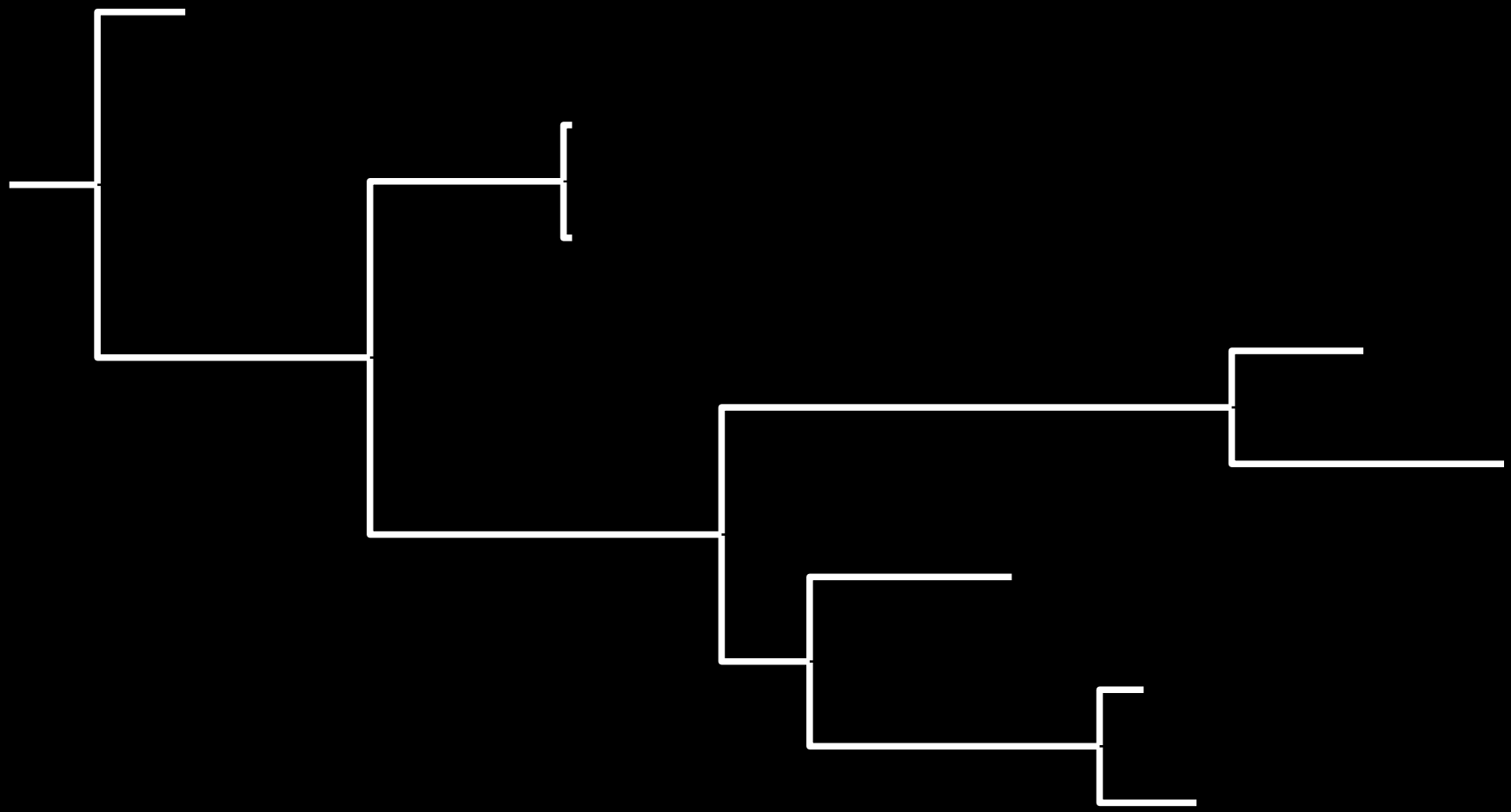
分子進化モデルと最尤系統推定法



東北大・院・生命科学　田邉晶史

まずはじめに,

最尤系統推定とは・・・

多重モデル選択

である.

最尤系統推定の手順

樹形を固定しての
分子進化モデルの選択

↓

分子進化モデルを固定しての
系統モデル(樹形)の選択

＝

多重モデル選択

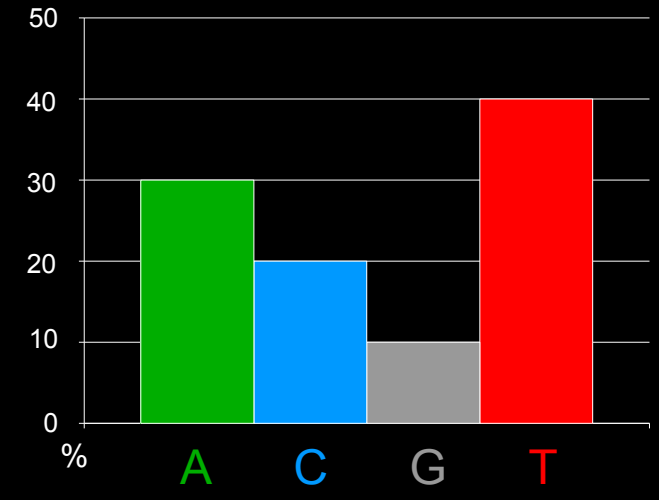
分子進化モデル超入門

とりあえず塩基置換モデルで

塩基置換モデルの3大要素

- 塩基置換確率行列 (nucleotide substitution rate matrix)
- 塩基平衡頻度 (nucleotide equilibrium frequencies)
- 座位間の速度の不均質性 (rate heterogeneity among sites)

	A	C	G	T
A		r_{AC}	r_{AG}	r_{AT}
C	r_{CA}		r_{CG}	r_{CT}
G	r_{GA}	r_{GC}		r_{GT}
T	r_{TA}	r_{TC}	r_{TG}	

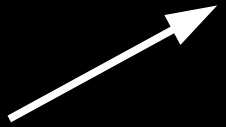


Taxon1	A	C	C	G	A	T	T
Taxon2	A	C	C	G	A	A	T
Taxon3	T	C	C	C	A	A	T
Taxon4	T	C	T	C	A	C	T
Taxon5	T	C	T	C	A	C	T
Taxon6	T	C	T	A	A	C	T
Taxon7	T	C	T	A	A	G	T
Taxon8	T	C	T	A	A	G	T
Taxon9	T	C	T	A	A	G	T

