最尤法による系統樹推定の理論と実際

下平英寿（東工大），曹纓（統数研）

1．尤度原理，確率モデル，ベイズ推定
2．系統樹推定の実際
3．推定の信頼性評価
   (a) ブートストラップ確率
   (b) Kishino-Hasegawa 検定
   (c) Shimodaira-Hasegawa 検定
   (d) マルチスケールブートストラップ

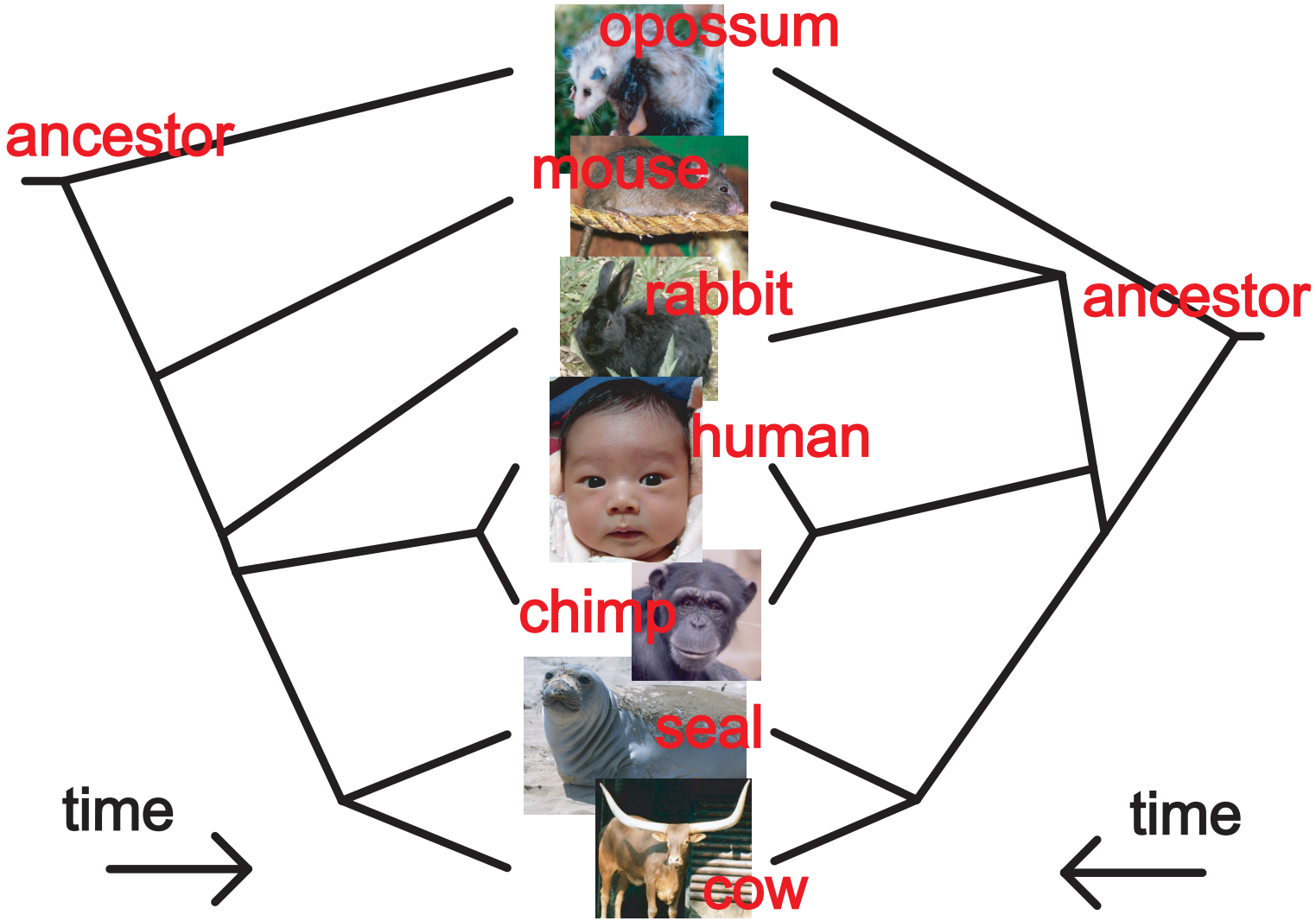
ohp-200208-sinka: 2002-08-04

最尤法による系統樹推定

```
human     GCCAACCTCCTACTCCTCATTGTACCCATTCTAATCGCAATGGCATTCCTAATGCTTACCGAACGAAAAATTCTAGGCTATATACAACTACGCAAAGGC
chimp     ACCAACCTCCTACTCCTCATTGTACCCATCCTAATCGCAATAGCATTCCTAATGCTAACCGAACGAAAAATTCTAGGCTACATACAACTACGCAAAGGT
seal      ATTAATATCATCTCACTAATTATCCCAATTCTCCTCGCCGTAGCTTTCCTAACATTAGTAGAACGGAAAGTACTAGGCTACATACAACTCCGAAAAGGA
cow       ATTAACATCTTAATACTAATTATTCCCATCCTATTGGCCGTAGCATTCCTTACGTTAGTGGAACGAAAAGTTCTAGGCTATATACAACTCCGAAAAGGT
rabbit    ATTAATACACTCCTTTTAATCCTACCTGTACTTTTAGCCATAGCATTCCTCACCTTAGTCGAACGAAAAATCTTAGGGTACATACAACTACGTAAAGGC
mouse     ATTAATATCCTAACACTCCTCGTCCCCATTCTAATCGCCATAGCCTTCCTAACATTAGTAGAACGCAAAATCTTAGGGTACATACAACTACGAAAAGGC
opossum   ATTAACTTATTAATATATATTATCCCTATCCTCCTAGCTGTAGCATTTTTAACTCTAGTAGAACGAAAAGTATTAGGCTATATACAATTCCGAAAAGGC
```

1. 進化のメカニズムを確率モデルで表現
2. 系統樹に関する仮説を立てる
3. データの確率を計算　＝　尤度
4. 尤度を最大にする系統樹を選ぶ
5. 推定の信頼性を評価

哺乳類の進化


opossum
ancestor
mouse
rabbit
ancestor
human
chimp
seal
time
time
cow

統計的推測　簡単な例

画びょうを１００回なげて落とし，表を向くか，裏を向くか？

1	2	3	4	5	6	7	8	9	10
表	裏	表	裏	裏	表	裏	裏	裏	表
11	12	13	14	15	16	17	18	19	20
裏	裏	裏	表	表	表	裏	裏	裏	表
91	92	93	94	95	96	97	98	99	100
裏	表	表	表	裏	表	裏	裏	表	裏

結果：　表＝43回，裏＝57回

$$\text{表を向いた頻度} = \frac{43}{100} = 0.43$$

画びょうが表を向く頻度を１０回毎に区切ってみる

	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	1-100
表	4	4	4	3	6	3	3	5	6	5	43
裏	6	6	6	7	4	7	7	5	4	5	57
表の割合	0.4	0.4	0.4	0.3	0.6	0.3	0.3	0.5	0.6	0.5	0.43

- 標本 ＝ 有限 $\Rightarrow$ 頻度のバラツキ

- 母集団 ＝ 実験回数が非常に多い (無限) のとき

  頻度の極限 $\Rightarrow$ 画びょうが表を向く確率

実際には有限回しか実験は行えない．同じ条件で実験を多数繰り返すのは大変難しい$\Rightarrow$ 確率というのは抽象的な概念 $\Rightarrow$ モデル

確率モデル： ２項分布

$n =$ 実験の回数

$x =$ 表を向いた回数 (データ)

$p =$ 表を向く確率 (パラメタ)

尤度

$$L(p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

対数尤度

$$\ell(p) = \log L(p)$$

最尤法

p	$L(p)$	$\log L(p)$
0.0	0.	$-\infty$
0.1	9.4×10^{-18}	-39.21
0.2	1.0×10^{-7}	-16.11
0.3	0.0018	-6.29
0.4	0.0667	-2.71
0.5	0.0300	-3.50
0.6	0.0002	-8.38
0.7	1.3×10^{-8}	-18.15
0.8	3.7×10^{-16}	-35.52
0.9	4.1×10^{-31}	-69.97
1.0	0.	$-\infty$

p	$L(p)$	$\log L(p)$
0.40	0.0667	-2.71
0.41	0.0740	-2.60
0.42	0.0787	-2.54
0.43	0.0804	-2.52
0.44	0.0788	-2.54
0.45	0.0741	-2.60
0.46	0.0670	-2.70
0.47	0.0582	-2.84
0.48	0.0486	-3.02
0.49	0.0390	-3.24
0.50	0.0301	-3.50

$p =$ 表を向く確率，$L(p) =$ 尤度

検定

最尤推定

$$\widehat{p} = \frac{x}{n}$$

推定値は近似的に正規分布に従う．

平均$= p$**，標準誤差**$= \sqrt{p(1-p)/n}$

仮説： $p = 0.5$**，データ：** $x = 43$, $n = 100$

確率値 (p**-value)**

$$\Pr\left\{\frac{X}{n} \leq \frac{x}{n}\right\} = 0.081 \quad \textbf{片側}$$

$$\Pr\left\{\left|\frac{X}{n} - p\right| \geq \left|\frac{x}{n} - p\right|\right\} = 0.162 \quad \textbf{両側}$$

最尤法の特徴

- 仮定や仮説が確率モデルで明示的に表現される．

  (良い点) いろいろなアイデアを統合的に組み込める

  (悪い点) 確率モデルが誤っていると推測の信頼性が失われる

  $\Rightarrow$ データの蓄積に伴い，モデルを更新して行く

- 評価基準が明確．

  最大化すべき関数（尤度）と，最大化をおこなう探索アルゴリズムの分離

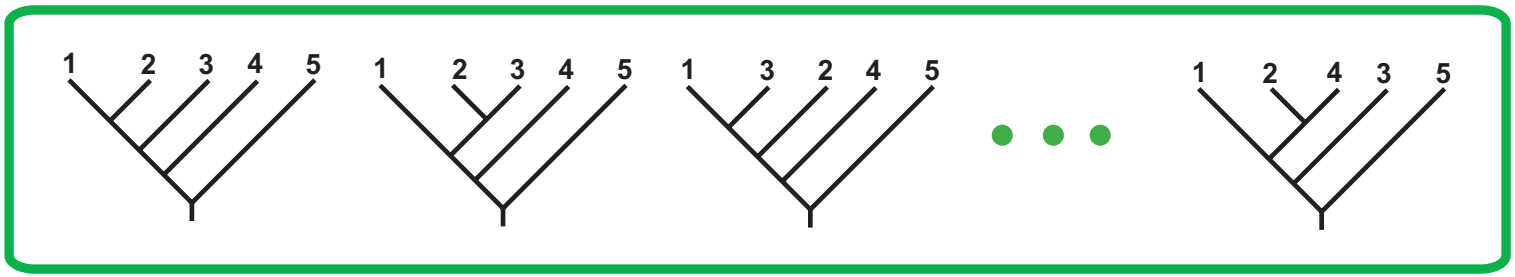
- 尤度原理は統計学の中心のひとつ．ベイズ法とも直接関係．

系統樹推定

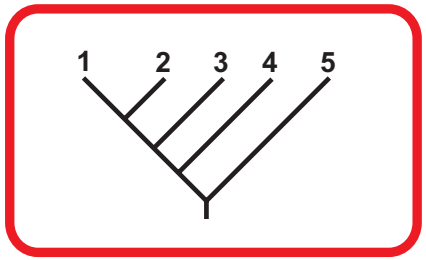
DATA

```
          2         3         4         5         6         7         8         9
          012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
human     ERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPATSTITLYITAPTLALTIALLLWAPLPMPNPLVNLNLGLL
seal      ERKVLGYMQLRKGPNIVGPYGLLQPIADAVKLFTKEPLRPLTSSTTMFIMAPILALALALTMWVPLPMPYPLINMNLGVL
cow       ERKVLGYMQLRKGPNVVGPYGLLQPIADAIKLFIKEPLRPATSSASMFILAPIMALGLALTMWIPLPMPYPLINMNLGVL
rabbit    ERKILGYMQLRKGPNIVGPYGLLQPIADAIKLFTKEPLRPLTSSPLLFIIAPTLALTLALSMWLPIPMPYPLVNLNMGIL
mouse     ERKILGYMQLRKGPNIVGPYGILQPFADAMKLFMKEPMRPLTTSMSLFIIAPTLSLTLALSLWVPLPMPHPLINLNLGIL
opossum   ERKVLGYMQFRKGPNVIGPYGILQPFADALKLFIKEPLRPMTSSISMFTIAPTLALTLAFTIWTPLPMPNALLDLNLGLL
```

