分子系統学の基礎

本日の講義資料

kiso3

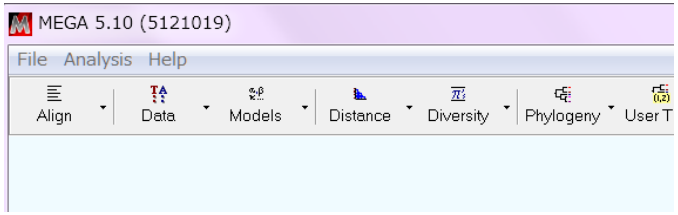
bacteria_16S.fas

bacteria_alaS.fas

bacteria_pgk.fas

本日の講義で使用するWebページへのリンクが載せてあります.

本日の講義では，MEGAを使います.



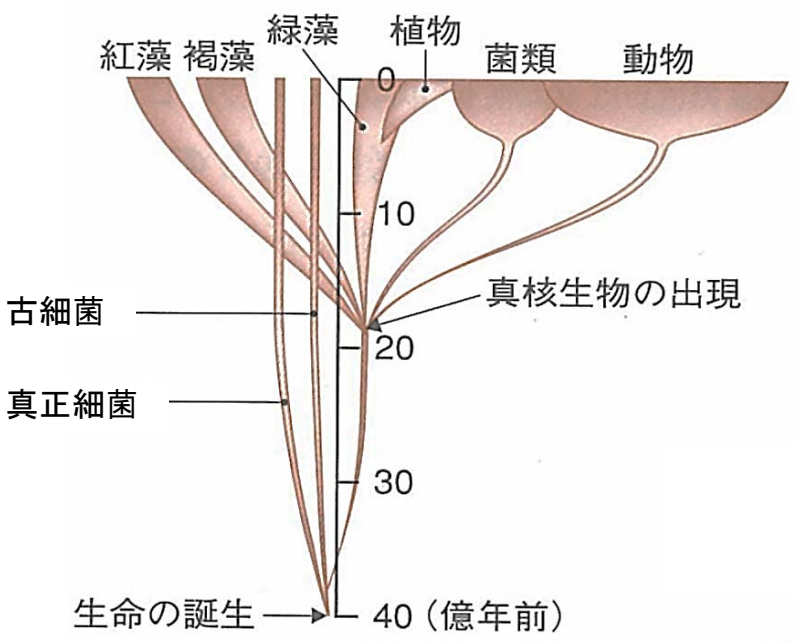
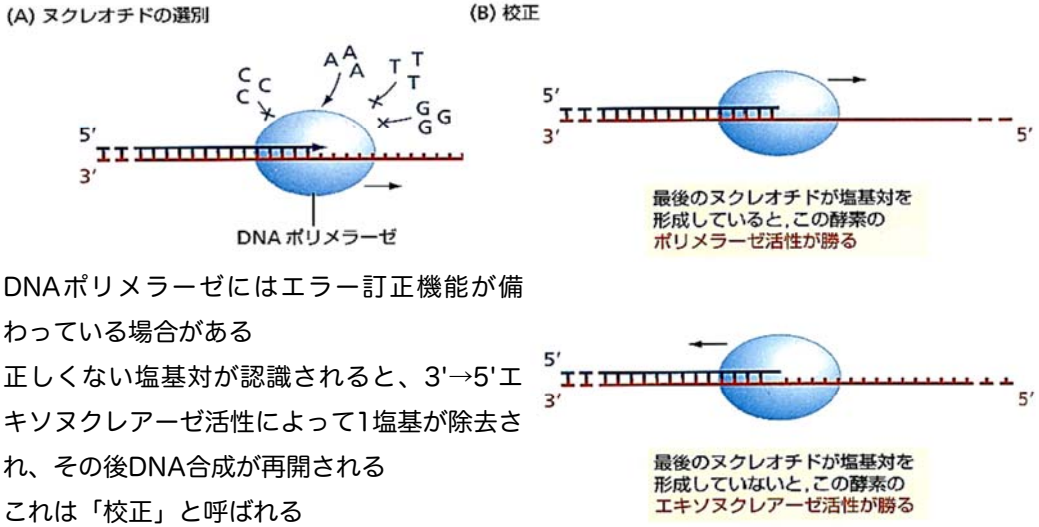


図2 地質年代

DNA修復

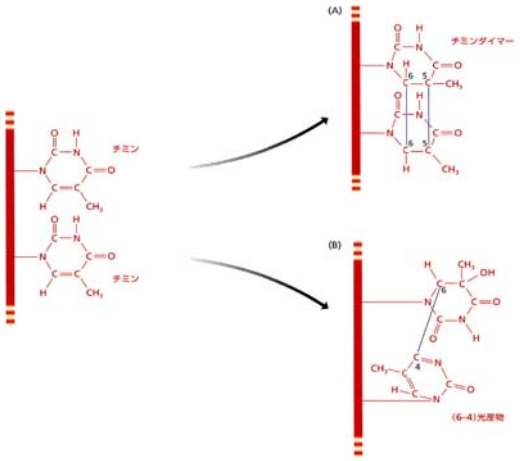
3

DNA分子の損傷は1日1細胞あたり最大50万回程度発生することが知られており、その原因は、正常な代謝活動に伴うもの（DNAポリメラーゼによるDNA複製ミス）と環境要因によるもの（紫外線など）がある



- DNAポリメラーゼにはエラー訂正機能が備わっている場合がある
- 正しくない塩基対が認識されると、3'→5'エキソヌクレアーゼ活性によって1塩基が除去され、その後DNA合成が再開される
- これは「校正」と呼ばれる

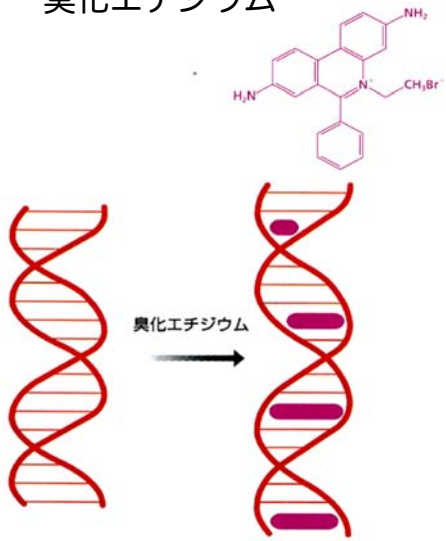
チミン二量体



チミン二量体は、通常生成することはなく、DNA配列の混乱、複製の中断、ギャップの生成、複製のミスを発生させる

臭化エチジウム

4



二本鎖DNAにインターカレートして、DNAの複製や転写を阻害することにより変異原性を示すと考えられている

(A) 直接修復

損傷したヌクレオチド

(B) 除去修復

損傷したヌクレオチド　除去部位　DNA の再合成

(C) 不適正塩基対修復(ミスマッチ修復)

不適正塩基対(ミスマッチ)　除去部位　DNA の再合成

(D) 組み換え修復

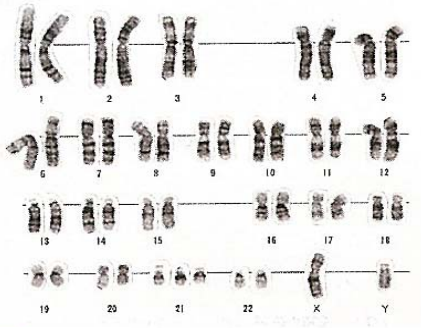
二本鎖切断

- 多くの生物にはDNA修復を行う機構が備わっており、これらをDNA修復系と呼ぶ

突然変異

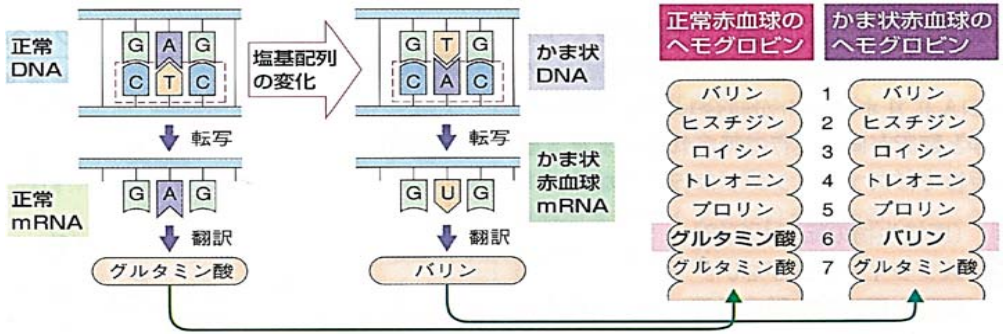
突然変異 { 遺伝子突然変異 { 欠失・逆位・転座・重複 / 染色体突然変異 { 異数性・倍数性

アルビノのカラス

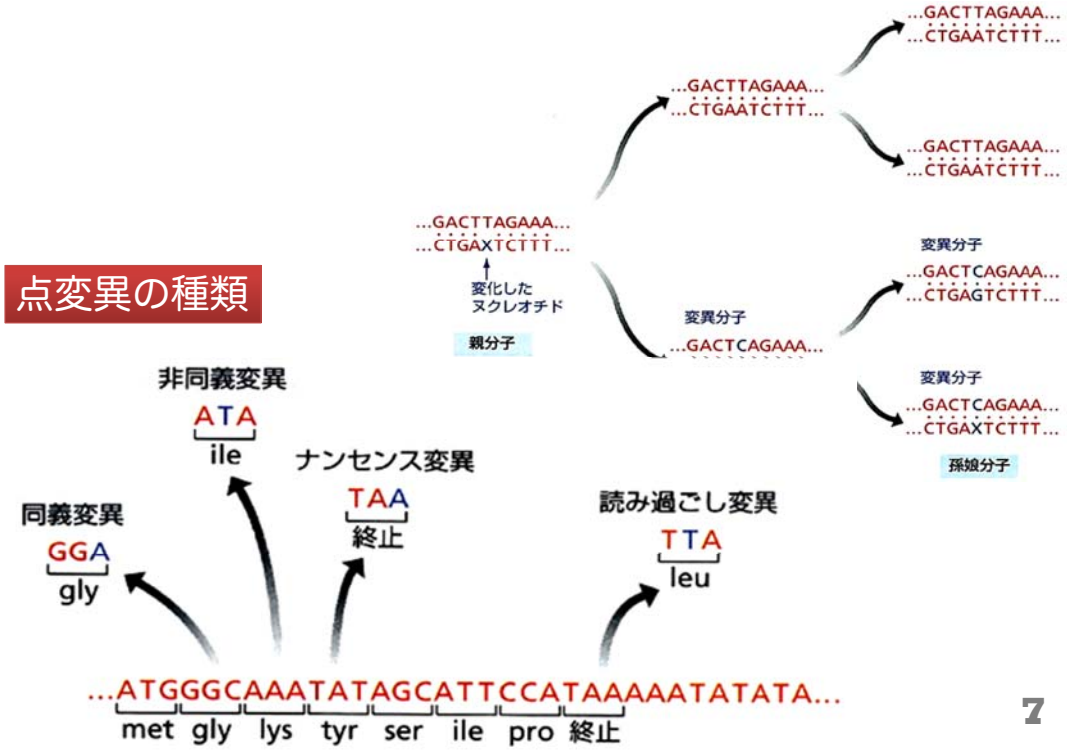


ダウン症の男子の染色体

1つのDNAに生じた突然変異によって鎌状赤血球貧血症になる



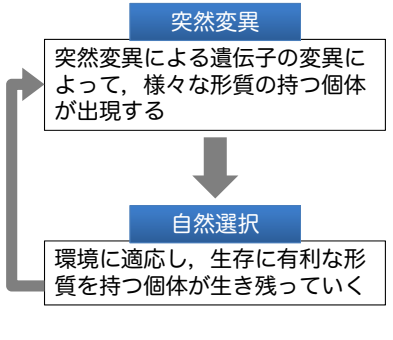
点変異の種類

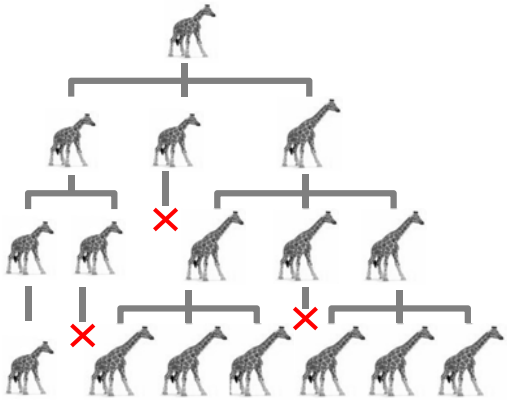


進化の総合説

現在、進化を説明する理論として最も支持されているのは進化の総合説と呼ばれるもので、自然選択説や突然変異説、隔離説、メンデルの遺伝子の理論、集団遺伝学の理論や中立進化説などを統合したものである。

