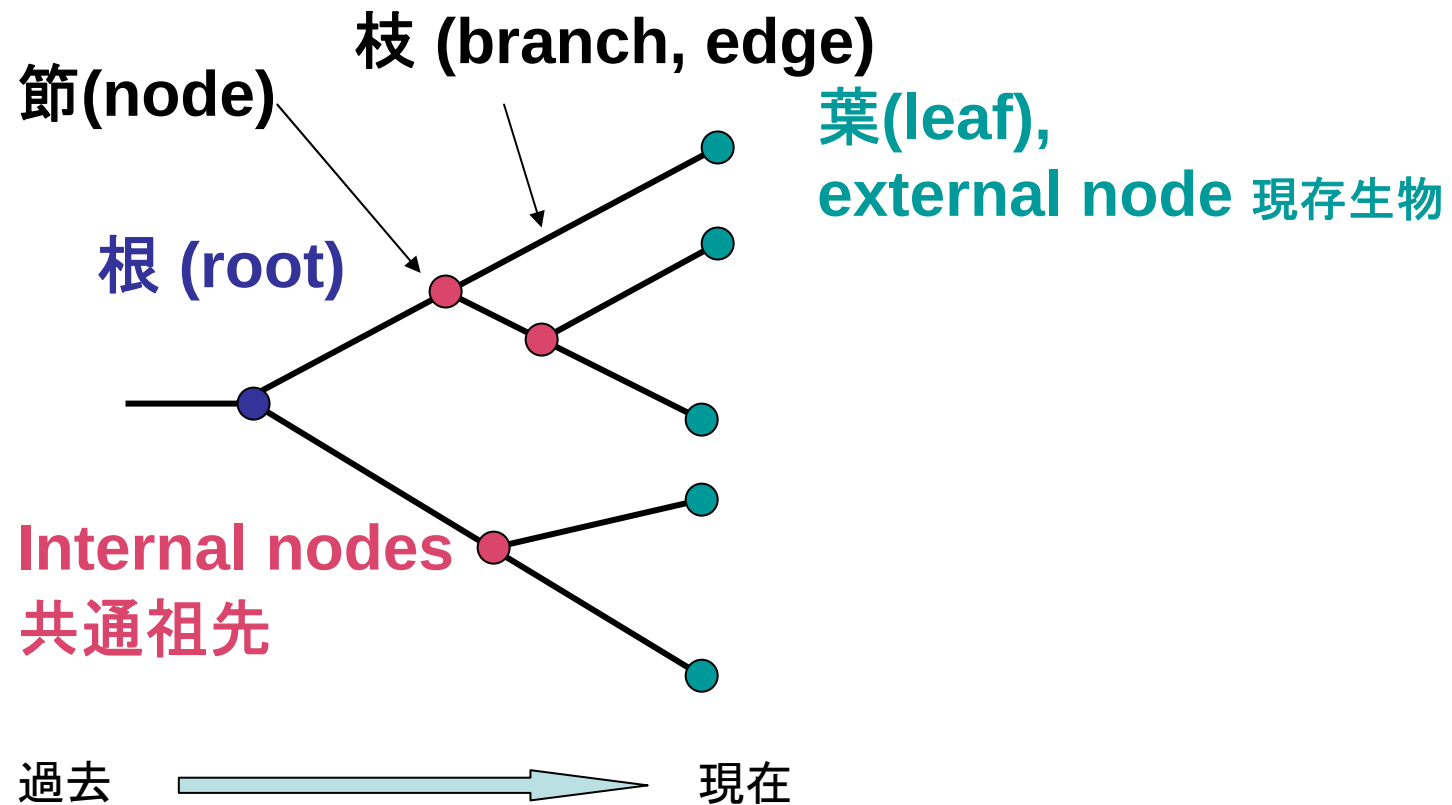


分子系統解析

分子系統解析と系統樹

- 分子系統解析：アミノ酸配列や塩基配列を使って、生物間または遺伝子の進化的道筋（系統）を解明する解析
- 全生物は共通祖先から進化した、という仮説に基づく
- よって、全生物には関連（系統）がある
 - 表現の1つ→系統樹（共通祖先からの分岐）
 - 形態の差異、遺伝子の違いなどをもとに作成する

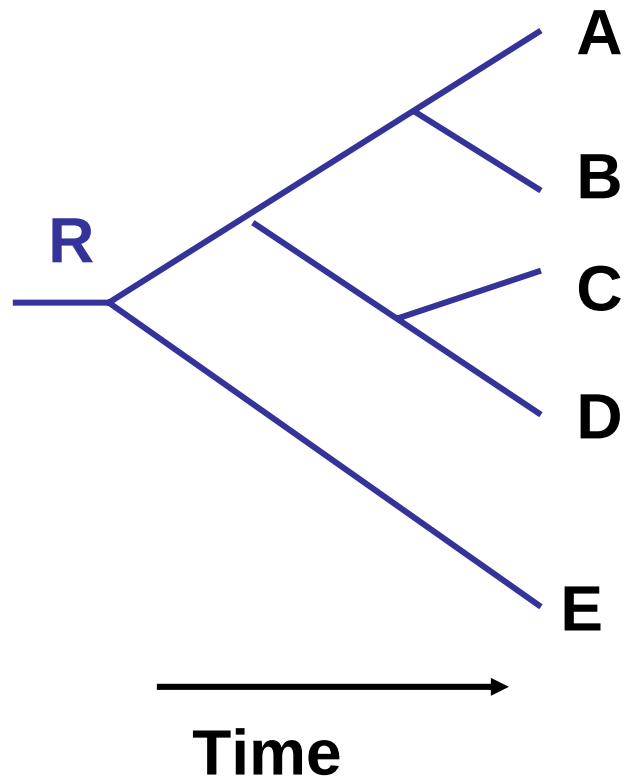
系統樹は節(node)と枝(branch, edge)からなる グラフ(graph)