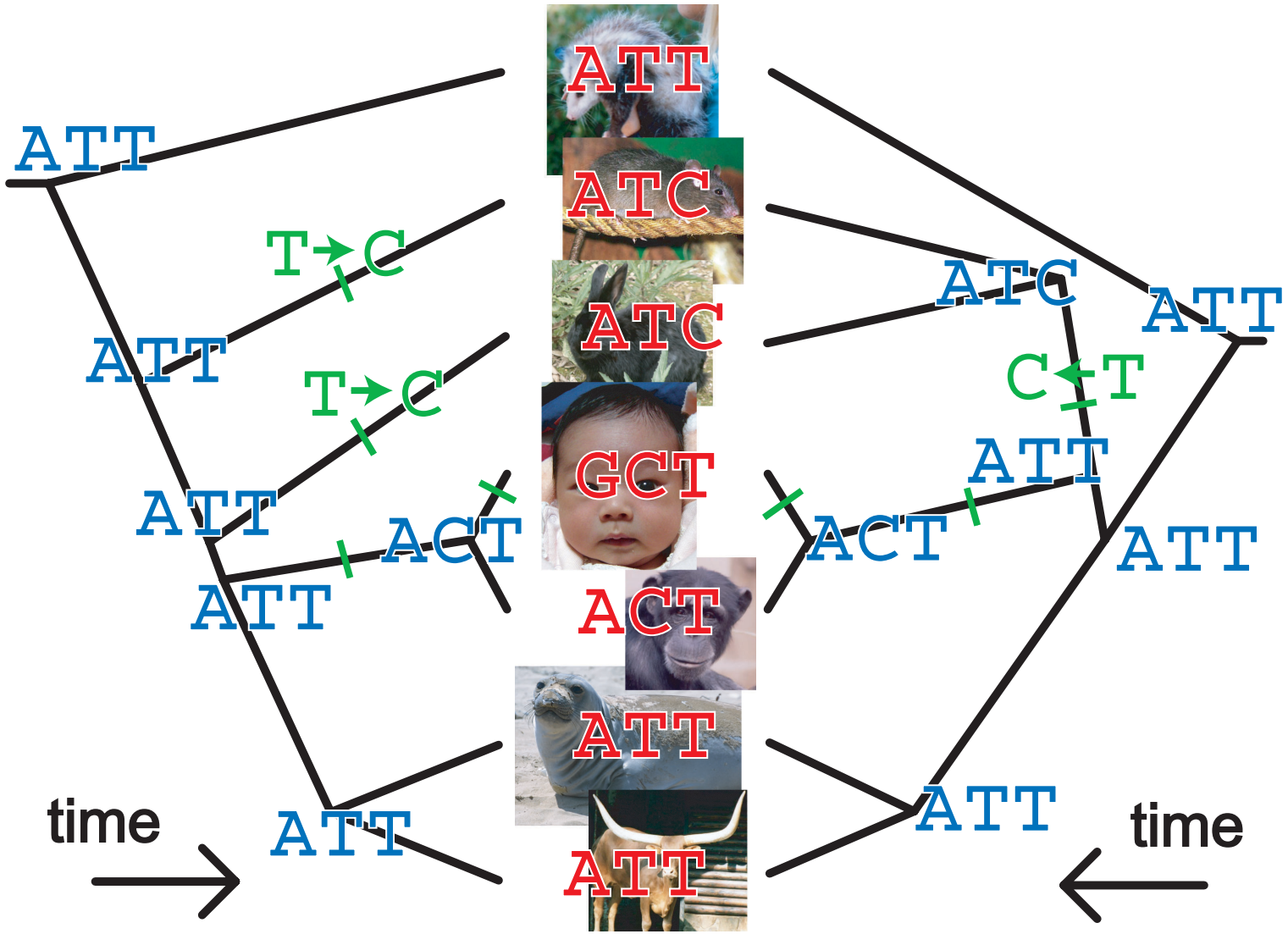
COMPARISON



SELECTION



塩基置換


ATT
ATC
ATC
GCT
ACT
ATT
ATT
ATT
ATT
ATT
ATT
ATT
ATT
ATC
ATT
ACT
ATT
ATT
T→C
T→C
C←T
ACT
time
time

分子時計

１００万年でＤＮＡの１%が変異すると仮定する

２００万年でＤＮＡの２%が変異する

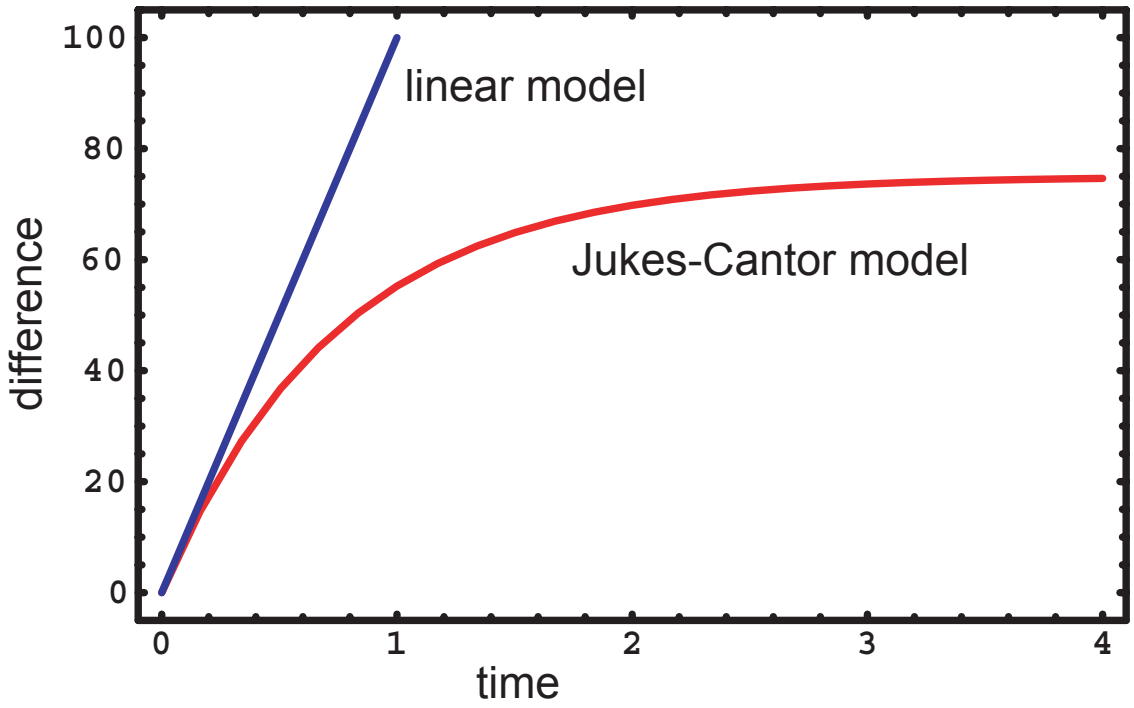
１０００万年でＤＮＡの９%が変異する

１億年でＤＮＡの５５%が変異する

time	dif	
0	0%	AAAAAAAAAA.....AAAAAAAAAA
0.01	1%	AAAATAAAAA.....AAAAAAAAAA
0.1	9%	AAAATAAAAA.....AAAAACAAAA
0.2	18%	AAGATAAAAA.....AATAACAAAA
0.3	25%	AAGAAAAAAA.....AATAACAAAG
1.0	55%	AAGTTAACAA.....AATAGCATGC
3.0	74%	TACTCGACTT.....CTATAGCAGC

置換過程

７５％までしか変わらない

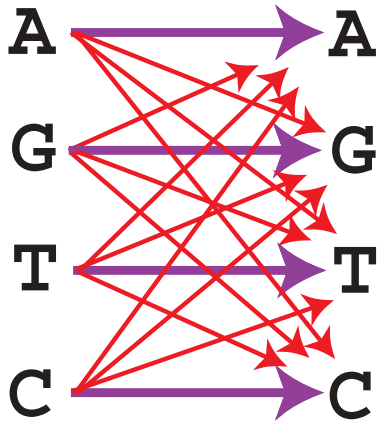


$$P(t) = \frac{3}{4}\left(1 - \exp\left(-\frac{4}{3}\lambda t\right)\right)$$

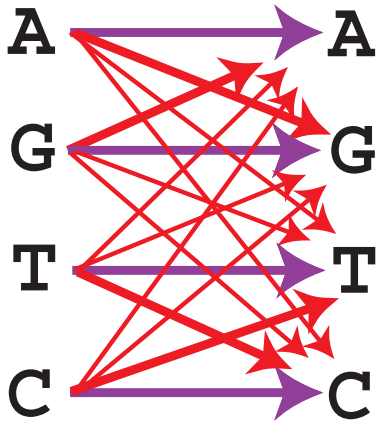
いろいろな置換モデル

確率モデルの一種である「マルコフ過程」でモデリング

- 塩基置換（A,T,G,Cの４種類)



JCモデル



HKYモデル

(トランジションとトランスバージョン，組成比)

- アミノ酸置換（２０種類)

Dayhoffモデル，Jonesモデルなど

Markov Process for Evolution

- **Probability of future state-b given current state-a after time-t**

$$P_{ba}(t)$$

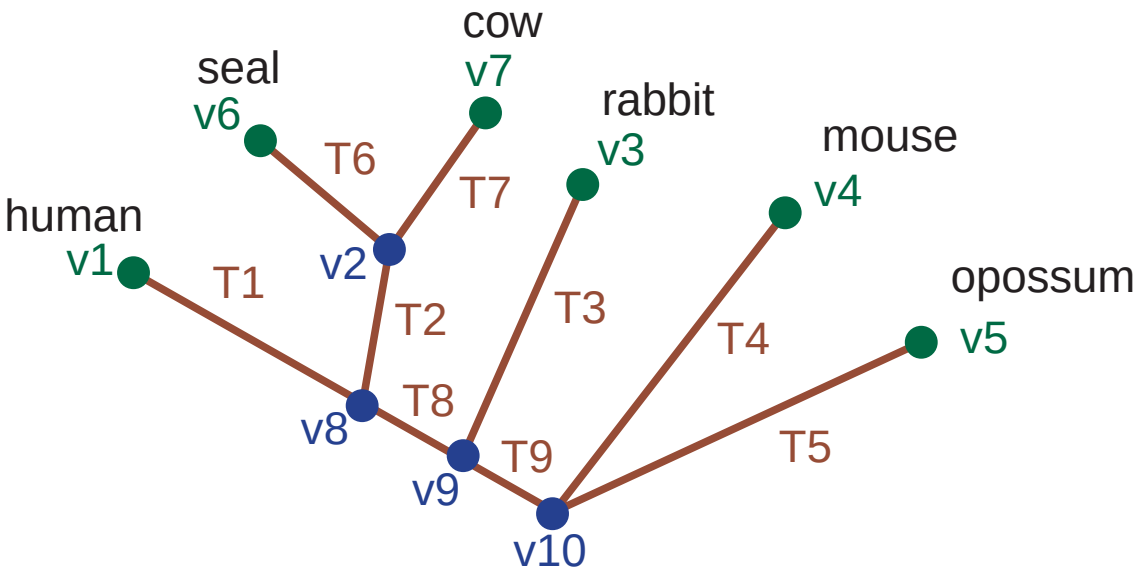
note: $a, b \in \{\mathsf{A}, \mathsf{T}, \mathsf{G}, \mathsf{C}\}$ **for nucleotide sequences.**

- **Transition probability matrix**

$$P(t) = \exp(tQ) = \sum_{k \geq 0} \frac{t^k}{k!} Q^k$$

note: $Q = 4 \times 4$ **matrix for nucleotide sequences, and** $Q = 20 \times 20$ **matrix for those of amino-acid.**

系統樹の確率モデル



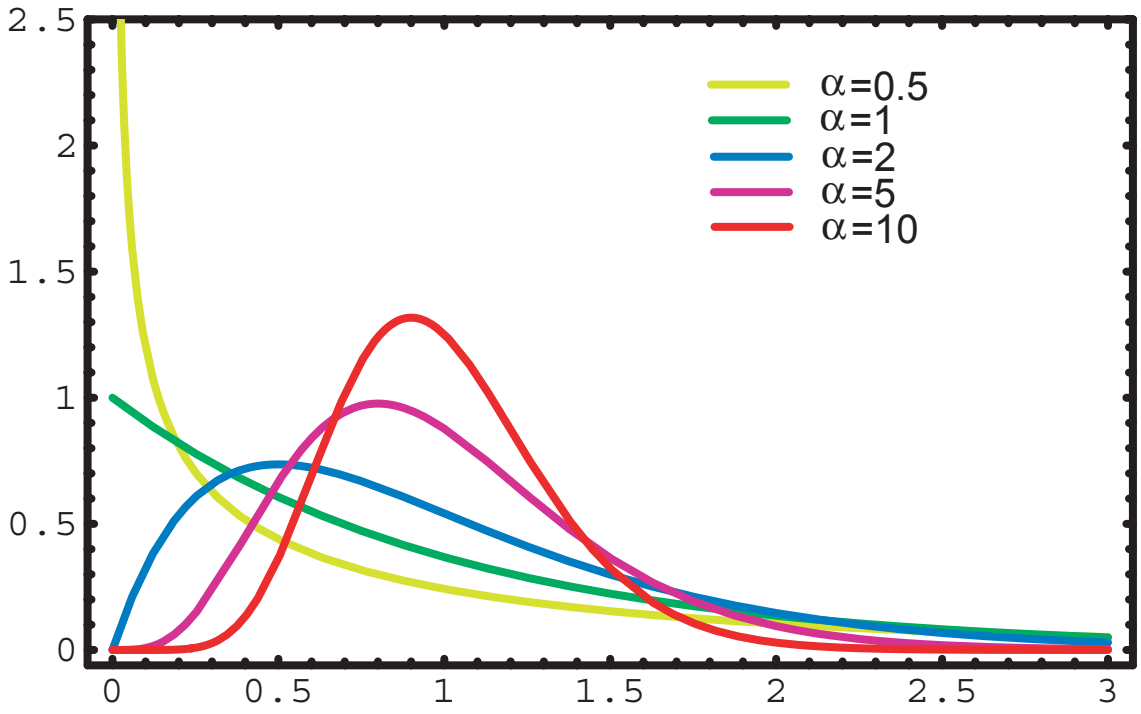
$$\theta = (T1, T2, T3, T4, T5, T6, T7, T8, T9, Q, \pi)$$

サイト h における配列パタン X_h の確率

$$P(X_h|\theta) = \sum_{v2,v8,v9,v10} P(v1|v8;T1)P(v2|v8;T2)P(v3|v9;T3) \times P(v4|v10;T4)P(v5|v10;T5)P(v6|v2;T6) \times P(v7|v2;T7)P(v8|v9;T8)P(v9|v10;T9)P(v10)$$

各枝でのマルコフ過程(つまり進化)は独立という仮定

進化速度の不均質性

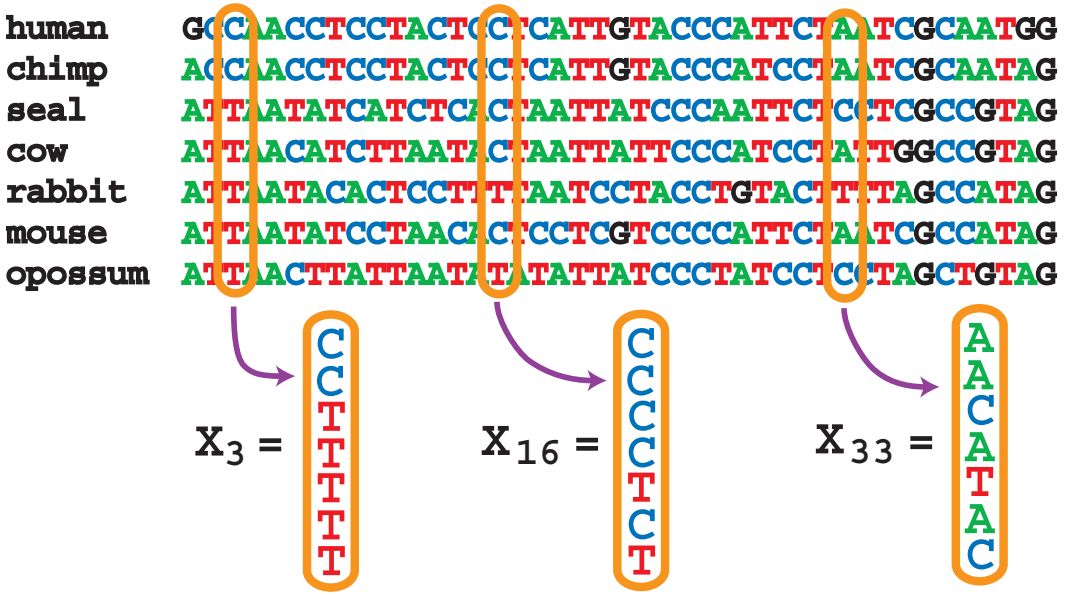


ガンマ分布の確率密度関数（平均＝１）

⇒ どのサイトが早いか遅いかが未知

$$P(X_h|\theta) = \int P(X_h|\theta,\lambda) f(\lambda|\alpha)\, d\lambda$$

配列データの確率モデル



各サイトは独立に進化すると仮定

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2)\cdots P(X_n)$$

$\Rightarrow$ **配列データを観測する確率**

尤度の計算と最尤法

系統樹トポロジ $T = 1, 2, \ldots, M$ **(例えば 105)**

配列データ X

パラメタベクトル θ **(枝の長さ，組成比など)**

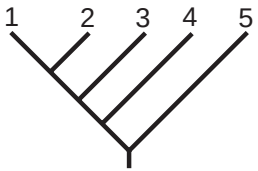
1．トポロジ T**，パラメタ** θ **の尤度** $L(\theta, T) = P(X|\theta, T)$

2．トポロジ T**の尤度** $L(T) = L(\hat{\theta}, T)$**．** $\hat{\theta}$**はパラメタの最尤推定**

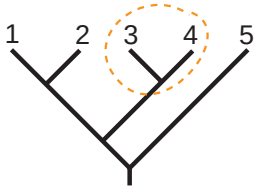
3． $L(1), L(2), \ldots, L(M)$**のうち最大値をとるトポロジを選ぶ**

系統樹の対数尤度

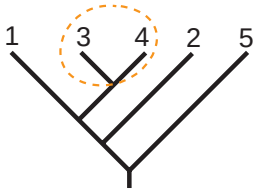
1=human, 2=(seal, cow), 3=rabbit, 4=mouse, 5=opossum



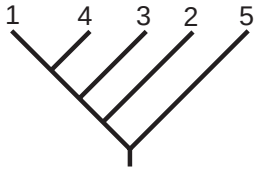
log L = 0



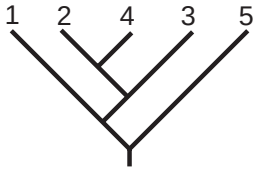
log L = 17.6



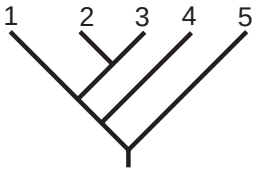
log L = 20.6



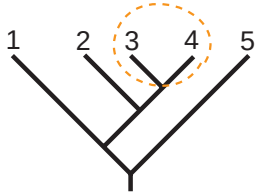
log L = 26.3



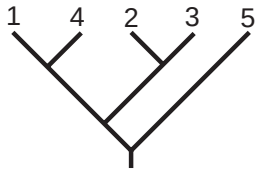
log L=31.7



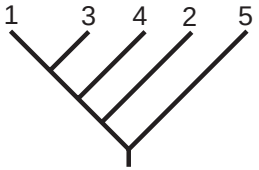
log L = 2.7



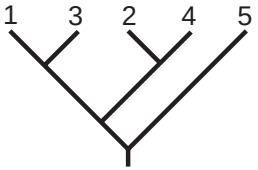
log L = 18.9



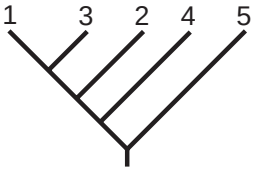
log L = 22.2



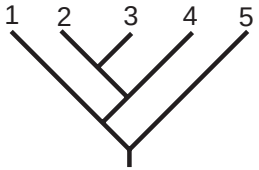
log L = 28.9



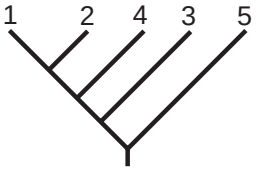
log L = 34.7



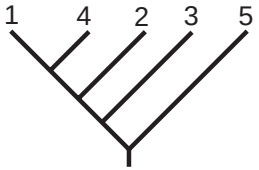
log L = 7.4



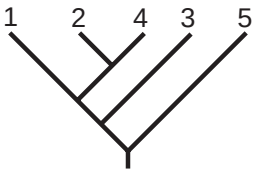
log L = 20.1



log L = 25.4



log L = 31.6



log L = 36.2

1=human, 2=(seal,cow), 3=rabbit, 4=mouse, 5=opossum

log L = the log-likelihood difference

ベイズ推測

ベイズの定理

$$P(T|X) \propto P(X|T)P(T)$$

系統樹推定においては近似的に

$$\textbf{トポロジ}\,T\,\textbf{の事後確率} \approx \frac{L(T)}{L(1) + L(2) + \cdots + L(M)}$$

