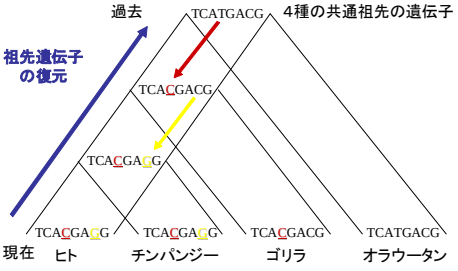
分子進化のモデリングと分子系統樹の推定

足立　淳　　データ科学研究系・リスク科学NOE ゲノム解析　准教授

2013年6月14日　統計数理研究所　オープンハウス

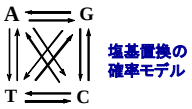
分子系統学とは

遺伝子の変異差から種の系統を推定



配列の違いは共通祖先から進化してきた結果生じたものであり、進化の歴史を反映している。

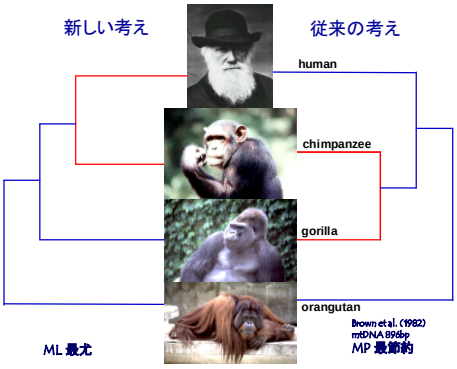
→ 最尤法による系統樹推定



1. Human　ヒト
2. Chimpanzee　チンパンジー
3. Gorilla　ゴリラ
4. Orangutan　オランウータン

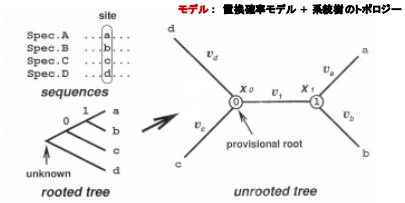
確率過程
・突然変異
・変異遺伝子の集団への固定

```
1 CTAGGCTATATACAACTACGCAAAGGCCCCAACGTTGTAGGCCCCTAC
2 CTAGGCTACATACAACTACGCAAAGGTCCCAACATTGTAGGTCCTTAC
3 TTAGGCTATATACAACTACGTAAAGGCCCCAACGTCGTAGGCCCCTAC
4 CTAGGCTATACACAACTACGCAAGGGACCTAACATCGTAGGCCCCTGC
```



最尤法と尤度(Likelihood)　$L = P(\text{data}|\text{model})$

モデル：置換確率モデル ＋ 系統樹のトポロジー



尤度は、与えられたモデルのもとで、データが実現する確率。
最尤法ではこれを最大化するモデルを選択。
モデルに含まれるパラメータも最尤推定。

Likelihood of a site

$$L = \sum_{x_0} \pi_{x_0} [P_{x_0 c}(v_c) \times P_{x_0 d}(v_d) \times \sum_{x_1} P_{x_0 x_1}(v_1) P_{x_1 a}(v_a) P_{x_1 b}(v_b)]$$

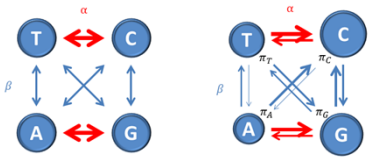
置換モデル

塩基置換やアミノ酸置換のマルコフモデル
Transition probability matrix $\mathbf{P}(t)$ during time t

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

- DNAモデル（塩基置換）
  - A ⇔ G,　T ⇔ C　トランジション
  - (A,G) ⇔ (T,C)　トランスバージョン
- タンパクモデル（アミノ酸置換）
  - 経験的に得られた 20×20 の置換配列
- コドンモデル
  - アミノ酸置換 と 同義塩基置換 の組み合わせ
  - 深い分岐の系統と浅い分岐の系統を同時推定

