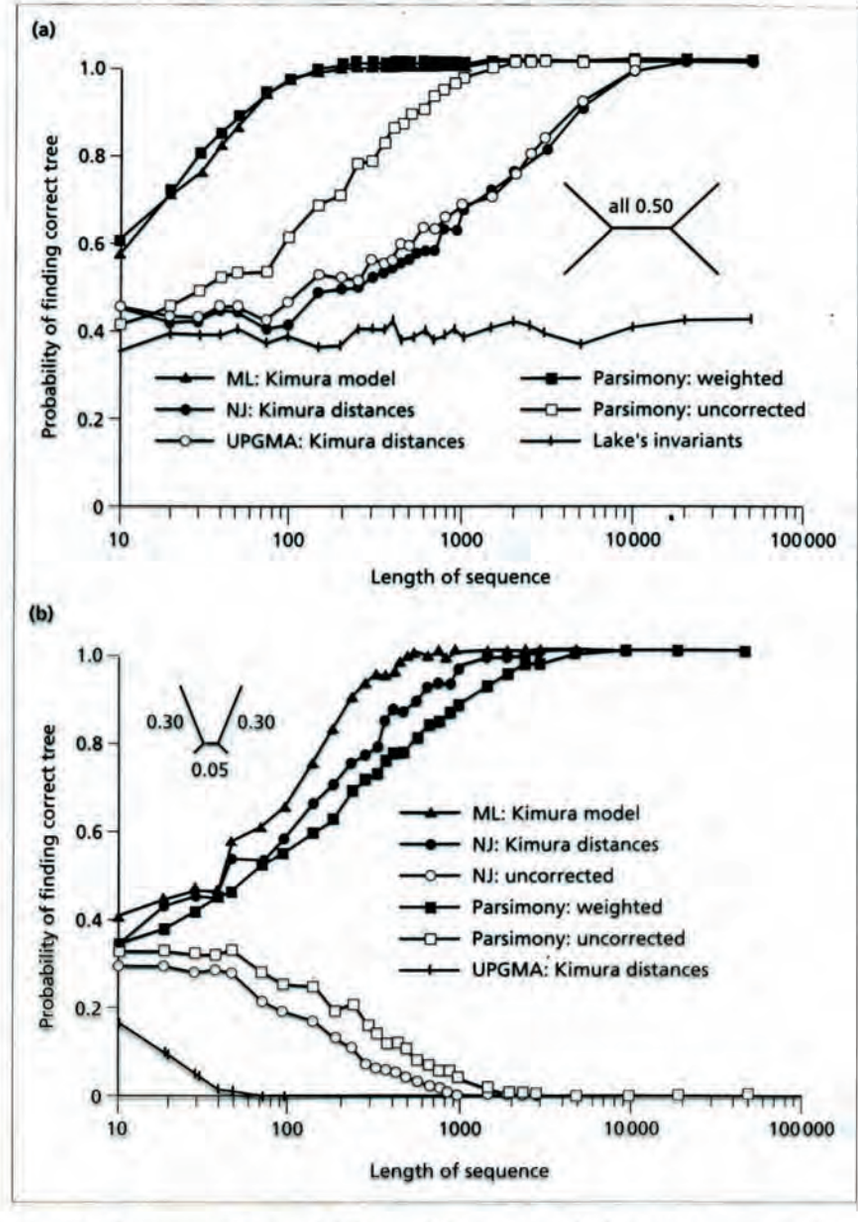



**Fig. 6.31** The accuracy of several different phylogenetic methods in reconstructing two four-taxon trees with (a) all edges equal in length and (b) with a short internal edge and two long terminal edges. In each graph the proportion of analyses that recovered the correct tree is plotted against the length of the simulated sequences. From Huelsenbeck *et al.* (1996: Figs 2 and 3).

