

2013年7月30日

## 統合データベース講習会 AJACS 琉球

# 配列比較解析の基礎 (相同性検索、アラインメント、系統樹構築)



バイオサイエンスデータベースセンター  
(National Bioscience Database Center ; NBDC)

大波純一

# 本日の講習会資料

「AJACS琉球」  
で検索

ホーム シンポジウム 講習会 展示会 連載

クリップ 0 チェック 0 +1 0 Share 0いいね! 0 ツイート 3

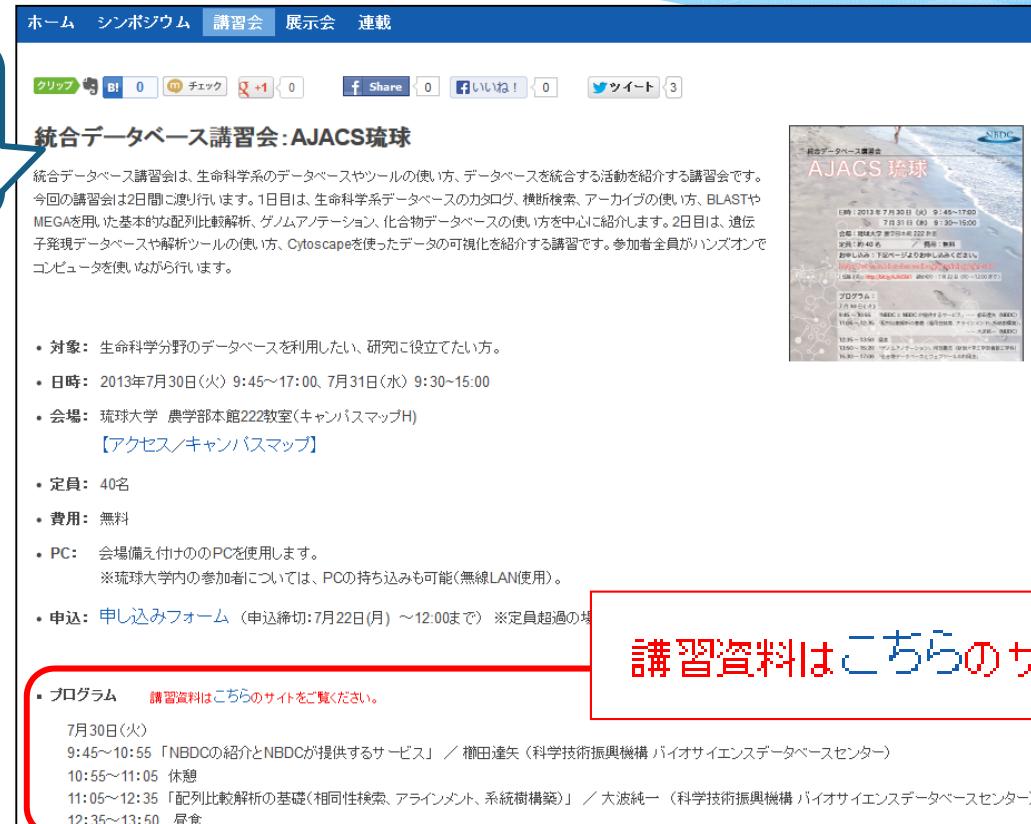
## 統合データベース講習会: AJACS琉球

統合データベース講習会は、生命科学系のデータベースやツールの使い方、データベースを統合する活動を紹介する講習会です。今回の講習会は2日間に渡り行います。1日目は、生命科学系データベースのカタログ、横断検索、アーカイブの使い方、BLASTやMEGAを用いた基本的な配列比較解析、ゲノムアノテーション、化合物データベースの使い方を中心に紹介します。2日目は、遺伝子発現データベースや解析ツールの使い方、Cytoscapeを使ったデータの可視化を紹介する講習です。参加者全員がハンズオンでコンピュータを使いながら行います。

- 対象: 生命科学分野のデータベースを利用したい、研究に役立てたい方。
- 日時: 2013年7月30日(火) 9:45~17:00、7月31日(水) 9:30~15:00
- 会場: 琉球大学 農学部本館222教室(キャンパスマップ)
- [【アクセス/キャンパスマップ】](#)
- 定員: 40名
- 費用: 無料
- PC: 会場備え付けのPCを使用します。  
※琉球大学内の参加者については、PCの持ち込みも可能(無線LAN使用)。
- 申込: [申し込みフォーム](#) (申込締切: 7月22日(月) ~12:00まで) ※定員超過の場合

■ プログラム [講習資料はこちらのサイトをご覧ください。](#)

7月30日(火)  
 9:45~10:55 「NBDCの紹介とNBDCが提供するサービス」 / 植田達矢 (科学技術振興機構バイオサイエンスデータベースセンター)  
 10:55~11:05 休憩  
 11:05~12:35 「配列比較解析の基礎(相同意性検索、アラインメント、系統樹構築)」 / 大波純一 (科学技術振興機構バイオサイエンスデータベースセンター)  
 12:35~13:50 昼食



講習資料はこちらのサイトをご覧ください。

# Index

1. 配列比較解析の目的
2. 統合解析環境MEGA
3. 相同性検索 ~MEGAから利用するNCBI BLAST
4. マルチプルアラインメント ~MUSCLEとWebLogo
5. 系統樹構築 ~近隣結合法と最尤法
6. 配列比較解析演習
7. 参考 ~文献情報とNCBI BLASTの詳細