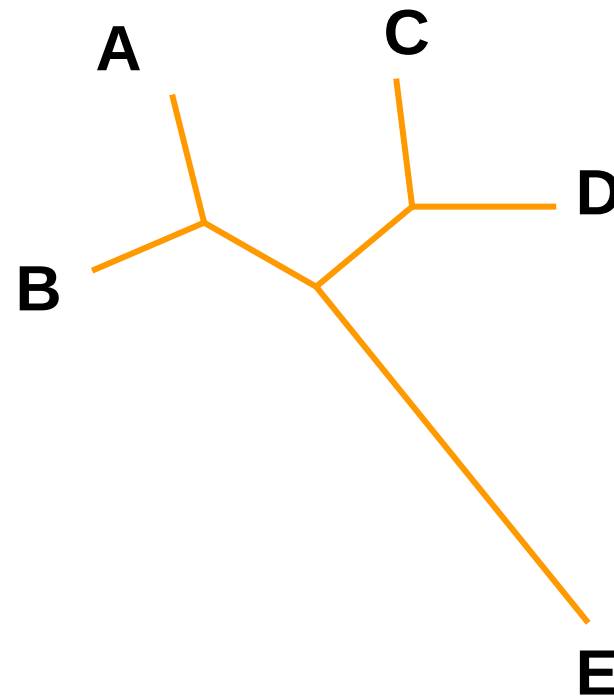
- 枝分かれのパターン(構造)を樹形(Topology)という
- 枝の長さは、進化的形質の違いの大きさを表す

有根系統樹 と 無根系統樹

有根系統樹



無根系統樹

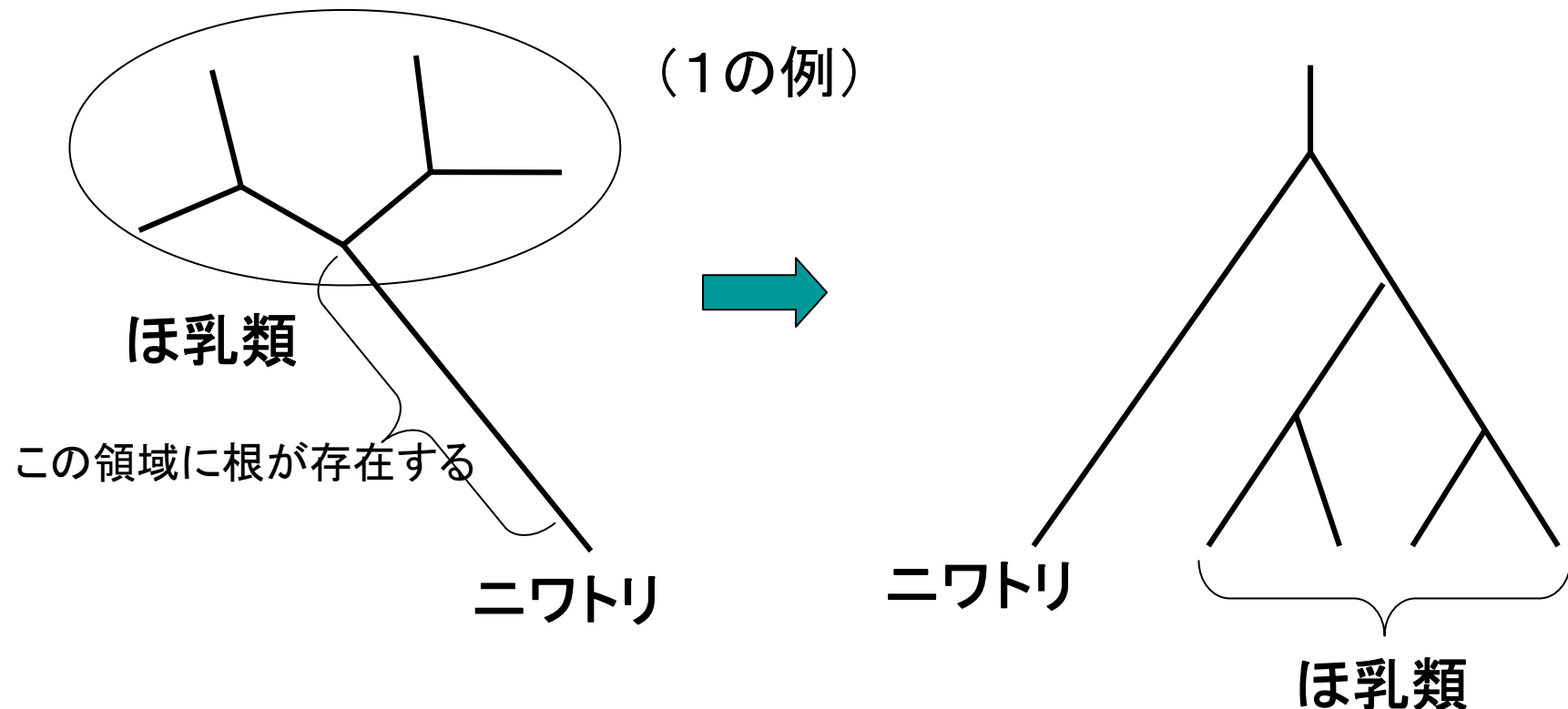


系統樹に含まれる操作単位(生物種、遺伝子などから成る部分集合)を
Operational taxonomic unit (OTU)という

無根系統樹に根をつける方法

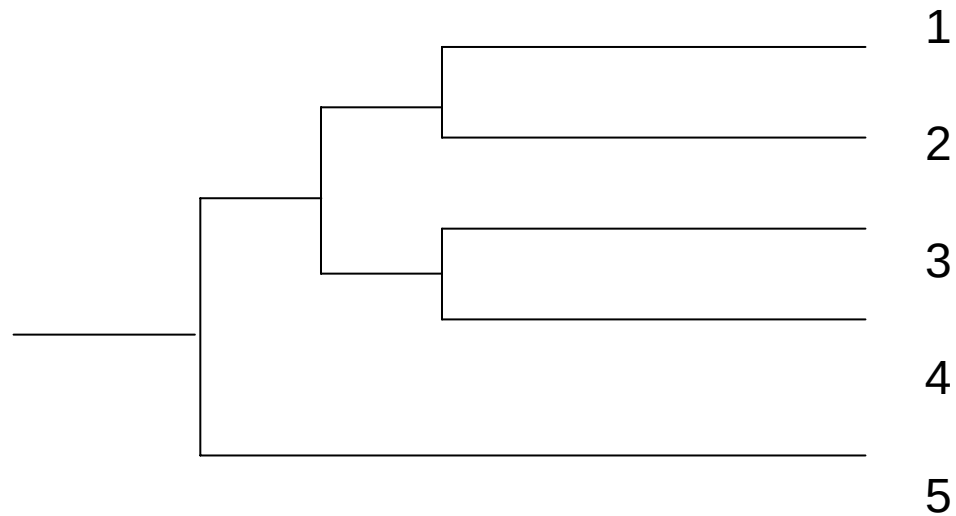
根をつける方法は2つ

1. 最も遠い関係にあると知られている生物種の配列(外群、outgroup)を1つ以上含める
2. 最も遠い関係にある2つの配列を結ぶ枝の中点をinternal nodeとする