塩基置換確率行列と塩基平衡頻度

	A	C	G	T
A		r_{AC}	r_{AG}	r_{AT}
C	r_{CA}		r_{CG}	r_{CT}
G	r_{GA}	r_{GC}		r_{GT}
T	r_{TA}	r_{TC}	r_{TG}	

	A	C	G	T
A		$\pi_C r_{AC}$	$\pi_G r_{AG}$	$\pi_T r_{AT}$
C	$\pi_A r_{CA}$		$\pi_G r_{CG}$	$\pi_T r_{CT}$
G	$\pi_A r_{GA}$	$\pi_C r_{GC}$		$\pi_T r_{GT}$
T	$\pi_A r_{TA}$	$\pi_C r_{TC}$	$\pi_G r_{TG}$	



$A : C : G : T = \pi_A : \pi_C : \pi_G : \pi_T$

$(\pi_A + \pi_C + \pi_G + \pi_T = 1)$

真の置換確率と塩基頻度に分けて見かけの塩基置換確率行列を表現することで非対称な行列を効率的に表現できる

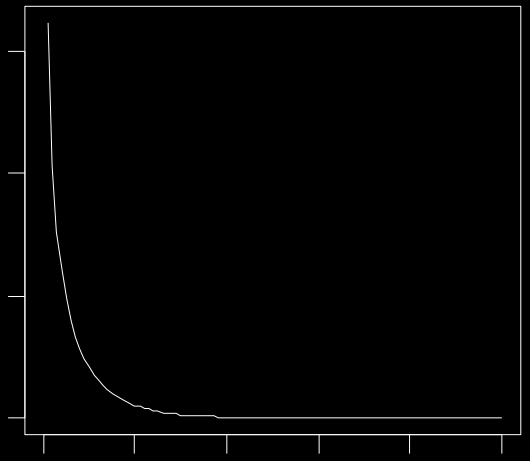
塩基置換確率行列と主なモデルの名称

塩基置換確率パラメータ数	等塩基頻度	不等塩基頻度
0	JC69	F81
1	K80(K2P)	HKY85
2	TN93ef	TN93
2	K81(K3P)	K81uf(K3Puf)
3	TIMef	TIM
4	TVMef	TVM
5	SYM	GTR

座位間の速度の不均質性

Taxon1	A	C	C	G	A	T	T
Taxon2	A	C	C	G	A	A	T
Taxon3	T	C	C	C	A	A	T
Taxon4	T	C	T	C	A	C	T
Taxon5	T	C	T	C	A	C	T
Taxon6	T	C	T	A	A	C	T
Taxon7	T	C	T	A	A	G	T
Taxon8	T	C	T	A	A	G	T
Taxon9	T	C	T	A	A	G	T

離散化ガンマ分布による近似
(Gamma site rate heterogeneity)



各座位の変異速度

パラメータはshapeだけで済む

(+G)

不変座位率 (Proportion of invariable sites)

変異のある座位の変異速度が一定なら
ガンマ分布による近似より，不変座位・
変異座位の2カテゴリに分ける方が良い

(+I)

不変座位・変異座位のカテゴリに分けた
上で変異座位をさらにガンマ分布に基づ
いて複数カテゴリに分ける併用も可

(+GI)

Site-Specific rate

各座位or座位群ごとに変異速度を推定
パラメータ数は座位群数-1

(+SS)

多数の遺伝子領域を取り扱う
model heterogeneity among sites

- 1遺伝子領域と同様にモデル選択 = Concatenate model
  - 分子進化モデルはただ1つ
  - 枝長パラメータ数はOTU数×2-3
- 各領域に異なるモデルを適用し，相対速度比を推定 = Proportional model
  - 分子進化モデルは領域毎に異なる
  - 枝長パラメータ数はOTU数×2-3
  - 領域毎の相対速度比パラメータ数は領域数-1
- 各領域に異なるモデルを適用し，対数尤度の和を採用 = Separate model
  - 分子進化モデルは領域毎に異なる
  - 枝長パラメータ数は(OTU数×2-3)×領域数

その他の分子進化モデル

- アミノ酸置換モデル

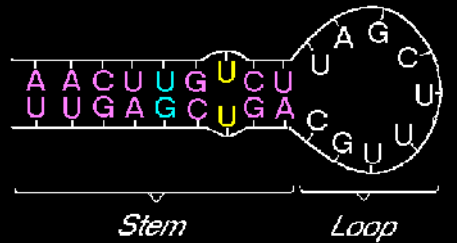
  アミノ酸間の置換速度を塩基置換モデルと同様にモデル化．ただ，アミノ酸は核酸よりも種類が多く，データ量に対してパラメータ数が増えすぎるので，既知の系統樹から求めた速度を近縁種の解析に用いることがほとんどである．

- コドン置換モデル

  同義置換と非同義置換を区別してそれぞれに異なるモデルを適用した上で同義置換/非同義置換速度比を導入したモデル．今後，モデルの改善と優れた実装ソフトウェアが登場すればタンパクコード領域データの解析で主流になると思われる．