# 1. 配列比較解析の目的

# Q.この配列は何でしょうか？

>sampleA

```
GAATTCCGGAAAGCGAGCAAGAGATAAGTCCTGGCATCAGATACAGTGGAGATAAGGACGGACGTGTGGCAGCTCCCCAGAGGAT  
TCACTGGAAGTGCATTACCTATCCCATGGGAGGCCATGGAGTTCGTGGCGCTGGGGGGCCGGATGCGGGCTCCCCACTCCGTTCC  
CTGATGAAGCCGGAGCCTCCTGGGCTGGGGGGCGAGAGGACGGAGGCAGGGCTGCTGGCCTCCTACCCCCCTCAGGC  
CGCGTGTCCCTGGTGCCGTGGCAGACACGGGTACTTGGGGACCCCCCAGTGGGTGCCGCCACCCAAATGGAGCCCCC  
CTACCTGGAGCTGCTGCAACCCCCCGGGCAGCCCCCCCCTACCCCTCCGGGCCCTACTGCCACTCAGCAGCAGGGCCCC  
CCTGCGAGGCCCGTGAGTGCCTATGCCAGGAAGAACTGCGGAGCGACGGCAACGCCGTGTGGGCCGGACGGCACGGGAT  
TACCTGTGCAACTGGCCTCAGCCTGCGGGCTTACCAACCGCCTAACGCCAGAACCGCCGCTCATCCGCCAAAAGCGCCT  
GCTGGTGAGTAAGCGCGAGGCACAGTGTGCAGCCACGCGTAAAACGCCAGACATCCACCACACTCTGTGGCGTCGCAGCC  
CCATGGGGACCCGTGCAACAACATTCACGCCCTGCCCTACTACAAACTGCACCAAGTGAACCGCCCCCTACGATGCGC  
AAAGACGGAATCCAACCCGAAACCGCAAAGTTCCAAAGGGTAAAAGCGGCCCGGGGGGGAAACCCCTCCGCCAC  
CGCGGGAGGGGGCGCTCTATGGGGGGAGGGGGGACCCCTCTATGCCCGGCCCGGGGGGGGGGGGGGGGGGGGGGGGG  
GCGACGCTCTGTACGCTCTGCCCGCTGGCCTTCCGGCCATTTCTGCCCTTGAAACTCCGGAGGGTTTTGGGGGGGG  
GCGGGGGGTTACCGGCCCGGGCTGAGCCCGCAGATTAAATAACTCTGACGTGGCAAGTGGCCTTGCTGAGAAGA  
CACTGTAACATAATAATTGCACCTCGCAATTGCAGAGGGTCGATCTCCACTTGGACACAACAGGGCTACTCGTAGGAC  
TAAGCACTTGCTCCCTGGACTGAAAAGAAAGGATTATCTGTTGCTTGCTGACAAATCCCTGTGAAAGGTAAAAGTCGA  
CACAGCAATCGATTATTCTGCCTGTGAAATTACTGTGAATATTGAAATATATATATATATATATCTGTATAGAAC  
AGCCTCGGAGGCCATGGACCCAGCGTAGATCATGCTGGATTGACTGCCGGAATT
```

The Lost World, by Michael Crichton (Ballantine Books, 1996)  
<ftp://ftp.ncbi.nih.gov/pub/FieldGuide/lostworld.txt>

# A. 恐竜ゲノムの一部(ただし空想)



生物アイコン © ライフサイエンス統合データベースセンター licensed under CC表示2.1 日本

もし恐竜の全ゲノム配列が分かったら、  
恐竜を復活させることができる？

# 残念ながら



生物アイコン © ライフサイエンス統合データベースセンター licensed under CC表示2.1 日本

配列情報だけで生命現象全てを再現することは、  
ほぼ不可能。

(さらに1億年以上昔の全ゲノム配列を入手することも、現状ほぼ不可能)

# 生命科学研究の実際

塩基・アミノ酸配列以外の  
形質関連因子が見つかっても…

エピジェネティックス、細胞質遺伝、環境の影響、短いRNAの存在、転移因子転写産物…  
どれほど多くのゲノムが読まれても…

2001:ヒトゲノム

2002:マウス、トラフグ、ゼブラフィッシュ

2004:ラット、ニワトリ、ミドリフグ

2005: チンパンジー、イヌ

2007:ハイイロジネズミオポッサム、メダカ、アカゲザル

2008:カモノハシ

…

配列解析は生命現象の一端を解き明かす  
重要な一手。

# 配列比較解析を使った研究のモデルケース

## ■環境メタゲノム研究



配列相同性検索

深海の泥中生物の  
メタゲノム解析

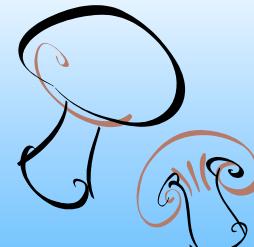
## ■生態や進化の研究



系統樹構築

短期間に分化した  
ある種の魚類間の関係

## ■毒性研究



マルチプル  
アライメント

毒分泌に関わる  
遺伝子の比較

最近の研究でも  
配列解析は有用なツールとして  
頻繁に利用されている。

# 配列解析を行う環境

- 配列相同性検索
- マルチプルアラインメント
- 系統樹構築



**MEGA** MOLECULAR EVOLUTIONARY  
GENETICS ANALYSIS  
Authors: Koichiro Tamura, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei and Sudhir Kumar

全て統合解析環境  
MEGAから実行可能

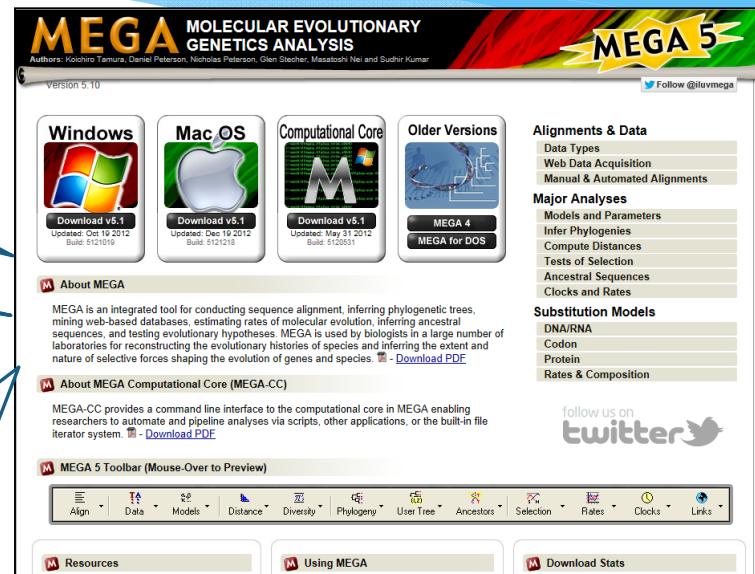
## 2. 統合解析環境MEGA

# MEGA(Molecular Evolutionary Genetics Analysis)

WindowsやMac上で動作する  
視覚的に分かりやすい  
GUIインターフェース  
(コマンドを打つ必要がない)

多彩なオプション

インターネットに接続せずに  
クローズドな環境で利用可能。  
未公開配列の解析に便利  
(配列相同性検索を除く)



<http://www.megasoftware.net/>

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011)  
MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood,  
Evolutionary Distance, and Maximum Parsimony Methods.  
Molecular Biology and Evolution 28: 2731-2739.