Newick format

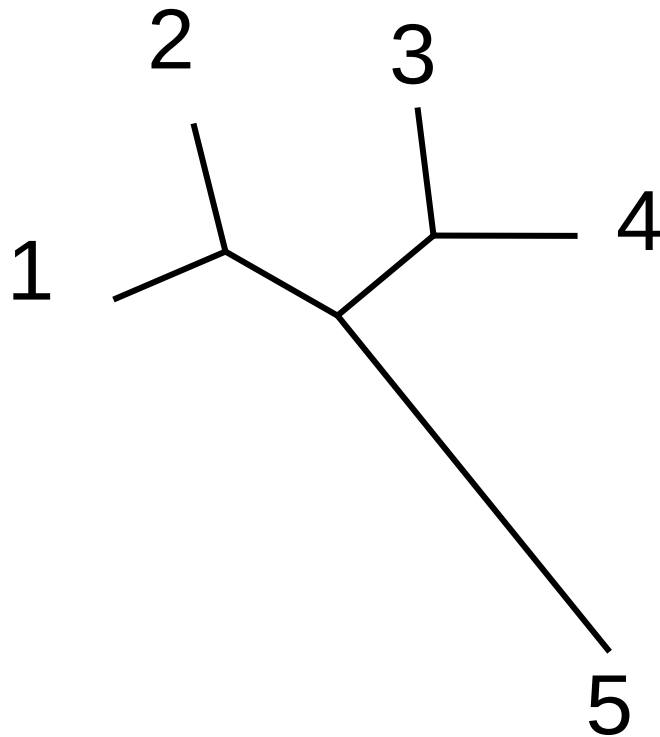
Newick format: テキスト形式での系統樹表現



上の例は、以下のように書ける

`((1,2),(3,4)),5`

同じ系統樹を表すNewick formatは複数



- $((1,2),(3,4),5)$
- $((1,2),5),3,4)$
- $(1,((3,4),5),2)$
-

進化距離

- 進化距離：配列の相違度を示す指標
- 「配列の分岐後の時間の長さと、正の相関がある」と想定する
- 進化距離は、枝の長さに反映される（系統樹作成方法に応じて異なる）

主な分子系統樹推定法

- 距離行列法 (distance matrix method)
 - 平均距離法 (UPGMA)
 - 近隣結合法 (neighbor joining method, N-J法)
- 形質状態法(character state method)
 - 最節約法 (maximum parsimony method, MP法)
 - 最尤法 (maximum likelihood method, ML法)
 - 事前確率を考慮したものは、特に「ベイズ法(Bayesian method)」という

系統樹作成の前に、 マルチプルアラインメントを作成する

HIT000015863_Homo.
XM_508814_Pan.
U93310_Macaca.
BC054436_Mus.
X03150_Rattus.
XM_844606_Canis.
XM_001503225_Equus.
BC104530_Bos.

```

21 KVTSLTACLVDDQSLRLDCRHENTSS-SPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTN 79
   KVTSLTACLVDDQSLRLDCRHENTSS-SPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTN
   KVTSLTACLVDDQSLRLDCRHENTSS-SPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTN
   KVTSLTACLVNQNLRLDCRHENNTKDNSIQHEFSLTREKKHVLSGTLGIPEHTYRSRVT
   RVISLTACLVNQNLRLDCRHENNTN-LPIQHEFSLTREKKHVLSGTLGVPEHTYRSRVN
   KVTSLTACLVDDQSLRLDCRHENATT-LPIQYEFMTREKKQHVIYGTGVGVPEHSYRSRTN
   KVTSLTACLADQSLRLDCRHENTTA-TPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTS
   KVTSLTACLVNQSLRLDCHHENTTT-TPIQYEFSLTRDTKKWVLSGSNGVTRPEYSSRTK
   :* *****.*.*****:*** : .**:*:*:*:*:..: *: *: *:.. * **..

```

HIT000015863_Homo.
XM_508814_Pan.
U93310_Macaca.
BC054436_Mus.
X03150_Rattus.
XM_844606_Canis.
XM_001503225_Equus.
BC104530_Bos.

```

80 FTSKYNMKVLYLSAFTSKDEGTYTCALHHS GHSPPISSQNVTVL RDKL VKCEGISLLAQN 139
   FTSKYNMKVLYLSAFTSKDEGTYTCALHHS GHSPPISSQNVTVL RDKL VKCEGISLLAQN
   FTSKYNMKVLYLSAFTXKDEGTYTCX LHSGHSPPISSQNVTVL RDKL VKCEGISLLAQN
   LSNQPYIKVLT LANFTTKDEGDYFCE LQVSGANPMSSNKSISVYRD KL VKCGGISLLVQN
   LFSDRFIKVL T LANFTTKDEGDYMC ELRVSGQNPTSSNKTINVIR DKL VKCGGISLLVQN
   FTSKYNMKVLYLSGFTTKDEGTYTC ELRLSGQPPITSSKNVTVL RDKL VKCEGISLLAQN
   FTSKYNMKVLYLSGFTTKDEGLYTCE LRLSGQPPVVSNTVSVLR DKL VKCEGISLLAQN
   FFSKYNLKVLYLSNFTTKDEGIYTC ELRLSGPNPSVSNKDVS VLR DKL VRCGGISLLIQN
   : .. :*** *: ** ***** * * *: ** * *: :. * *****.* ***** **