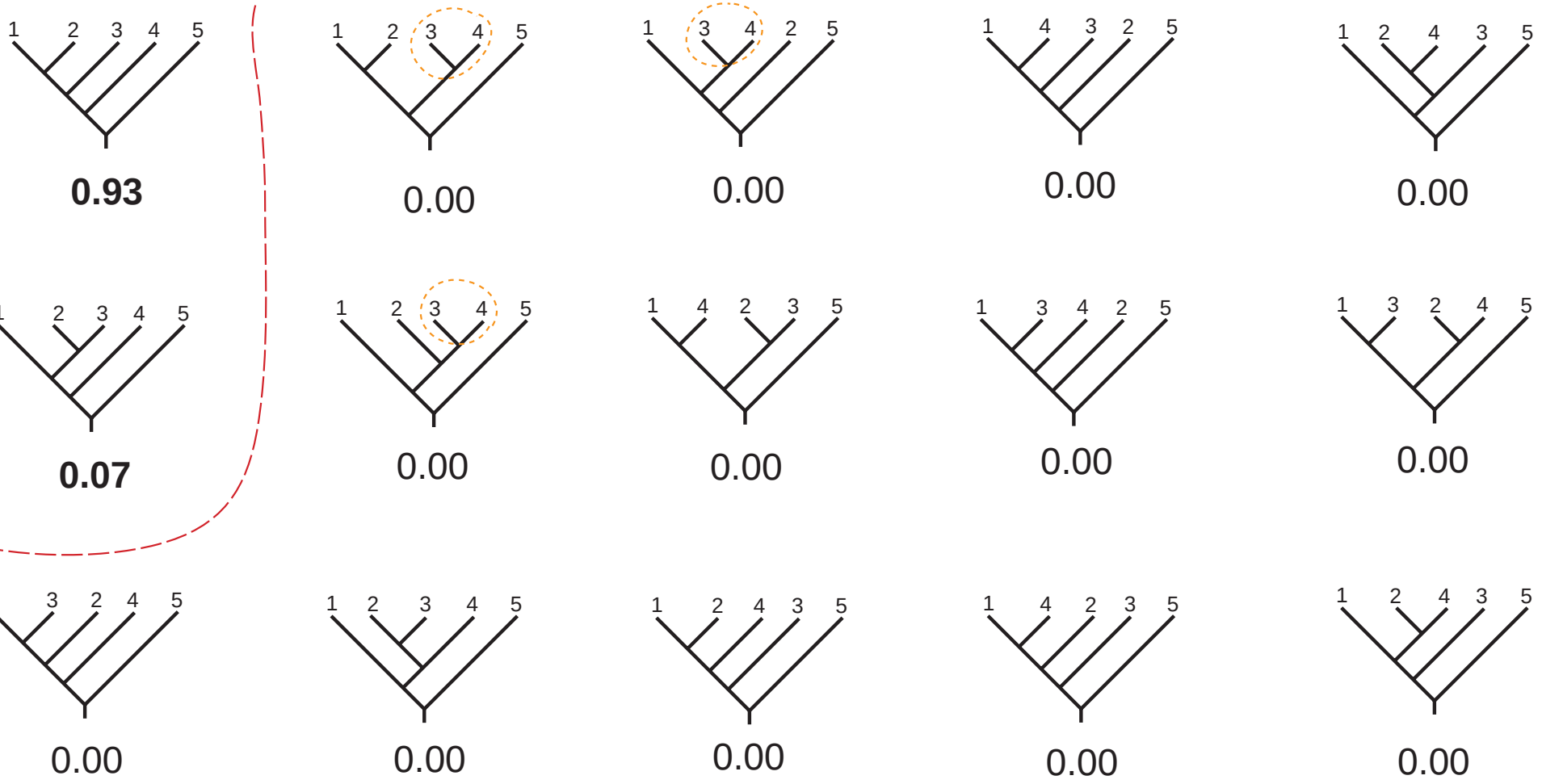
Schwarz (Annals of Statistics 1978)

$$\log P(X|T) \approx \log P(X|\hat{\theta}, T) - \frac{\dim \theta}{2} \log n$$

$$P(T|X) \propto L(T)P(T)$$

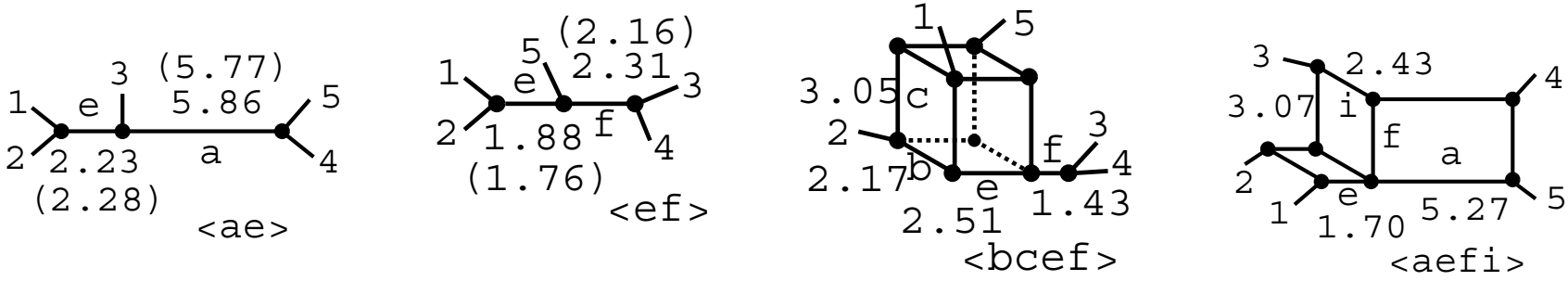
系統樹の事後確率

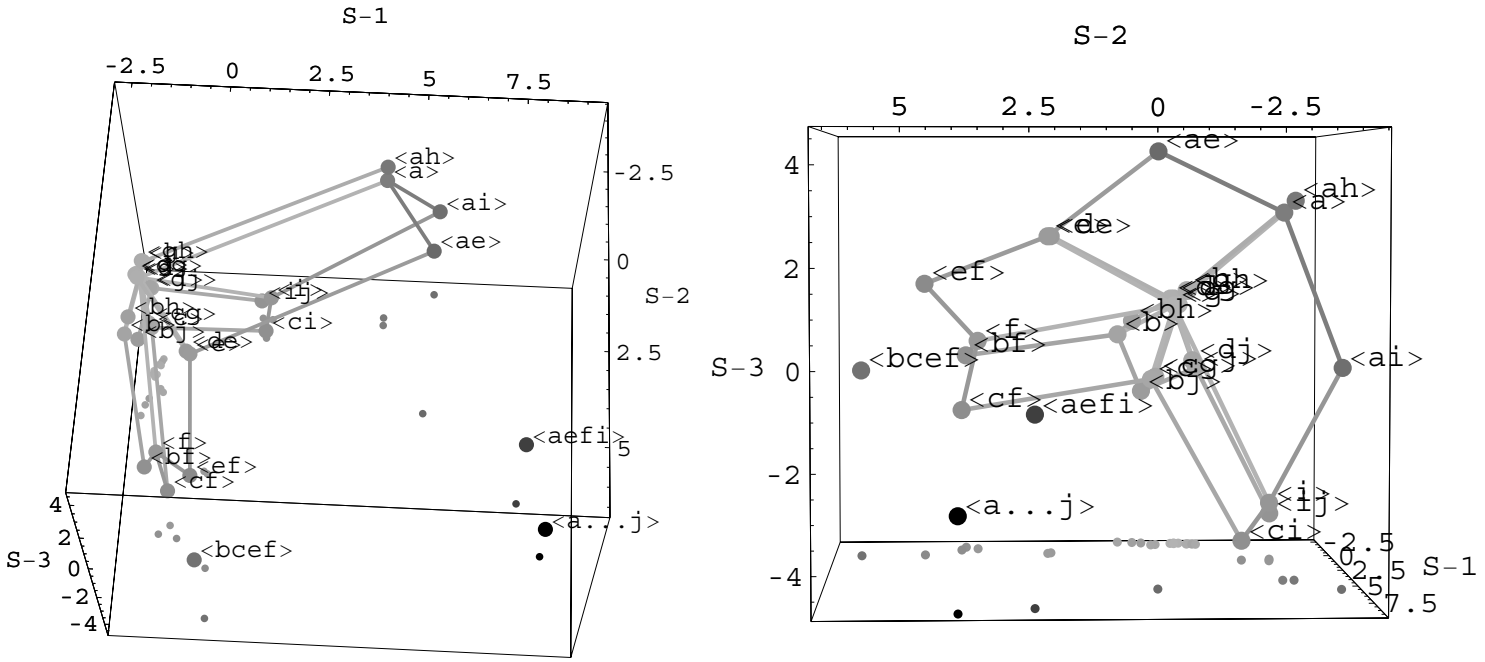
1=human, 2=(seal, cow), 3=rabbit, 4=mouse, 5=opossum