遺伝子レベル、タンパク質  
モジュールレベルの  
配列解析に便利

無料  
(ダウンロードの際に  
メールアドレスと名前を入力)

1993年にVersion1が公開  
されて以来、バージョンアップ  
継続中。

強力な機能を持つ統合配列解析環境のGUIソフトウェア

# MEGAの利用例

## ■シーラカンスのゲノム解析

**IN FOCUS NEWS**

**GENETICS**

**'Living fossil' genome unlocked**

The genes of an ancient fish, the coelacanth, have much to reveal about our distant past.

BY CHRIS WOOLSTON

The South African fisherman who pulled a prehistoric-looking blue creature out of his net in 1938 had unwittingly missed one of the zoological finds of the century. The coelacanth, a type of fish that had been thought to have become extinct 70 million years earlier.

Since then, scientists have identified two species of coelacanth, one found in Indonesia. While the fish are lobed fins, complete with bones and joints — and round, paddle-like tails, they look strikingly similar to the coelacanths that lived during the Devonian period, when the first land mammals Earth.

Now an international team of scientists has sequenced and analyzed the genome of the African coelacanth, *Latimeria chalumnae*, the findings of which are published online in *Nature*.

Like lungfish, the other surviving lineage of lob-finned fishes, coelacanths are actually more closely related to humans and other mammals than to the fish we eat, such as salmon and trout. Ancient lob-fins were the first vertebrates to break the land, and the coelacanth genome is expected to tell us much about the origins of tetrapods, the group of vertebrates that gave rise to amphibians, reptiles, birds and mammals, says lead author Chris Amemiya, a biologist at the University of Washington in Seattle. "The coelacanth is a cornerstone for our attempt to understand tetrapod evolution," he says.

Ending one long-standing argument, analysis of the coelacanth genome shows that it is not the closest living fishy relative to tetrapods, Amemiya says. That honor goes to the lungfish.

However, he adds, the lungfish genome is unlikely to be sequenced any time soon because it is much more complex than that of the coelacanth.

Although coelacanths are often called "living fossils," these fish have not been static from an evolutionary perspective. Karen Lindblad-Toh, a comparative genomics at Uppsala University in Sweden. Comparison of protein-coding genes in coelacanths with those of cartilaginous fish, bony fish, and tetrapods shows that change has been steadily accruing DNA changes.

For the rate of change has been remarkably slow. The latest analysis shows that the genes of modern coelacanths can themselves be considered living fossils, says Janice Nooman, a geneticist at Yale University in New Haven, Connecticut.

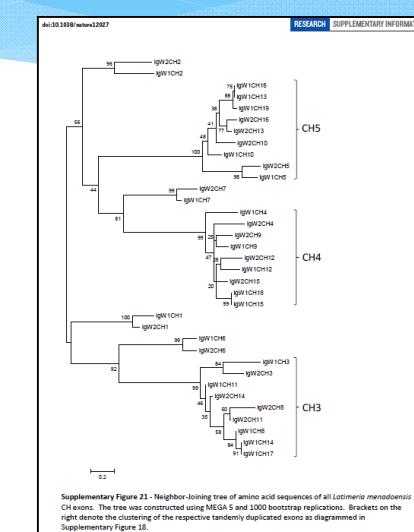
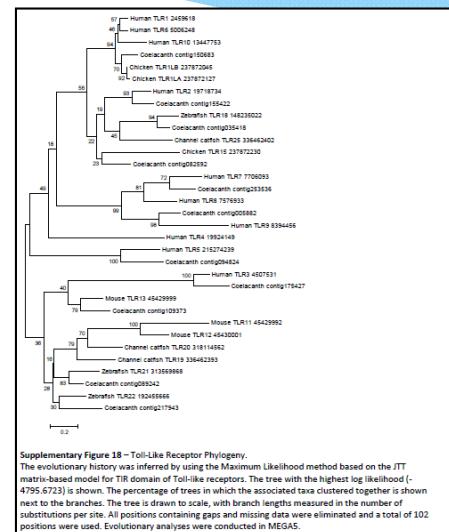
Scientists already had hints of the coelacanth's stages of evolution. In a 2008 study, Nooman and her colleagues compared the DNA of the African and Indonesian coelacanths. Specifically they looked at HOX genes, which help to get embryos to develop correctly. In the coelacanth, the HOX genes are clustered together, like their ancestors. For these similar genes, the difference between the two species of coelacanth was about 11 times smaller than that between the HOX genes of humans and chimps, two species that parted ways perhaps 6 million years ago.

**SLOW TO CHANGE**

It is possible that the slow rate of coelacanth evolution could be due to a lack of natural-selection pressure, Lindblad-Toh says. Modern coelacanths, like their ancestors, "live far down in the ocean, where life is pretty stable," she says. "There is not a lot of selection pressure." But it is possible that the slow genetic change explains why the fish show such a striking resemblance to their fossilized ancestors.

Further analysis of the genome is bound to reveal more about the coelacanth's past, says Nooman. "It will allow us to identify the genetic drivers of tetrapod evolution, the genes and regulatory elements that are responsible for the vertebrate land transition," he says.

© 2013 Macmillan Publishers Limited. All rights reserved



Per site, all positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGAS.