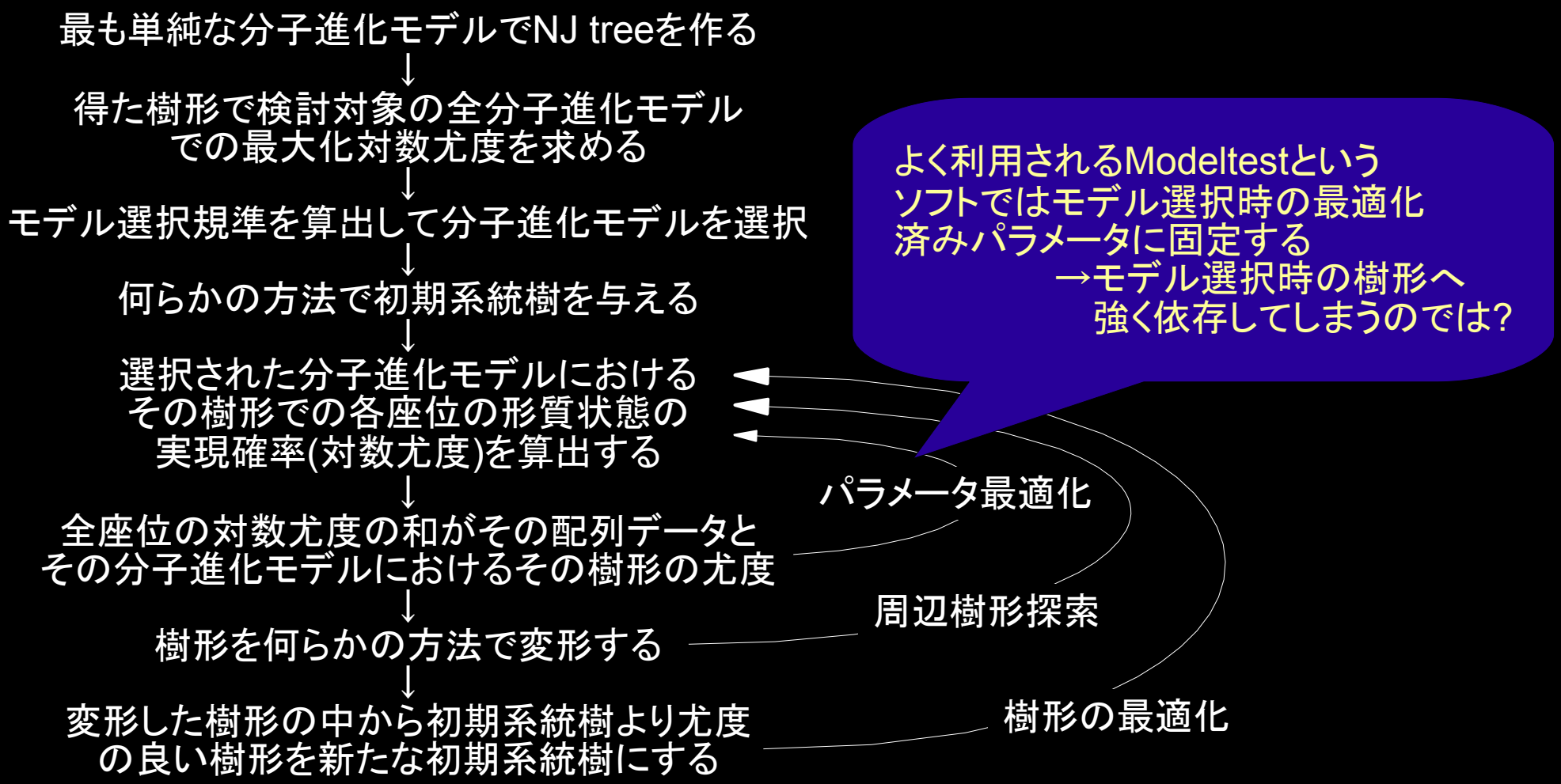
- rRNA遺伝子stem領域用モデル

  rRNA遺伝子のstem領域はmismatch，UG-pair，Watson-Click pair間で置換速度が異なり，Watson-Click pair内でも異なることを考慮したモデル．データ量に対してパラメータ数が増えすぎるので，既知の系統樹から求めた速度を近縁種の解析に用いることもある．

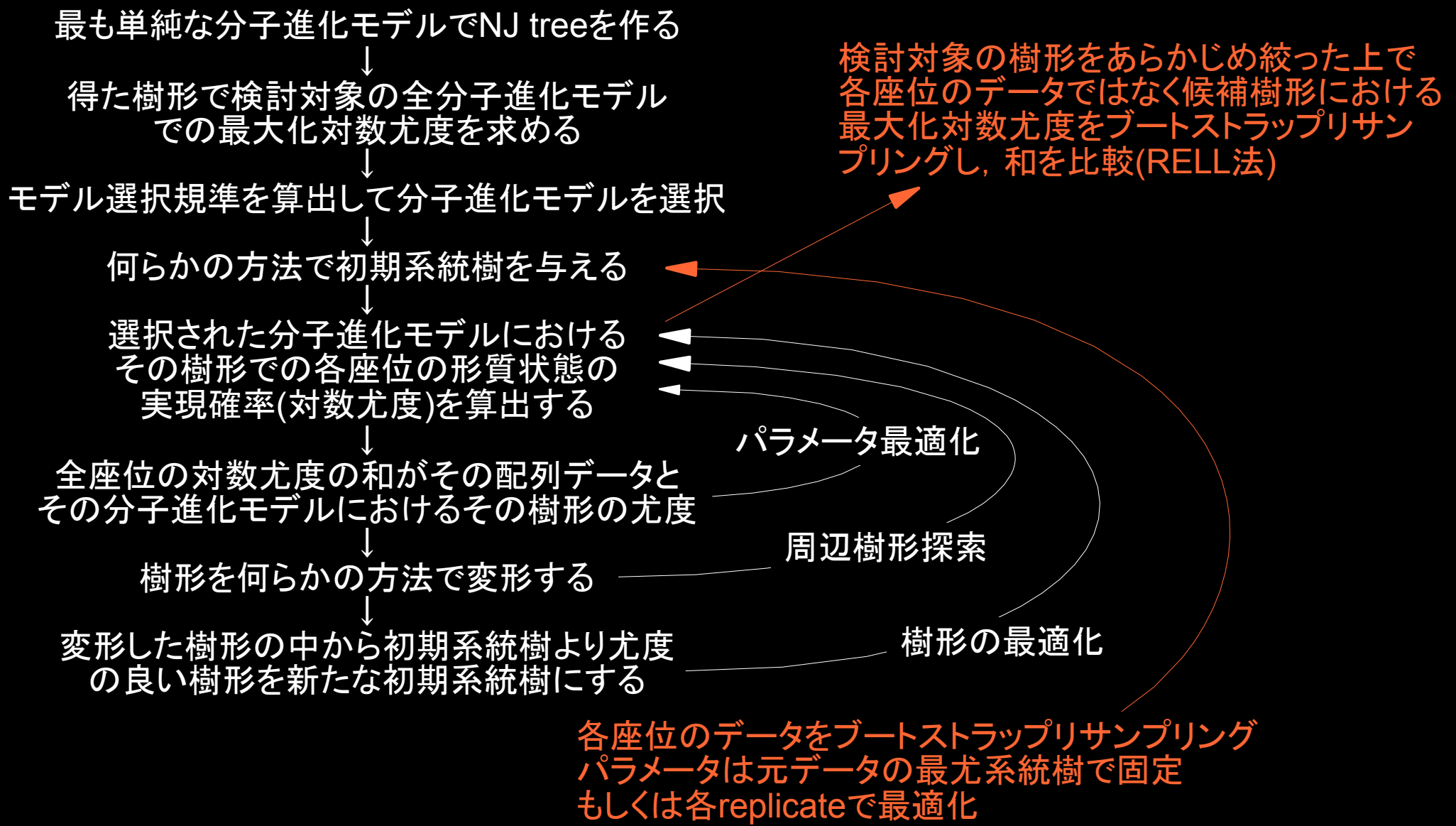


最尤系統推定法の現在

最尤系統推定の手順(発見的探索)