1=human, 2=(seal,cow), 3=rabbit, 4=mouse, 5=opossum

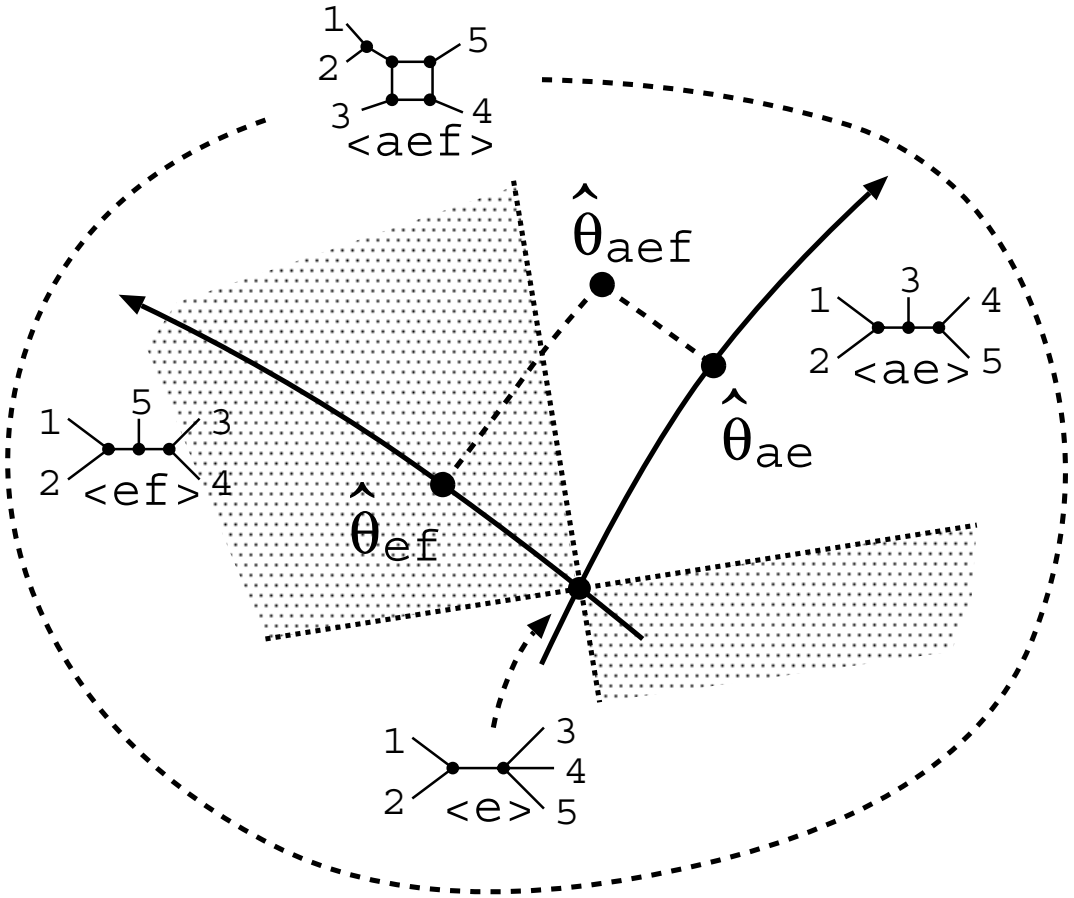
確率モデルの誤り（misspecification）

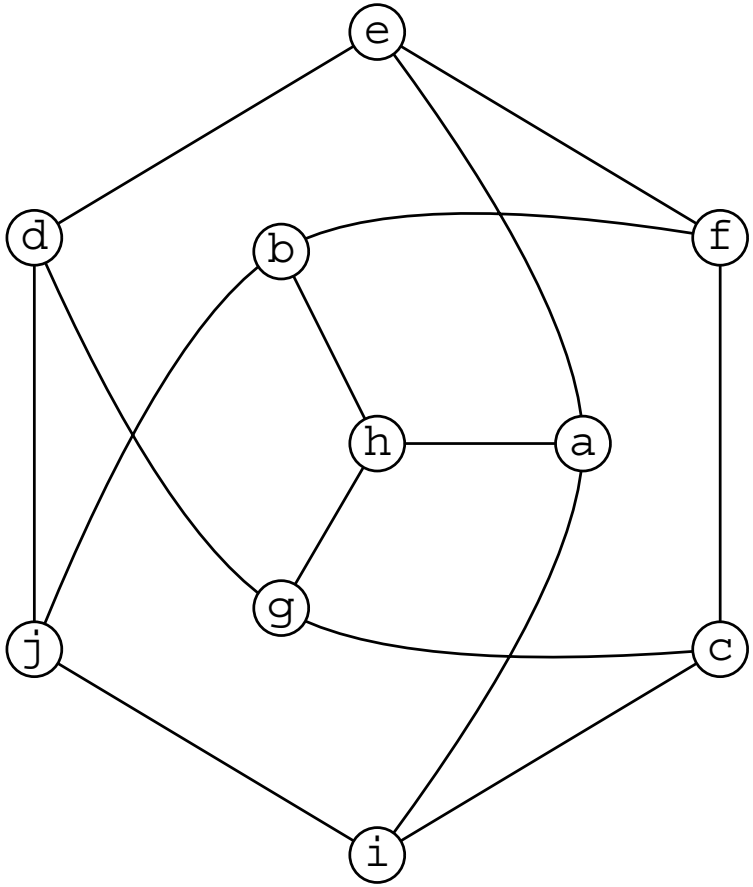




Shimodaira (2001) *Comm. Stat. A Theory Meth.* **30, 1751–1772.**

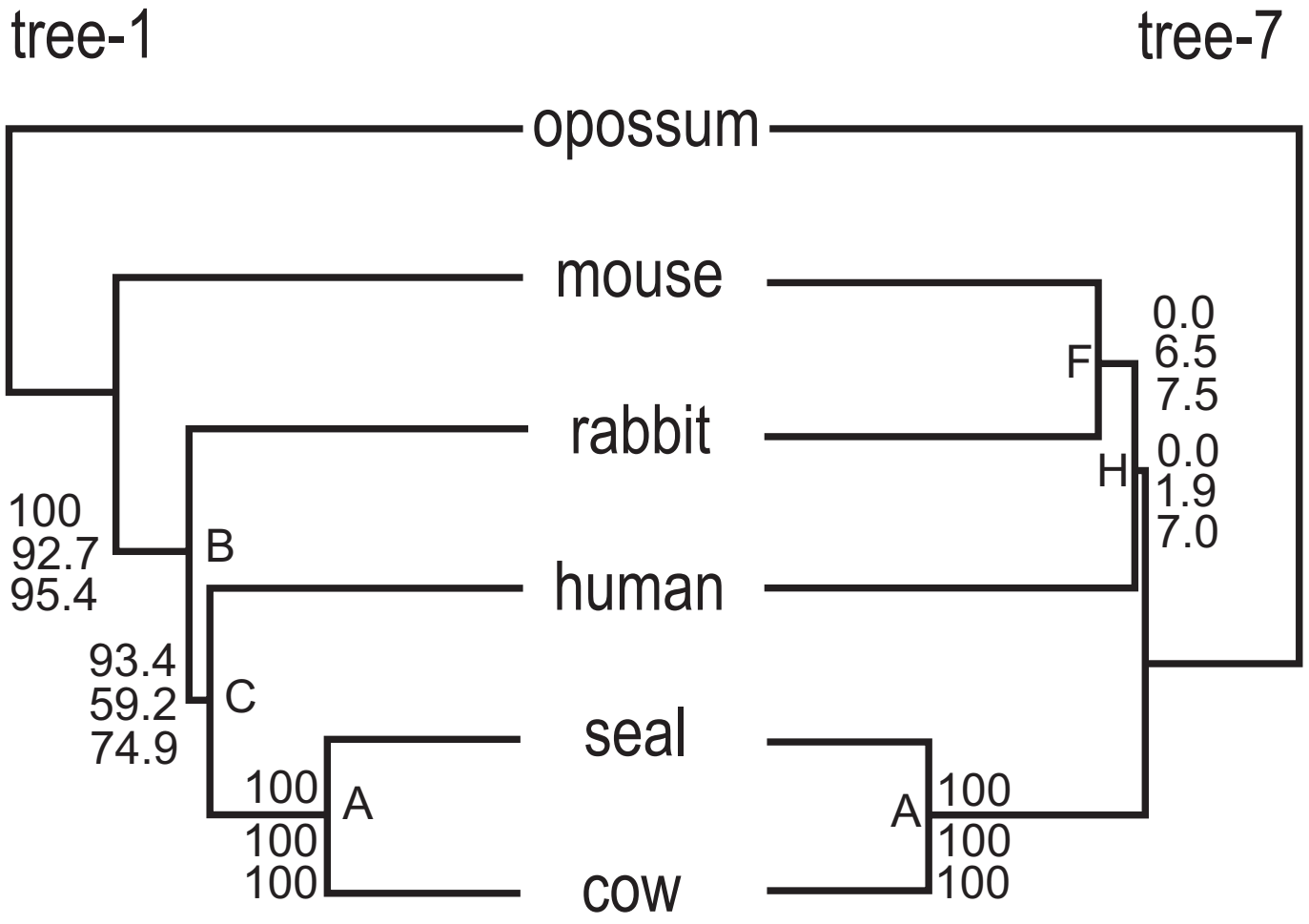
確率分布の空間





$a=\{1,2,3\}$, $e=\{1,2\}$, $f=\{3,4\}$

信頼性評価



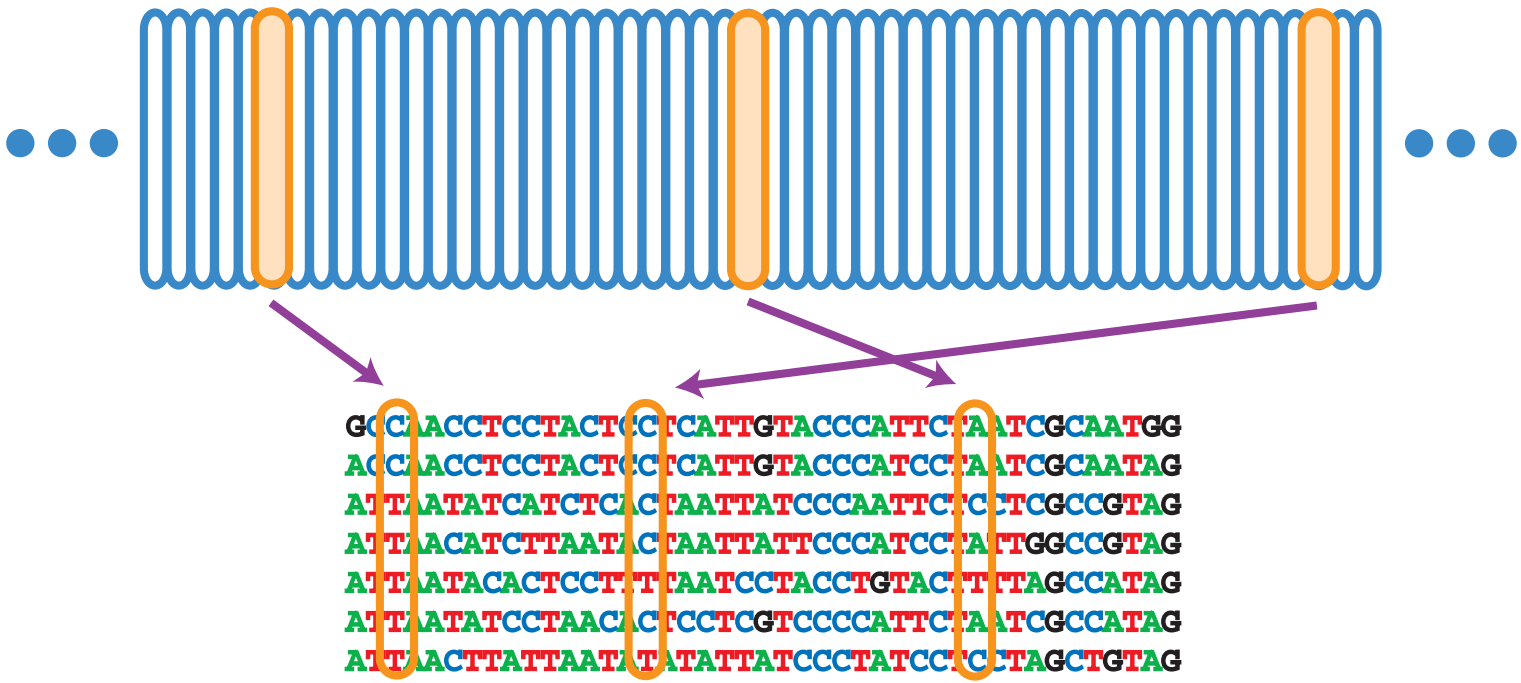
ベイズ事後確率，ブートストラップ確率，近似的に不偏な確率値

推定の信頼性評価

1. データのバラツキ
2. ブートストラップ確率
3. Kishino-Hasegawa検定
4. Shimodaira-Hasegawa検定
5. マルチスケールブートストラップ法（AU検定）

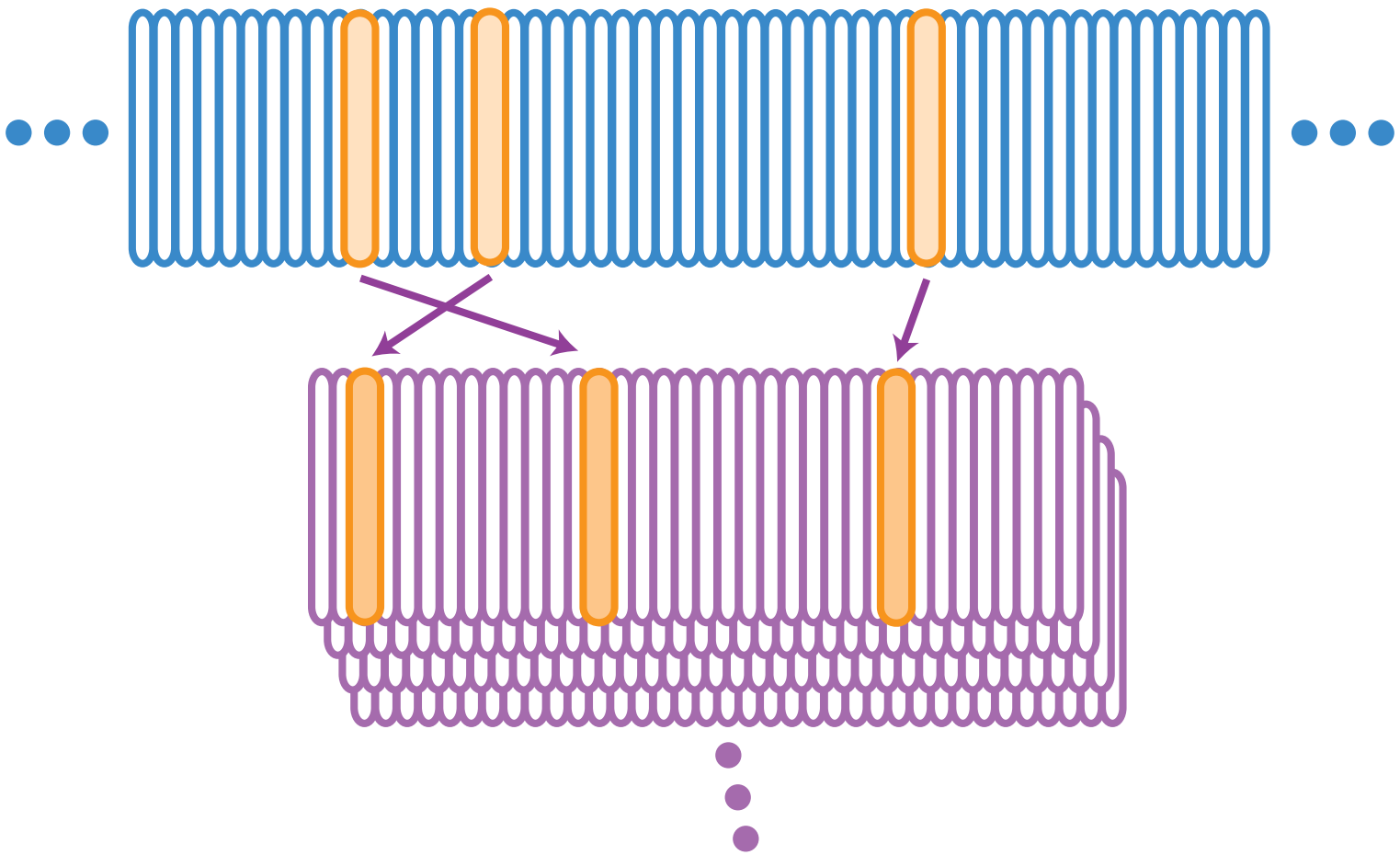
検定のバイアスという観点

データのバラツキ



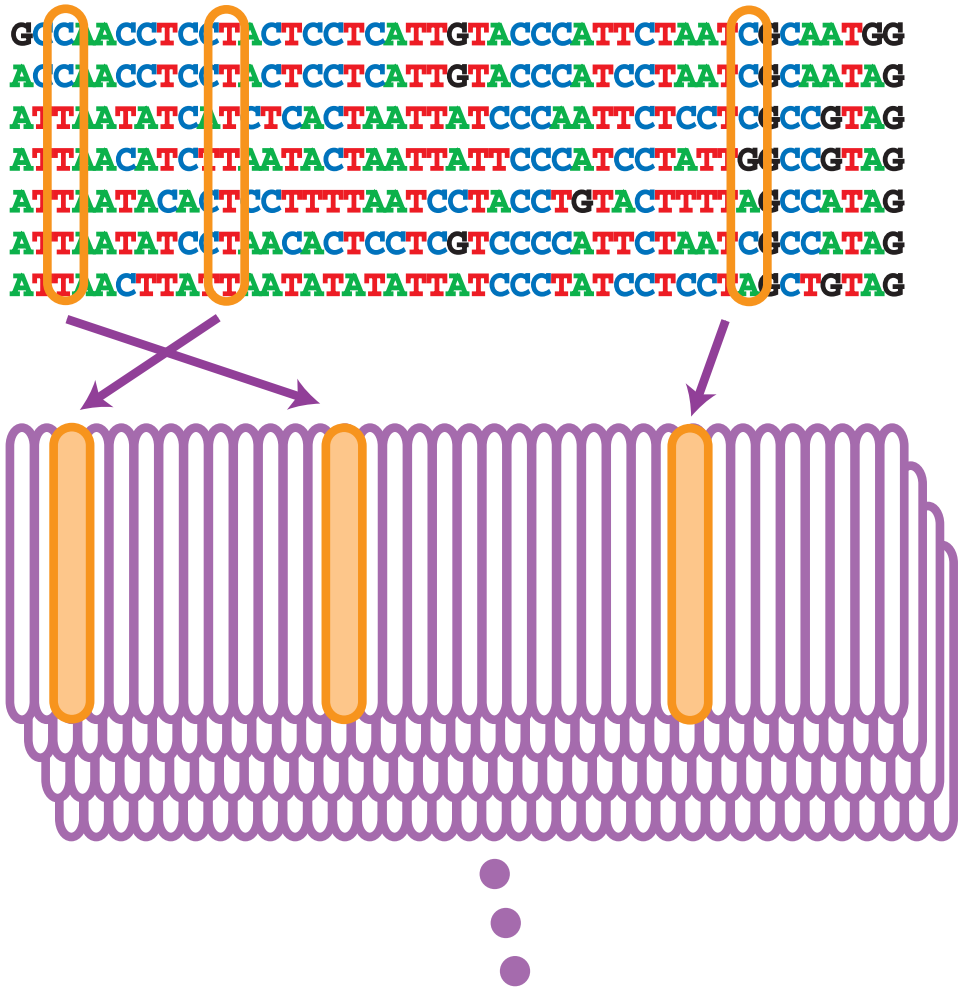
数学的には，確率モデルで生成される無限長の配列データからサンプルを取り出しているのと同じこと