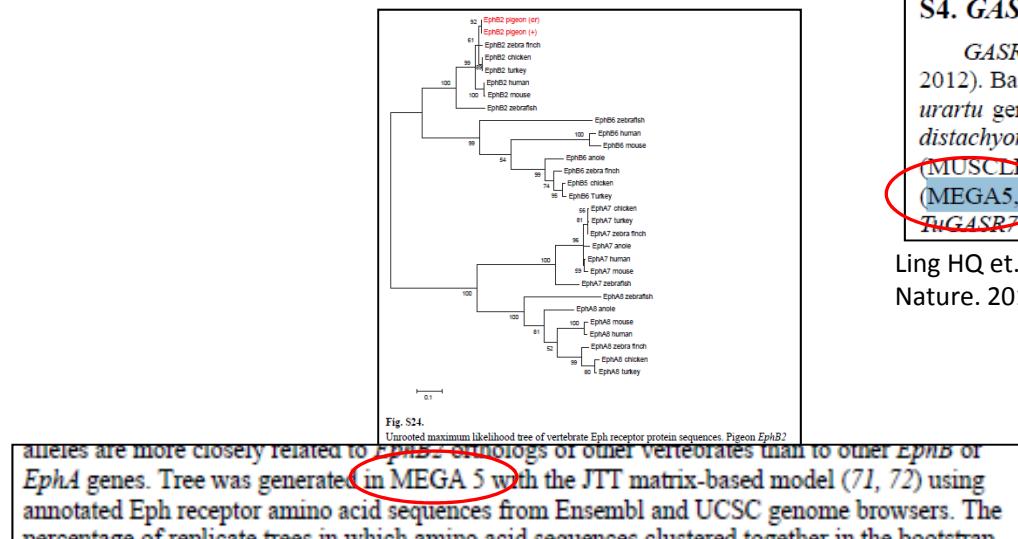
15. The tree was constructed using MEGA 5 and 1000 bootstrap replications. Brackets on the right denote the clustering of the respective tandemly duplicated exons as diagrammed in Supplementary Figure 18.

'Living fossil' genome unlocked. Woolston C. Nature. 2013 Apr 18;496(7445):283. doi: 10.1038/496283a.

Amemiya CT et. al., The African coelacanth genome provides insights into tetrapod evolution. Nature. 2013 Apr 18;496(7445):311-6. doi: 10.1038/nature12027.

# MEGAの利用例

## ■カワラバトのゲノム解析



Shapiro MD et. al., Genomic diversity and evolution of the head crest in the rock pigeon. Science. 2013 Mar 1;339(6123):1063-7. doi: 10.1126/science.1230422. Epub 2013 Jan 31.

## ■ウラルツコムギのゲノム解析

### S4. *GASR7*, a gene-related to yield traits

*GASR7* is a gibberellin-regulated gene that controls grain length in rice (Huang et al. 2012). Based on a BLASTN search, we identified one *GASR7* homolog (*TuGASR7*) in the *T. urartu* genome (TRIUR3\_28594), and compared it to orthologous counterparts in rice, *B. distachyon*, maize, sorghum, barley, and bread wheat by multiple sequence alignment (MUSCLE v1.8, <http://www.ebi.ac.uk/Tools/msa/muscle/>) and phylogenetic analysis (MEGA5, Tamura et al. 2011). After sequencing a portion of the 5' genomic region of *TuGASR7* (417 bp from the start codon ATG) in 92 *T. urartu* accessions, we discovered two

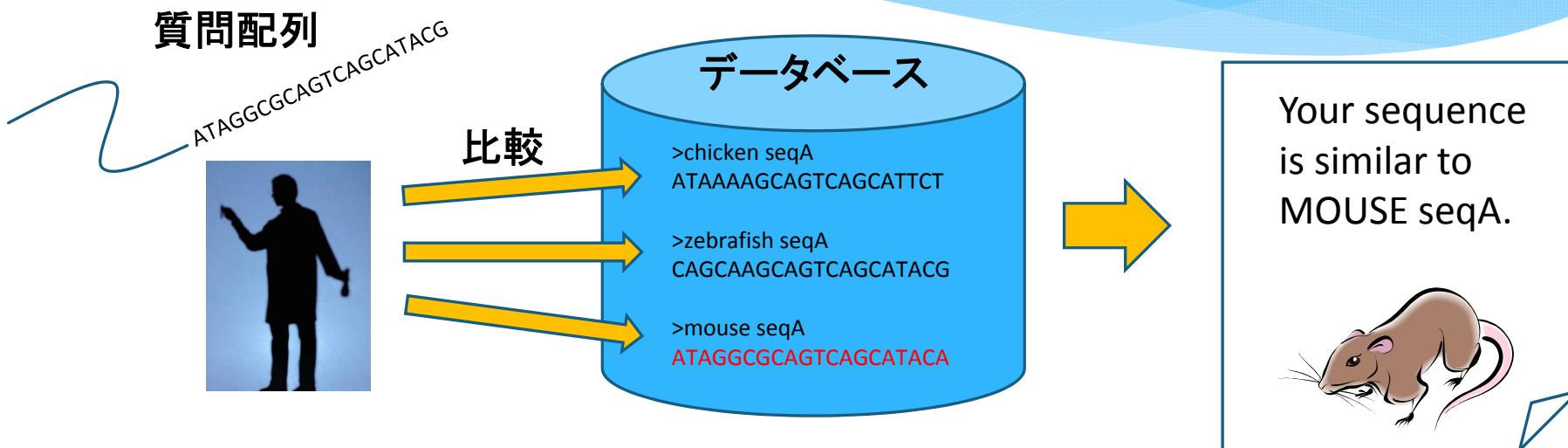
Ling HQ et. al., Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature. 2013 Apr 4;496(7443):87-90. doi: 10.1038/nature11997. Epub 2013 Mar 24.

## 今年のNature, Scienceのゲノム解析論文で 遺伝子ごとの系統解析の実績あり。

### 3. 相同性検索

~MEGAから利用するNCBI BLAST

# 相同性検索とは



質問配列に対し、データベースから  
類似性の高い配列(=相同な可能性の高い配列)を  
返す検索手法

# 相同性検索の利用

- 未知の配列を、既知の配列と比較して機能予測
- サンプルの生物種判別
- ある遺伝子のゲノム内の位置を検索
- etc...



配列解析の初手として利用

# 相同性検索の2大メソッド

- BLAST(Basic Local Alignment Search Tool)

"Basic local alignment search tool.", Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. J Mol Biol. 1990 Oct 5;215(3):403-10.