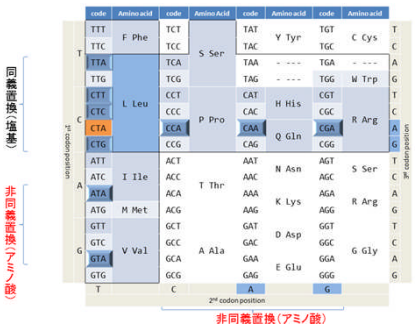
塩基置換モデル



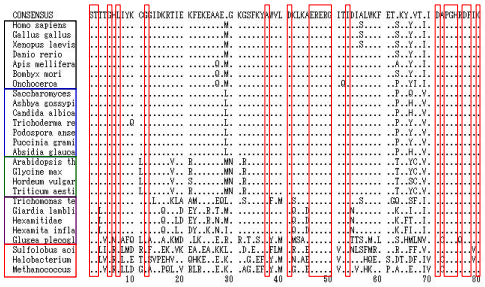
Transition / Transversion model
K80
Kimura's 2 parameter (1980)

Ts / Tv + Proportional model
HKY85
Hasegawa, Kishino & Yano 1985

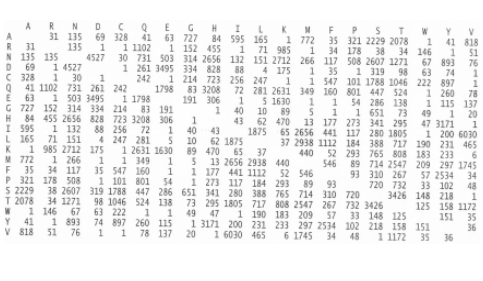
コドンモデル　CTA へ9通りの置換



EF1 αのアミノ酸配列：動物、菌類、植物、原生生物、細菌で共通の配列が見られ

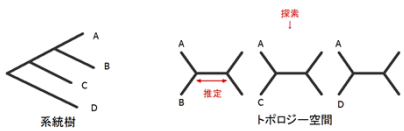


アミノ酸置換モデル　mtREV



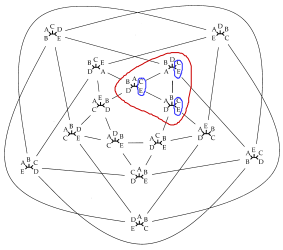
系統樹 ＝ トポロジー ＋ 距離

- トポロジー　＝　枝分かれの順番
  トポロジー空間から探索
- 距離　＝　枝の長さ（推定置換数）
  置換確率モデルから計算



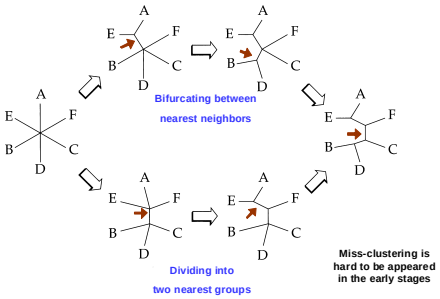
トポロジー空間と探索法

最近隣木間の関係

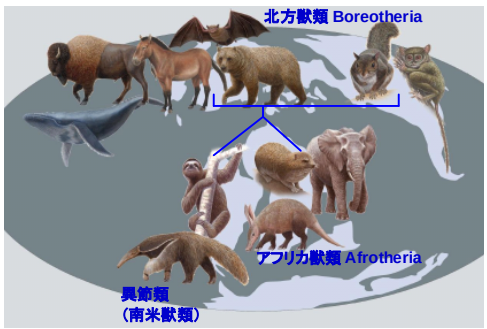
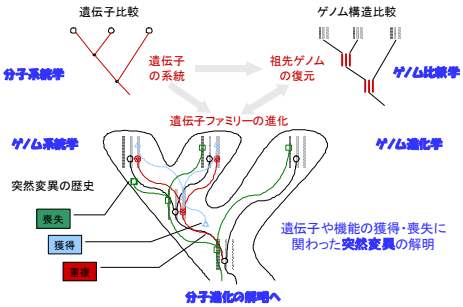


The space of all 15 possible unrooted trees with 5 tips

New star decomposition



1億年前の大陸配置と真獣類の進化



遺伝子系統×ゲノム構造比較＝進化ダイナミクス