パラメトリックブートストラップ



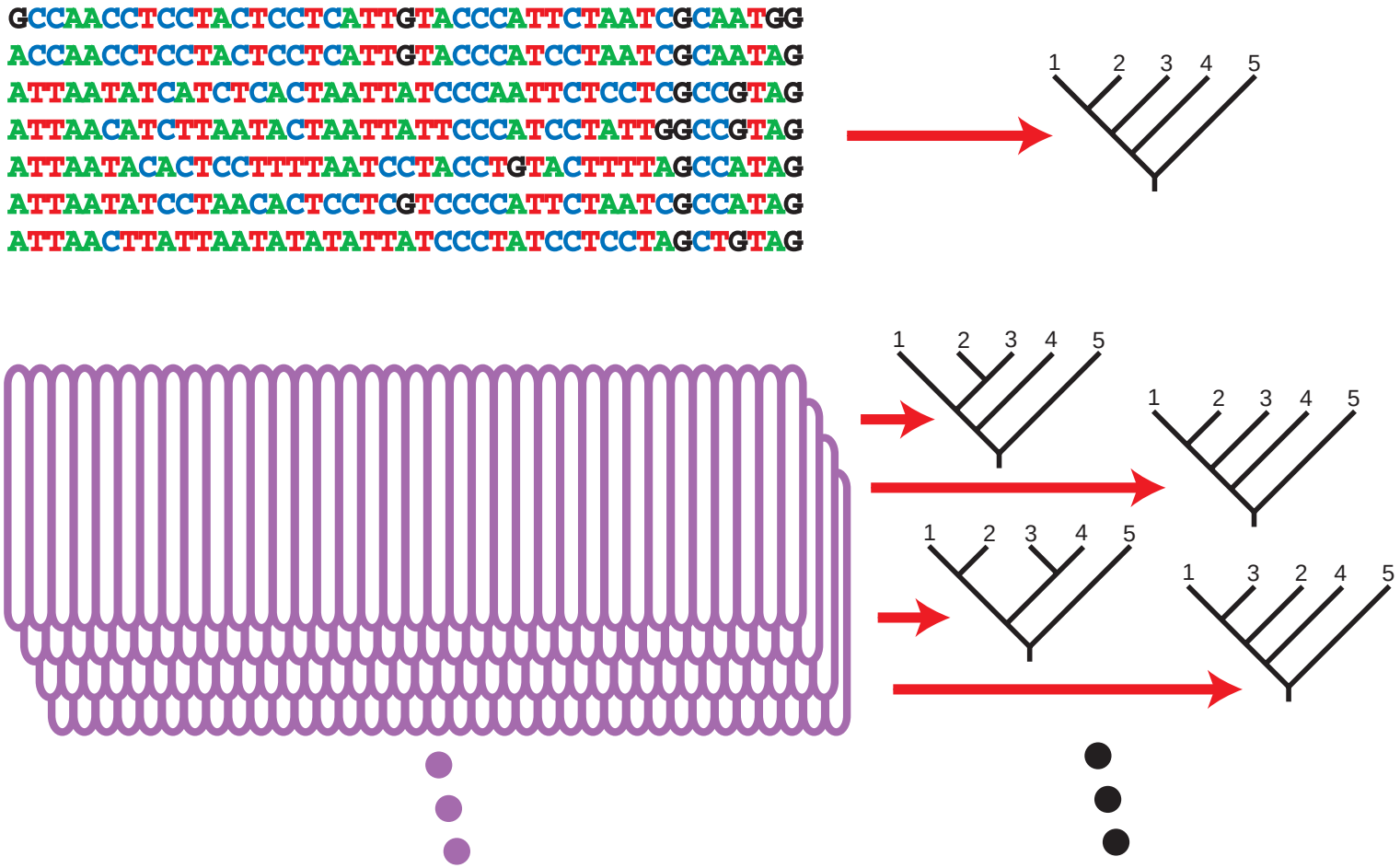
無限長の配列データからのサンプリング．ただしパラメタは推定値を使う．

(ノンパラメトリック)ブートストラップ



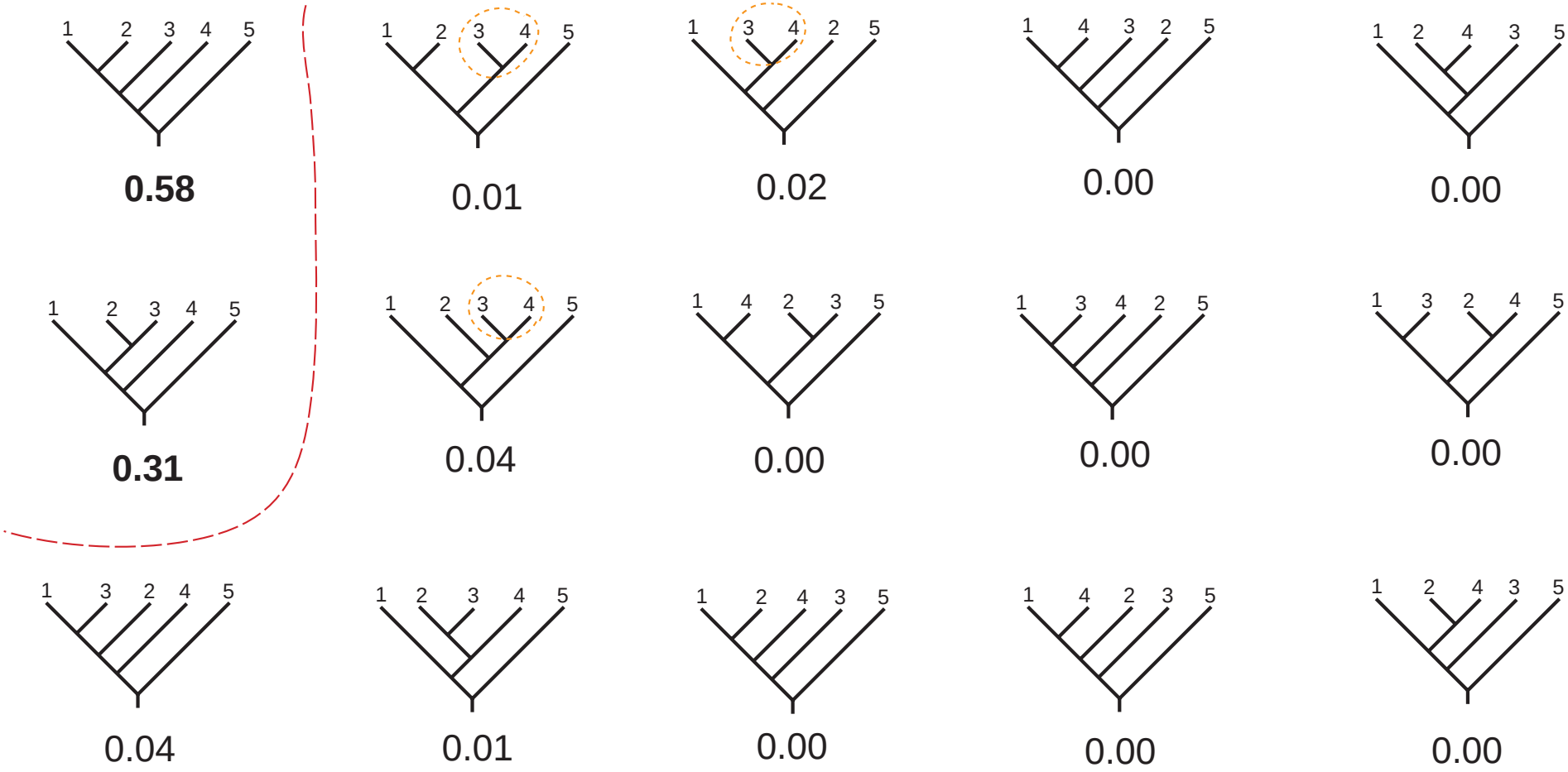
有限長の配列データからのリサンプリング

ブートストラップ確率


GCCAACCTCCTACTCCTCATTGTACCCATTCTAATCGCAATGG
ACCAACCTCCTACTCCTCATTGTACCCATCCTAATCGCAATAG
ATTAATATCATCTCACTAATTATCCCAATTCTCCTCGCCGTAG
ATTAACATCTTAATACTAATTATTCCCATCCTATTGGCCGTAG
ATTAATACACTCCTTTTAATCCTACCTGTACTTTTAGCCATAG
ATTAATATCCTAACACTCCTCGTCCCCATTCTAATCGCCATAG
ATTAACTTATTAATATATATTATCCCTATCCTCCTAGCTGTAG
1 2 3 4 5
1 2 3 4 5
1 2 3 4 5
1 2 3 4 5
1 2 3 4 5
1 3 2 4 5

系統樹のブートストラップ確率

1=human, 2=(seal, cow), 3=rabbit, 4=mouse, 5=opossum



1=human, 2=(seal,cow), 3=rabbit, 4=mouse, 5=opossum

Bootstrap probability of Felsenstein (1985)

p-values for fifteen trees

tree	$\Delta\ell$	PP_i	BP_i	KH_i	AU_i	SH_i	WSH_i	tree form
1	0.0	**0.934**	**0.579**	**0.639**	**0.789**	**0.944**	**0.948**	(((1(23))4)56)
2	2.7	**0.065**	**0.312**	**0.361**	**0.516**	**0.799**	**0.791**	((1((23)4))56)
3	7.4	0.001	0.036	**0.122**	**0.114**	**0.575**	**0.422**	(((14)(23))56)
4	17.6	0.000	0.013	0.044	**0.075**	**0.178**	**0.210**	((1(23))(45)6)
5	18.9	0.000	0.035	**0.066**	**0.128**	**0.149**	**0.299**	(1((23)(45))6)
6	20.1	0.000	0.005	0.049	0.029	**0.114**	**0.105**	(1(((23)4)5)6)
7	20.6	0.000	0.017	**0.051**	**0.101**	**0.112**	**0.252**	((1(45))(23)6)
8	22.2	0.000	0.001	0.032	0.009	**0.073**	**0.050**	((15)((23)4)6)
9	25.4	0.000	0.000	0.003	0.000	0.032	0.015	(((1(23))5)46)
10	26.3	0.000	0.003	0.019	0.028	0.034	**0.124**	(((15)4)(23)6)
11	28.9	0.000	0.000	0.010	0.003	0.018	**0.069**	(((14)5)(23)6)
12	31.6	0.000	0.000	0.003	0.001	0.006	0.033	(((15)(23))46)
13	31.7	0.000	0.000	0.003	0.001	0.006	0.034	(1(((23)5)4)6)
14	34.7	0.000	0.000	0.001	0.005	0.003	0.013	((14)((23)5)6)
15	36.2	0.000	0.000	0.001	0.002	0.002	0.009	((1((23)5))46)

1=human, 2=seal, 3=cow, 4=rabbit, 5=mouse, 6=opossum

p-values for ten edges

edge	PP_e	BP_e	KH_e	AU_e	SH_e	WSH_e	clade	trees
1	**1.000**	**0.927**	**0.956**	**0.954**	**0.994**	**0.991**	{1234}	1, 2, 3
2	**0.934**	**0.592**	**0.639**	**0.749**	**0.910**	**0.921**	{123}	1, 4, 9
3	**0.065**	**0.318**	**0.361**	**0.469**	**0.754**	**0.735**	{234}	2, 6, 8
4	0.001	0.036	**0.122**	**0.111**	**0.567**	**0.411**	{14}	3, 11, 14
5	0.000	**0.065**	0.044	**0.075**	**0.177**	**0.253**	{45}	4, 5, 7
6	0.000	0.040	**0.066**	**0.088**	**0.147**	**0.277**	{2345}	5, 6, 13
7	0.000	0.019	**0.051**	**0.070**	**0.112**	**0.227**	{145}	7, 10, 11
8	0.000	0.004	0.032	0.016	**0.072**	**0.113**	{15}	8, 10, 12
9	0.000	0.000	0.003	0.000	0.032	0.031	{1235}	9, 12, 15
10	0.000	0.000	0.003	0.000	0.006	0.032	{235}	13, 14, 15

1=human, 2=seal, 3=cow, 4=rabbit, 5=mouse, 6=opossum

仮説の確率値と信頼集合

- **仮説** $i = 1, 2, \ldots$ **(例えば系統樹トポロジやクレード)**
- **確率値 (**p**-value)** $P_1, P_2, \ldots$