FASTAより高速で動作(デフォルト設定の場合)

- FASTA(FAST-All)

"Improved tools for biological sequence comparison." Pearson WR, Lipman DJ. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444-8.

BLASTより検出感度が高い(デフォルト設定の場合)

それぞれ独自の統計理論を採用し、  
高速な配列検索を実現している。



本講義では広く利用されている**BLAST**を使用する。

※BLASTとFASTAの採用検討については、ゲノムネットのデータベース利用法(金久実 共立出版 2002)を参照

# 配列資料について

本講義では主にFASTA形式のファイルを使用します。  
配列ファイルは「AJACS琉球ページ」の講義資料から  
取得してください。

配列.zip



解凍

sampleA.txt	2013/07/24 15:41	テキスト ドキュメント	2 KB
sampleB.txt	2013/07/24 15:42	テキスト ドキュメント	2 KB
sampleC.fasta	2013/07/24 15:42	FASTA File	16 KB
sampleD.fasta	2013/07/24 15:47	FASTA File	22 KB
T1R2amino.fas	2013/07/24 15:49	FAS File	18 KB
T1R3amino.fas	2013/07/24 15:44	FAS File	6 KB
T1R3nuc30.fas	2013/07/24 15:45	FAS File	1 KB

※もし今回取得できない場合でも、後日ご自宅でお試しいただけます。

# 相同性検索用サンプル配列

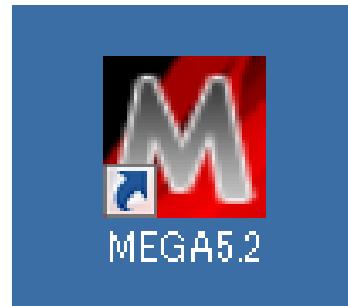
## SapmpleB.txt

```
>sampleB
ATGACCCCAACACGAAAATTAAACCCACTAATAAAATTAAATCACTCATTATCGACCTCCCACCCCATCCAAACATTCCGCATGATGGAACCTCGGCTCC
CTTCTGGCGCCTGCCTAACCTCAAATCACCAAGGATTATTCTAGCCATACACTACTCACCGAGCGCTCAACCGCCTCTCGTCGATGCCACATCACCC
GAGACGTAAACTATGGCTGGATCATCCGCTACCTCACGCCAATGGCGCCTCAATATTTTATCTGCCTCTCCTACACATGGTCGAGGCCTATATTACGGCT
CATTCTCTACCTAGAAACCTGAAACATTGGCATCATCCTTGTCTACAACCATAAGCAACAGCCTTATGGGCTATGTCCTCCATGAGGCCAAATATCCTTCT
GAGGAGGCCACAGTAATTACAAACCTACTATCGGCCACCCGTATATCGGAACAGACCTGGTCCAATGAGTCTGAGGAGGCTACTCAGTAGACAGCCCTACCC
CACACGATTCTCACCTTCACTTCATCTTACCCCTCATTATCACAGCCCTAACACACTTCATCTCCTATTCTTACACGAAACAGGATCAAATAACCCCTAGGA
ATCACCTCCACTCCGACAAAATTACCTTCCACCCCTACTACACAATCAAAGATATCCTGGTTATTCCCTTCTCTTATCCTAATGACATTAACGCTATTCTC
ACCAGACCTCTAGGCGATCCAGACAACACTACCCCTAGCTAACCCCTAAACACCCCCACCCCCACATTAAACCCGAGTGATACTTCTATTGCCTACACAATTCT
CCGATCCATCCCAACAAACTAGGAGGTGTCTCGCCTACTGCTATCTACCTGATCCTAGCAGCAATCCCTGTCCTCCATACATCCAAACAACAAAGCATAAT
ATTCCGCCACTAAGCCAATGCTTACTGACTCCTAGCCGAGACCTCCTATCCTAACCTGAATGGAGGGCAACCAGTAAGCTACCCCTCATCACCATCGG
ACAAGTAGCATCCGTATTATACTTCACAACAATCTAACCAATTACCTCCCTAACGAAAACAAAATCTGAATGAACC
```

※参考:FASTA形式

- ・1行目に「大なり(greater-than)記号」と配列名(例:>sampleB)
- ・2行目に塩基もしくはアミノ酸配列を記述します。(上記の例では塩基配列が何行にも渡っていますが途中に改行は入ません)
- ・ファイルに保存するときはテキスト形式で。拡張子は便宜的に.fastaったり.fastaたりするがアプリによる。
- ・アライメントファイルは名前の行と配列の行を交互に連続して記述する(後述)。
- ・配列名は基本何でもよい(空白や縦棒も可)が、特殊記号\$や¥を入れるとプログラム内で問題が発生する場合があるため、使用する記号は"\_"や"\_"にしておくのが無難。