説の名称	提唱者	おもな内容
用不用説	ラマルク	よく使う器官は発達し，使わない器官は退化する。
自然選択説	ダーウィン	自然選択…環境に適した形質が子孫に受け継がれる。
隔離説	ワグナーら	地理的隔離・生殖的隔離により種が分化する。
突然変異説	ド フリース	突然変異が進化の要因。
中立説	木村資生	有利でも不利でもない偶然で決まる。





ヘモグロビンのアミノ酸配列（一部）のアラインメント

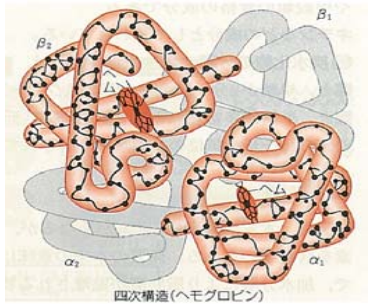
```
                  *         20        *         40        *         60        *         80
human     : MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSAL : 84
monkey    : MVLSPADKSNVKAAWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTLAVGHVDDMPQALSAL : 84
mouse     : MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHGKKVADALANAAGHLDDLPGALSAL : 84
cat       : MVLSAADKSNVKACWGKIGSHAGEYGAEALERTFCSFPTTKTYFPHFDLSHGSAQVKAHGQKVADALTQAVAHMDDLPTAMSAL : 84
dog       : MVLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHGKKVADALTTAVAHLDDLPGALSAL : 84
pig       : MVLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHGQKVADALTKAVGHLDDLPGALSAL : 84
chicken   : MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHFDLSHGSAQIKGHGKKVVAALIEAANHIDDIAGTLSKL : 84
alligator : MVLSMEDKSNVKAIWGKASGHLEEYGAEALERMFCAYPQTKTYFPHFDMSHNSAQIRAHGKKVFSALHEAVNHIDDLPGALCRL : 84
frog      : MHLTADDKKHIKAIWPSVAAHGDKYGGEALHRMFMCAPKTKTYFPDFDFSEHSKHHLAHGKKVSDALNEACNHLDNIAGCLSKL : 84
```

ヘモグロビンのアミノ酸配列の類似度（ 相同性 ）

	ヒト	サル	マウス	ネコ	イヌ	ブタ	ニワトリ	ワニ	カエル
ヒト	100%	97%	92%	93%	90%	90%	79%	79%	71%
サル		100%	92%	92%	90%	92%	80%	80%	70%
マウス			100%	89%	88%	90%	81%	81%	69%
ネコ				100%	92%	90%	80%	81%	71%
イヌ					100%	88%	78%	79%	69%
ブタ						100%	78%	79%	71%
ニワトリ							100%	82%	70%
ワニ								100%	72%
カエル									100%

- 近縁 な生物同士はDNAやアミノ酸配列が似ている
- 遠縁になるほどDNAやアミノ酸配列の相同性は 低く なる

中立変異




```
ゴリラ      CCGGCGGCTGGCTAGGGATGAAGAATAAAAGGAAGCACCCTTCCAGCAGTTCCACACACTC
チンパンジー CCGGCGGCTGGCTAGGGATGAAGAATAAAAGGAAGCACCCTTCCAGCAGTTCCACACACTC
ヒト        CCGGCGGCTGGCTAGGGATGAAGAATAAAAGGAAGCACCCTTCAGCAGTTCCACACACTC
            ********************************************|*********************

ゴリラ      GCTTCTGGAACGTCTGAGGTTATCAATAAGCTCCTAGTGCAGACGCCATGGGTCATTTCA
チンパンジー GCTTCTGGAACGTCTGAGGTTATCAATAAGCTCCTAGTCCAGACGCCATGGGTCATTTCA
ヒト        GCTTCTGGAACGTCTGAGGTTATCAATAAGCTCCTAGTCCAGACGCCATGGGTCATTTCA
            ************************************************************

ゴリラ      CAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATGCTG
チンパンジー CAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATGCTG
ヒト        CAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATGCTG
            ************************************************************

ゴリラ      GAGGAGAAACCCTGGGAAGGTAGGCTCTGGTGACCAGGACAAGGGAGGGAAGGAAGGACC
チンパンジー GAGGAGAAACCCTGGGAAGGTAGGCTCTGGTGACCAGGACAAGGGAGGGAAGGAAGGACC
ヒト        GAGGAGAAACCCTGGGAAGGTAGGCTCTGGTGACCAGGACAAGGGAGGGAAGGAAGGACC
            ************************************************************

ゴリラ      CTGTGCCTGGCAAAAGTCCAGGTCGCTTCTCAGGATTTGTGGCACCTTCT
チンパンジー CTGTGCCTGGCAAAAGTCCAGGTCACTTCTCAGGATTTGTGGCACCTTCT
ヒト        CTGTGCCTGGCAAAAGTCCAGGTTGCTTCTCAGGATTTGTGGCACCTTCT
            ***********************||*********************
```

ヘモグロビン遺伝子の比較

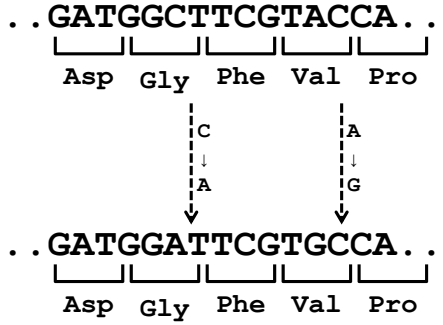
生物に対する影響の度合いで突然変異を分類するのなら，それは，

① **生物に有利な変異**
② **生物にとって不利な変異**
③ **生物にとって有利でも不利でもない変異**

の三つにわけられる。突然変異がある塩基で起こるのは，偶然によるものであるから，その突然変異が生物にとって有利である可能性は，非常に少ない。そして，生物に対して有利でも不利でもない変異は，**中立変異**（ちゅうりつへんい）と呼ばれる。

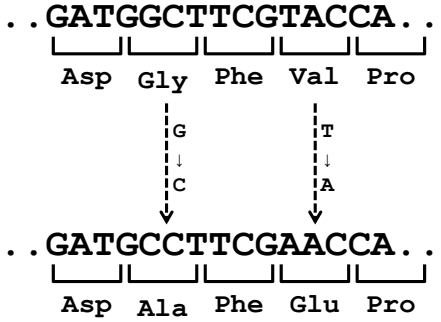
分子進化の中立説

分子レベルの進化学において，重要な理論の一つに木村資生による「**分子進化の中立説**」（1968年）がある。木村は分子レベルにおける進化的変化は自然選択に中立である，あるいはほとんどが中立な突然変異遺伝子の偶然的な浮動（ふどう）で起こると述べた。この説は，それ以前には，生物の進化をもっぱら表現型レベルにおける自然選択で説明していたのに対して，大きな変革を迫るものであった。



2ヵ所の突然変異が生じたが，アミノ酸配列は変わらない → 中立変異

このような変異は害がないので，次世代に伝わりやすい



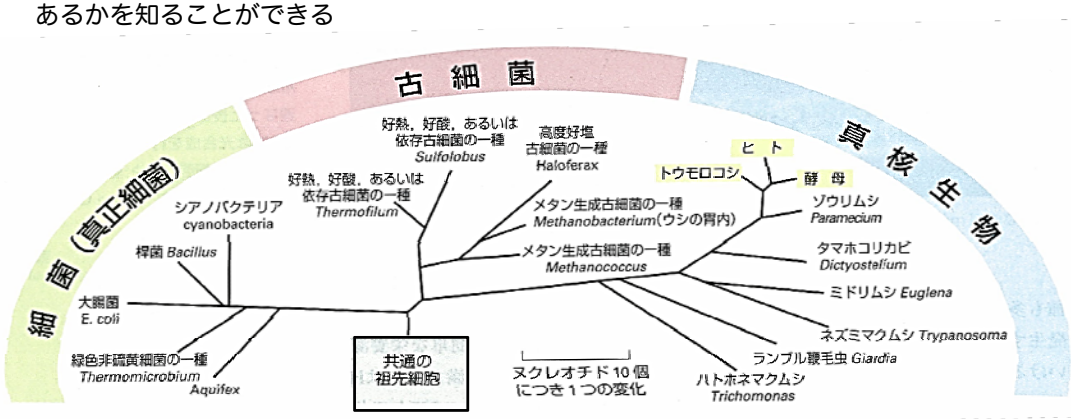
突然変異によって，アミノ酸配列が変化した

このような変異は，多くの場合，有害であるので，次世代に伝わりにくい

分子系統樹

DNAに蓄積する変異は一定の割合で起こっており、そのほとんどが自然選択とは無関係な中立の変異である

DNAの配列がどのくらい似ているかを調べることによって、進化的にどの程度近縁であるかを知ることができる



分子時計　アミノ酸の変化した数と，化石から知られる2つの系統の生物が分かれた時期とから，アミノ酸が1個変わるのにどれだけ時間がかかるかを計算できる。

系統樹の作成法

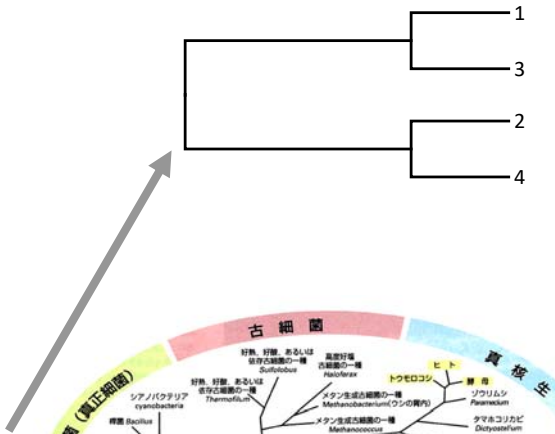
- DNA配列のアラインメント(並行配列)から，系統樹の推定に必要な比較データを得る。
- 比較データを変換して，系統樹を推定する。
- 系統樹の信頼度を検定する。