```

配列ペアごとに進化距離を計算し、距離行列が得られる

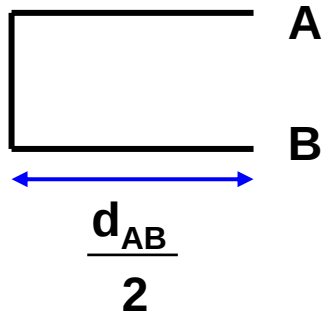
UPGMA法

スタート

配列ペアの距離行列

	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

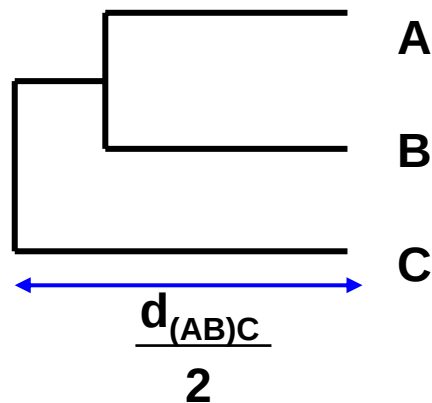
最小



新しい距離行列

	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

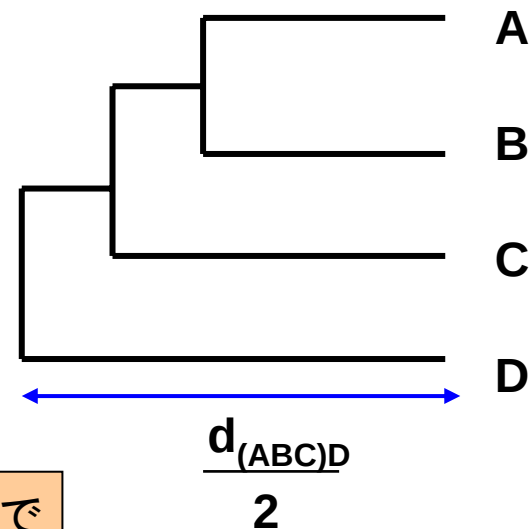
最小



繰り返して



完成

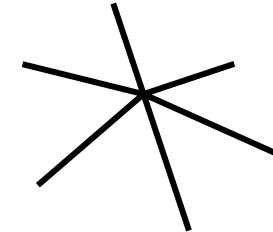


UPGMA法(概要)

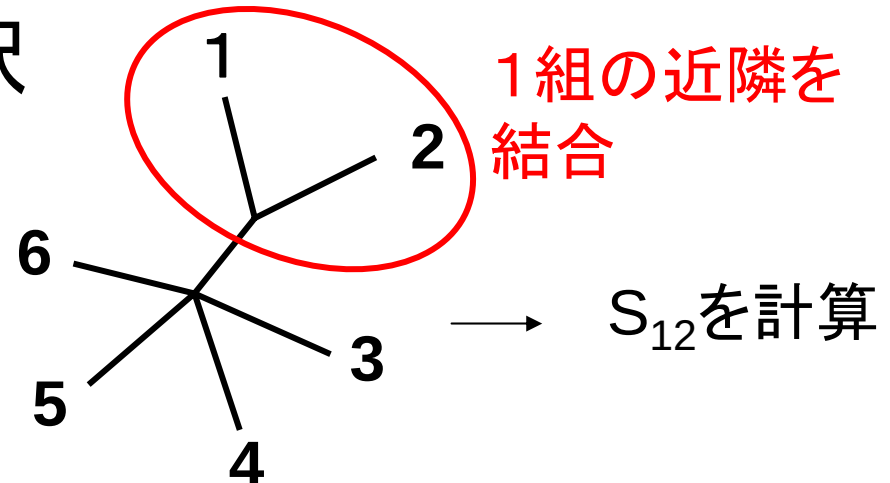
1. 配列の全ペアから距離行列の計算
 2. 距離行列の要素中の最小値をとり、それに対応する組み合わせ(A、Bとする)を1つのOTU(AB)とする
 3. 上記OTU(AB)についての系統樹を作成する
枝の長さは「AB間の距離 d_{AB} 」の半分
1. 距離行列の要素が1つなら終了
 2. 上記のOTUと他の配列との距離行列を構築
 3. 新規な距離行列を元に、ステップ1に戻る

近隣結合法

1. 配列の全ペアから距離行列の計算
2. 樹形を星状樹に設定
3. 星の中心にある近隣ペアを選び、結合（星から分離）した樹形を作成し（下図）、各枝の長さの総和 S_{ij} を計算（全ペアについて実行）

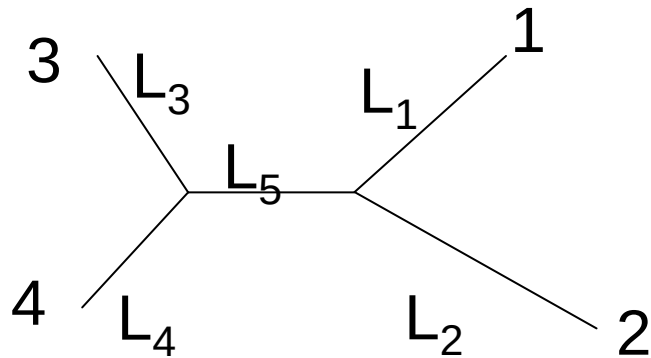


4. S_{ij} が最小の樹形を選択
5. OTUが3つなら終了
6. step3に戻る



枝の長さ L_i の計算

1, 2 (または 3, 4) を結合
した場合の図



距離行列からとる

- $L_1 + L_2 = d_{12}$
- $L_1 + L_3 + L_5 = d_{13}$
- $L_1 + L_3 + L_4 = d_{14}$
- ...
- 枝の長さの総和
 $S_{12} = L_1 + L_2 + L_3 + L_4 + L_5$

