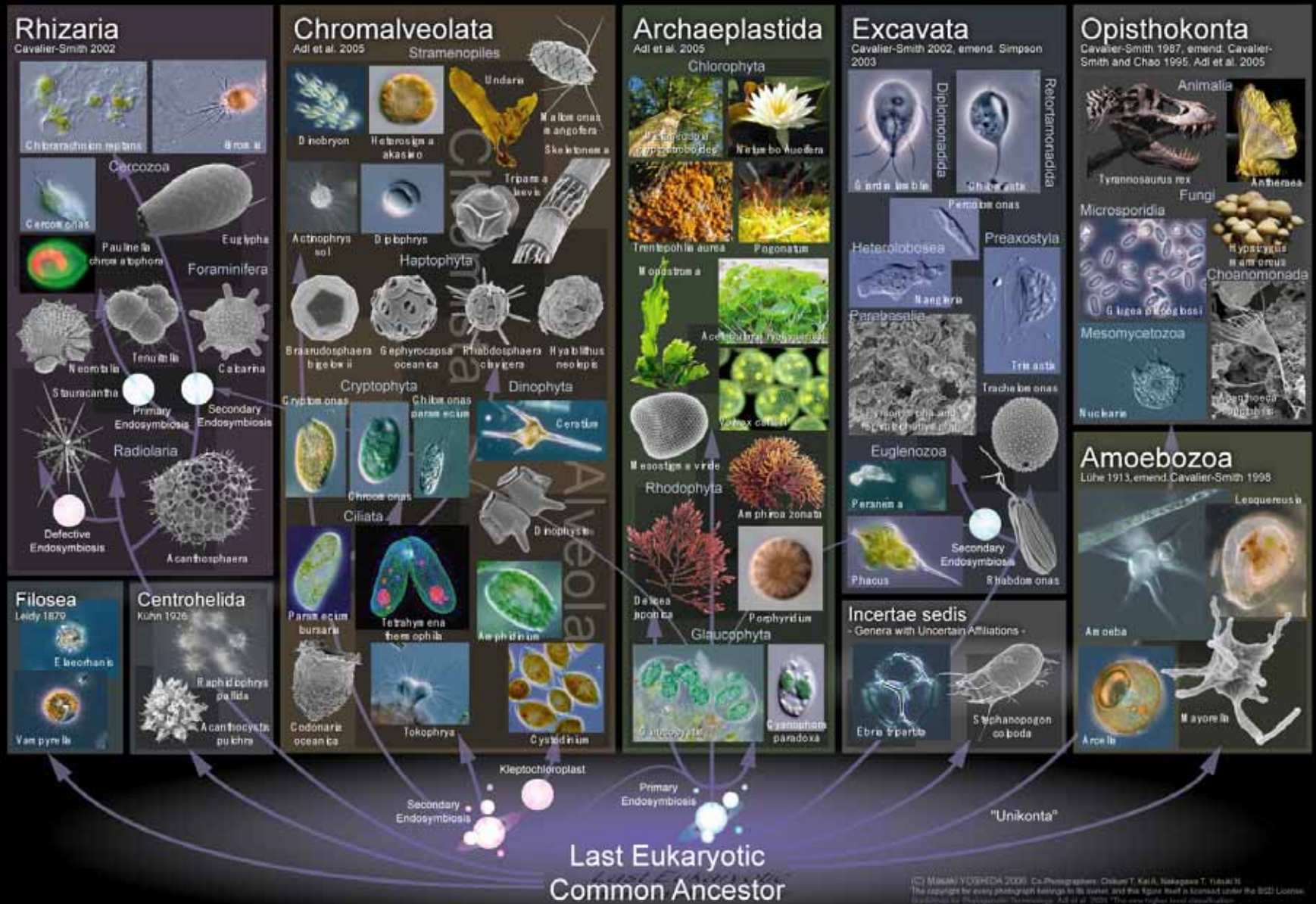


Tracing back our origin

大規模遺伝子データ解析による 真核生物分子系統

分子進化研究グループ

真核生物の多様性



研究目標

- 真核生物大グループ間の系統関係の推定
- もっとも原始的な真核生物の推定
 - 複数の遺伝子配列データを統合
 - 最尤 (Maximum-likelihood) 法
- 新奇生物群からの遺伝子データの取得
 - 細胞の培養、分子生物学実験
- 解析条件・頑健な推定法の検討
 - モデルとなる現存データによるテスト
 - シミュレーションデータによる解析