# なお続く，
# 「方法」をめぐる論争
# （五十年戦争か）

## Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous

**Bryan Kolaczkowski**[1] **& Joseph W. Thornton**[2]

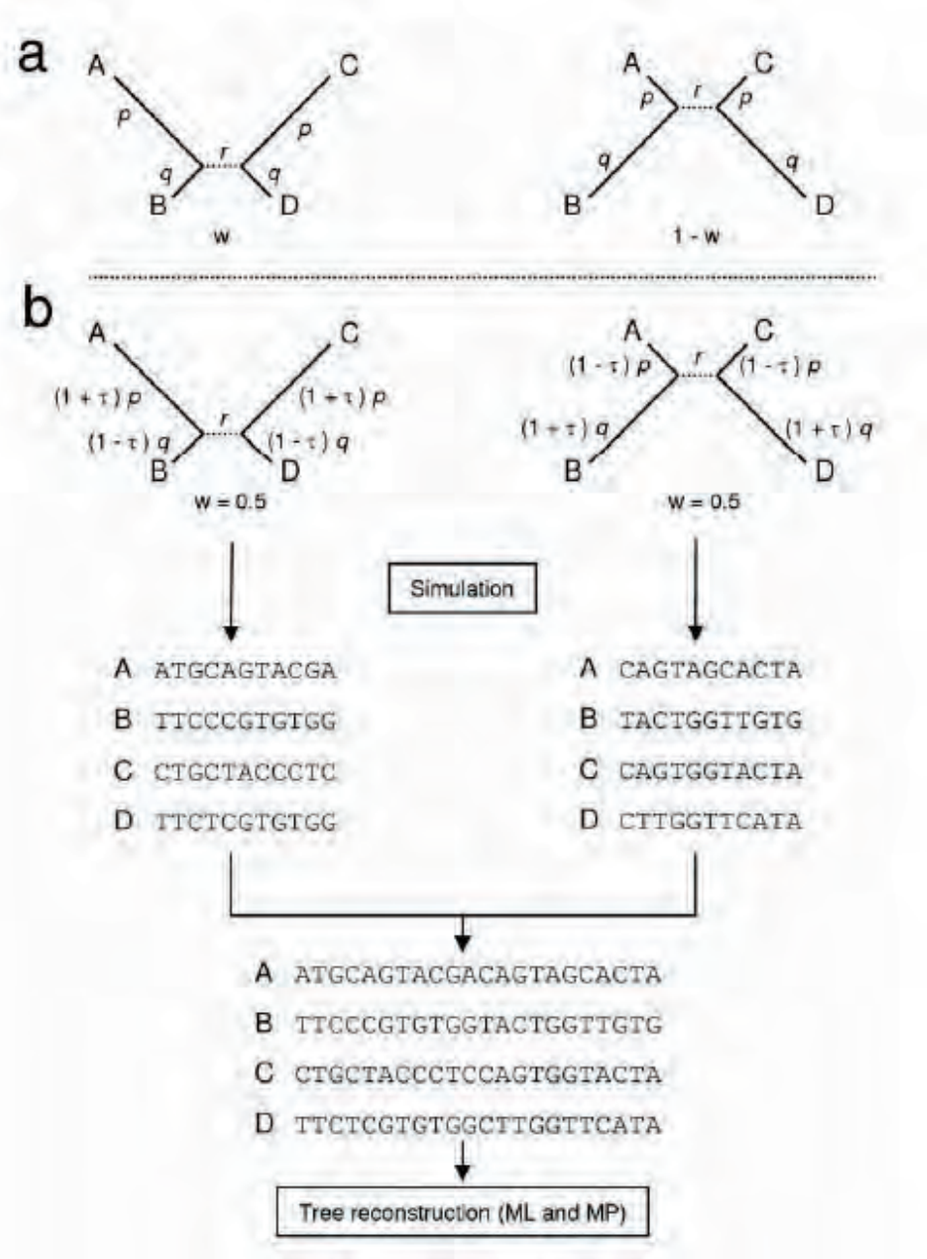
[1]*Department of Computer and Information Science and* [2]*Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon 97403, USA*

All inferences in comparative biology depend on accurate estimates of evolutionary relationships. Recent phylogenetic analyses have turned away from maximum parsimony towards the probabilistic techniques of maximum likelihood and bayesian Markov chain Monte Carlo (BMCMC). These probabilistic techniques represent a parametric ap… atistical phylogenetics, because their … pology—the probability … —is calculated with reference to an …olutionary model from which the data are assumed to be identically distributed. Maximum parsimony can be considered nonparametric, because trees are evaluated on the basis of a general metric—the minimum number of character state changes required to generate the data on a given tree—without assuming a specific distribution[1]. The shift to parametric methods was spurred, in large part, by studies showing that although both approaches perform well most of the time[2], maximum parsimony is strongly biased towards recovering an