最尤系統推定とブートストラップ解析

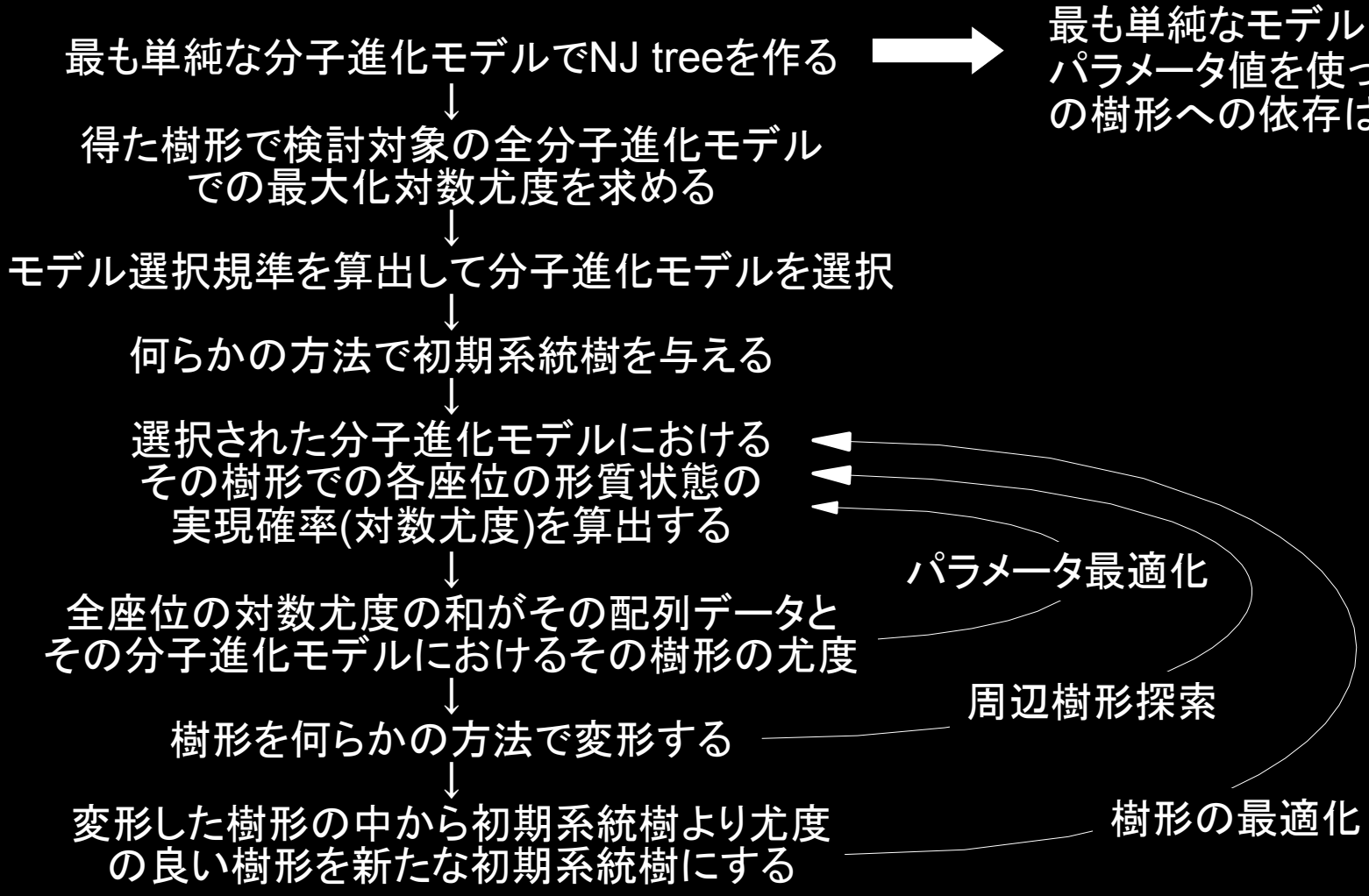


モデル依存性をいかに回避するか?

モデル依存性を抑制する方法

- weightの大きい分子進化モデルを全て検討
- モデル平均化 (model averaging)
- 最尤系統樹で再度分子進化モデル選択する
- ブートストラップ解析