多重アラインメント

```
1 AGGCCAAGCCATAGCTGTCC
2 AGGCAAAGACATACCTGACC
3 AGGCCAAGACATAGCTGTCC
4 AGGCAAAGACATACCTGTCC
```

距離行列

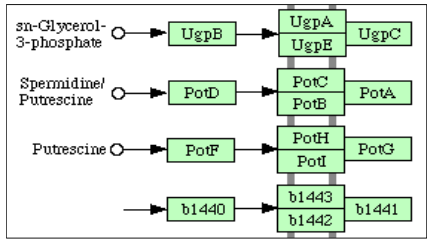
	1	2	3	4
1	–	0.20	0.05	0.15
2		–	0.15	0.05
3			–	0.10
4				–



オーソログ遺伝子とは？

・オーソログ（Ortholog）
種分岐の際に同じ遺伝子だったもの．通常同じ機能を持つ．

・パラログ（Paralog）
遺伝子重複によってできた類似遺伝子
（例えば一つの生物種が持っている様々なトランスポーター遺伝子など）



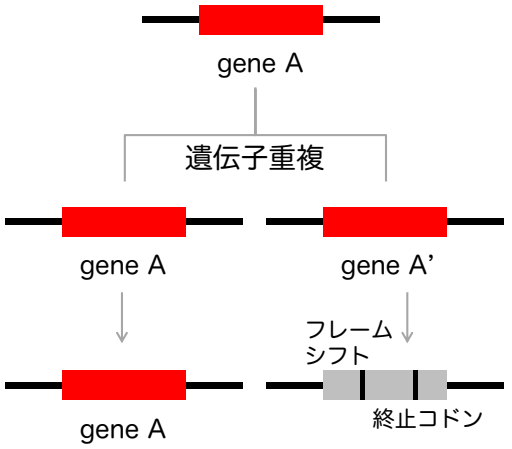
http://www.genome.jp/tools/blast/

「tmk_eco」
あるいは
「znuC_eco」
をコピー＆ペースト

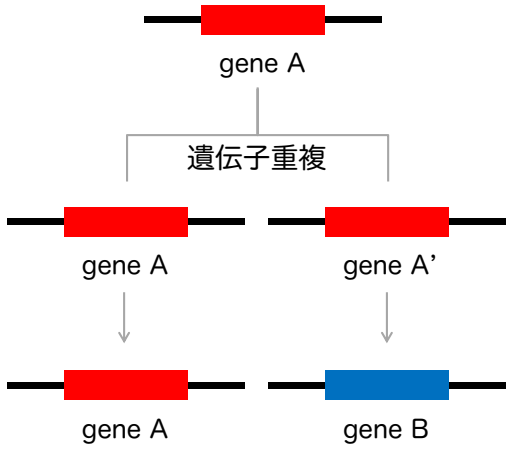
「Favorite」
を選択

「eco bsu hpy sau pae」
と入力

遺伝子重複

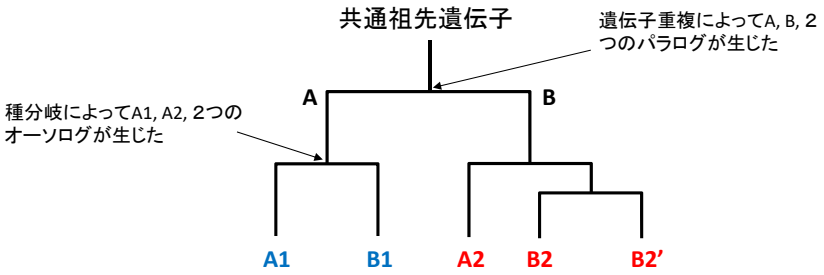


片方の遺伝子が機能を失っても生存には影響しないので、通常は一方が偽遺伝子になっていく

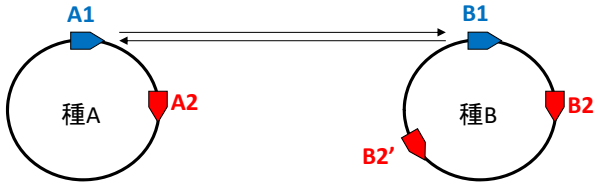


ごく稀に、一方の遺伝子が新たな機能を獲得し、元とは異なる遺伝子へと進化する

オーソログ遺伝子とは？



2つのゲノム間のオーソログは，双方向ベストヒットを調べることで同定できる．

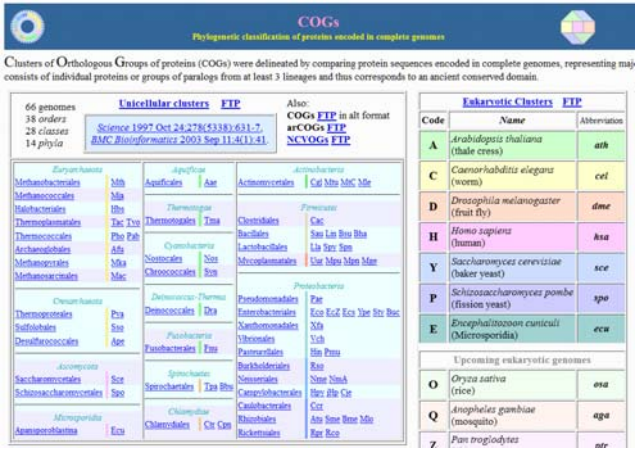


COGデータベース

http://www.ncbi.nlm.nih.gov/COG/

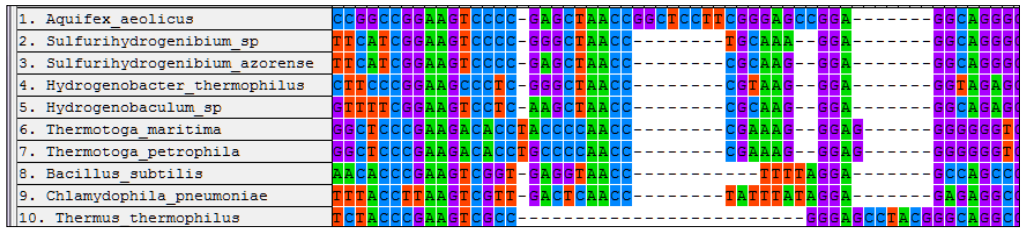
Clusters of orthologous groups of proteins

ゲノムが解読された生物について、アミノ酸配列のオーソログ関係を双方向ベストヒットによって整理したデータベース．



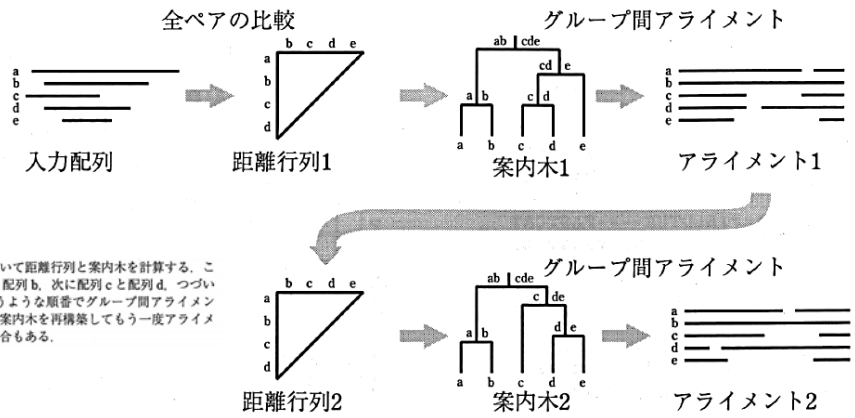
多重アラインメント

- 配列中の残基同士を対応づけたものがアラインメント
- アラインメント中で挿入や欠失により対応させることのできない残基には空白文字を対応させ、連続した空白文字のことをギャップと呼ぶ
- ２配列間でのアラインメントをペアワイズアラインメント、３配列以上でのアラインメントをマルチプルアラインメントという
- 主要なマルチプルアラインメントアルゴリズムとして、「累進法」と「反復改善法」が挙げられる



累進法

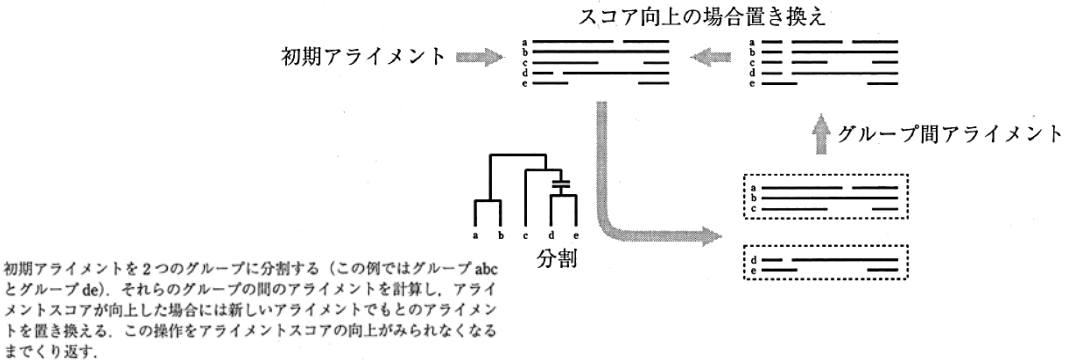
- ペアワイズアラインメントからはじめて、徐々にアラインメントを組み上げて最終的なマルチプルアラインメントを得る方法
- 高速に計算可能であるものの、途中段階のアラインメントに生じたエラーを取り除くことが出来ないため、最終的な精度は反復改善法に比べて低下することが多い
- 累進法の代表的なアルゴリズム：ClustalW、T-Coffee など



入力配列の間の全通りの比較に基づいて距離行列と案内木を計算する．この案内木にしたがってまず配列 a と配列 b，次に配列 c と配列 d，つづいて配列グループ cd と配列 e，というような順番でグループ間アライメントをくり返す．下段に示すように，案内木を再構築してもう一度アライメントをくり返すと精度が向上する場合もある．

反復改善法

- 累進法などで得られたアラインメントを２つに分割し、それらのアラインメントから再び１つのアラインメントを計算する、ということを繰り返す方法
- アラインメント中のエラーを取り除くことが可能
- 系統樹の枝を切断する形でアラインメントを２つに分割する
- 累進法に比べて計算量が多いものの、高精度なアラインメントを得ることができる
- 代表的なアルゴリズム： Prrn、MAFFT、MUSCLE　など



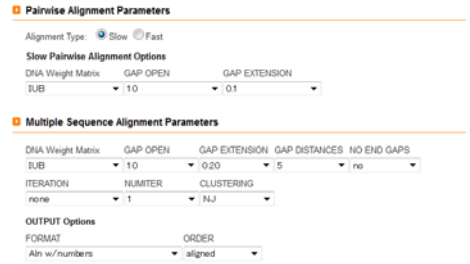
初期アライメントを２つのグループに分割する（この例ではグループ abc とグループ de）．それらのグループの間のアライメントを計算し，アライメントスコアが向上した場合には新しいアライメントでもとのアライメントを置き換える．この操作をアライメントスコアの向上がみられなくなるまでくり返す．