NATURE | VOL 431 | 21 OCTOBER 2004 | www.nature.com/nature

B. Kolaczkowski & J. W. Thornton (2004)

第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法


a
A
C
p
p
r
q
q
B
D
w
1 - w
b
(1 + τ) p
(1 - τ) p
(1 - τ) q
(1 + τ) q
w = 0.5
w = 0.5
Simulation
A ATGCAGTACGA
B TTCCCGTGTGG
C CTGCTACCCTC
D TTCTCGTGTGG
A CAGTAGCACTA
B TACTGGTTGTG
C CAGTGGTACTA
D CTTGGTTCATA
A ATGCAGTACGACAGTAGCACTA
B TTCCCGTGTGGTACTGGTTGTG
C CTGCTACCCTCCAGTGGTACTA
D TTCTCGTGTGGCTTGGTTCATA
Tree reconstruction (ML and MP)

MP > ML, Bayes

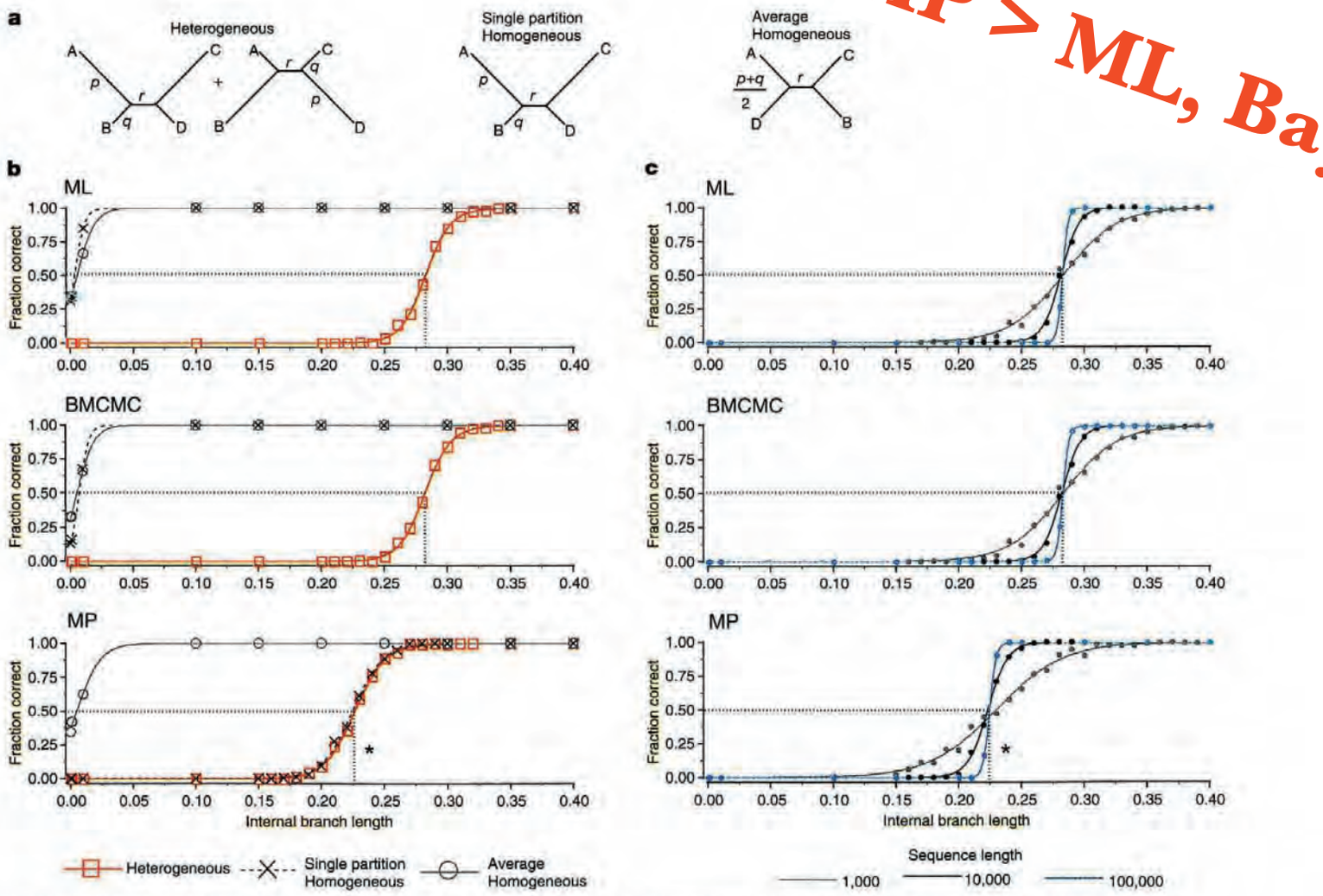



**Figure 1** Likelihood-based methods are less accurate than maximum parsimony (MP) under heterogeneous conditions. **a**, Trees on which heterogeneous and control sequences were simulated. **b**, Heterotachy reduces the accuracy of likelihood methods. Accuracy is plotted against internal branch length for sequences with and without strong heterotachy. Dotted lines, $BL_{50}$ for each method (asterisk: maximum parsimony < maximum likelihood (ML) and BMCMC, $P < 0.001$). **c**, Likelihood methods are inconsistent below the $BL_{50}$ under strong heterotachy, recovering the incorrect tree with increasing frequency as the amount of data increases.