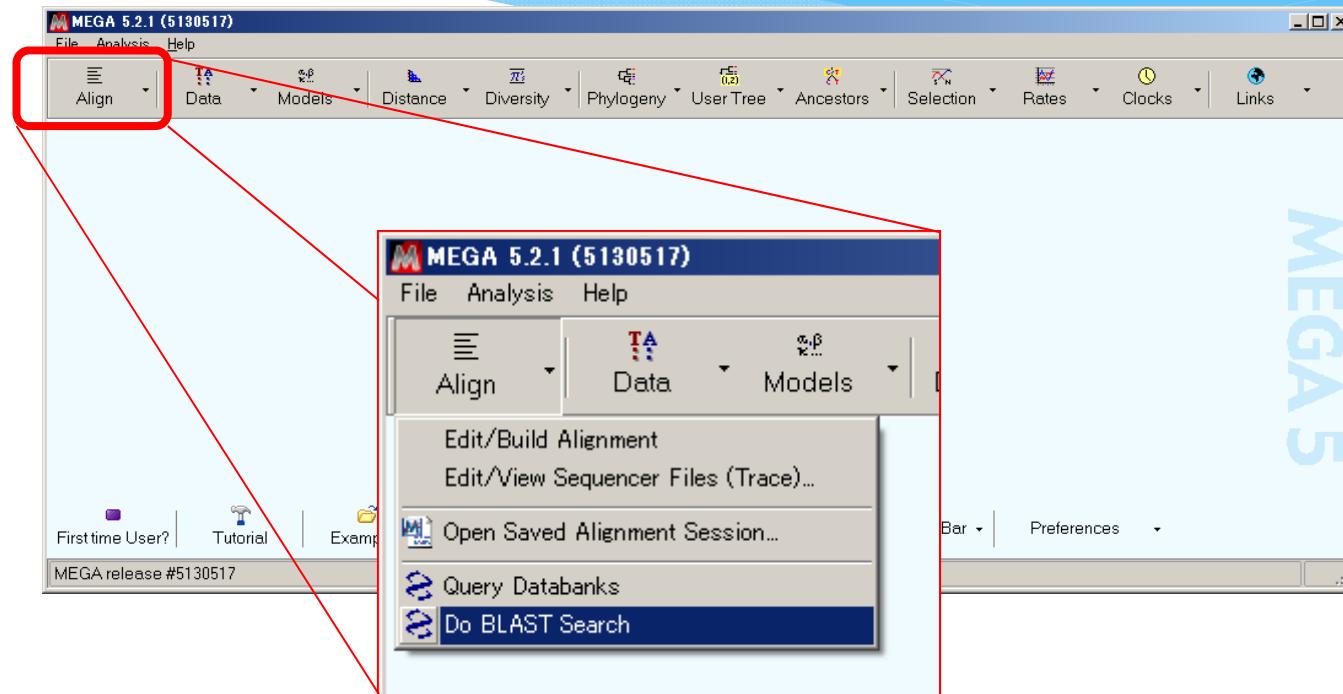
# MEGA起動



すでにMEGAが導入されている場合は、  
ショートカットやプログラムメニューから起動してください。

※もし未導入の場合、MEGAサイト(<http://www.megasoftware.net/>)から  
お名前とメールアドレスを入力し、ダウンロードを行ってください。

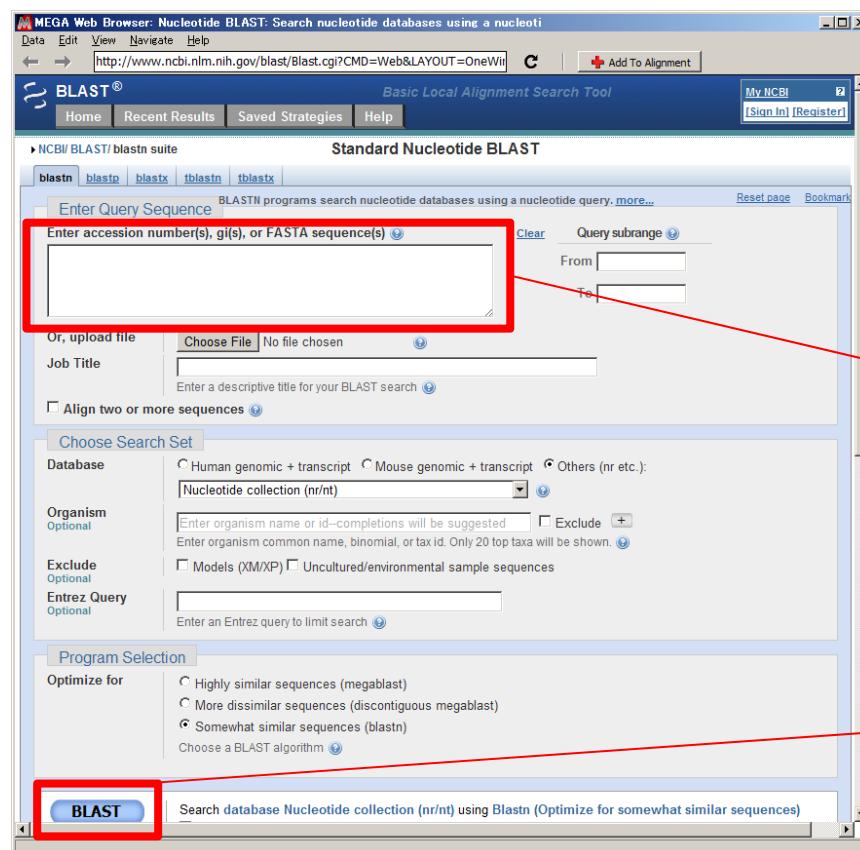
# BLASTの実行



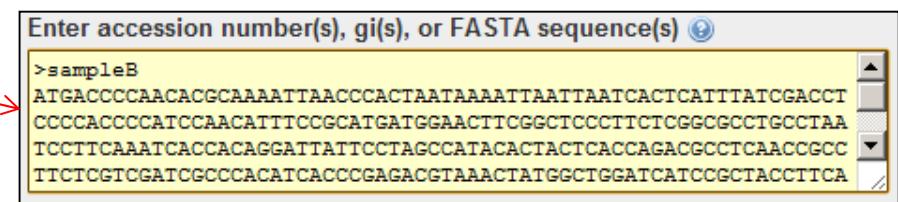
MEGAメインウィンドウから  
「Align」ボタン→「Do BLAST Search」を選択