最尤法

- 「尤度」の計算
 - 置換モデルに基づき、現存の配列データの分岐関係（樹型）が実現化する確率
 - 尤度が最大になるよう樹長を決定する
- 最大の尤度をもつ樹型＝「最尤系統樹」
- 最尤法の利点
 - 最尤法による推測は比較的頑健である
 - モデルを改良しうる
 - 良いモデルを使用し、十分に大きなデータを与えれば、高い確率で真の系統樹を復元する

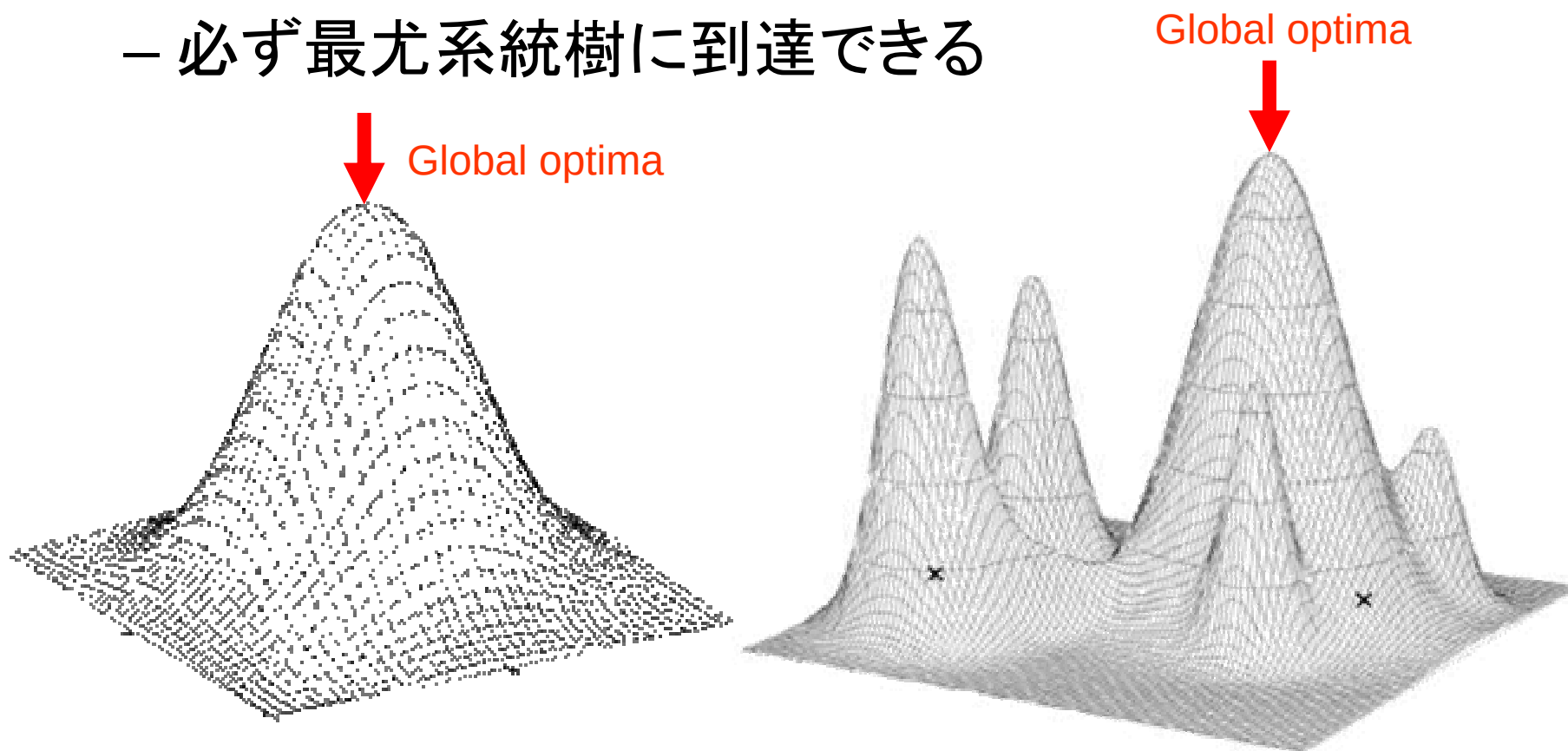
網羅的な最尤系統樹の探索(1)