最大節約法

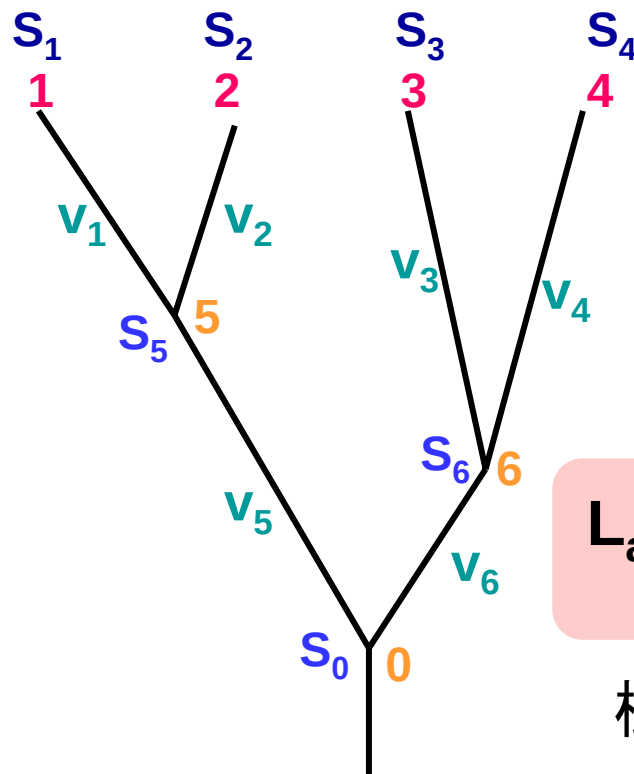
Maximum Parsimony Method (MP法)

- 原理: 残基置換数を最小にする系統樹を選ぶ
 1. 情報を持つサイト(2残基以上が2配列以上)を全て見つける
 2. 情報を持つサイトそれぞれについて、各樹形における置換数を求める
 3. 情報を持つサイト全てについて、置換数の和が最小な樹形を選択する

最尤法 (Maximum likelihood method, ML法)

アライメント中のある座位kを考える。座位kで、下図の樹形aの
尤度 L_{ak} (与えられた塩基データが得られる確率)を求める。

(4本の塩基配列の場合)



$S_0, S_1, \dots, S_6$: 塩基(A, C, G, T)

$v_1, \dots, v_6$: アライメントから求めた各枝の塩基置換率

g_{s0} : ノード0で、塩基が s_0 である確率

$P_{ij}(v)$: 塩基iがvの置換率で塩基jとなる確率

祖先配列 S_0, S_5, S_6 は全ての可能性
(A, C, G, T)について足し合わせる。

$$L_{ak} = \sum_{S_0} \sum_{S_5} \sum_{S_6} g_{s0} P_{s_0s_5}(v_5) P_{s_5s_1}(v_1) P_{s_5s_2}(v_2) \\ \cdot P_{s_0s_6}(v_6) P_{s_6s_3}(v_3) P_{s_6s_4}(v_4)$$

樹形aの全座位(1からnまで)の尤度は、

$$L_a = L_{a1} \times L_{a2} \times \dots \times L_{an}$$

全樹形のうち、最大尤度の樹形を解とする

系統樹推定プログラム

- MEGA (<http://evolgen.biol.metro-u.ac.jp/MEGA/>)
 - 最節約法、距離行列法など
- PHYLIP (<http://evolution.genetics.washington.edu/phyli.html>)
 - 最尤法、最節約法、距離行列法など
- PAUP* (<http://paup.csit.fsu.edu>)、有償
 - 最尤法、最節約法、距離行列法など
- Molphy (http://www.ism.ac.jp/ismlib/softother_j.html)
 - 最尤法
- PAML (<http://abacus.gene.ucl.ac.uk/software/paml.html>)
 - 最尤法、統計的検定機能が豊富
- phyML (<http://www.atgc-montpellier.fr/phyml/>)
 - 最尤法(高速)
- MrBayes (<http://mrbayes.csit.fsu.edu/index.php>)

同義置換と非同義置換(1)

- **同義置換**: 遺伝子DNAのコード配列に生じる塩基置換のうち、アミノ酸に変異を生じないもの
- **非同義置換**: 遺伝子DNAのコード配列に生じる塩基置換のうち、アミノ酸に変異を生じるもの

コドン表

1文字目	2文字目				3文字目
	U	C	A	G	
U	フェニルアラニン	セリン	チロシン	システイン	U
	フェニルアラニン	セリン	チロシン	システイン	C
	ロイシン	セリン	終 止	終 止	A
	ロイシン	セリン	終 止	トリプトファン	G
C	ロイシン	プロリン	ヒスチジン	アルギニン	U
	ロイシン	プロリン	ヒスチジン	アルギニン	C
	ロイシン	プロリン	グルタミン	アルギニン	A
	ロイシン	プロリン	グルタミン	アルギニン	G
A	イソロイシン	トレオニン	アスパラギン	セリン	U
	イソロイシン	トレオニン	アスパラギン	セリン	C
	イソロイシン	トレオニン	リシン	アルギニン	A
	メチオニン(開始)	トレオニン	リシン	アルギニン	G
G	バリン	アラニン	アスパラギン酸	グリシン	U
	バリン	アラニン	アスパラギン酸	グリシン	C
	バリン	アラニン	グルタミン酸	グリシン	A
	バリン	アラニン	グルタミン酸	グリシン	G

同義置換と非同義置換(2)