# 検索開始

MEGAのWebブラウザから、NCBI BLASTを行う



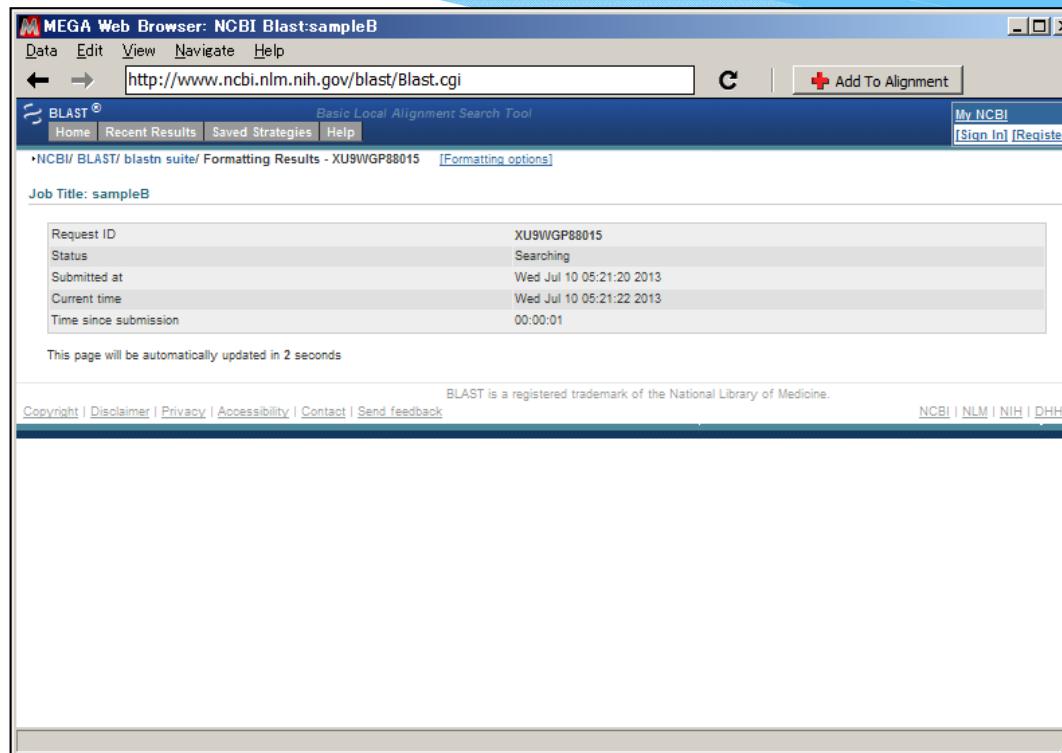
①Enter Query Sequence(質問配列入力)に、先ほどの「SampleB.txt」の配列をコピーして貼り付け



②BLASTボタンをクリック

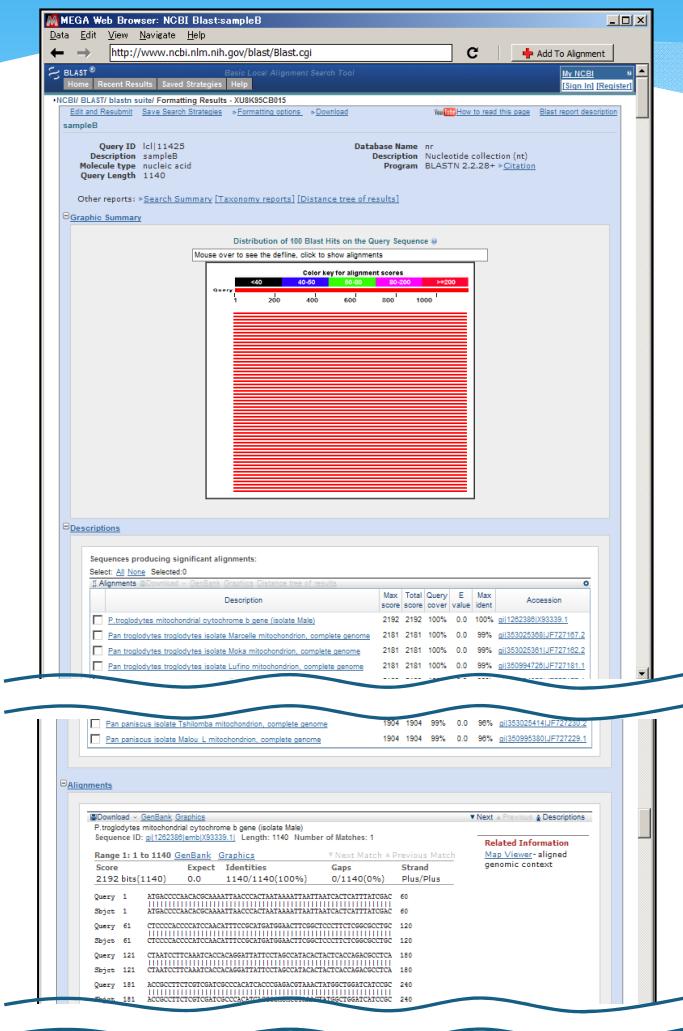
**BLAST**

# 検索中…



※NCBIネットワークの混雑状況により30秒以上かかる場合があります。  
結果は本スライドで確認できますので、そのままお待ちください。

# NCBI BLAST結果画面



## ■検索条件

## ■Graphic Summary

上位100件のスコアとマッチ領域を色で表示

## ■Descriptions

データベース内でヒットした配列100件の説明とスコア、NCBIアクセス番号

## ■Alignments

質問配列とヒット配列のアライメント  
画面を一番下にスクロールすると追加読み込

# NCBI BLAST結果画面

## ■検索条件

The screenshot shows the NCBI BLAST search results page. At the top, there's a navigation bar with links for Data, Edit, View, Navigate, Help, and a URL bar showing <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>. Below the bar is the BLAST logo and the text "Basic Local Alignment Search Tool". On the right, there's a "My NCBi" section with "Sign In" and "Register" buttons.

The main content area displays search parameters and database details:

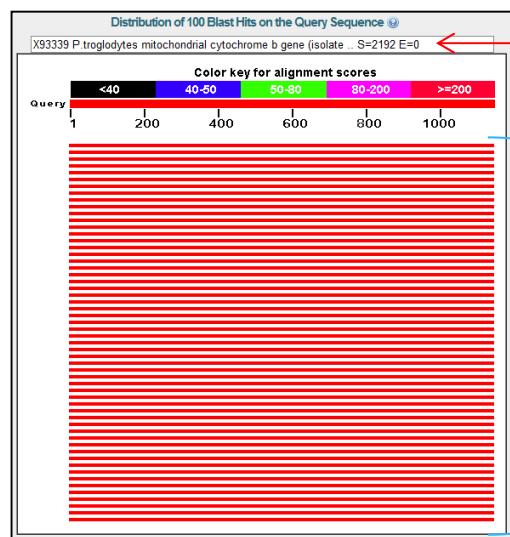
- sampleB**
- Query ID:** lcl|21223 (①)
- Description:** sampleB (②)
- Molecule type:** nucleic acid (③)
- Query Length:** 1140 (④)
- Database Name:** nr (⑤)
- Description:** Nucleotide collection (nt) (⑥)
- Program:** BLASTN 2.2.28+ (⑦)

At the bottom, there are links for "Edit and Resubmit", "Save Search Strategies", "Formatting options", "Download", "YouTube How to read this page", and "Blast report description".