- 生成可能なすべての樹形を探索するべき
 - 配列(OTU)数の増加→樹形数は爆発的に増加

OTU数	可能な樹型数
3	1
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
11	34459425
12	654729075
13	15058768725
14	376469218125
20	8200794532637891559375

網羅的な最尤系統樹の探索(2)

- 生成可能なすべての樹形を探索する
 - すべての樹形の対数尤度を計算
 - 必ず最尤系統樹に到達できる



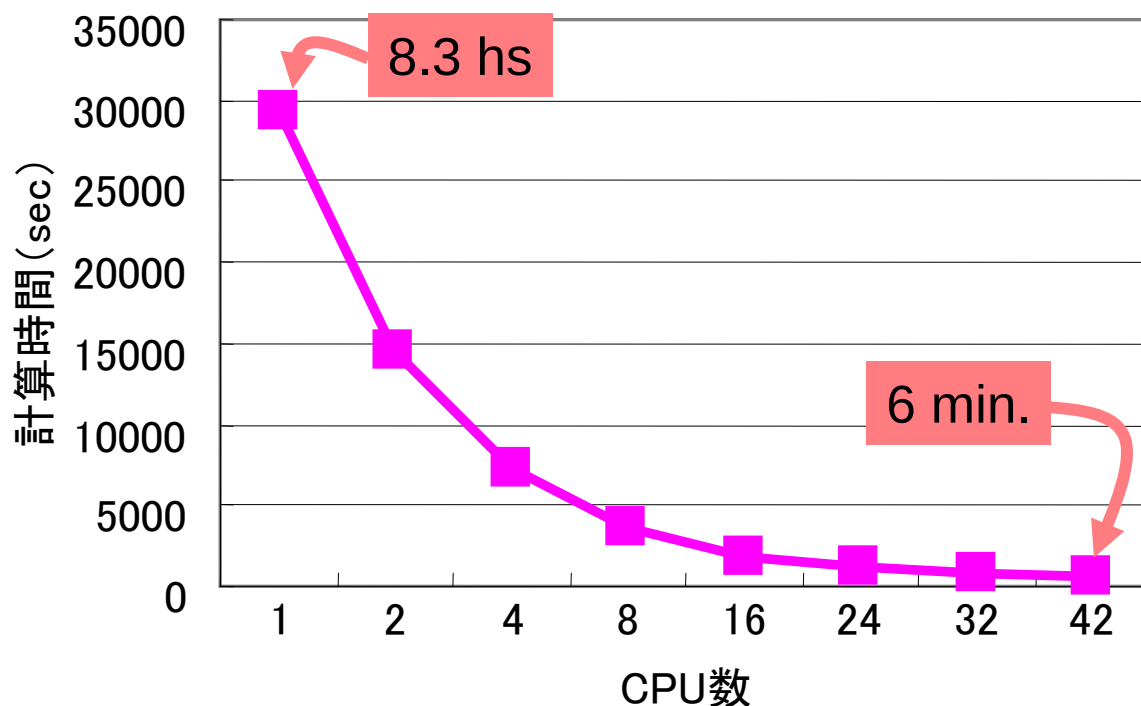
網羅的探索法・・・連結データ解析の場合

- 最尤系統樹：平均的に最も良く支持される樹形

	T_1	T_2	T_3	...	T_X	...	T_{N-1}	T_N
Gene ₁	InL	InL	InL	...	InL	...	InL	InL
Gene ₂	InL	InL	InL	...	InL	...	InL	InL
Gene ₃	InL	InL	InL	...	InL	...	InL	InL
⋮	⋮	⋮	⋮		⋮		⋮	⋮
Gene ₁₀	InL	InL	InL	...	InL	...	InL	InL
⋮	⋮	⋮	⋮		⋮		⋮	⋮
Gene _{M-1}	InL	InL	InL	...	InL	...	InL	InL
Gene _M	InL	InL	InL	...	InL	...	InL	InL
Total	InL ₁	InL ₂	InL ₃	...	InL _X	...	InL _{N-1}	InL _N

連結データ解析の並列化

- 並列計算による効率化の例
 - 「ホトギス」グリッド @長浜バイオ大学
 - Intel Pentium4 2GHz (8* G flops) x42
 - 292座位
 - 10395樹形
 - 250樹形
/processor
 - 最大42倍の
効率化



網羅的探索法の問題点(1)