- $P_i < 0.05$ **なら仮説**i**は棄却される**

- **棄却されない仮説の集合** $\Rightarrow$ **信頼集合**

$$\mathcal{T} = \{i : P_i \geq 0.05\}$$

- **被覆確率 (coverage probability)**
  真の仮説が信頼集合に含まれる確率

$$P^* = \mathsf{Pr}\{i^* \in \mathcal{T}\}$$

不偏な検定

真の仮説を棄却する確率 ≤ 0.05

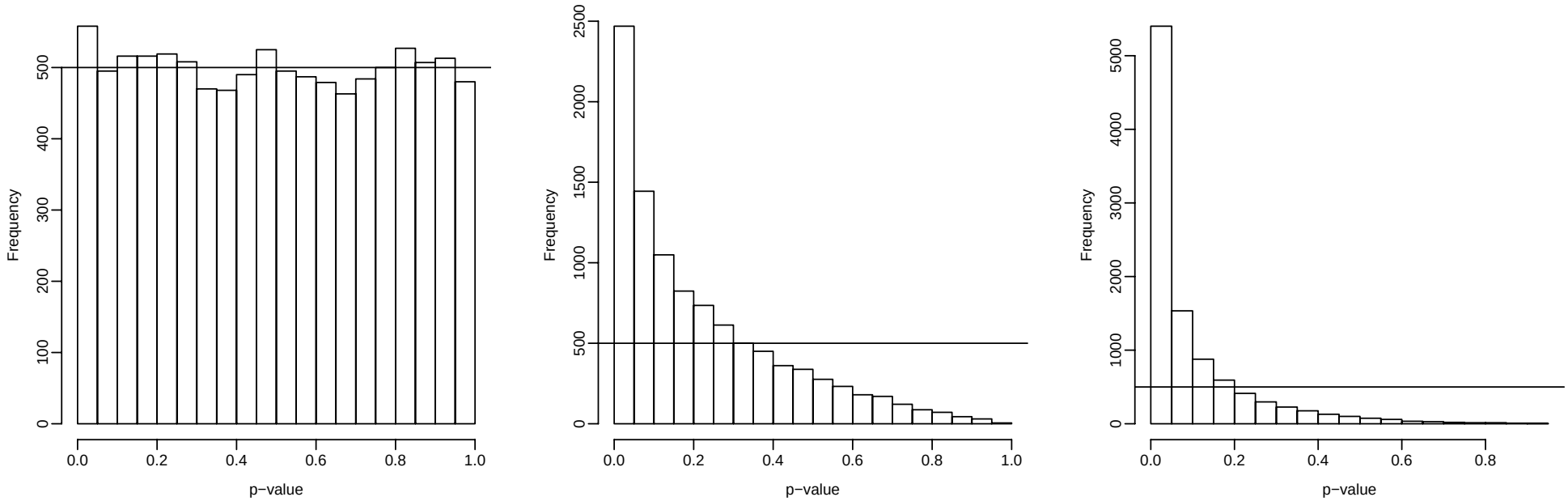
偽の仮説を棄却する確率 ≥ 0.05

真のパラメタがちょうど仮説の境界上のときに」

仮説を棄却してしまう確率 $= 0.05$

被覆確率 $= 0.95$

ブートストラップ確率のバイアス



真の系統樹のブートストラップ確率のバラツキ

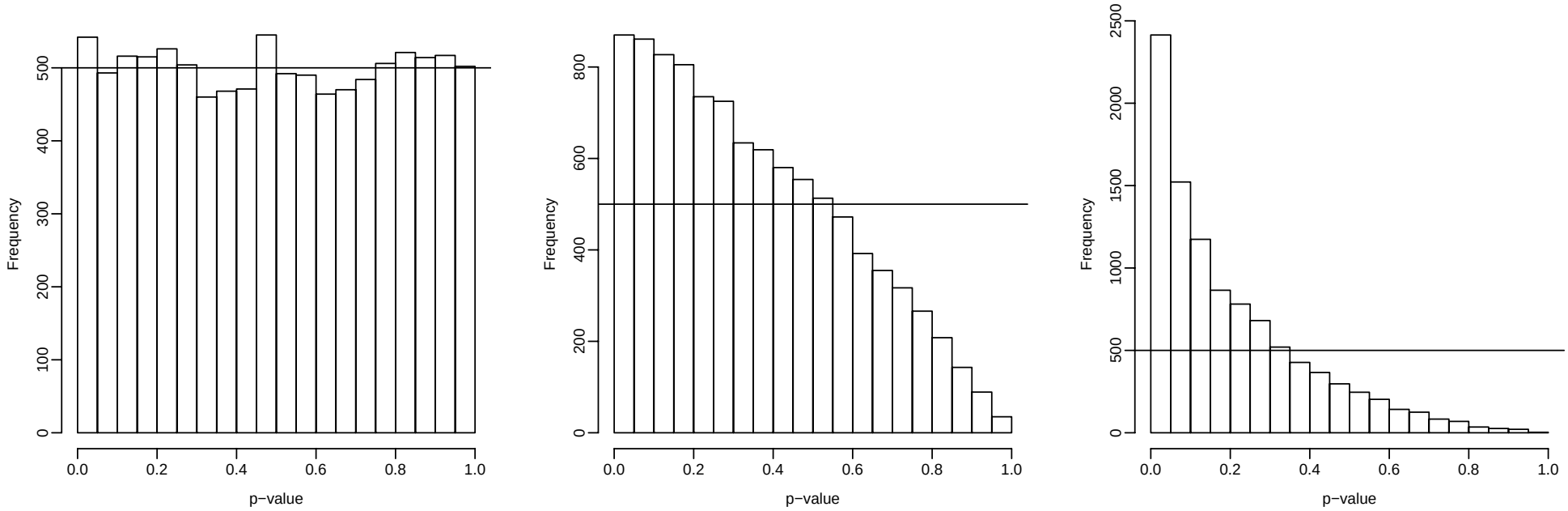
１０個の系統樹を比較したシミュレーションで

１番と２番の良さを同じにしてある（仮説の境界）

左：３番目以降がとても悪い，中：すこし悪い，右：同じ良さ

table

Kishino-Hasegawa検定のバイアス



左：３番目以降がとても悪い，中：すこし悪い，右：同じ良さ

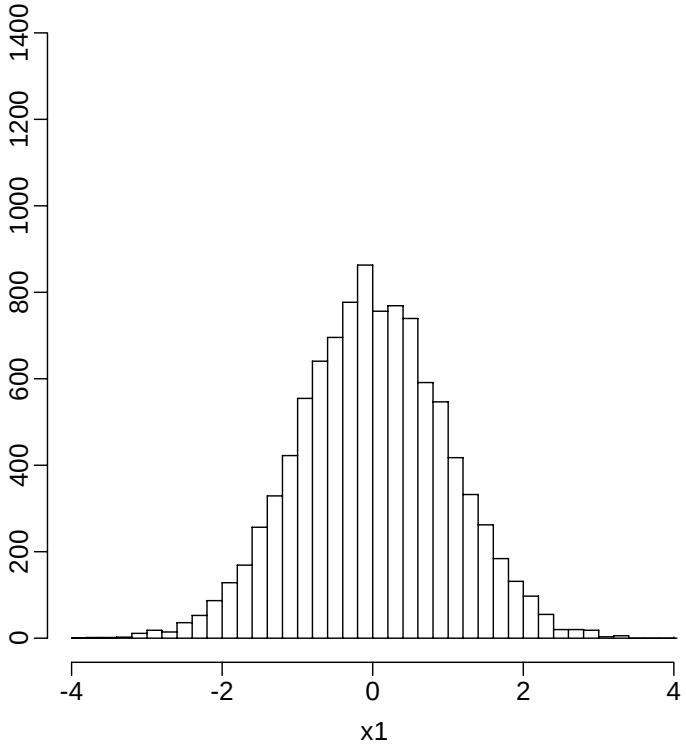
帰無仮説: 二つの系統樹の対数尤度の期待値が等しい

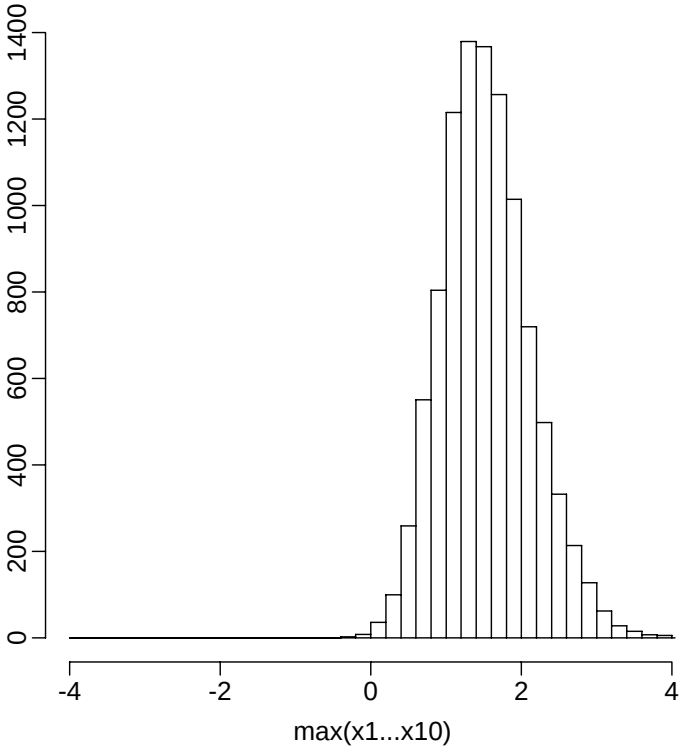
$\ell(1) - \ell(T)$が有意に大きければトポロジTを棄却

KH-testには選択バイアスがある

table

Selection Bias

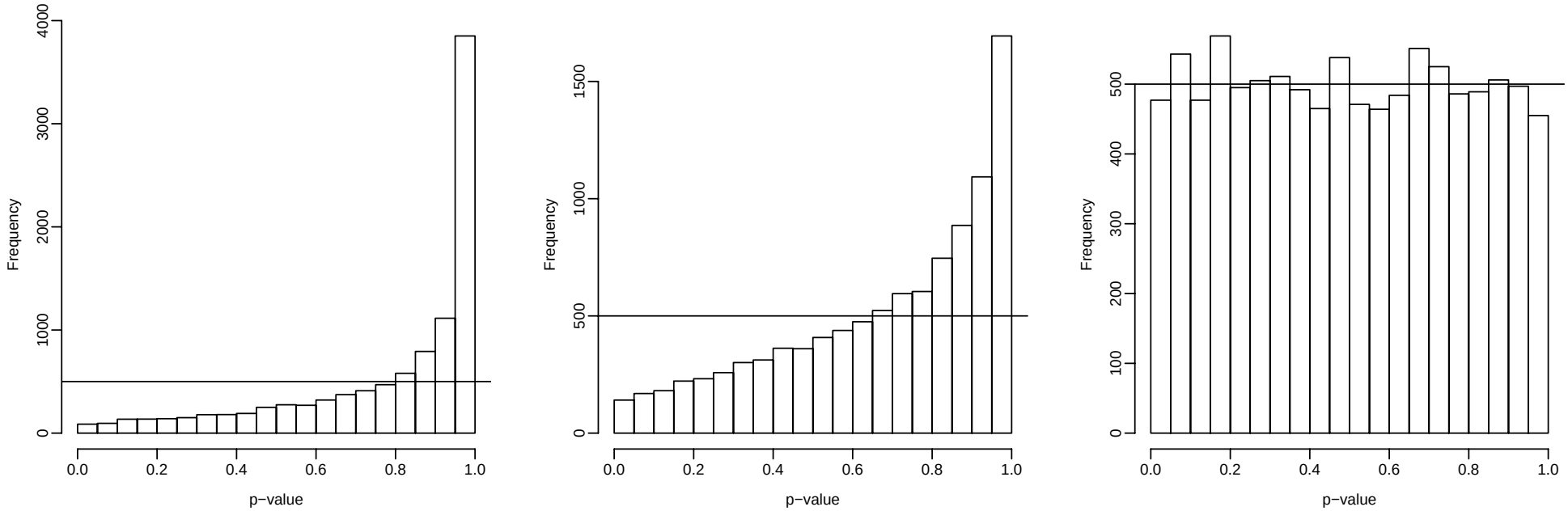




ℓ_1 $\max(\ell_1, \ldots, \ell_{10})$

$(\ell_1, \ldots, \ell_{10}) \sim N_{10}(0, I)$, **10000 samples**

Shimodaira-Hasegawa検定のバイアス



左：３番目以降がとても悪い，中：すこし悪い，右：同じ良さ

帰無仮説: 一番良い系統樹と比べて対数尤度の期待値が等しい

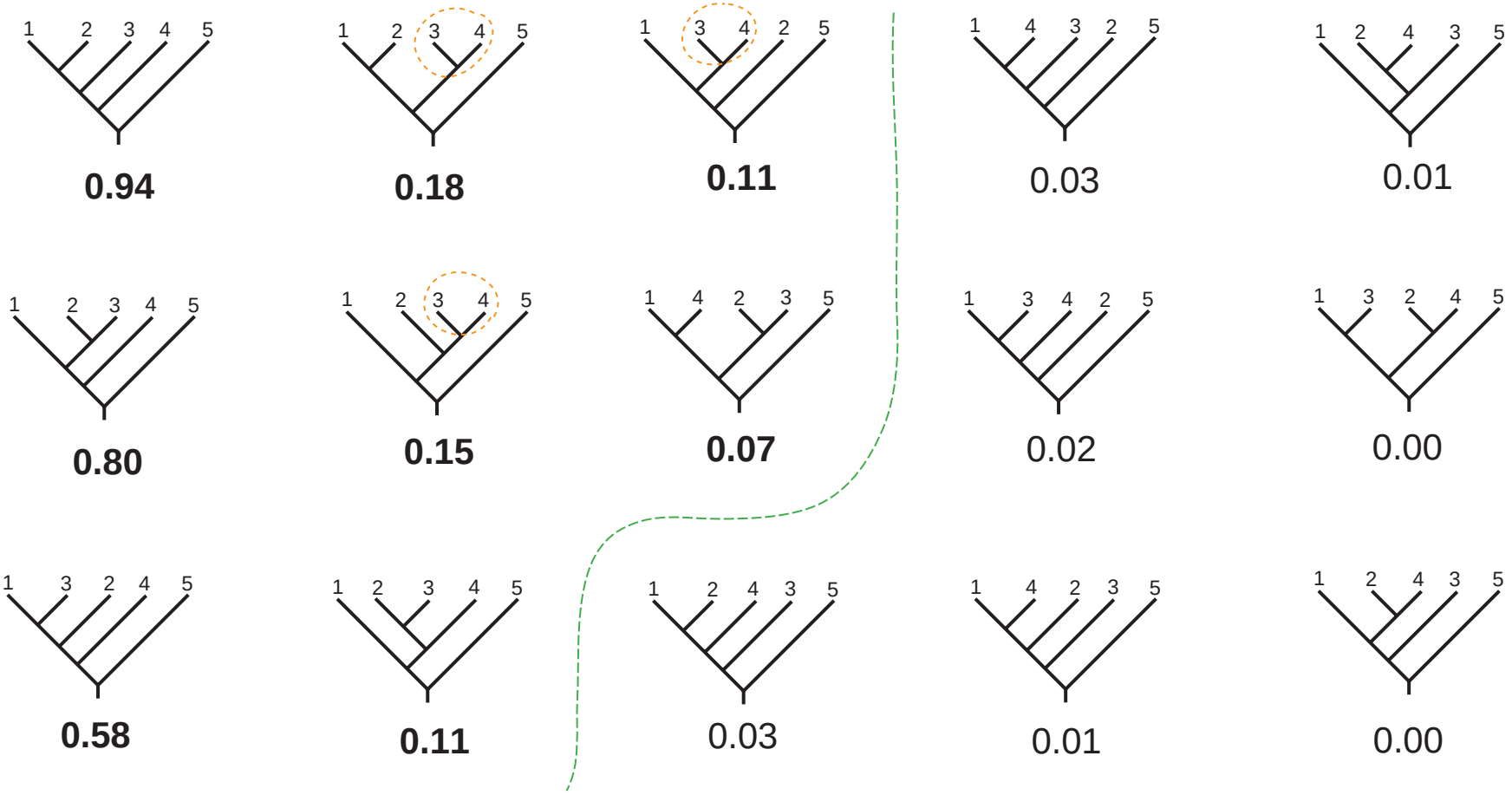
多重比較法

SH-testは保守的な傾向

table

系統樹のSH-testの確率値

1=human, 2=(seal, cow), 3=rabbit, 4=mouse, 5=opossum



1=human, 2=(seal,cow), 3=rabbit, 4=mouse, 5=opossum

Shimodaira (1998) *Annals of Institute of Statistical Mathematics* — **method**
Shimodaira and Hasegawa (1999) *Molecular Biology and Evolution* — **application**
Goldman, Anderson, and Rodrigo (2000) *Systematic Biology* — **review**
Shimodaira (2001) *Communications in Statistics A* — **review**

近似的に不偏な検定（AU検定）

真のパラメタがちょうど仮説の境界上のときに」

仮説を棄却してしまう確率 ≈ 0.05

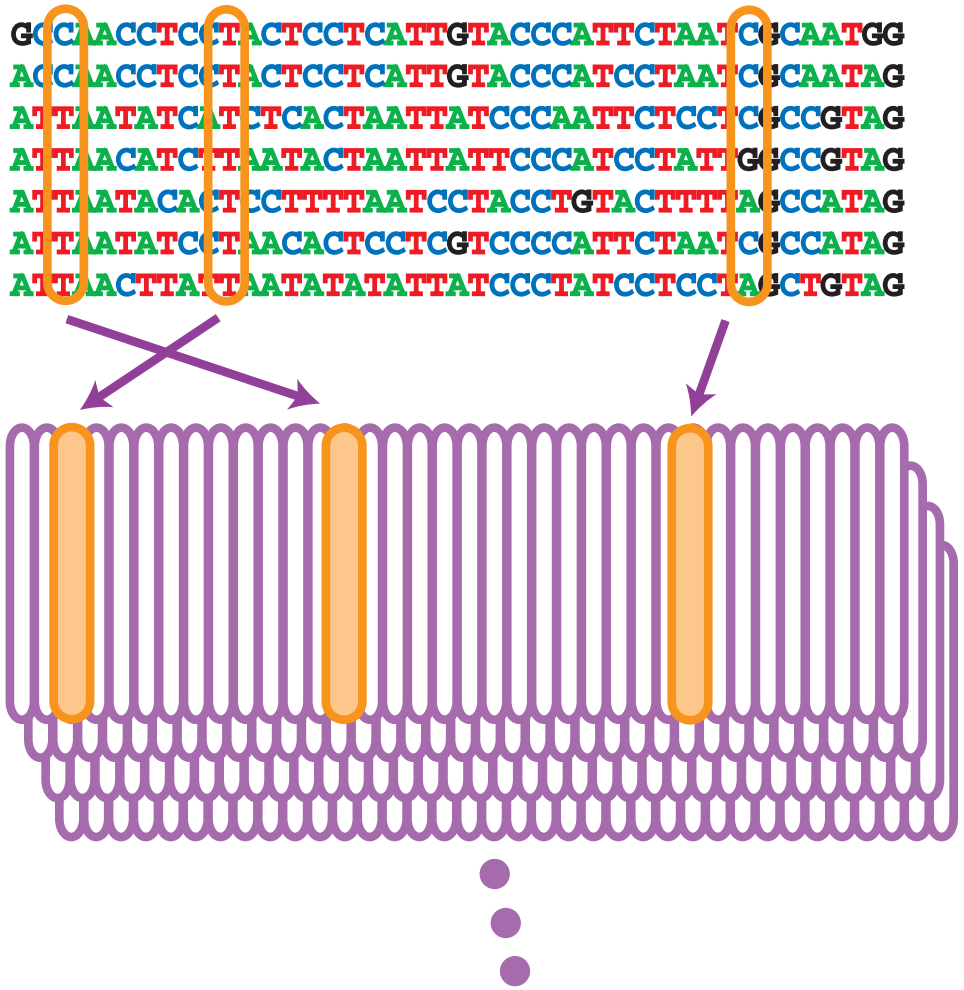
被覆確率 ≈ 0.95

通常のブートストラップ法 $=$ **1次の精度** $=$ **誤差** $O(n^{-1/2})$

two-level bootstrap法 $=$ **2次の精度** $=$ **誤差** $O(n^{-1})$

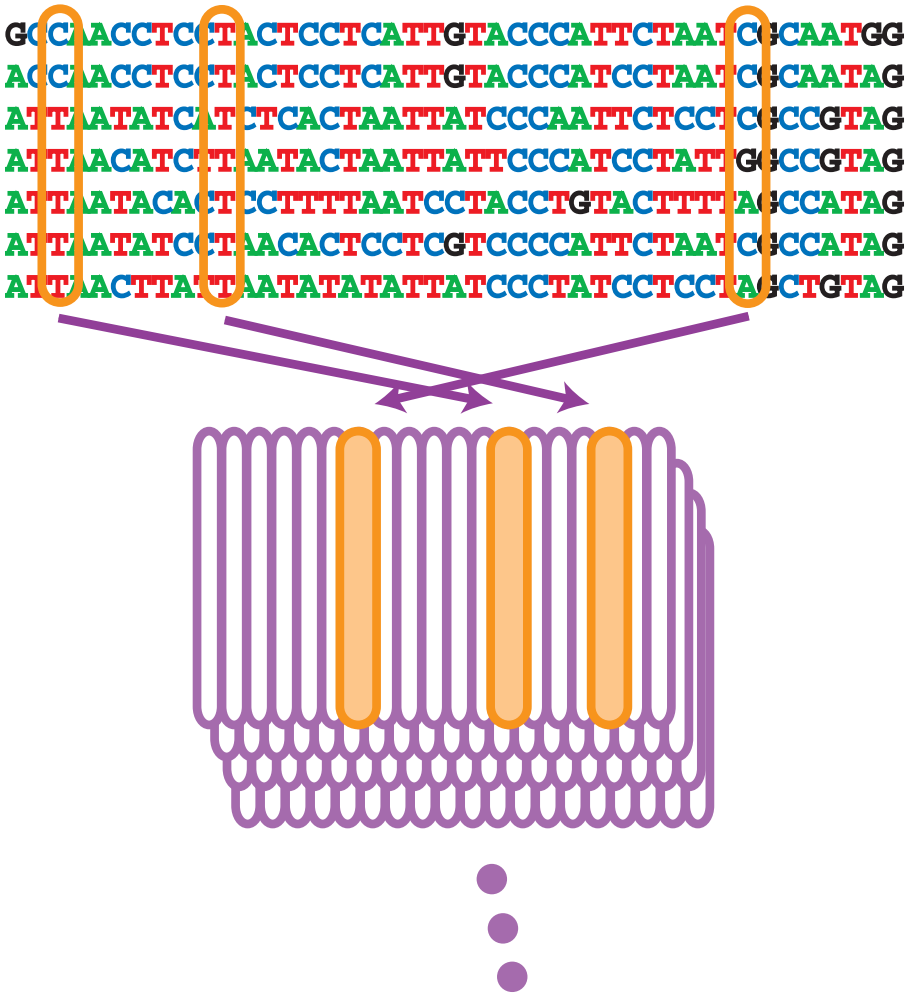
マルチスケールブートストラップ法 $=$ **3次の精度** $=$ **誤差** $O(n^{-3/2})$