ClustalW

http://clustalw.ddbj.nig.ac.jp/index.php?lang=ja

「bacteria_16S.fas」の中身をコピー＆ペースト

設定確認してクリック



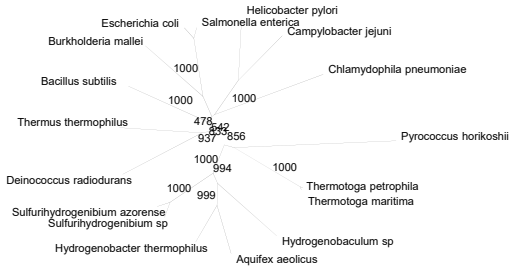
マルチプルアラインメントの結果が表示される

```
Thermotoga_maritima             CCTAACACATGCAAGTCGAGCGGGGGAA--------ACTCCCTTCGGGGA 88
Thermotoga_petrophila           CCTAACACATGCAAGTCGAGCGGGGGAA--------ACTCCCTTCGGGGA 89
Aquifex_aeolicus                CCTAACACATGCAAGTCGTGCGCAGGGTCGG------CCCCTTTTGGGGC 92
Hydrogenobacter_thermophilus    CCTAACACATGCAAGTCGTGCG--GGGT-GG------CTCT--------- 80
Hydrogenobaculum_sp             CCTAACACATGCAAGTCGTACGGAGAGTGGGGCA--ACTCA--------- 78
Sulfurihydrogenibium_sp         CCTAACACATGCAAGTCGTG-GGGCAGCAGGCTACTACCTTCGGGTAGTA 88
Sulfurihydrogenibium_azorense   CCTAACACATGCAAGTCGTG-GGGCAGCAGGCTATTACCTTCGGGTAATA 97
Escherichia_coli                CCTAACACATGCAAGTCGAACGGTAACAGGAAG-AAGCTTGCTTCTTT-- 93
Salmonella_enterica             CCTAACACATGCAAGTCGAACGGTAACAGGAAG-CAGCTTGCTGCTTC-- 83
Burkholderia_mallei             CCTTACACATGCAAGTCGAACGGCAGCACGG-----GCTT-CGGCCT--- 62
Helicobacter_pylori             CCTAATACATGCAAGTCGAACG--ATGAAGCTTCTAGCTTGCTAGAGT-- 92
Campylobacter_jejuni            CCTAATACATGCAAGTCGAACG--ATGAAGCTTTTAGCTTGCTAGAA--- 93
Chlamydophila_pneumoniae        GATGAGGCATGCAAGTCGAACGGAATAATGACTTCGG-TTGTTAT----- 86
Bacillus_subtilis               CCTAATACATGCAAGTCGAGCGGACAGGTGGG---AGCTTGCTCCCT--- 92
Deinococcus_radiodurans         CTTAAGACATGCAAGTCGAACG-----------CGGTCTTCGGAC---- 80
Thermus_thermophilus            CCTAAGACATGCAAGTCGTGCGGGCCGCGGGGT----TTTACTCCGT--- 90
Pyrococcus_horikoshii           ACTAAGCCATGCGAGTCAAGGGGGCGTC---------CCTTCTGGGAC-- 81
```

Download Bootstrapped Tree File

Tree Fileをダウンロードして TreeViewで開けば、系統樹を表示させることもできる

```
(
(
(
(
(
(
(
Thermotoga_maritima:0.00318,
Thermotoga_petrophila:0.00387)
1000:0.08738,
Pyrococcus_horikoshii:0.18451)
856:0.01173,
(
(
(
```



16S rRNA　の塩基配列による進化系統解析

1987 年，米国イリノイ大学のカール・ウース（C. R. Woese）はタンパク質合成にかかわるリボソームの小サブユニットを構成する RNA，すなわち 16S rRNA の塩基配列を用いて原核生物の系統分類を行った．この研究により原核生物は真正細菌（Bacteria）と古細菌（Archaea → Chapter 9）の二つに分かれることを示した．

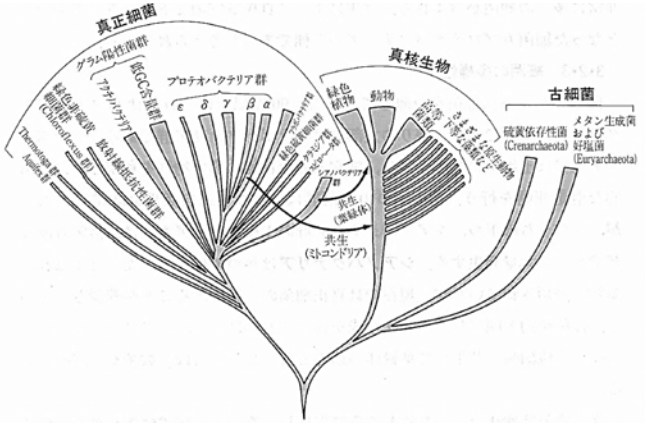


図 3･5　リボソーム RNA の構造に基づいて推定された生物の進化

リボソーム

iv） リボソーム　細菌は細胞内に 70 S のリボソームを持っており，これは低濃度の Mg^{2+} 緩衝液中で 50 S と 30 S のサブユニットに可逆的に解離する．50 S サブユニットには 23 S と 5 S の RNA，および 35 種類の蛋白がそれぞれ一分子ずつ含まれており，また 30 S サブユニットには 16 S RNA と 21 種の蛋白がそれぞれ一分子ずつ含まれている（図 21）．

リボソームの役割は菌体が生存と増殖のために必要とする数千種類の蛋白質を合成することである．

- 70 S リボソーム
  - 50 S サブユニット
    - 23 S RNA（MW：1.1×10^6）1 分子
    - 5 S RNA（MW：4.0×10^4）1 分子
    - 35 種類の蛋白質[a]（MW：10,000～40,000）各 1 分子
  - 30 S サブユニット
    - 16 S RNA（MW：6×10^5）1 分子
    - 21 種類の蛋白質[b]（MW：10,000～40,000）各 1 分子

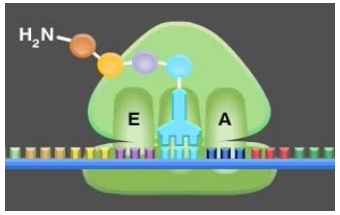


図 21　細菌リボソームの構成成分
a）と b）の間に共通なものはない．

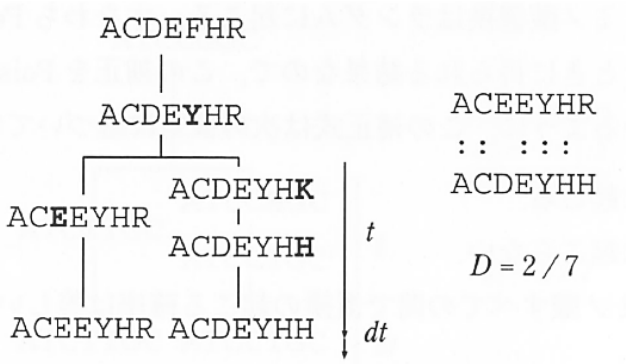
進化距離の計算

- 共通の祖先に由来する２つの配列において、それらが分岐してから現在までに蓄積した置換の数を置換数といい、置換数を座位数で割ったものを進化距離という
- 進化距離の推定値は、距離行列法による系統樹の計算などに用いられる
- 進化距離は、基本的に２つの配列の類似性に基づいて推定されるが、そのためのいろいろな方法が存在する

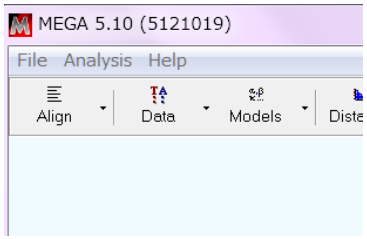
- 類似性の指標としてもっとも単純なものは相違度である
- 相違度Dは、２本の配列の間で一致しない座位の数を、比較した座位の数で割ったものである

$$D = \frac{\text{一致しない座位数}}{\text{比較する座位数}}$$

- 実際には、「相違度」と「進化距離」とは比例関係にはない
- 相違度を見ただけでは、
  ・多重置換：同一の座位に複数回の置換が起こること
  ・復帰置換：祖先型に戻るような置換
  を無視してしまうからである

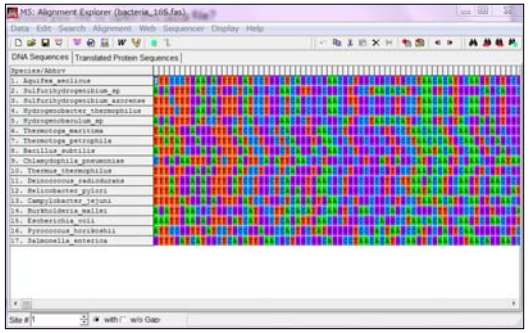


MEGAなどの系統樹作成ソフトを使用して計算する



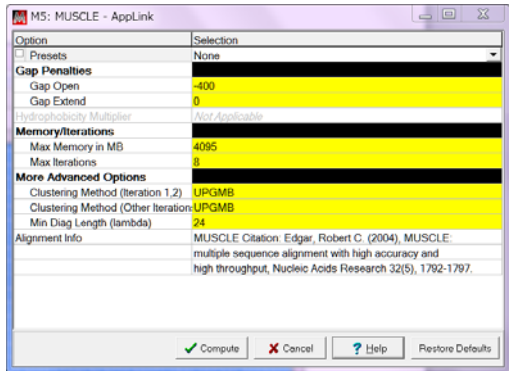
ファイルを開く

- メニュー File → Open A File
- 「bacteria_16S.fas」を選ぶ
- How would you like to open this fasta file?　と聞かれるので
- Align　を選択



アラインメントを作成

- Alignment Explorerが開く
- メニュー Alignment
- Align by Muscle
- Select all? と聞かれるので
- OK を押す
- Muscleの設定ウインドウが開く
- Compute
- アラインメントの結果が表示される



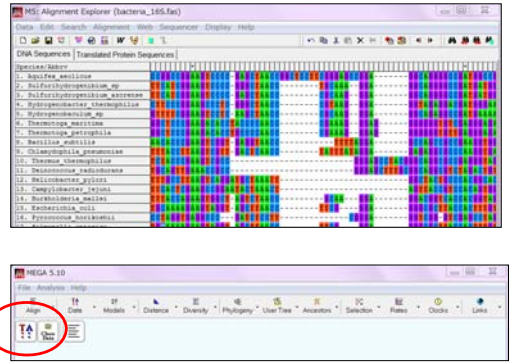
MEGA形式で保存する

- メニュー Data
- Export Alignment
- MEGA format
- 保存
- Title：入力しなくても大丈夫
- Protein-coding nucleotide data? と聞かれるので　No を選択

メインウィンドウに移行する

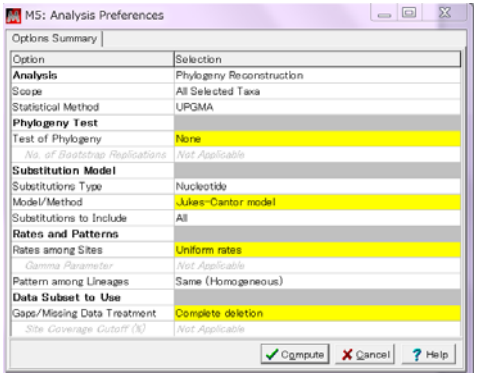
- メニュー Data
- Phylogenetic analysis
- Protein-coding nucleotide data? と聞かれるので　No を選択
- メインウィンドウに戻る