- 20 OTUの場合
 - 8.2×10^{22} 樹形の探索
 - P4 2GHz 10^3 個のグリッドを仮定 (8 T flops)
 - 1プロセッサあたり 8.2×10^{19} 樹形の尤度計算
 - 1プロセッサあたり 10^4 樹形の計算に 8.3 時間
 - 計算には 2.8×10^{15} 日が必要 ☹
- たった 13 OTUの場合
 - 1.5×10^{11} 樹形の探索
 - 1プロセッサあたり 1.5×10^8 樹形の尤度計算
 - 計算には 5000 日以上が必要 ☹

もしも1ペタ・フロップスなら

- たった13 OTUの場合
 - 1.5×10^{11} 樹形の探索
 - P4 2 GHz $\times 10^3$ グリッド (8 T flops)
 - 計算には5000日以上が必要☹
- もし1 P flopsのスーパーコンピューターなら
 - 計算速度は125倍
 - 1.5×10^{11} 樹形の尤度計算は約40日で終了☺
- 真核生物の主要大グループは12-13程度であると予想される
 - 最尤系統樹の推定のみなら実現可能か？

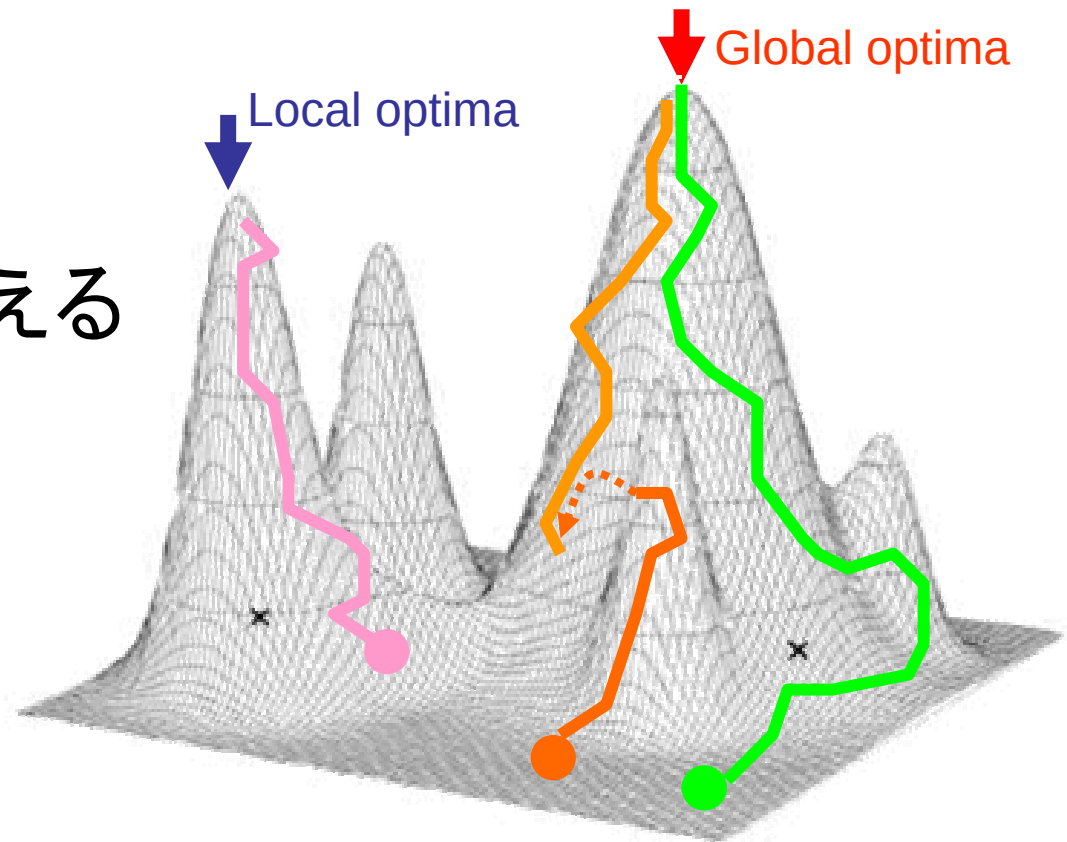
網羅的探索法の問題点(2)

- 結果ファイルサイズの「爆発」
- 9 OTU $\cdot 1.3 \times 10^5$ 樹形の場合: 10^4 残基
 - 結果ファイル = **約** 13.0 GB
- 13 OTU $\cdot 1.5 \times 10^{11}$ 樹形の場合: 10^4 残基
 - 結果ファイル = 1 PB 以上... ☹
- 計算プログラムの改良が必須
 - ファイルのフォーマットの変更
 - 結果ファイルの逐次処理