マルチスケールブートストラップ法（1）



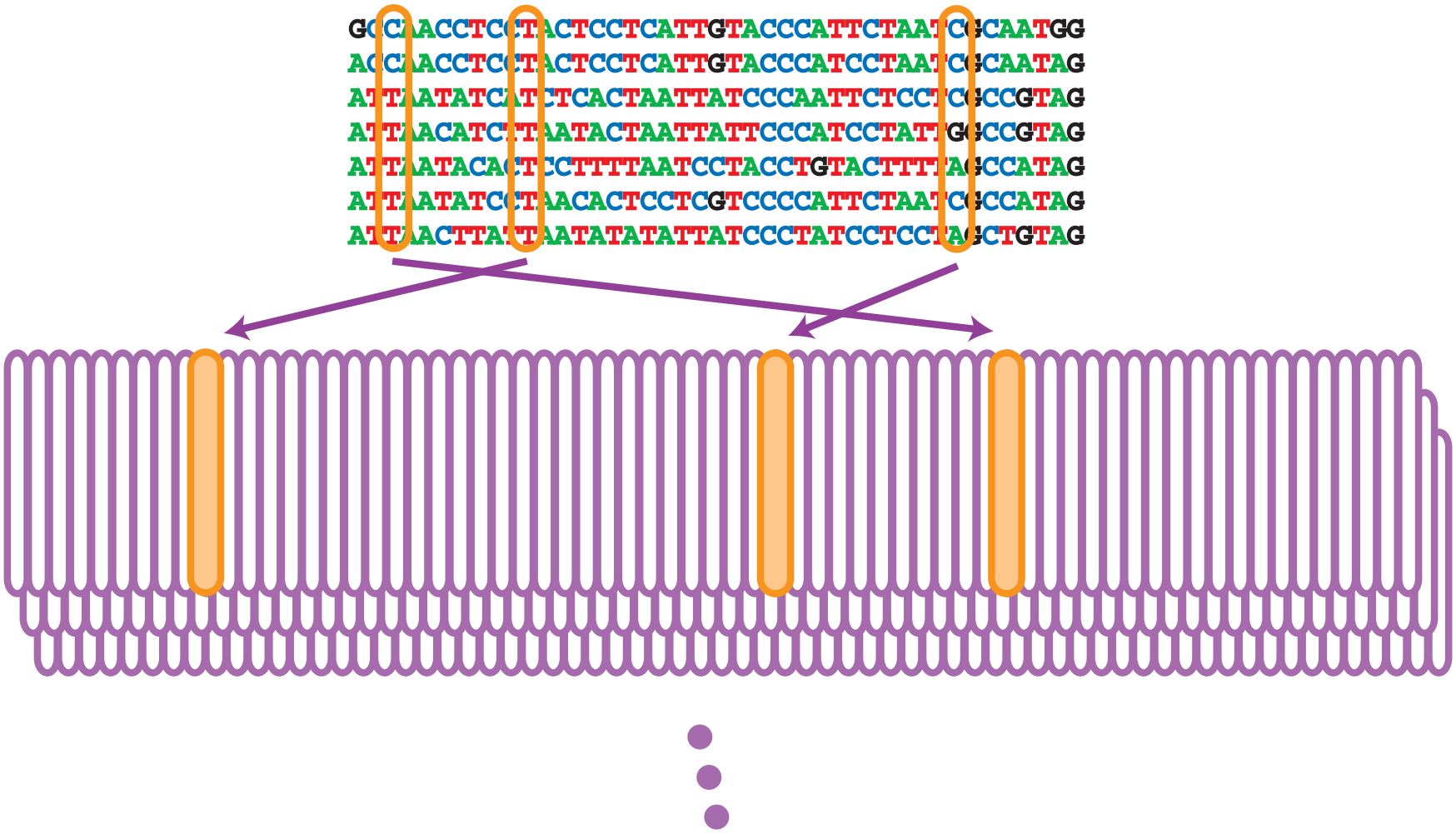
通常のブートストラップ

マルチスケールブートストラップ法（2）



通常より短い配列

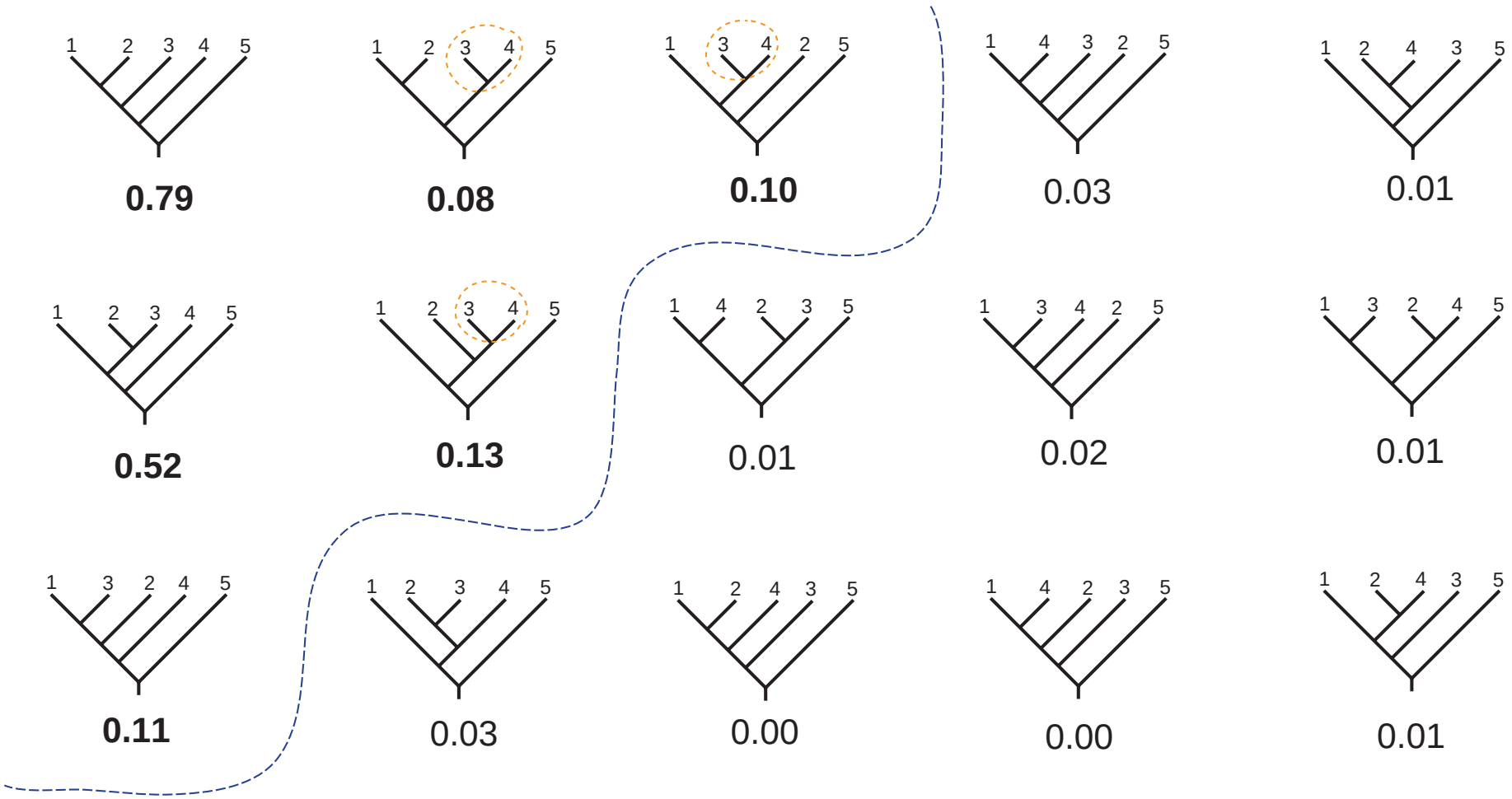
マルチスケールブートストラップ法（3）



通常より長い配列

系統樹の近似的に不偏な確率値（AU-test）

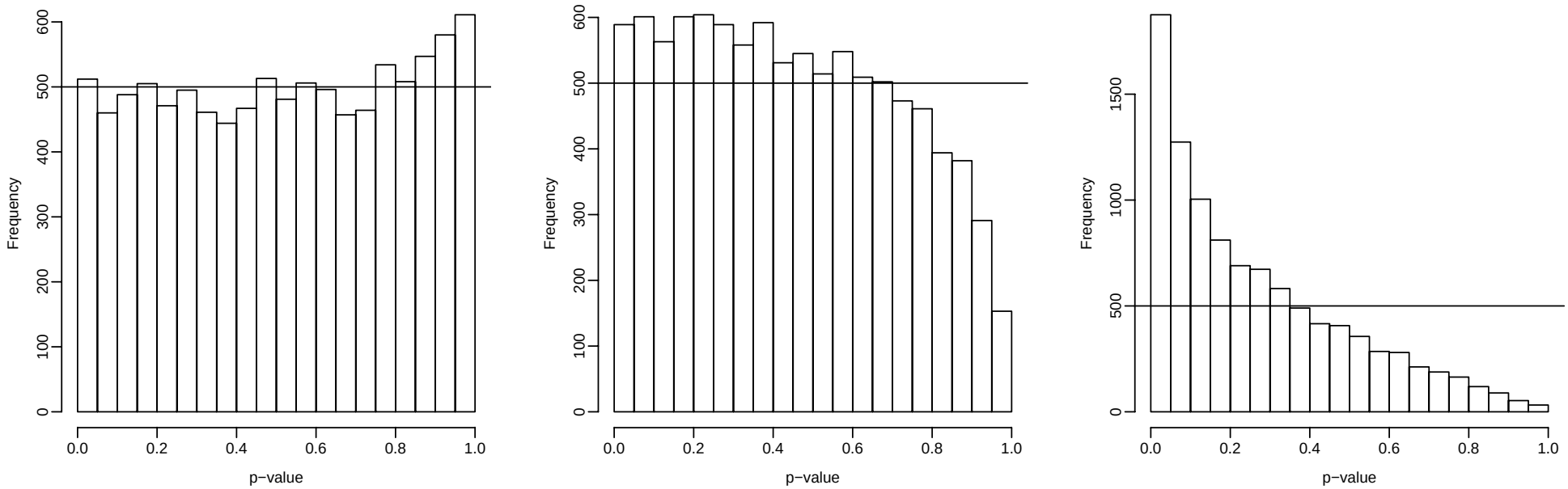
1=human, 2=(seal, cow), 3=rabbit, 4=mouse, 5=opossum



1=human, 2=(seal,cow), 3=rabbit, 4=mouse, 5=opossum

Shimodaira and Hasegawa (2001) *Bioinformatics* **— computer program**
Shimodaira (2002) *Systematic Biology* **— theory and method (multiscale)**
Shimodaira (submitted) — theory and method (multistep-multiscale)

AU検定のバイアス



左：３番目以降がとても悪い，中：すこし悪い，右：同じ良さ

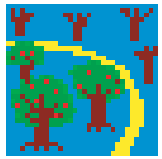
つねに「不偏」にできるわけではない

table

Software for Multiscale Bootstrap

H. Shimodaira and M. Hasegawa (2001). "CONSEL: for assessing the confidence of phylogenetic tree selection," *Bioinformatics*, 17, 1246–1247.

- Computer Software written in C language (Unix/Dos)
- Assessing phylogenetic tree selection



- Available at http://www.ism.ac.jp/~shimo/
  Also available at http://www.is.titech.ac.jp/~shimo/
- Implements Shimodaira-Hasegawa test, Multiscale bootstrap, and the other standard methods (bootstrap probability, Bayesian posterior probability, Kishino-Hasegawa test)
- Work with ML phylogeny programs (MOLPHY, PAML, PAUP)

系統樹の検定に関するまとめ

- **被覆確率** ≤ 0.95 **となる傾向**

  ベイズの事後確率，ブートストラップ確率，KH-test

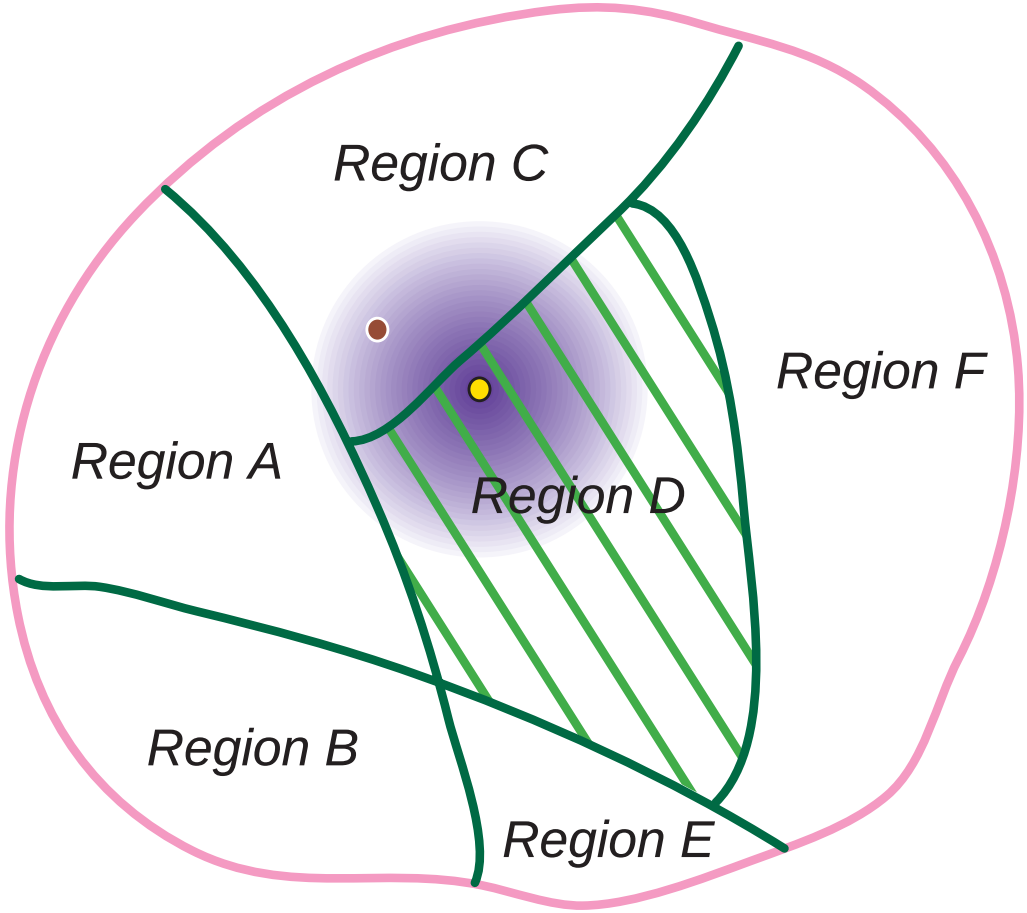
- **被覆確率** ≥ 0.95 **となる傾向**

  SH-test

- **被覆確率** ≈ 0.95 **となる傾向**

  AU-test（マルチスケールブートストラップ法)

Problem of Regions

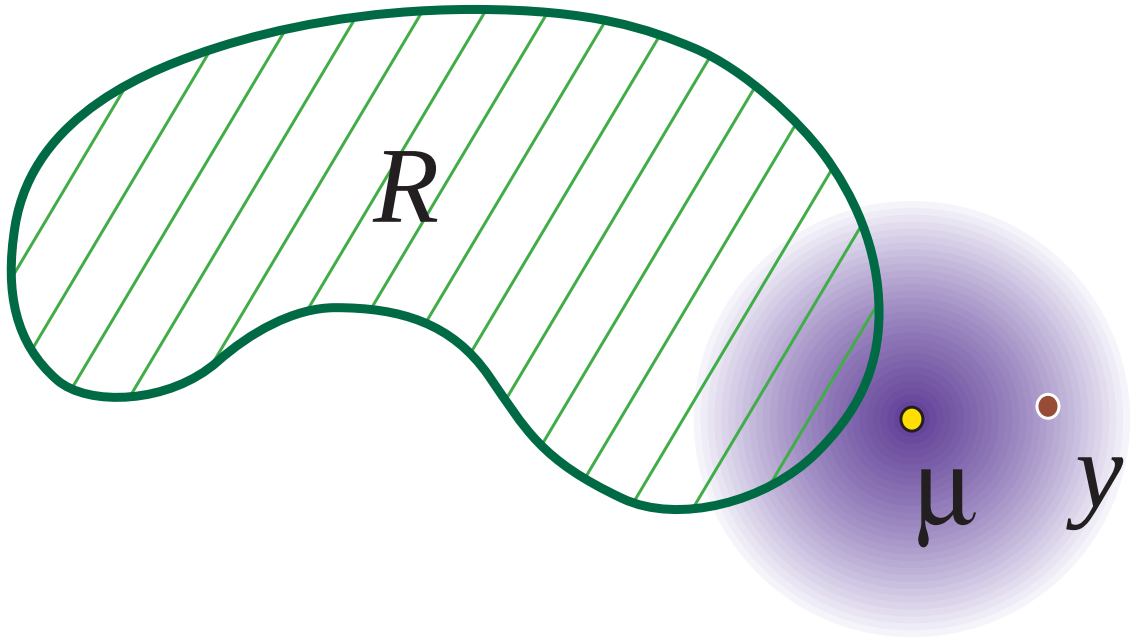


Efron et al. (1996), Efron and Tibshirani (1998)

Simplified Working Model

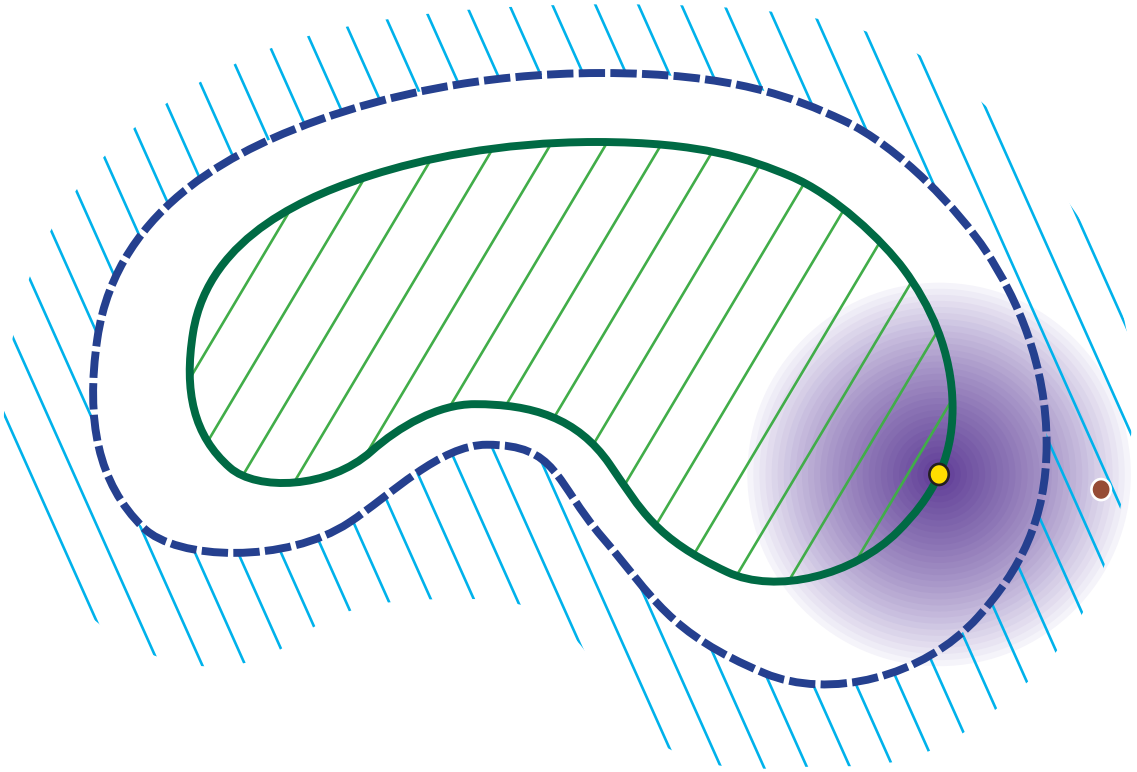
Multivariate normal distribution

$$y \sim N(\mu, I)$$



Null hypothesis: $\mu \in \mathcal{R}$

Unbiased Test



$$\Pr\{\hat{\alpha}(y) < \alpha \mid \mu\} \leq \alpha, \quad \mu \in \mathcal{R}$$