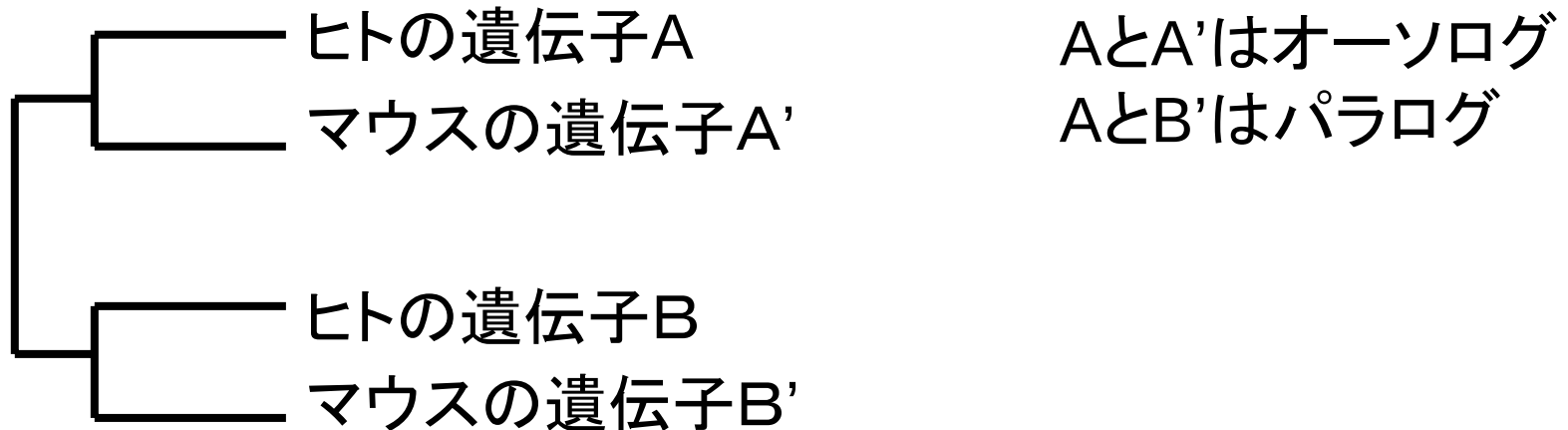
- ホモログのDNAを比較することによって、同義置換と非同義置換の数を推定できる
- 一般には同義置換の方が非同義置換より多い
 - 1世代での塩基の変異はコドンの位置に寄らない
 - アミノ酸に変異のない置換(同義置換)は排除されない傾向
 - アミノ酸に変異のある置換(非同義置換)は有害なため排除される傾向
- 同義置換率(同義置換数/同義サイト数)は多くの遺伝子で共通であるが、非同義置換率は遺伝子によってバラツキがある

相同性と類似性

- 相同性(ホモロジー, homology)
 - 2つの配列(遺伝子/タンパク質)があり、それらが共通祖先から進化して出来たものであるなら、それらには「相同性(homology)がある」または「それらは相同である(homologous)」という
 - ホモログス(homologous)な配列AとBを、ホモログ(homologs/homologues)という
- 相同性(homology)と類似性(similarity)は違う
 - 2つの遺伝子に類似性があっても、相同とは限らない
 - 例: 収斂進化(convergent evolution): 祖先が異なるが、配列が似た(機能が同じ)遺伝子が出来ること
 - 相同性は質的性質(ある、なし)
 - 「相同性が高い」は不適切
 - 類似性は量的性質(「2つの遺伝子の配列類似性は70%」)

オーソログとパラログ

- ・ ホモログは、オーソログとパラログの2種類に分けられる
- ・ オーソログ・オルソログ(orthologs):
 - 種分化の際に分岐したホモログ
- ・ パラログ(paralogs):
 - 遺伝子重複によって生じたホモログ



実習

マルチプル・アラインメントの作成法
CLUSTALWを例に

マルチプルアラインメントの例

```

HIT000015863_Homo.      21 KVTSLTACLVLDQSLRLDCRHENTSS-SPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTN 79
XM_508814_Pan.          KVTSLTACLVLDQSLRLDCRHENTSS-SPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTN
U93310_Macaca.          KVTSLTACLVLDQSLRLDCRHENTTS-SPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTN
BC054436_Mus.           KVTSLTACLVNQNLRLDCRHENNTKDNSIQHEFSLTREKKHVLSGTLGIPEHTYRSRVT
X03150_Rattus.          RVISLTACLVNQNLRLDCRHENNTN-LPIQHEFSLTREKKHVLSGTLGVPEHTYRSRVN
XM_844606_Canis.        KVTSLTACLVLDQSLRLDCRHENATT-LPIQYEFSMTREKKQHVIYGTGVGVPEHSYRSRTN
XM_001503225_Equus.     KVTSLTACLADQSLRLDCRHENTTA-TPIQYEFSLTRETKKHVLFGTVGVPEHTYRSRTS
BC104530_Bos.           KVTSLTACLVNQSLRLDCHHENTTT-TPIQYEFSLTRDTKKWVLSGSNGVTRPEYSSRTK
                        :* *****:* *****:*** :   .**:*:*:*:*:..: *:*:*:*..
                        :

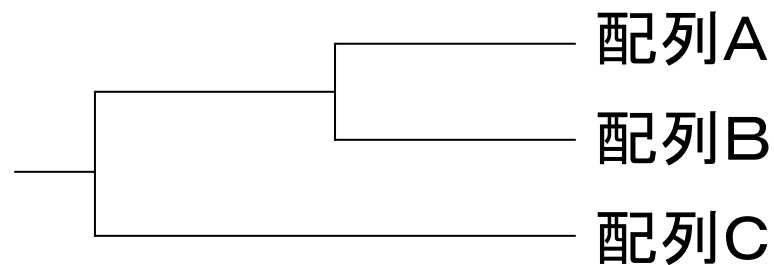
HIT000015863_Homo.      80 FTSKYNMKVLYLSAFTSKDEGTYTCALHHSGHSPPISSQNVTVLRDKLVKCEGISLLAQN 139
XM_508814_Pan.          FTSKYNMKVLYLSAFTSKDEGTYTCALHHSGHSPPISSQNVTVLRDKLVKCEGISLLAQN
U93310_Macaca.          FTSKYNMKVLYLSAFTXKDEGTYTCXHLHSGHSPPISSQNVTVLRDKLVKCEGISLLAQN
BC054436_Mus.           LSNQPYIKVLTLANFTTKDEGDYFCELVSGANPMSSNKSISVYRDKLVKCGGISLLVQN
X03150_Rattus.          LFSDRFIKVLTLANFTTKDEGDYMCELRVSGQNPTSSNKTINVIRDKLVKCGGISLLVQN
XM_844606_Canis.        FTSKYNMKVLYLSGFTTKDEGTYTCELRLSGQPPITSSKNVTVLRDKLVKCEGISLLAQN
XM_001503225_Equus.     FTSKYNMKVLYLSSFTTKDEGLYTCELRLSGQPPVVSNTVSVLRDKLVKCEGISLLAQN
BC104530_Bos.           FFSKYNLKVLYLSNFTTKDEGIYTCELRLSGPNPSVSNKDVSVLRDKLVRCGGISLLIQN
                        : .. :*** *: ** ***** * * *: ** * * : : * *****:* ***** **