モデル平均化



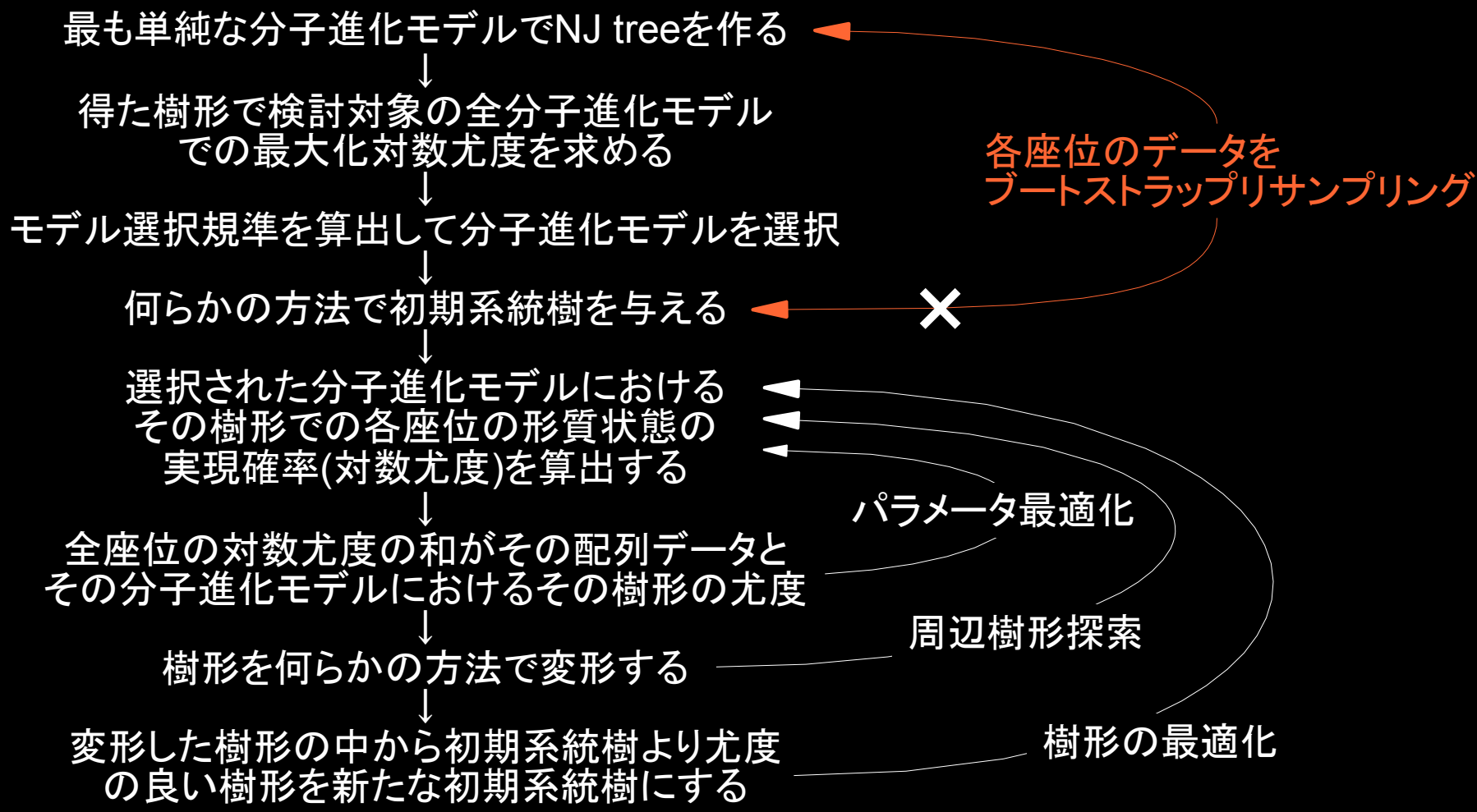
最も単純なモデルにおけるNJ treeでのパラメータ値を使ったモデル平均化はその樹形への依存は残るのでは?

最尤系統樹で再度分子進化モデル選択する

最も単純なモデルにおけるNJ tree
で分子進化モデルを選択
↓
選択された分子進化モデルで樹形選択
↓
選択された樹形で再度モデル選択

やらないよりはマシ
マズいとは言えるが疑い無しとは言えない
計算量から言えば現実的な対処法

ブートストラップ解析を用いた分子進化モデルと系統モデル依存性の抑制



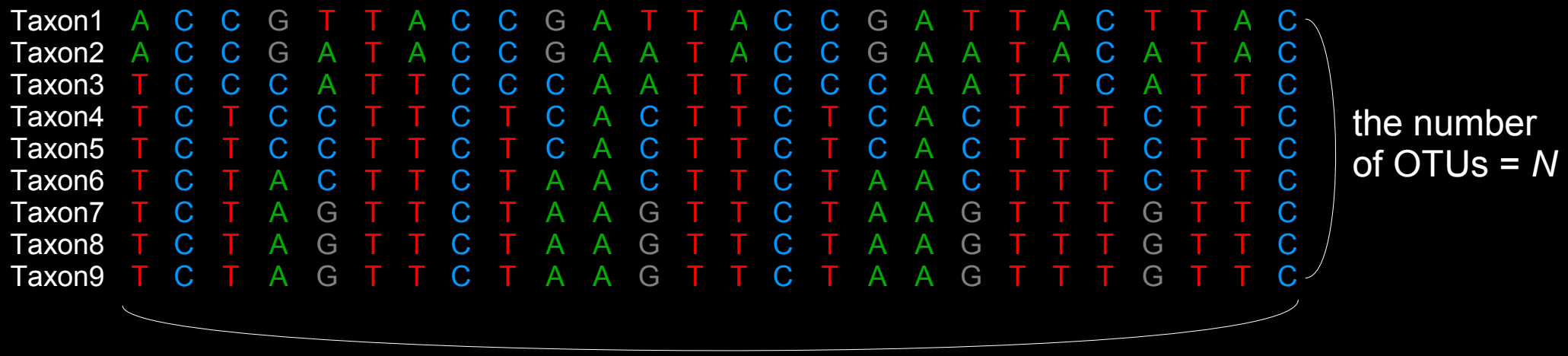
計算量を考えると現時点では非現実的か

モデル選択規準は何を使うべきか?

その前に・・・

サンプルサイズ(標本数)って何?