これが表示されるようになる



系統樹を作成

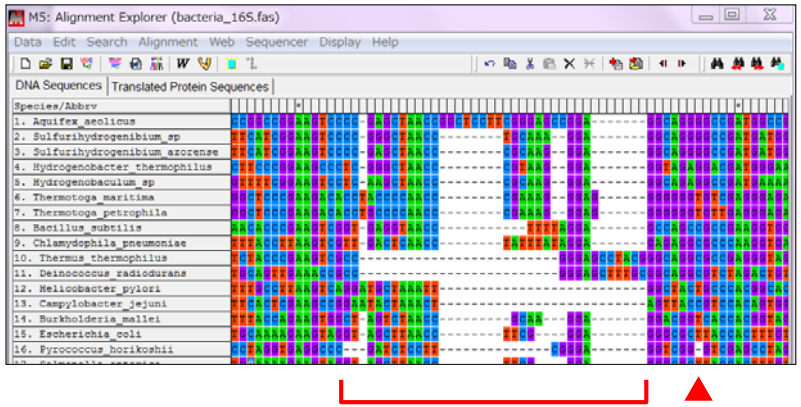
- Phylogeny
- Construct UPGMA Tree
- 設定画面が表示される



- Test of Phylogeny : None
- Model/Method : Jukes-Cantor
- Rates among Sites : Uniform rates
- Gaps/Missing Data Treatment : Complete deletion

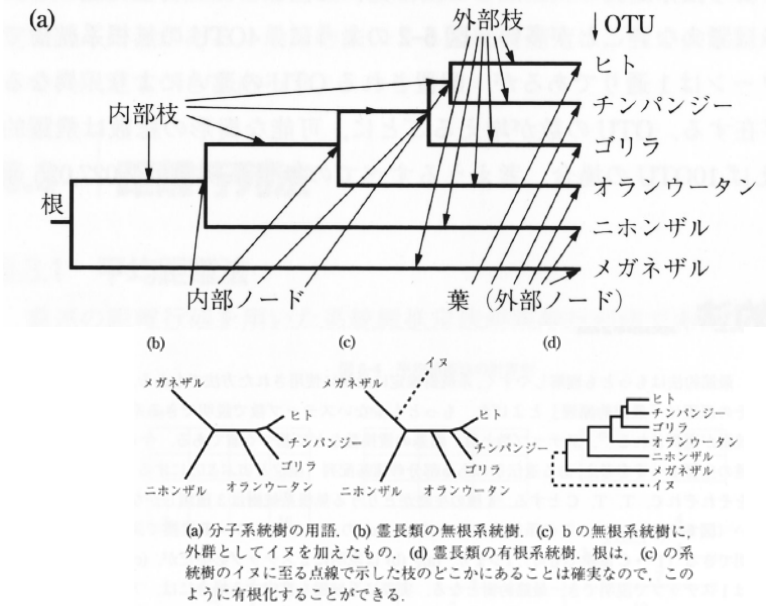
- Compute

- Gaps/Missing Data Treatment : Complete deletion



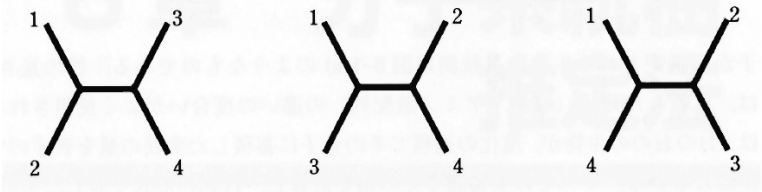
◆通常は、ギャップを含むサイトを無視して系統樹を作成する

クラスター分析 cluster analysis
2つ以上の対象を，それらの間の類似度あるいは非類似度を手がかりにして似たものを集め，いくつかのグループ（クラスター）に分類する方法



(a) 分子系統樹の用語．(b) 霊長類の無根系統樹．(c) b の無根系統樹に，外群としてイヌを加えたもの．(d) 霊長類の有根系統樹．根は，(c) の系統樹のイヌに至る点線で示した枝のどこかにあることは確実なので，このように有根化することができる．

図 6–2　4OTU の場合にとりうる 3 つの無根系統樹



10 OUTの場合は、考えうるすべての無根系統樹は2,027,025通り存在する

系統樹を作成するためのクラスタリング法には

- UPGMA
- 近隣結合法
- 最尤法
- ベイズ法

など、様々な方法がある

UPGMA （平均距離法、非加重結合法）

Sokal & Michener（1985）

Unweighted Pair Group Method with Arithmetic mean

- 距離の算術平均の小さなものから結合することにより得られる樹形を選ぶ、段階的探索法の一種
- 進化速度の一定性が仮定されるため，有根系統樹が得られる
- 一番簡単な方法で計算も容易であるが，進化速度一定の仮定が必要であるため，進化速度が系統間で異なるときは誤った推定を行いやすい.

平均距離法の計算例

(a)

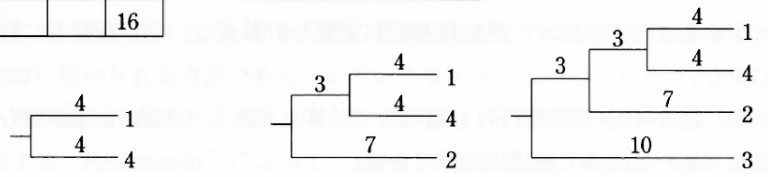
	2	3	4
1	12	16	8
2		24	16
3			16

(b)

	2	3
(1, 4)	14	16
2		24

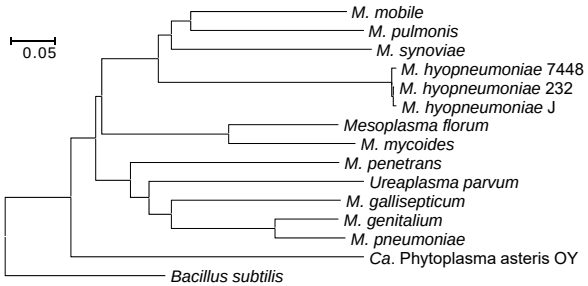
(c)

	3
((1, 4), 2)	20

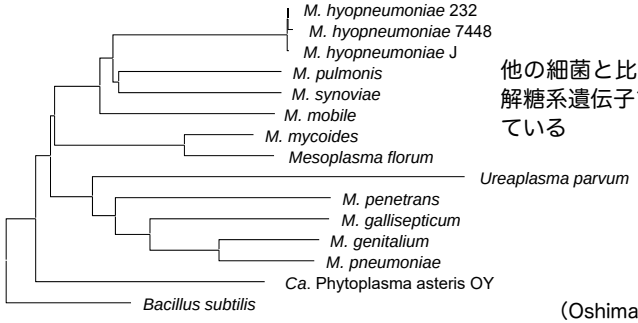


進化速度が系統間で異なる例

◆UvrAのアミノ酸配列を用いた系統樹

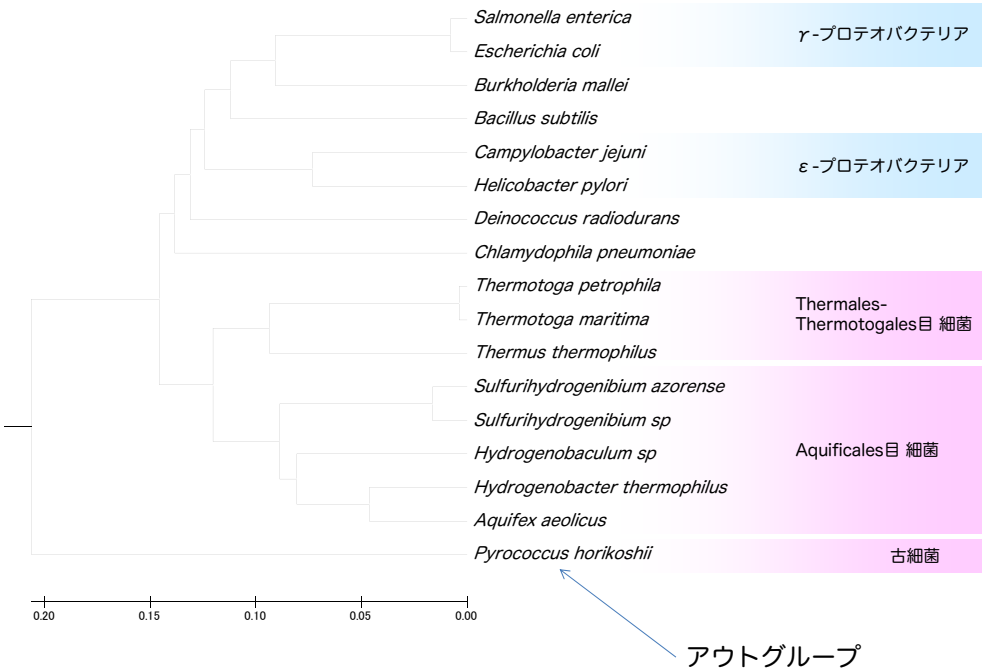


◆Enolaseのアミノ酸配列を用いた系統樹



他の細菌と比較して *Ureaplasma parvum* の解糖系遺伝子では高い置換頻度で変異が蓄積している

(Oshima & Nishida, *J. Mol. Evol.*, 2008)



原始生命は熱水中で生まれた？

イエローストーン国立公園の間欠泉

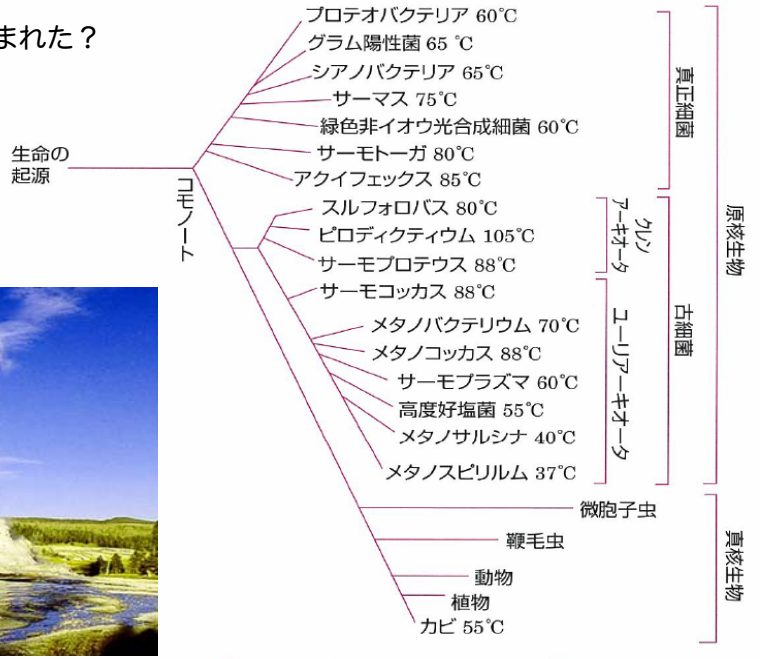
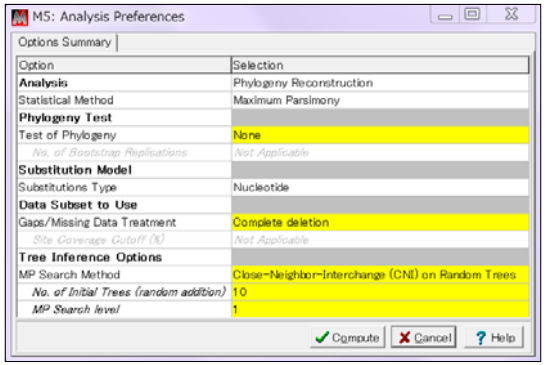