```

ペアワイズ・アラインメントよりも多くの情報が得られる
保存性の高いアミノ酸サイトや領域など

ペアワイズ・アラインメントと マルチプルアラインメントの違い

- 3本以上の配列を同時に比較するアルゴリズムは時間がかかり、現実的ではない
- 類似性の高い順に、2本ずつ配列を整列していくのが効率的かつ高精度である



CLUSTALWの基本アルゴリズム

1. 配列データの読み込み
2. ペアワイズの配列間距離の計算
3. 近隣結合法(NJ法、Neighbor Joining)による
ガイド系統樹の作成
4. ガイド系統樹上で近接している順に配列を
2本ずつ整列
5. まだ配列が残っているか？ Yesなら4へ
6. マルチプルアラインメントと進化系統樹の出力

DDBJにおけるCLUSTALW

CLUSTALW WWW System

CLUSTALW
version 1.83

Search

- FASTA [Help](#)
- BLAST [Help](#)
- PSI-BLAST [Help](#)
- SSEARCH [Help](#)
- HMMPFAM [Help](#)

Analysis

- CLUSTALW [Help](#)
- LIBRA! [Help](#)

Utility

- Traffic
- Result Viewer

Others

- What's New
- Statistics
- Documents
 - Q and A
- About
 - References
- Top
- DDBJ Top Page

相溶性検索システムの機能更新(2007年6月11日)
(6/19 18:00-19:00) getentry 保守のため一時サービス休止(2007年6月11日)
BLASTの検索結果の Text View 画面での結果表示不備のお詫び(2007年5月24日)
2007年2月のシステム移行に伴うお知らせ (重要) (2007年3月23日)

DDBJオリジナルオプション

- DOTSINOUTPUT(アラインメントの一致文字をドットで表示)
- DISTANCE(系統樹作成時の塩基置換数推定法)

解析配列データ

※SEQNO_RANGEオプションを使用する場合は、配列名に「/」を付加しないで下さい。

File Upload

選択...

Copy & Paste

解析結果 ※系統樹描画ツールTreeViewダウンロードページへのリンク

WWW E-Mail

入力内容の送信 入力項目の中止

より詳細な設定 DDBJオリジナルオプション

<http://clustalw.ddbj.nig.ac.jp/top-j.html>

データ 入力画面

①配列を入力

②結果の受信方法

③解析実行ボタン

④配列の種類
DNA/Protein

⑤アラインメントの
詳細

⑥系統樹作成の
詳細

⑦ブートストラップ
確率計算

⑧解析実行ボタン

CLUSTALW WWW System

DDBJ DNA Data Bank of Japan

Japanese English

Search

- FASTA 配列 Help
- BLAST 配列 Help
- PSI-BLAST Help
- SSEARCH 配列 Help
- HMMPFAM Help

Analysis

- CLUSTALW Help
- UBRA1 Help

Utility

- Traffic
- Result Viewer

Others

- What's New
- Statistics
- Documents
 - Q and A
- About
 - References
- Top
- DDBJ Top Page

解析配列データ

※SEQNO_RANGEオプションを使用する場合は、配列名に「/」を付加しないで下さい。

File Upload

選択...

Copy & Paste

①

解析結果

※系統樹描画ツールTreeViewダウンロードページのリンク

WWW E-Mail ②

③ 入力内容の送信 入力項目の中止

より詳細な設定 ☐ DDBJオリジナルオプション

TYPE 解析する配列種類の設定

- TYPEがAUTOMATICの場合、プログラムが配列の種類を自動判別して全てデフォルト値で解析を行います。自動判別後のDNA、PROTEINのデフォルト値につきましては、[こちら](#)をご覧ください。
- TYPEを選択し直すと、オプションはデフォルト値で上書きされます。

TYPE

AUTOMATIC

ALIGN 配列のアラインメントのパラメータ変更

DOTSINOUTPUT オプションは、ALIGNの解析のみご利用いただけます。

⑤ ON OFF SHOW ALIGNMENT SCORE ☐

OUTPUT	OUTORDER	MATRIX	SEQNO_RANGE	SHOW SEQNO_RANGE SCORE
clustal	aligned	blosum	OFF START LENGTH	
GAOPEN	GAPEXT	GAPDIST	MAXDIV	ENDGAPS
DEF	DEF	8	40	OFF
QUICKTREE	ON OFF			
KTUPLE	WINDOW	SCORE	TOPDIAGS	PAIRGAP
DEF	DEF	percent	DEF	DEF

⑥ TREE TREEの作成パラメータの変更

ON OFF ※DISTANCEでDDBJオリジナルオプションを使用する場合は、TYPEをDNAに設定して下さい。

DISTANCE	TOSSGAPS	OUTTREETREE
Kimura	ON	phylip

BOOTSTRAP BOOTSTRAP-TREEの作成パラメータの変更

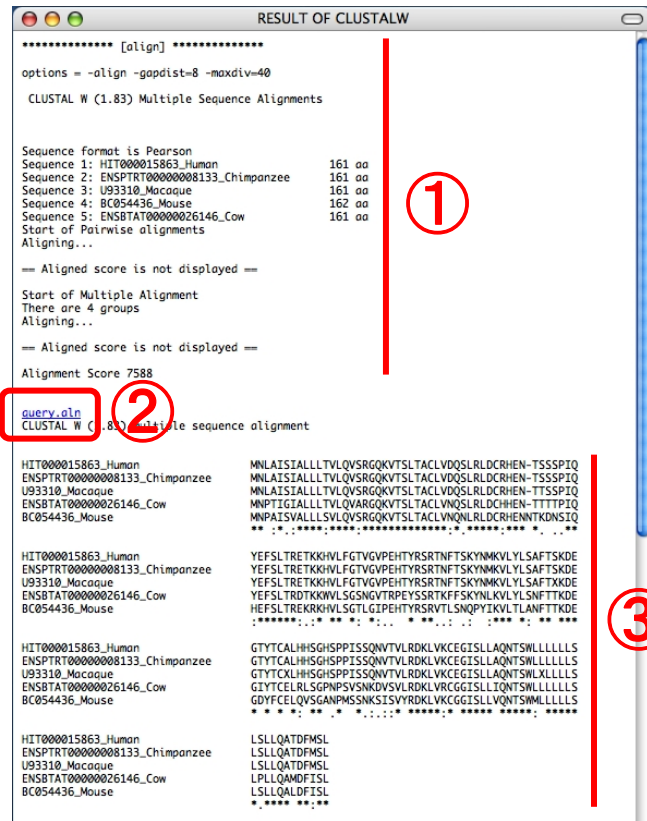
ON OFF ※DISTANCEでDDBJオリジナルオプションを使用する場合は、TYPEをDNAに設定して下さい。