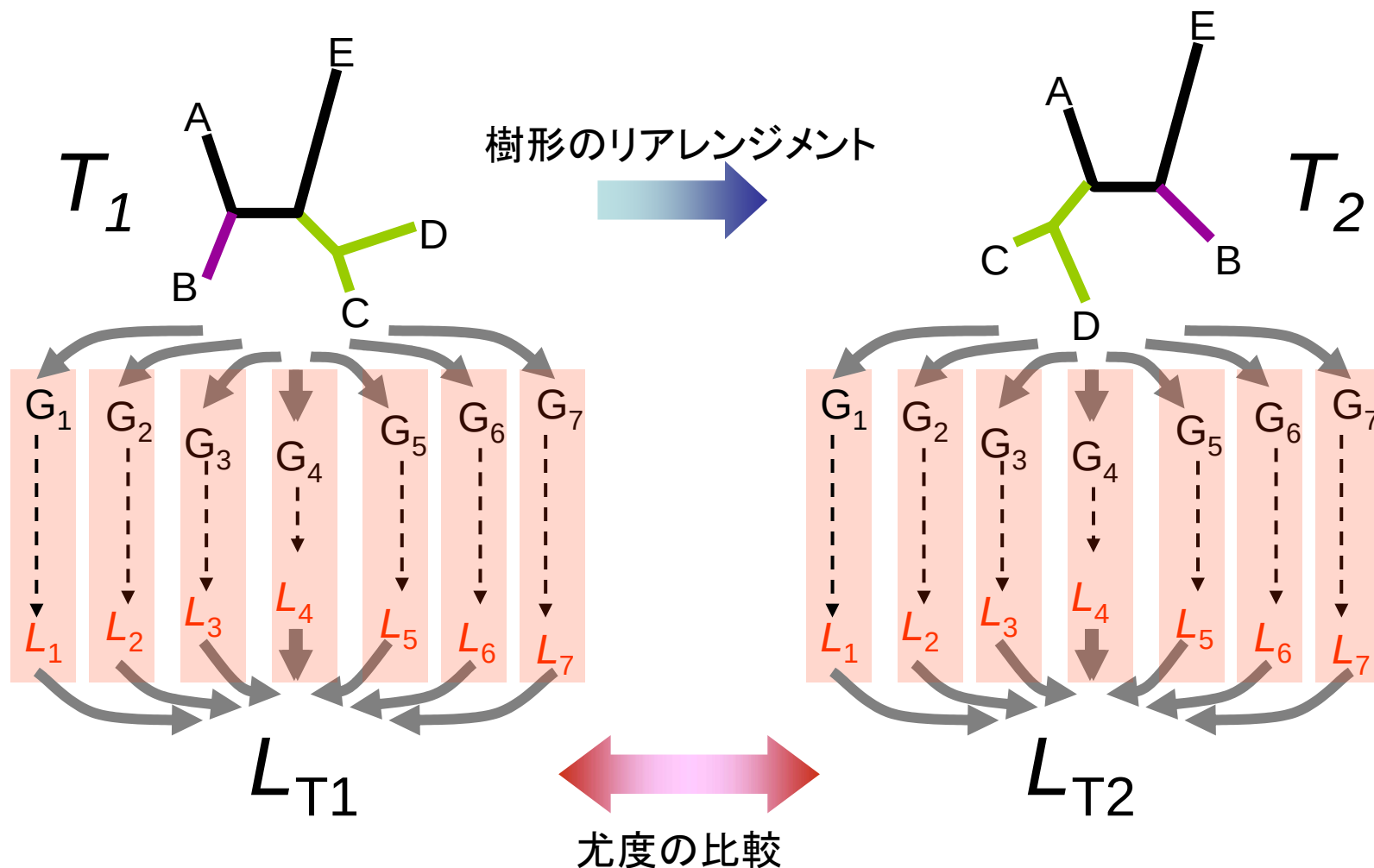
連結データ解析—最尤系統樹の自発的探索

- すべての樹形を探索しない
 - Local optimaにトラップされる可能性がある

- スタート樹形を変える
- 「賢い」樹形探索
 - SPR
 - TBR



連結データにおける自発的系統樹の探索



- 遺伝子毎の尤度計算を並列化、集計する