図 9.1　全生物界の分子系統樹

最節約法による系統樹を作成

- Phylogeny
- Construct Maximum parsimony
- 設定画面が表示される

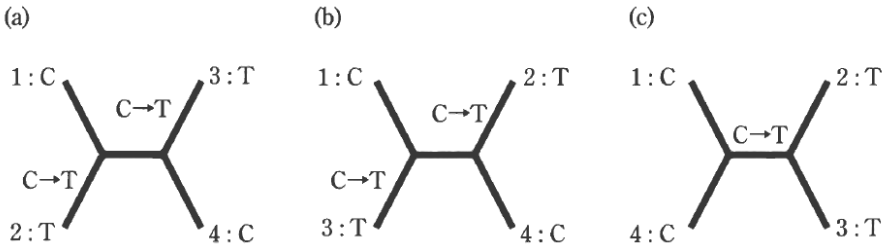


- Test of Phylogeny : None
- Gaps/Missing Data Treatment : Complete deletion

- Compute

最節約法（最大節約法、Maximum parsimony）

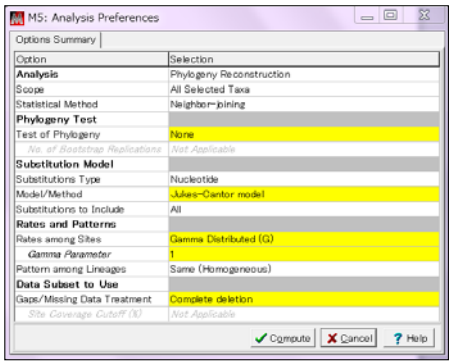
- 考えられるすべての系統樹の中から、形質の変化の数が最も少ないものを選択する方法
- 配列の座位毎に必要な変異の数を数え、 総変異数が最小の系統樹を採用する



- 最節約法では、同じ配列の同じ部位に２回以上の置換が起きると（多重置換）、推定を誤ることがある
- 最節約法による系統樹推定は、互いの配列が比較的類似したものに限定すべき

近隣結合法による系統樹を作成

- Phylogeny
- Construct Neighbor-Joining Tree
- 設定画面が表示される



- Test of Phylogeny : None
- Model/Method : Jukes-Cantor
- Rates among Sites : Gamma Distributed
- Gamma Parameter : 1
- Gaps/Missing Data Treatment : Complete deletion

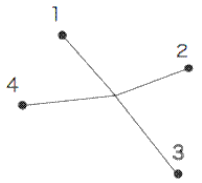
- Compute

近隣結合法　（neighbor-joining method、略してNJ法）

Saitou N. and Nei M. (1987)

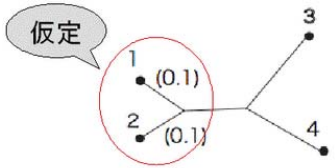
- 系統樹を段階的に構成するアルゴリズムのクラスタ解析法
- アルゴリズムの各段階で「全ての枝の長さの合計が最小」となるようなトポロジーが望ましいという基準に基づいている
- 最終的に全枝長を最小にするトポロジーが得られるとは限らない（が、多くの場合、最適なものに非常に近い系統樹が得られることが分かっている）

配列	1	2	3	4
1				
2	0.2			
3	0.1	0.3		
4	0.4	0.4	0.5	



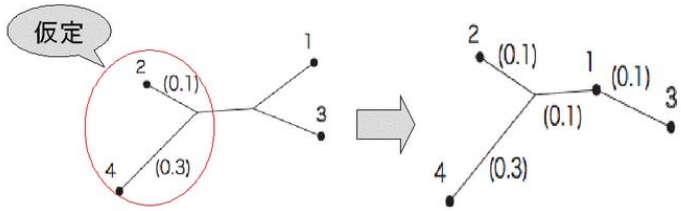
✓ まず、放射状の系統樹を仮定します

✓ 次に特定の配列が、二本の枝でつながる近隣関係にあると仮定して、系統樹全体に必要な枝の長さを計算します



配列１と２が近隣関係にあると仮定していますが、これでは配列１と３、配列２と３の距離関係をうまく表すことができません

✓ そこで今度は配列２と４が近隣関係にあると仮定してみると、全ての距離関係をうまく表すことができます



- 計算効率が高く、ほかの系統解析法（最大節約法、最尤法、ベイズ法など）では計算能力的に不可能なほどの大量のデータセットも扱うことが可能である
- UPGMAと異なり、近隣結合法はすべての系統が同じ速度で進化する（分子時計の仮説）ことを仮定せずに無根系統樹を作ることができる

JC69

JukesとCantorが1969年に提案したモデル。どの塩基サイト(A,G,C,T)でも同じの変化率αで互いに変化すると仮定している。

	A	G	C	T
A	−	α	α	α
G	α	−	α	α
C	α	α	−	α
T	α	α	α	−

K80

Kimura が1980年に提案したモデル。サイト当たりの変化率を次のように仮定している。

	A	G	C	T
A	−	α	β	β
G	α	−	β	β
C	β	β	−	α
T	β	β	α	−

トランジションとトランスバージョン

塩基置換には、プリン間 (A ⇔ G) やピリミジン間 (C ⇔ T) の置換である転位 (トランジション) と，それ以外の 置換である転換 (トランスバージョン) とがある．一般に，トランジションの方が起こりやすい．

T92

Tamuraが1992年にK80モデルを一般化したモデル。塩基頻度(GとCの頻度、あるいは含量) $\theta = \pi_G + \pi_C$ を用いる。

	A	G	C	T
A	−	$\alpha\theta$	$\beta\theta$	$\beta(1-\theta)$
G	$\alpha(1-\theta)$	−	$\beta\theta$	$\beta(1-\theta)$
C	$\beta(1-\theta)$	$\beta\theta$	−	$\alpha(1-\theta)$
T	$\beta(1-\theta)$	$\beta\theta$	$\alpha\theta$	−

TN93

Tamura と Neiが1993年に提案したモデル。A⇔C,C⇔Tが異なる比率 α_R, α_Y で置換すると仮定している。

	A	G	C	T
A	−	$\pi_G\omega$	$\beta\pi_C$	$\beta\pi_T$
G	$\pi_A\omega$	−	$\beta\pi_C$	$\beta\pi_T$
C	$\beta\pi_A$	$\beta\pi_G$	−	$\pi_T\psi$
T	$\beta\pi_A$	$\beta\pi_G$	$\pi_C\psi$	−

$\omega = \alpha_R/\pi_R + \beta$

$\varphi = \alpha_Y/\pi_Y + \beta$

$\alpha_R/\alpha_Y = \pi_R/\pi_Y$ であるとHasegawa, Kisino and Yanoが1985年に提案したモデル（HKY）に等しい。

一般時間反転可能モデル(general time-reversible : GTR)

From ＼ To	A	C	G	T
A	-	$Rate_{AC}Freq_C$	$Rate_{AG}Freq_G$	$Rate_{AT}Freq_T$
C	$Rate_{AC}Freq_A$	-	$Rate_{CG}Freq_G$	$Rate_{CT}Freq_T$
G	$Rate_{AG}Freq_A$	$Rate_{CG}Freq_C$	-	$Rate_{GT}Freq_T$
T	$Rate_{AT}Freq_A$	$Rate_{CT}Freq_C$	$Rate_{GT}Freq_G$	-

RateXY は塩基*Y*から塩基*X*への移行速度、*FreqX* は塩基*X*の頻度
ただし、*RateXY* = *RateYX* とする (これを「時間反転可能」[time-reversible]と言う)

(Tavar´e, 1986, Posada and Crandall, 1998)

Bootstrap

- 作成した系統樹の信頼性を評価する方法
- 系統樹の作成に用いたアミノ酸配列を大量に複製(リサンプリング)し、それぞれのリサンプルデータから推定される系統樹が元データの系統樹を支持する確率を求める

元のデータ

```
data1:MNDRQAALDQALKQIEKQFG
data2:MACGGEKKTEANPETYPDKP
data3:MSENNQSNQNNQILKIIKST
data4:MTAEKSKALAAALAQIEKQF
```

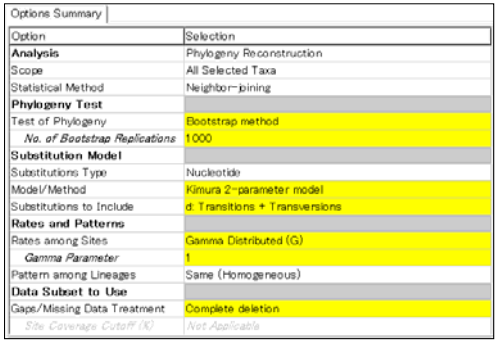
リサンプルデータ

```
data1:MDAIMKAGDADGKRQFQQAA
data2:MTETMPKPSKCPPGEKEDAK
data3:MQQKMISTESETINLSNKNS
data4:MLSQMEKFAKAFLEAQAKAK
```

- 元の配列データの長さと同じになるまで、データ行列の各列をランダムに抜き出して生成
- 同じ列を何回使っても良い

課題１

近隣結合法による系統樹を作成２



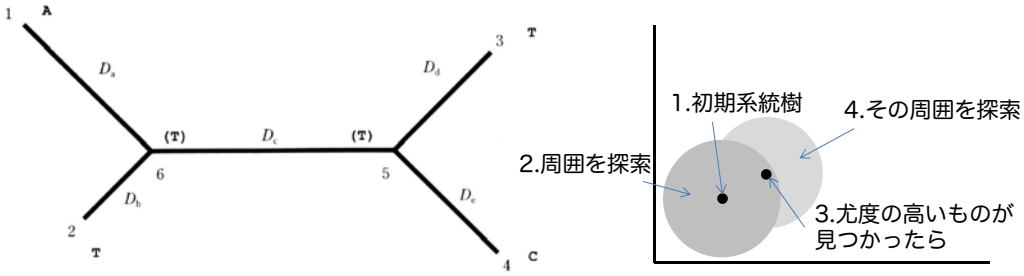
Options Summary	
Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	1000
Substitution Model	
Substitutions Type	Nucleotide
Model/Method	Kimura 2-parameter model
Substitutions to Include	d: Transitions + Transversions
Rates and Patterns	
Rates among Sites	Gamma Distributed (G)
Gamma Parameter	1
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*

- Phylogeny
- Construct Neighbor-Joining Tree
- 設定画面が表示される
  - Test of Phylogeny : Bootstrap
  - No of Bootstrap Replications : 1000
  - Model/Method : Kimura 2-parameter model
  - Substitutions to Include : Transitions + Transversions
  - Rates among Sites : Gamma Distributed
  - Gamma Parameter : 1
  - Gaps/Missing Data Treatment : Complete deletion
- Compute
- File → Save Current Session → 「kadai1」というファイル名で保存

最尤法 （Maximum likelihood estimation）

(Felsenstein, J. Mol. Evol., 1981))