サンプルサイズの数え方

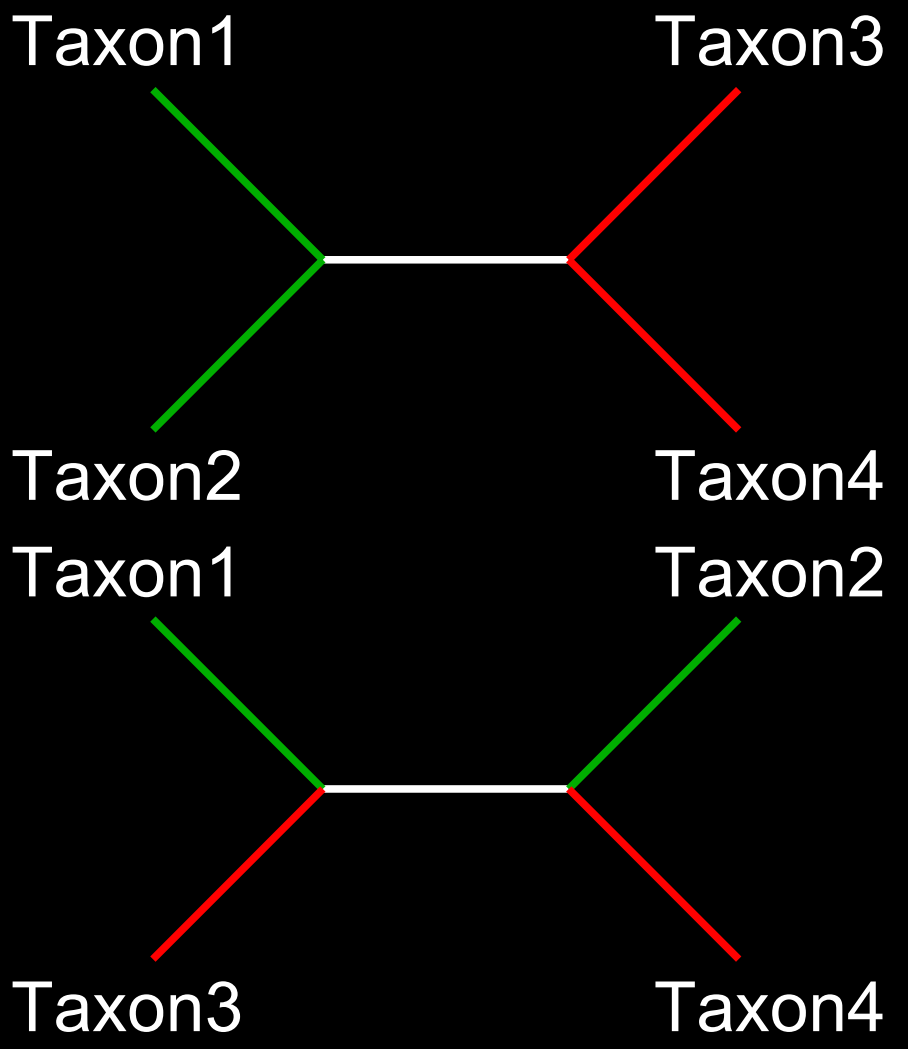


- 塩基平衡頻度(0~3) ･･･ $N \times L$
- 塩基置換確率行列(0~5)
  ･･･ 各座位における置換数の和? or 変異座位数?
- 座位間の速度の不均質性(0~) ･･･ L
- 枝長($2N$-3) ･･･ 各座位における置換数の和? or 変異座位数?

･･･全体としてはサンプルサイズの少ないものにあわせるべき

各座位における置換数は系統モデルによって変化する

Taxon1 A
Taxon2 A
Taxon3 T
Taxon4 T



しかし各座位の置換数の和にしろ，変異座位数にしろ，パラメータ数の40倍を下回ることは現実のデータ解析ではかなり多い → AICcが良い?

分子進化速度進化モデル選択はすべきか?

第3のモデル選択

分子進化速度進化モデルと
樹形を固定しての
分子進化モデルの選択

⬇

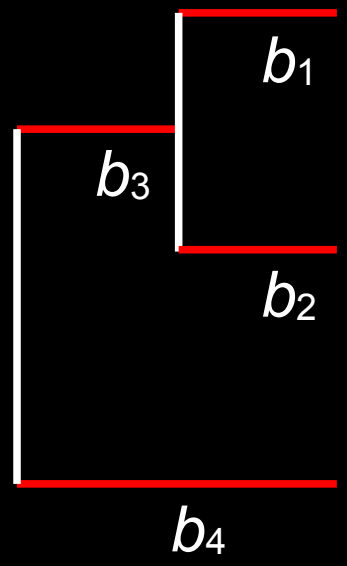
分子進化速度進化モデルと
分子進化モデルを固定しての
系統モデル(樹形)の選択

⬇

分子進化モデルと樹形を固定しての
分子進化速度進化モデル選択

分子進化一定の検証法

- No-Clock ML tree
  - 枝長パラメータ数はOTU数×2-3
- Enforce-Clock ML tree
  - 枝長パラメータ数はOTU数-1