DISTANCE	TOSSGAPS	OUTTREETREE	COUNT	SEED
Kimura	ON	phylip	1000	111

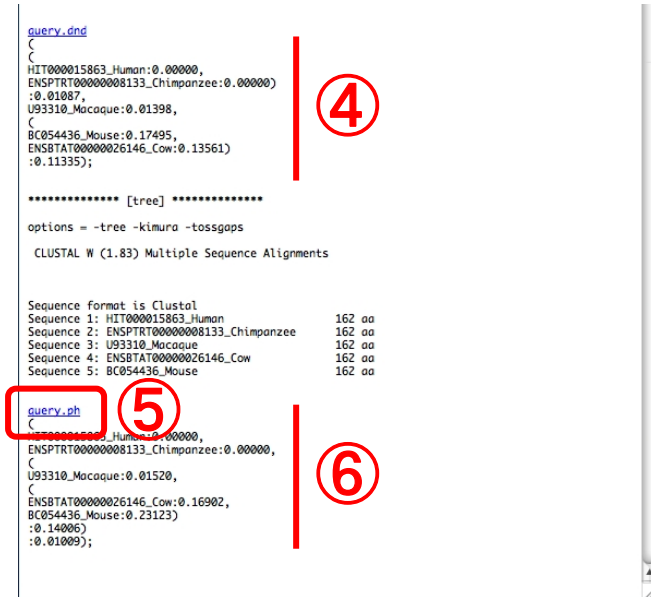
⑦

⑧ 入力内容の送信 入力項目の中止

結果表示画面

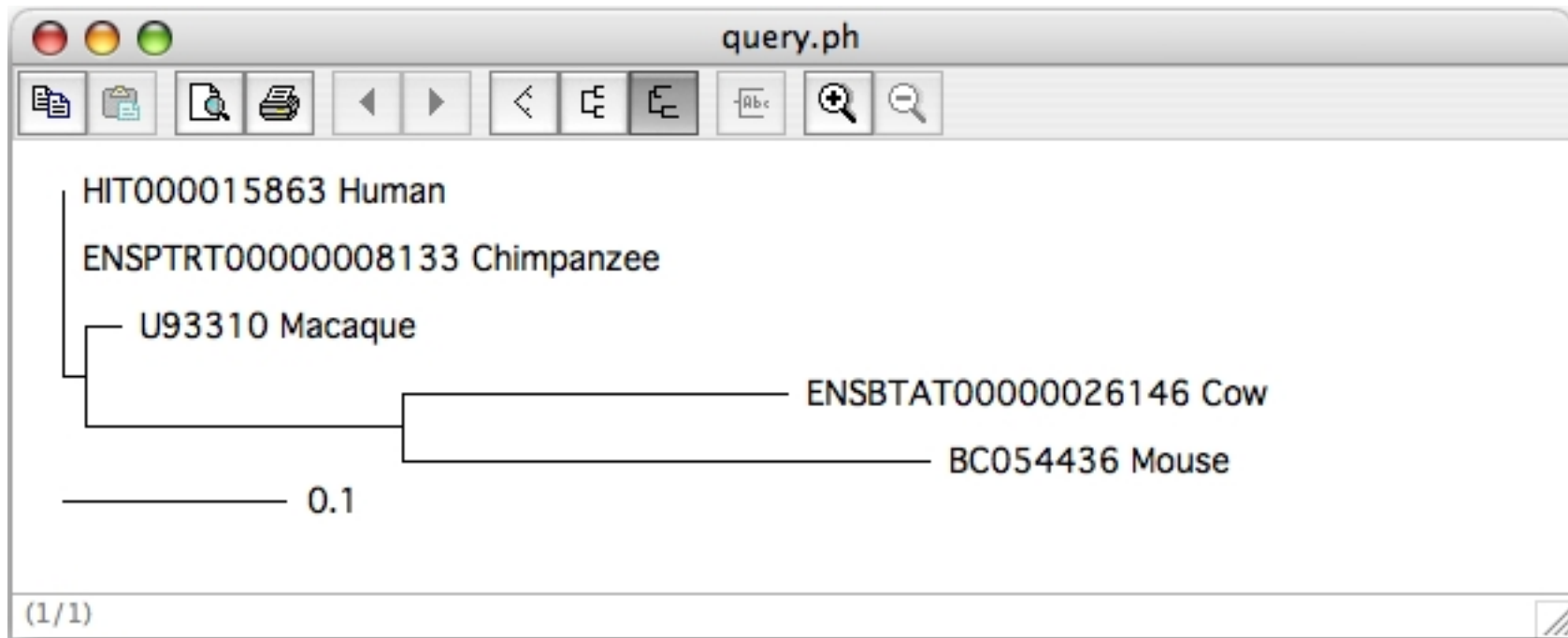


- ①出力メッセージ
- ②結果ダウンロードボタン
- ③アラインメントの表示



- ④ガイド系統樹の表示
- ⑤進化系統樹ダウンロード
- ⑥進化系統樹の表示

TreeViewを用いた分子系統樹表示



Njprotを利用 (<http://pbil.univ-lyon1.fr/software/njplot.html>)

前スライドの⑥に該当する部分をNjprotへ入力する。