- 想定される樹形ごとに手持ちの配列が得られる「尤度」を求め、最も尤度の高い樹形を採用する方法
- 塩基やアミノ酸配列の置換に関する確率モデルを仮定した上で、尤度を計算する
- 難点は計算量が多いこと　→　NJ法などで生成した初期系統樹と、それを枝交換して改変した系統樹の尤度を計算し、比較することを繰り返す
  → 発見的探索（heuristic search）



$$l_D(i) = \sum_{s_5, s_6} f(s_5) P_{s_5\mathrm{T}}(D_\mathrm{d}) P_{s_5\mathrm{C}}(D_\mathrm{e}) P_{s_5 s_6}(D_\mathrm{c}) P_{s_6\mathrm{A}}(D_\mathrm{a}) P_{s_6\mathrm{T}}(D_\mathrm{b})$$

尤度とは？

- あるモデルが正しいと仮定した状況で手元のデータが得られる確率
- データに対するモデルの当てはまりの良さを表す

10 回のコイントスを行って表が1 回、裏が9 回出たとき

- モデル１

「このコインを使ったコイントスでは表と裏が 1：9 の比率で出る」

尤度L_1 = (1/10) x $(9/10)^9$ = 0.0387

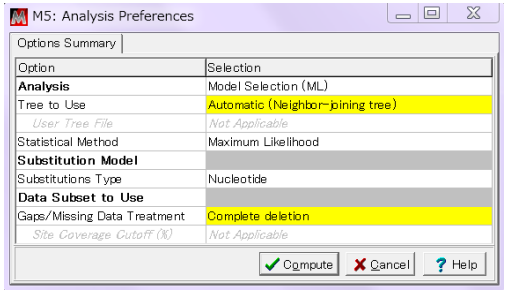
- モデル２

「このコインを使ったコイントスでは表と裏が等確率で出る」

尤度L_2 = $(1/2)^{10}$ = 0.000977

L1 > L0 であることから、前者の方が当てはまりが良い（尤もらしいモデル）ということになります

最適なモデル選択



- Models
- Find Best DNA/Protein Models
- 設定画面が表示される
  - Tree to Use : Automatic
  - Gaps/Missing Data Treatment : Complete deletion
- Compute

Table. Maximum Likelihood fits of 24 different nucleotide substitution models

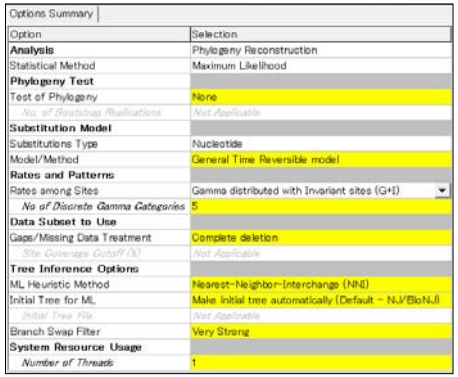
Model	Parameters	BIC	AICc	*lnL*	(+*I*)	(+*G*)	*R*	*f*(A)	*f*(T)	*f*(C)	*f*(G)	*r*(AT)	*r*(AC)	*r*(AG)	*r*(TA)	*r*(TC)	*r*(TG)	*r*(CA)	*r*(CT)	*r*(CG)	*r*(GA)	*r*(GT)	*r*(GC)
GTR+G	40	22573.995	22253.367	-11086.610	n/a	0.41	1.30	0.240	0.180	0.247	0.333	0.053	0.030	0.133	0.071	0.198	0.071	0.029	0.144	0.078	0.096	0.038	0.058
TN93+G	37	22574.717	22278.127	-11102.001	n/a	0.41	1.30	0.240	0.180	0.247	0.333	0.038	0.053	0.134	0.051	0.198	0.071	0.051	0.144	0.071	0.096	0.038	0.053
T92+G	34	22577.549	22304.997	-11118.445	n/a	0.40	1.29	0.210	0.210	0.290	0.290	0.045	0.062	0.165	0.045	0.165	0.062	0.045	0.120	0.062	0.120	0.045	0.062
GTR+G+I	41	22581.131	22252.491	-11085.169	0.14	0.56	1.29	0.240	0.180	0.247	0.333	0.053	0.030	0.133	0.071	0.198	0.071	0.029	0.144	0.078	0.096	0.039	0.058
TN93+G+I	38	22582.156	22277.553	-11100.710	0.13	0.55	1.29	0.240	0.180	0.247	0.333	0.038	0.053	0.134	0.051	0.198	0.071	0.051	0.144	0.071	0.096	0.038	0.053
T92+G+I	35	22584.811	22304.246	-11117.067	0.14	0.55	1.29	0.210	0.210	0.290	0.290	0.045	0.062	0.165	0.045	0.165	0.062	0.045	0.120	0.062	0.120	0.045	0.062
HKY+G	36	22622.193	22333.615	-11130.748	n/a	0.40	1.34	0.240	0.180	0.247	0.333	0.039	0.053	0.189	0.052	0.141	0.072	0.052	0.102	0.072	0.137	0.039	0.053

L : 最大尤度

lnL : 対数尤度

AIC (赤池情報量規準) = -2 *lnL* + 2k　(k :自由パラメータ数)
BIC (ベイズ情報量規準) = -2 *lnL* + k ln(n)　(n : 標本数)
統計モデルの良さを評価するための指標

最尤法による系統樹を作成



- Phylogeny
- Construct Maximum Likelihood Tree
- 設定画面が表示される
  - Test of Phylogeny : None
  - Model/Method : General Time Reversible model
  - Rates among Sites : Gamma Distributed with Invariant sites (G+I)
  - No of Discrete Gamma Parameter : 5
  - Gaps/Missing Data Treatment : Complete deletion
  - ML Heuristic Method : Nearest-Neighbor-interchange (NNI)
  - Initial Tree for ML : Make initial tree automatically
  - Branch Swap Filter : Very Strong
- Compute

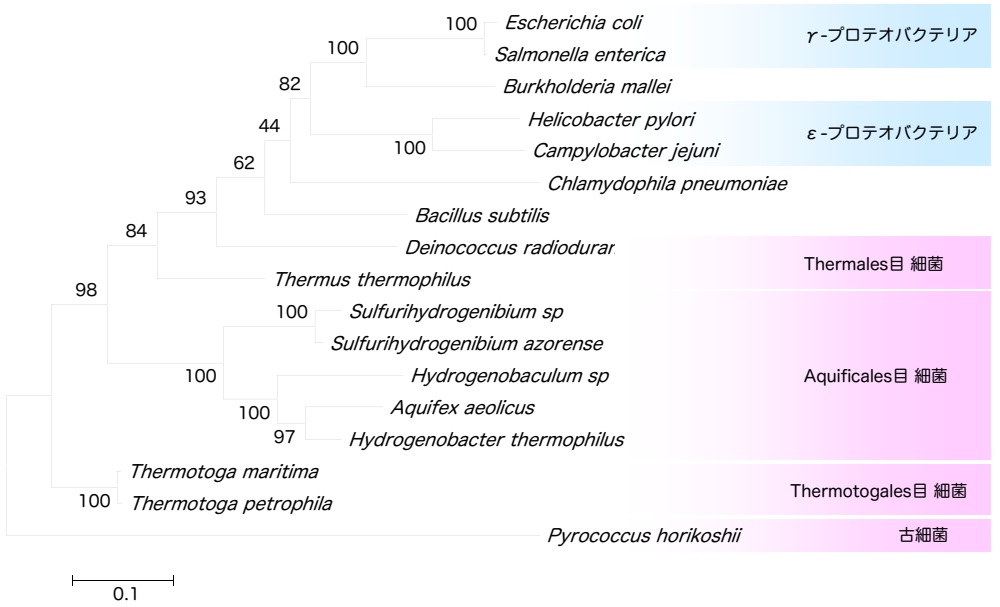
Rates among Sites : Gamma Distributed with Invariant sites

- 塩基配列、アミノ酸配列の中では、置換が滅多に起こらない座位がほとんどであり、置換が頻発する座位は限られている
- そこで、各座位を「ガンマ分布に基づいたカテゴリ」に分類するモデル(Yang, 1994) がよく利用される
- このようなモデルを使用する場合、+ G と表記される
- カテゴリ数としては、任意の数が設定できる（通常は５つくらいに設定しておくと良い）

Rates among Sites : Gamma Distributed with Invariant sites

- 置換の起きない座位(invariable site) と置換が起きる座位(variable site) の2つにカテゴリ分けするモデル(+ I と表記)

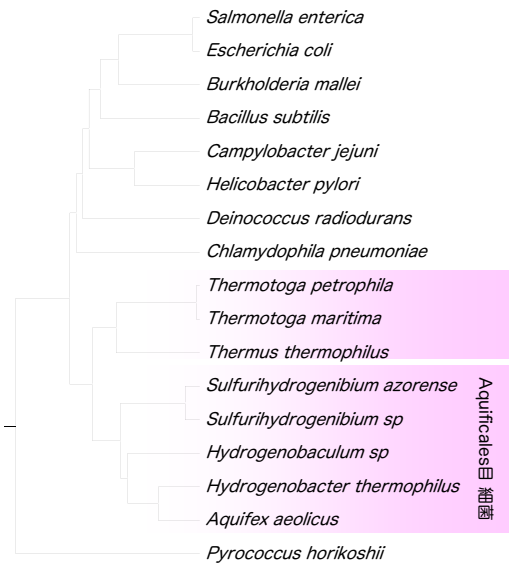
最尤法による系統樹



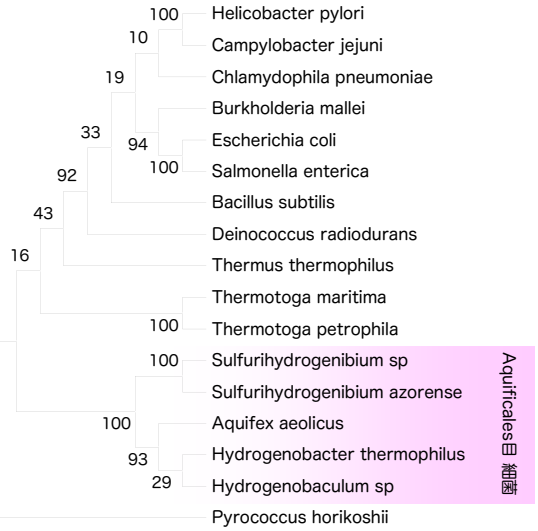
生命の起源に最も近い細菌はどれか？

他の方法で作成した系統樹と比べてみよう

UPGMA



最節約法



最尤法（アミノ酸配列）による系統樹

- Phylogeny
- Construct Maximum Likelihood Tree
- 設定画面が表示される

Option	Selection
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Amino acid
Model/Method	WAG model
Rates and Patterns	
Rates among Sites	Gamma distributed with Invariant sites (G+I)
No. of Discrete Gamma Categories	5
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - N.J/BioN.J)
Initial Tree File	*Not Applicable*
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

- Test of Phylogeny : None
- Model/Method : WAG model
- Rates among Sites : Gamma Distributed with Invariant sites (G+I)
- No of Discrete Gamma Categories : 5
- Gaps/Missing Data Treatment : Complete deletion
- ML Heuristic Method : Nearest-Neighbor-Interchange (NNI)
- Initial Tree for ML : Make initial tree automatically
- Branch Swap Filter : Very Strong

- Compute

alanyl-tRNA synthetaseのアミノ酸配列を用いた系統樹を作成する

- メニュー File → Open A File
- 「bacteria_alaS.fas」を選ぶ
- How would you like to open this fasta file? と聞かれるので
- Align を選択