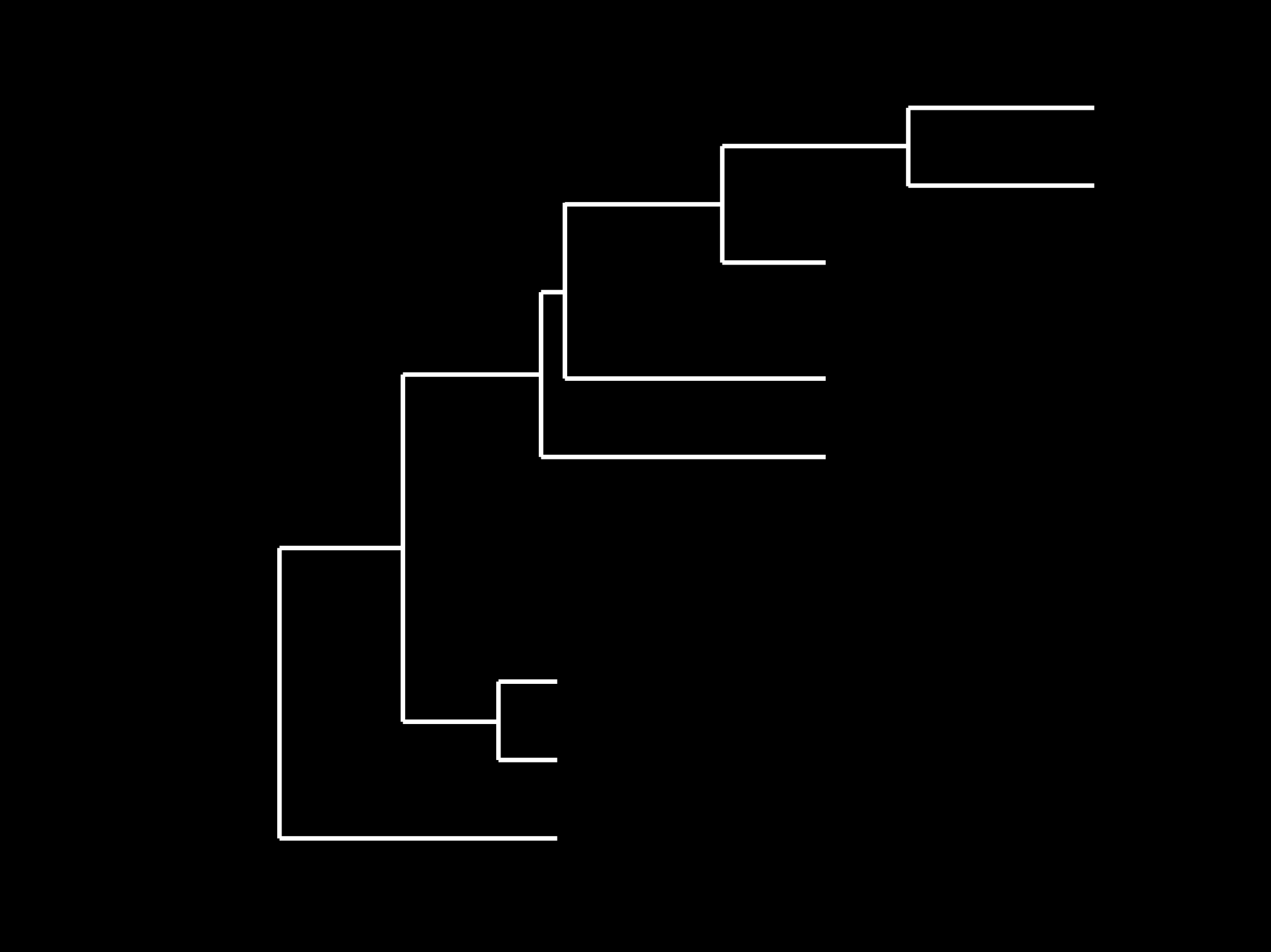
$b_1=b_2$

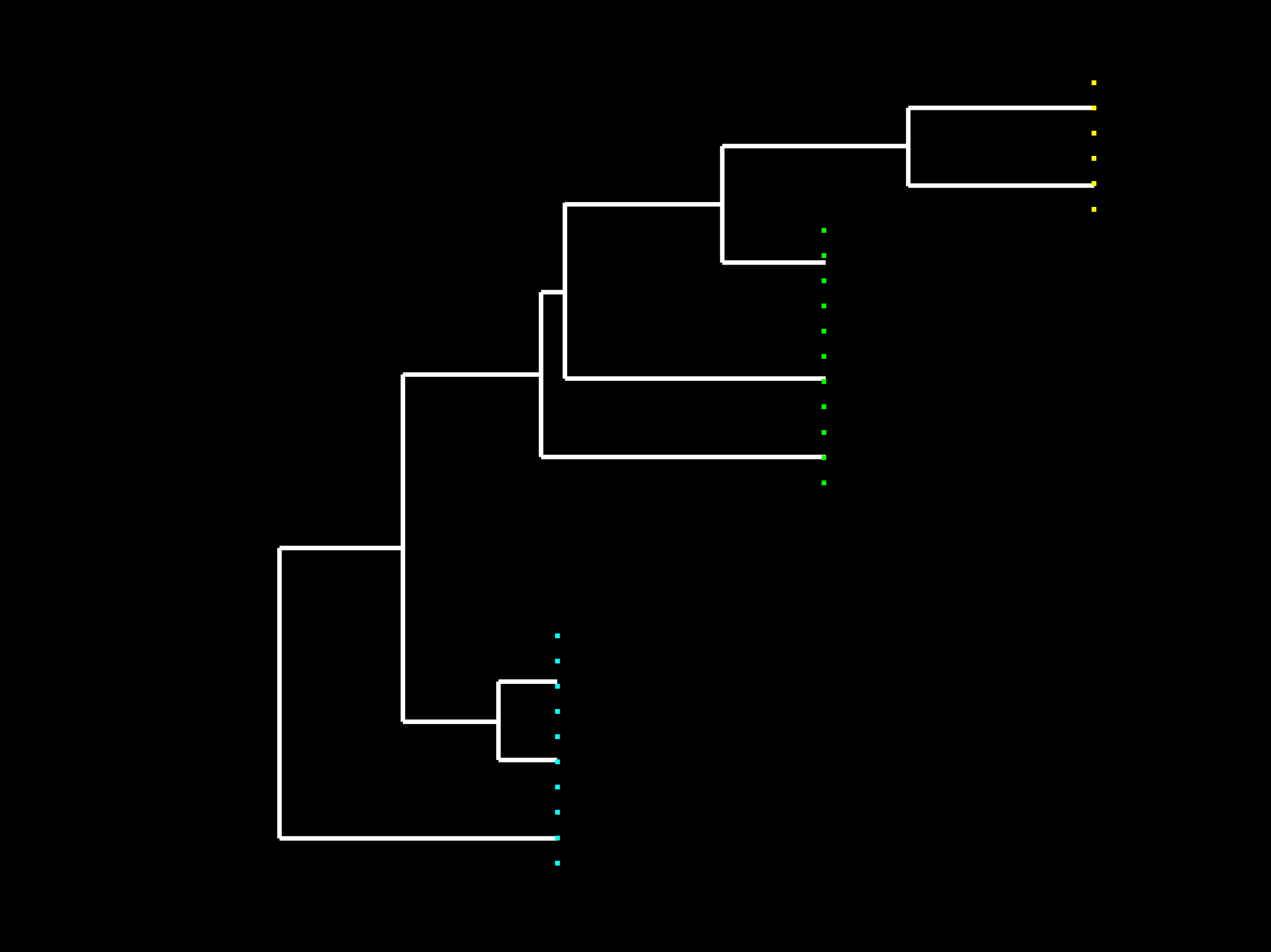
$b_1+b_3=b_4$

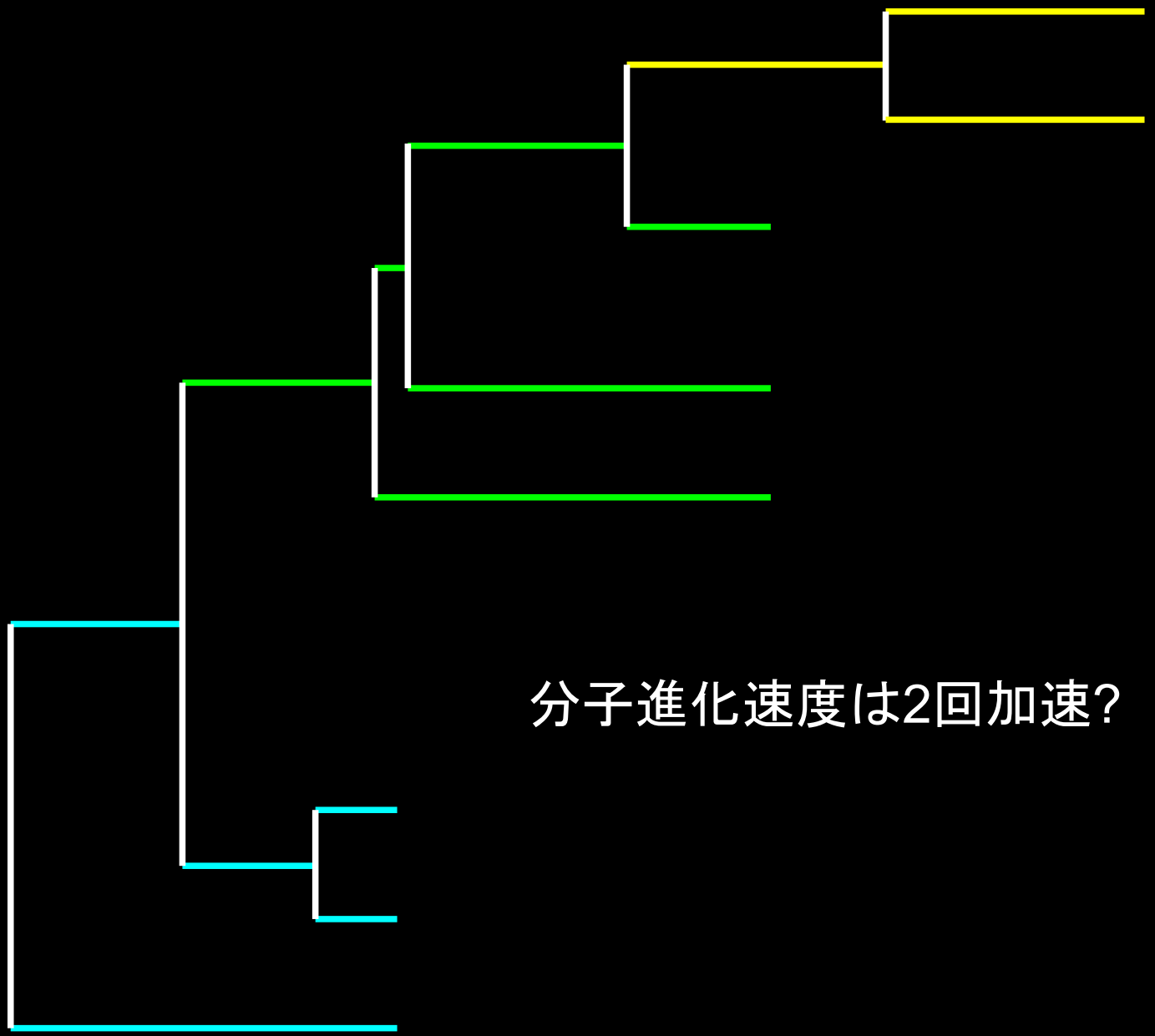
パラメータ数の差はOTU数-2

No-Clock vs Enforce-Clock

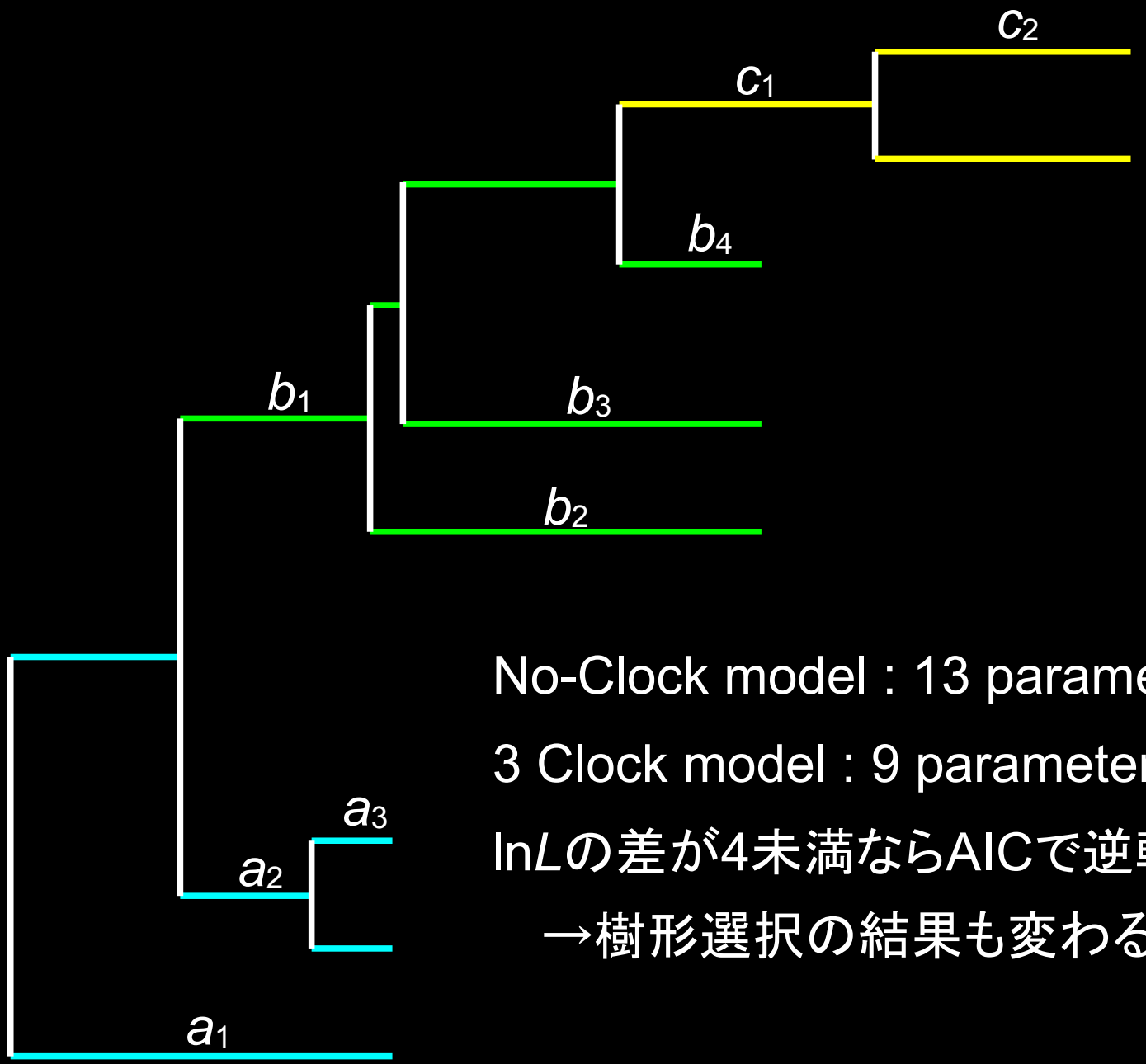
全部違う，と，全部同じ，の
二者択一







分子進化速度は2回加速?



No-Clock model : 13 parameters

3 Clock model : 9 parameters

ln*L*の差が4未満ならAICで逆転

→樹形選択の結果も変わるかも

分子進化速度進化モデル選択

- 利点
  - 分岐年代推定への応用可能
  - 系統モデル(樹形)選択の改善できる
  - 外群の無い系統解析での外群特定への応用可能
- 欠点
  - 膨大な計算量 → 既存技術を用いた仮説の限定が必要
  - 複雑なパラメータ推定 (絶望的?)
  - long branch attractionを助長?