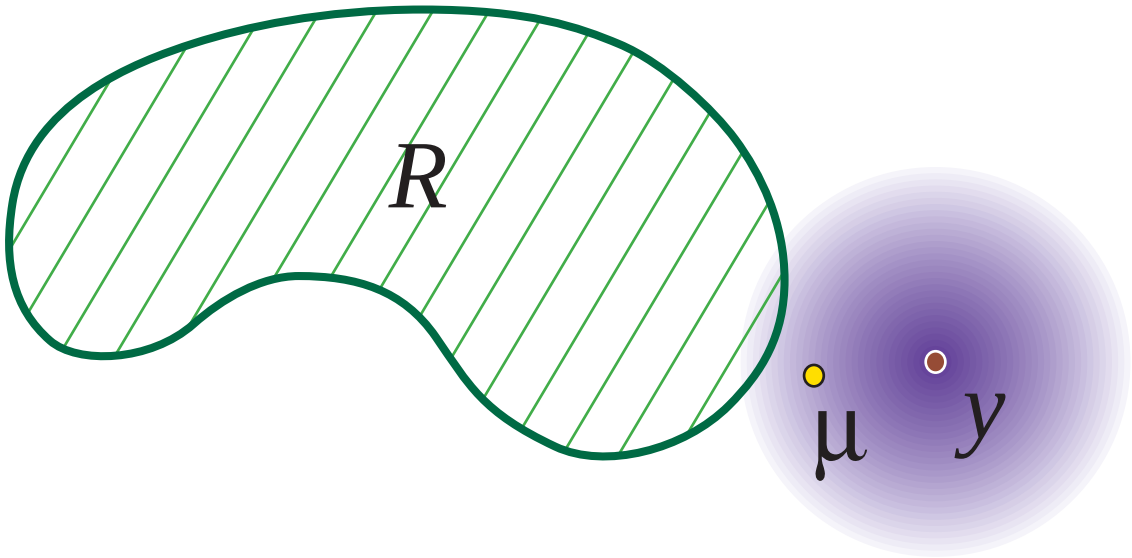
$$\Pr\{\hat{\alpha}(y) < \alpha \mid \mu\} = \alpha, \quad \mu \in \partial\mathcal{R}$$

$$\Pr\{\hat{\alpha}(y) < \alpha \mid \mu\} \geq \alpha, \quad \mu \notin \mathcal{R}$$

Generating Replicates

Multivariate normal distribution

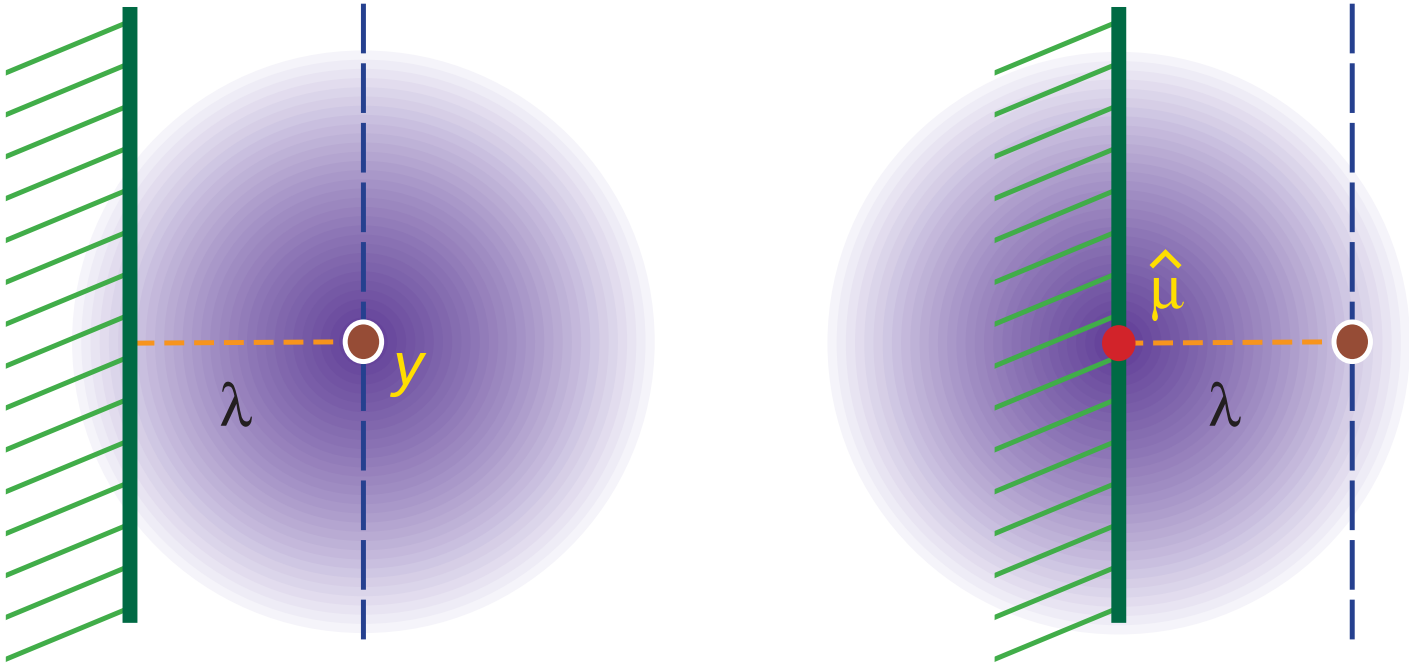
$$y^* \sim N(y, I)$$



$$\hat{\alpha}_{\mathsf{BP}} = \frac{\#\{y^{*1}, \ldots, y^{*10000} \in \mathcal{R}\}}{10000}$$

Flat Boundary

$\lambda = 2,\ c = 0$



$\hat{\alpha}_{\mathsf{BP}} = 0.0228$

$= 1 - \Phi(\lambda)$

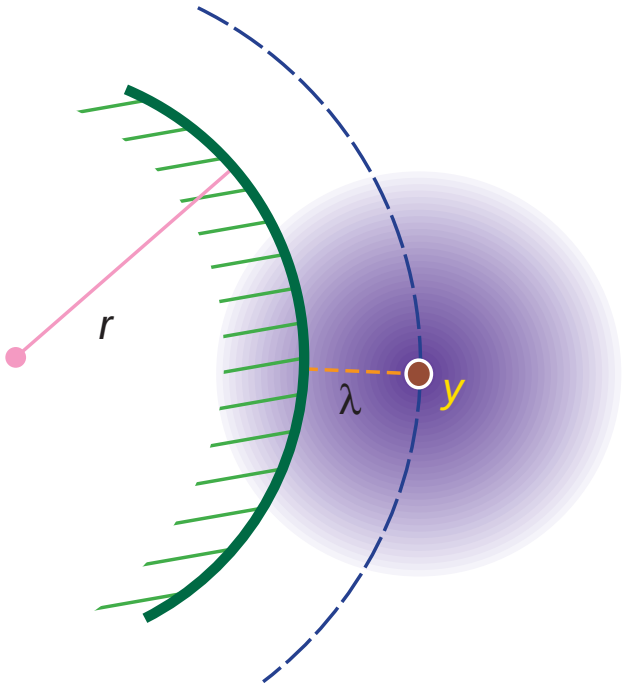
$\hat{\alpha}_{\mathsf{EX}} = 0.0228$

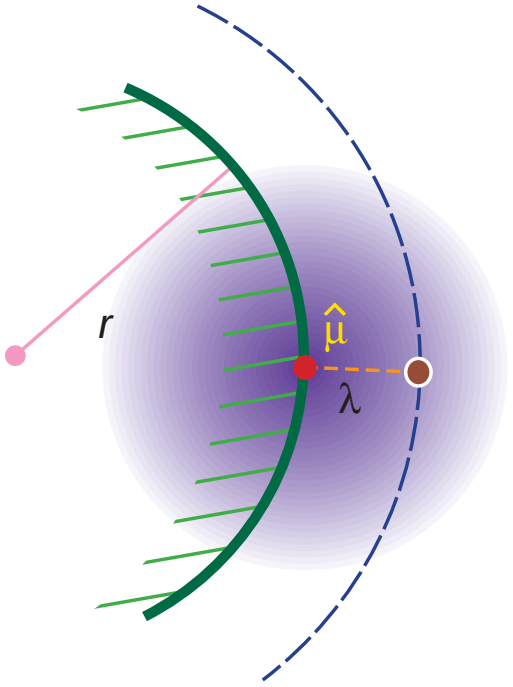
$= 1 - \Phi(\lambda)$

$$\Phi(\lambda) = \int_{-\infty}^{\lambda} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

Curved Boundary

$\lambda = 2,\ c = 0.24$





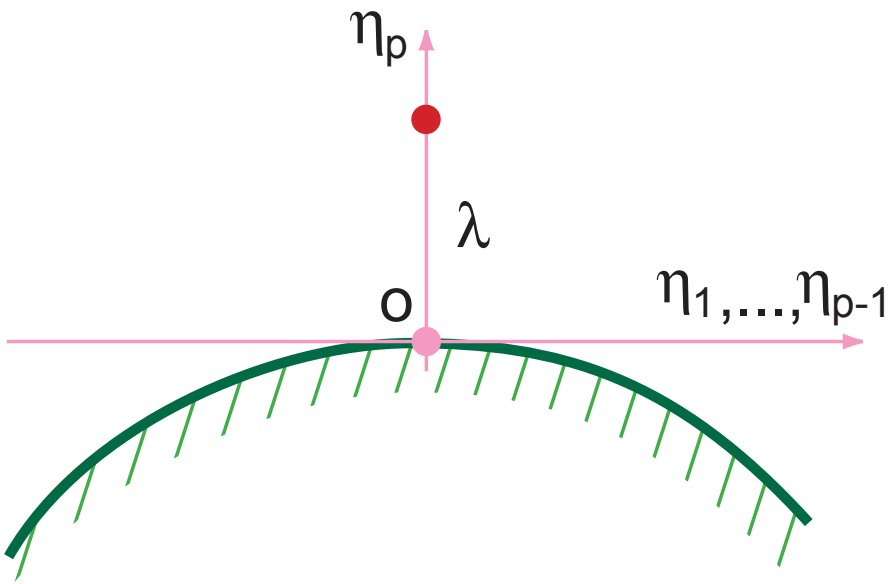
$\hat{\alpha}_{\mathsf{BP}} = 0.0121$

$\approx 1 - \Phi(\lambda + c)$

$\hat{\alpha}_{\mathsf{EX}} = 0.0404$

$\approx 1 - \Phi(\lambda - c)$

Signed distance and Curvature



$$\eta_p \leq -d^{ab}\eta_a\eta_b - e^{abc}\eta_a\eta_b\eta_c$$

$$\boxed{c = d^{aa} - \lambda d^{ab}d^{ab}}$$

note: a, b, c **run through** $1, \ldots, p-1$**.**

Multiscale Bootstrap Resampling

Original Data $X = (x_1, x_2, \ldots, x_n)$

Replicated Data $X^* = (x_1^*, x_2^*, \ldots, x_{n'}^*)$

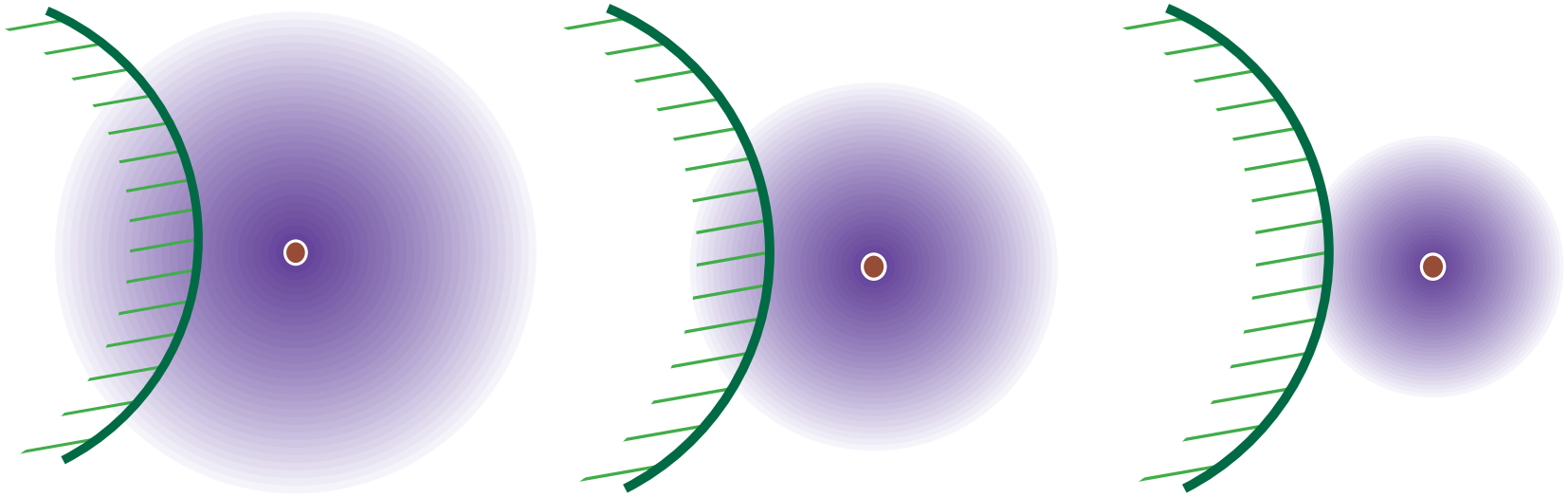
Choose n' items from $x_1, x_2, \ldots, x_n$ with replacements

$$\textbf{Scale} \quad \tau = \sqrt{\frac{n}{n'}}$$

Calculates BP's with several scales $\tau_1, \tau_2, \ldots, \tau_K$

$\Rightarrow$ **Estimates geometric quantities**

Change of Scale



$$\frac{n}{n'} = 2 \qquad \frac{n}{n'} = 1 \qquad \frac{n}{n'} = \frac{1}{2}$$

$$\hat{\alpha}_{\mathsf{BP}} = 0.0380 \qquad \hat{\alpha}_{\mathsf{BP}} = 0.0121 \qquad \hat{\alpha}_{\mathsf{BP}} = 0.0013$$

$$\hat{\alpha}_{\mathsf{BP}}(\tau) = 1 - \Phi(\lambda/\tau + c\tau); \qquad \tau = \sqrt{n/n'}$$

$\Rightarrow$ **Curve Fitting**

Estimating the Geometric Quantities

Estimate: signed distance $\hat{\lambda} = 1.998$, curvature $\hat{c} = 0.256$

true value: signed distance $\lambda = 2$, curvature $c = 0.24$

Multiscale bootstrap calculates very accurate p-value

$$\hat{\alpha}_{\mathrm{MS}} = 1 - \Phi(\hat{\lambda} - \hat{c}) = 0.0408$$

$$\hat{\alpha}_{\mathrm{EX}} = 0.0404$$

Bootstrap probability is less accurate

$$\hat{\alpha}_{\mathrm{BP}} = 0.0121$$

Accuracy of Approximately Unbiased Tests

k-th order asymptotic accuracy

$$\Pr\{\widehat{\alpha}(y) < \alpha \mid \mu\} = \alpha + O(n^{-k/2}), \quad \mu \in \partial\mathcal{R}.$$

- **Bootstrap probability (Efron 1979, Felsenstein 1985)**
  is first-order accurate ($k = 1$)
- **Two-level bootstrap (Efron et al. 1996)**
  is second-order accurate ($k = 2$)
- **Multiscale bootstrap (Shimodaira 2002)**
  is third-order accurate ($k = 3$)