アラインメントを作成

- Alignment Explorerが開く
- メニュー Alignment
- Align by Muscle
- Select all? と聞かれるので
- OK を押す
- Muscleの設定ウインドウが開く
- Compute
- アラインメントの結果が表示される

Option	Selection
Presets	None
Gap Penalties	
Gap Open	-2.9
Gap Extend	0
Hydrophobicity Multiplier	1.2
Memory/Iterations	
Max Memory in MB	4095
Max Iterations	8
More Advanced Options	
Clustering Method (Iteration 1,2)	UPGMB
Clustering Method (Other Iteration	UPGMB
Min Diag Length (lambda)	24
Genetic Code (when using cDNA)	Standard
Alignment Info	MUSCLE Citation: Edgar, Robert C. (2004), MUSCLE:
	multiple sequence alignment with high accuracy and
	high throughput, Nucleic Acids Research 32(5), 1792-1797

メインウィンドウに移行する

- メニュー Data
- Phylogenetic analysis
- Protein-coding nucleotide data? と聞かれるので No を選択
- メインウィンドウに戻る

進化モデルの選択

- 塩基置換速度行列は4x4 の行列でしたが、アミノ酸置換速度行列は20x20 の行列となるため、RateXY とFreqX の数は時間反転可能モデルでも190 + 20 = 210 となり膨大です
- そこで、既に系統関係の分かっている分類群間の系統樹において、大量のデータを用いてあらかじめ推定されたRateXYFreqX の値を用いたモデルをアミノ酸置換モデルとして用います
- これらは、実際のデータから観測された「経験的な」ものなので、empirical model と呼ばれます

From ＼ To	A	C	G	T
A	-	$Rate_{AC}Freq_C$	$Rate_{AG}Freq_G$	$Rate_{AT}Freq_T$
C	$Rate_{AC}Freq_A$	-	$Rate_{CG}Freq_G$	$Rate_{CT}Freq_T$
G	$Rate_{AG}Freq_A$	$Rate_{CG}Freq_C$	-	$Rate_{GT}Freq_T$
T	$Rate_{AT}Freq_A$	$Rate_{CT}Freq_C$	$Rate_{GT}Freq_G$	-

進化モデルの選択

- 様々なモデルが提案されています
- RateXY は既存のempirical model の値を用い、アミノ酸頻度FreqX はデータから推定するモデルも+ F モデルと呼ばれて広く用いられています
  - Dayhoff : 核 (Dayhoff et al., 1978)
  - JTT : 核 (Jones et al., 1992)
  - WAG :核 (Whelan and Goldman, 2001)
  - mtREV24 :ミトコンドリア(Adachi and Hasegawa, 1996)
  - rtREV :レトロウィルス(Dimmic et al., 2002)
  - cpREV :葉緑体(Adachi et al., 2000)

Table. Maximum Likelihood fits of 48 different amino acid substitution models

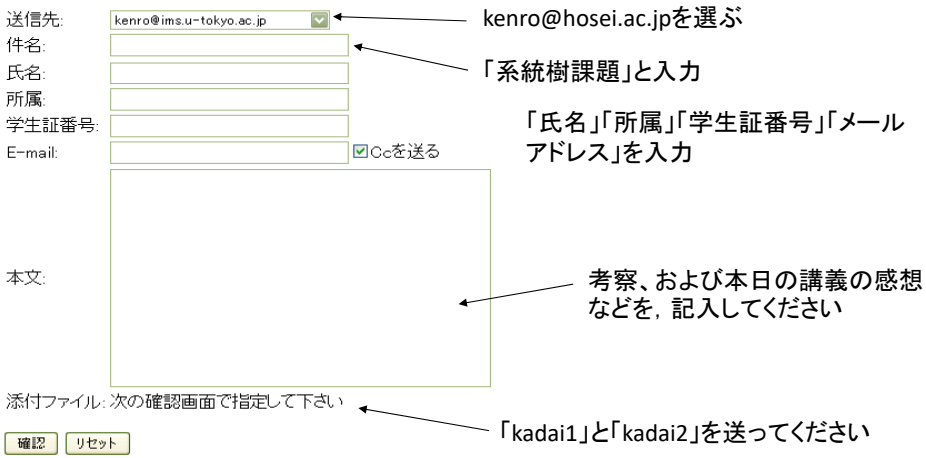
Model	Parameters	BIC	AICc	*lnL*	*(+I)*	*(+G)*	*f(A)*	*f(R)*	*f(N)*	*f(D)*	*f(C)*	*f(Q)*	*f(E)*	*f(G)*	*f(H)*	*f(I)*	*f(L)*	*f(K)*	*f(M)*	*f(F)*	*f(P)*	*f(S)*
WAG+G+I	27	29915.356	29718.300	-14832.081	0.14	1.98	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024	0.048	0.086	0.062	0.020	0.038	0.046	0.070
rtREV+G+I+F	46	29920.010	29584.446	-14746.025	0.13	1.47	0.072	0.063	0.033	0.056	0.007	0.029	0.100	0.081	0.023	0.058	0.093	0.072	0.021	0.052	0.030	0.050
WAG+G+I+F	46	29931.832	29596.268	-14751.936	0.14	1.82	0.072	0.063	0.033	0.056	0.007	0.029	0.100	0.081	0.023	0.058	0.093	0.072	0.021	0.052	0.030	0.050
rtREV+G+F	45	29937.535	29609.258	-14759.440	n/a	0.84	0.072	0.063	0.033	0.056	0.007	0.029	0.100	0.081	0.023	0.058	0.093	0.072	0.021	0.052	0.030	0.050
WAG+G	26	29946.621	29756.859	-14852.365	n/a	1.01	0.087	0.044	0.039	0.057	0.019	0.037	0.058	0.083	0.024	0.048	0.086	0.062	0.020	0.038	0.046	0.070
WAG+G+F	45	29957.705	29629.428	-14769.524	n/a	0.96	0.072	0.063	0.033	0.056	0.007	0.029	0.100	0.081	0.023	0.058	0.093	0.072	0.021	0.052	0.030	0.050
rtREV+G+I	27	30066.203	29869.148	-14907.505	0.13	1.55	0.065	0.045	0.038	0.042	0.011	0.061	0.061	0.064	0.027	0.068	0.102	0.075	0.015	0.029	0.068	0.049
cpREV+G+I	27	30082.145	29885.089	-14915.475	0.14	1.84	0.076	0.062	0.041	0.037	0.009	0.038	0.050	0.084	0.025	0.081	0.101	0.050	0.022	0.051	0.043	0.062

課題２

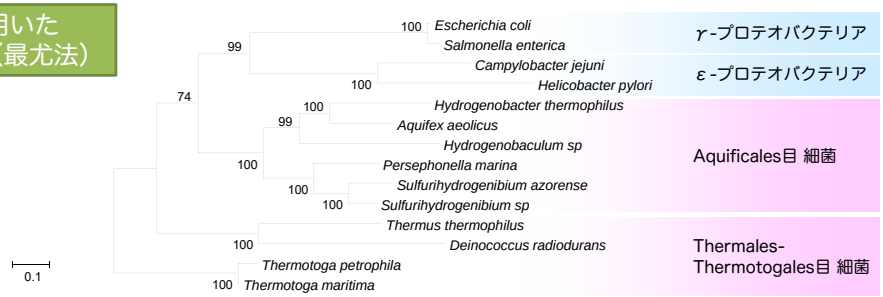
- 「bacteria_pgk.fas」には、14種のバクテリア由来のphosphoglycerate kinaseのアミノ酸配列が入っている
- alanyl-tRNA synthetaseのアミノ酸配列を用いたのと同様に、以下の設定を用いて最尤法の系統樹を作成し、ファイル名「kadai2」として保存
  - Test of Phylogeny : None
  - Model/Method : WAG model
  - Rates among Sites : Gamma Distributed with Invariant sites (G+I)
  - No of Discrete Gamma Categories : 5
  - Gaps/Missing Data Treatment : Complete deletion
  - ML Heuristic Method : Nearest-Neighbor-Interchange (NNI)
  - Initial Tree for ML : Make initial tree automatically
  - Branch Swap Filter : Very Strong
- AlaS、Pgkの２つの系統樹を比較して、トポロジーが異なる部分について考察しなさい

＜課題の提出方法＞

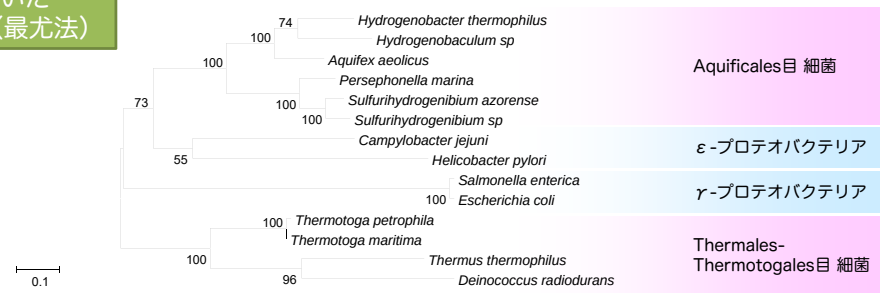
「受講生の方へ」のページ
↓
「課題提出用Web mailページへ（講義室のみからアクセス可）」



AlaSを用いた系統樹（最尤法）



Pgkを用いた系統樹（最尤法）

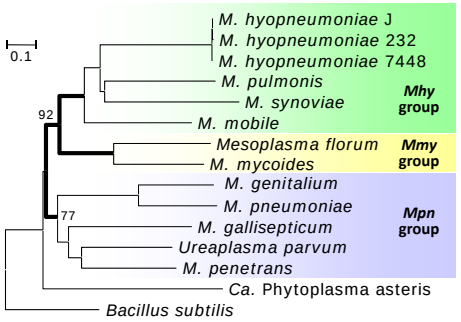


過去に、Aquificales目 細菌とε-プロテオバクテリアとの間で、大規模な遺伝子の水平移動が生じた可能性が考えられている

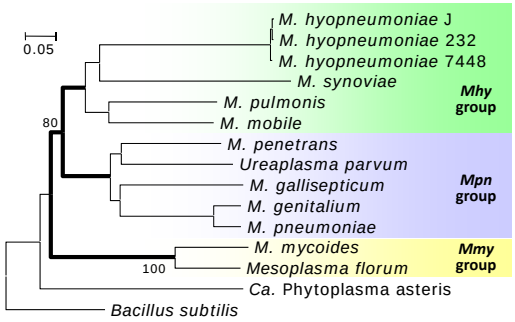
遺伝子によって系統樹のトポロジーが異なる場合がある

- 異なる遺伝子を用いて系統樹を作成した場合に，トポロジーが一致しないことがよくある
- これには，遺伝子の水平移動，分岐年代の近さ，塩基・アミノ酸置換の飽和，個々の遺伝子にかかる選択圧の違いなど，様々な原因が考えられる
- 従って，生物の進化を解析するためには，全ゲノム配列を用いるなど，なるべく多くの情報量を用いるのが望ましいと考えられている

DnaE (DNA polymerase III)



GyrB (DNA gyrase)



Phylogenetic Relationships Among Mycoplasmas Based on the Whole Genomic Information

(Oshima & Nishida, *J. Mol. Evol.*, 2007)