第 3 回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法

# Likelihood, Parsimony, and Heterogeneous Evolution

*Matthew Spencer,*† Edward Susko,* and Andrew J. Roger†*

*Department of Mathematics and Statistics and †Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

**No!**

**M. Spencer *et al.* (2005), MBE, 22(5)**

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

# Heterotachy and long-branch attraction in phylogenetics

Hervé Philippe*[1], Yan Zhou[1], Henner Brinkmann[1], Nicolas Rodrigue[1] and Frédéric Delsuc[1,2]

Address: [1]Canadian Institute for Advanced Research, Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada and [2]Laboratoire de Paléontologie, Phylogénie et Paléobiologie, Institut des Sciences de l'Evolution, UMR 5554-CNRS, Université Montpellier II, France

Email: Hervé Philippe* - herve.philippe@umontreal.ca; Yan Zhou - y.zhou@umontreal.ca; Henner Brinkmann - henner.brinkmann@umontreal.ca; Nicolas Rodrigue - nicolas.rodrigue@umontreal.ca; Frédéric Delsuc - delsuc@isem.univ-montp2.fr

* Corresponding author

H. Philippe *et al.* (2005), BMC Evol. Biol., 5(50)

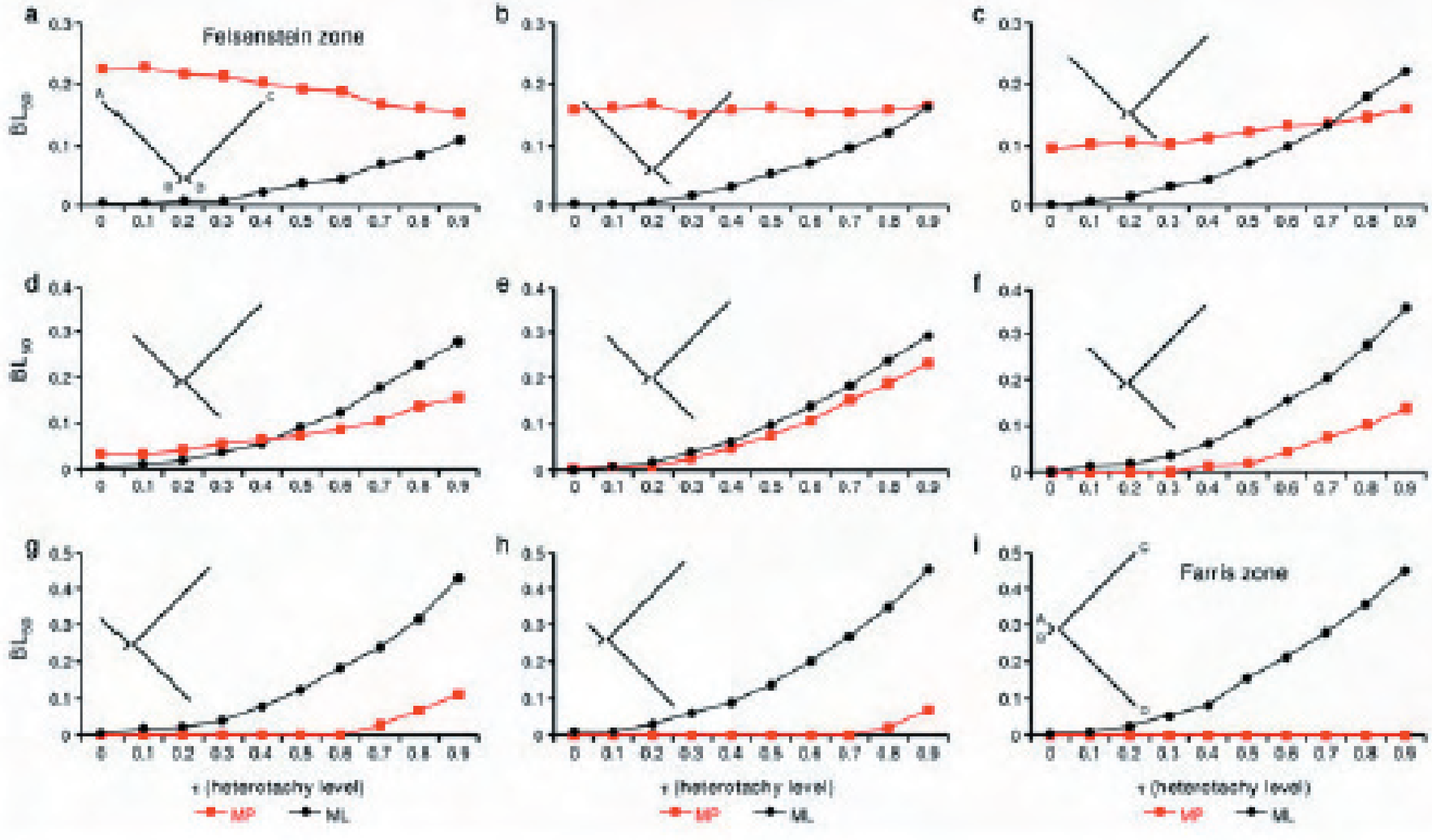
Published: 06 October 2005

*BMC Evolutionary Biology* 2005, 5:50 doi:10.1186/1471-2148-5-50

Received: 21 July 2005
Accepted: 06 October 2005

This article is available from: http://www.biomedcentral.com/1471-2148/5/50

No!


a
Felsenstein zone
b
c
d
e
f
g
h
i
Farris zone
τ (heterotachy level)
MP
ML

Misleading Bayesian MCMC

# Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees

Elchanan Mossel[1] and Eric Vigoda[2]

Markov chain Monte Carlo (MCMC) algorithms play a critical role in the Bayesian approach to phylogenetic inference. We present a theoretical analysis of the rate of convergence of many of the widely used Markov chains. For *N* characters generated from a uniform mixture of two trees, we prove that the Markov chains take an exponentially long (in *N*) number of iterations to converge to the posterior distribution. Nevertheless, the likelihood plots for sample runs of the Markov chains deceivingly suggest that the chains converge rapidly to a unique tree. Our results rely on novel mathematical understanding of the log-likelihood function on the space of phylogenetic trees. The practical implications of our work are that Bayesian MCMC methods can be misleading when the data are generated from a mixture of trees. Thus, in cases of data containing potentially conflicting phylogenetic signals, phylogenetic reconstruction should be performed separately on each signal.

E. Mossel & E. Vigoda (2005)

Bayesian inference is one of the most popular methods in phylogeny [illegible] Many widely [illegible] PAML (*4*), rely [illegible] chain Monte Carlo (MCMC) methods. These algorithms are often known as BMCMC. Part of the appeal of BMCMC is that they are supposed to be more robust and faster than standard maximum likelihood approaches. Our results show that these appealing features are overly optimistic in some settings.

The basis of the MCMC algorithms is a Markov chain whose stationary distribution is the desired posterior distribution. Reliable phylogenetic estimates depend on the Markov chain converging to the posterior distribution



第３回最先端育種セミナー（17 ～ 19 November 2005，神戸）
〈家畜・水産動物における遺伝的多様性研究―その目的と新しい方法論―〉
三中信宏：生物進化の歴史を推定する―分子系統樹を構築する理論と方法

# どの系統推定法を使えばいいか？

- 距離法　計算速い／単一解／情報損失（よくまちがう）／距離尺度が重要
- 最節約法　速度ほどほど／形質なんでもオッケー／複数解／まちがうことあり（Felsenstein Zone）
- 最尤法　速度遅い／分子データのみ／進化モデル依存／まちがうことあり（Farris Zone）
- ベイズ法　成長株／速度遅い（最尤法よりマシ）／事前確率あやしい／収束の判断に苦しむ

## 本日の噺をまとめると ―

- 事物の系譜や歴史をデータに基づいて推論する系統学は，既知の現在から未知の過去を復元するという一般歴史科学としての性格を帯びている．直接的な観察や実験的な観察が不可能であるという点で自然科学の範疇からはみ出している．
- 系統学の図形言語は離散数学（グラフ理論）に基づいている．グラフィック言語（チェイン，ツリー，ネットワークなど）を用いて，事物の多様性とその系譜は可視化されてきた．
- 利用可能な形質データによっては確率統計モデルに基づく系統推定論が適用できるだろう．ただし，確率論や統計学を系統推定に用いる際の前提条件は十分に理解する必要がある．