- |                 |             |                 |                |
|-----------------|-------------|-----------------|----------------|
| ①Query ID:      | 検索ID        | ⑤Database Name: | データベース名        |
| ②Description:   | 質問配列の名称     | ⑥Description:   | データベースの説明      |
| ③Molecule type: | 核酸/アミノ酸     | ⑦Program:       | 検索に用いたBLASTの種類 |
| ④Query Length:  | 質問配列の長さ(bp) |                 |                |

# NCBI BLAST結果画面

## ■ Graphic Summary



カーラーバー上にマウスを移動させると、配列アクセッショ番号、配列詳細、S={スコア}、E={E-value}が表示されます。

スコアの大きさを5段階(40未満、40以上60未満、60以上80未満、80以上200未満200以上)で色分けして表示しています。

全域(1bp～1140bp)において200以上のスコアの相同性のある領域が100件表示されています。

※スコア：質問配列とのアライメントスコア。高いほど相同配列となる可能性が高い。

E-value：アライメントの確からしさ。大きい場合はアライメントが偶然起こった可能性が高い。

参考：<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>

# NCBI BLAST結果画面

## ■ Description

**Sequences producing significant alignments:**

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

①	② Description	Max score	Total score	Query cover	E value	Max ident	Accession	⑨
<input type="checkbox"/>	<a href="#">P.troglodytes mitochondrial cytochrome b gene (isolate Male)</a>	2192	2192	100%	0.0	100%	<a href="#">gil1262386 X93339.1</a>	
<input type="checkbox"/>	<a href="#">Pan troglodytes troglodytes isolate Marcelle mitochondrion, complete genome</a>	2181	2181	100%	0.0	99%	<a href="#">gil353025368 JF727167.2</a>	
<input type="checkbox"/>	<a href="#">Pan troglodytes troglodytes isolate Moka mitochondrion, complete genome</a>	2181	2181	100%	0.0	99%	<a href="#">gil353025361 JF727162.2</a>	
<input type="checkbox"/>	<a href="#">Pan troglodytes troglodytes isolate Lufino mitochondrion, complete genome</a>	2181	2181	100%	0.0	99%	<a href="#">gil350994726 JF727181.1</a>	
<input type="checkbox"/>	<a href="#">Pan troglodytes troglodytes isolate Golfi mitochondrion, complete genome</a>	2169	2169	100%	0.0	99%	<a href="#">gil350994672 JF727177.1</a>	
<input type="checkbox"/>	<a href="#">Pan troglodytes troglodytes isolate Botsomi mitochondrion, complete genome</a>	2169	2169	100%	0.0	99%	<a href="#">gil350994645 JF727175.1</a>	

- ①チェックボックス: 上部AlignmentやDownloadに利用
- ②Description: 配列の詳細
- ③Max score: 配列内断片の最大スコア
- ④Total score: 配列内断片の合計スコア
- ⑤Query cover: 質問配列がどれだけカバーされているか

- ⑥E value: 配列内断片の最小のE-value
- ⑦Max ident: 配列内断片の相同性割合の最大値
- ⑧Accession: アクセッション番号
- ⑨カラム設定: 表示するカラムを選択する

参考: [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)

# NCBI BLAST結果画面

## ■ Alignments

The screenshot shows the NCBI BLAST results page for a mitochondrial cytochrome b gene search. The search query is "P.troglodytes mitochondrial cytochrome b gene (isolate Male)" with Sequence ID gil1262386[emb|X93339\_1]. The length of the query is 1140, and there is 1 match found.

Key data points shown in the alignments table:

Range	Score	Expect	Identities	Gaps	Strand	Length
1: 1 to 1140	2192 bits(1140)	0.0	1140/1140(100%)	0/1140(0%)	Plus/Plus	1140
Query 1	ATGACCCCT	CACGCAAAATTAAACCCACTAATAAAATTAAAT	AATCACTCATTTTGAC			60
Sbjct 1	ATGAC	ACCGCAAAATTAAACCCACTAATAAAAT	AICACTCA			60
Query 61	CTCCCCACCCCATC	ACCAACATTCCGCATGATGGAACCTCGGCTCCCTCTGGCGCTGC				120
Sbjct 61	CTCCCCACCCCATC	ACCAACATTCCGCATGATGGAACCTCGGCTCCCTCTGGCGCTGC				120
Query 121	CTAACCTCAAATCAC	CAGGATATTCCCTAGCCATACTACTCACCAGACGCC				180
Sbjct 121	CTAACCTCAAATCAC	CAGGATATTCCCTAGCCATACTACTCACCAGACGCC				180
Query 181	ACCGCCTTCTCGTC	GATGCCACATCACCGAGACGTAAACTATGGCTGGATCATCCG				240
Sbjct 181	ACCGCCTTCTCGTC	GATGCCACATCACCGAGACGTAAACTATGGCTGGATCATCCG				240

Callouts numbered 1 through 13 point to specific elements of the interface and results:

- Download: FASTA/Genbank形式で配列をダウンロード
- Genbank: Genbankの配列のページへ。
- Graphics: Genbankの配列のGraphics表示ページへ
- Next/Previous: 次/前の配列断片へ
- Descriptions: 上部Descriptionリストへ
- 配列名
- アクセスション番号
- Number of Matches: 配列内でヒットした断片の数
- Score: アライメントスコア ※()内は配列の長さ
- Expect: E-value
- Gaps: マッチしない塩基数/全体の塩基数(その割合)
- Strand: 質問配列とデータベース配列の向き。  
登録されているものと同じ: Plus、Complement: Minus
- アライメント: Query: 質問配列、Sbjct: データベースの配列

①Download: FASTA/Genbank形式で配列をダウンロード

②Genbank: Genbankの配列のページへ。

③Graphics: Genbankの配列のGraphics表示ページへ

④Next/Previous: 次/前の配列断片へ

⑤Descriptions: 上部Descriptionリストへ

⑥配列名

⑦アクセスション番号

⑧Number of Matches: 配列内でヒットした断片の数

⑨Score: アライメントスコア ※()内は配列の長さ

⑩Expect: E-value

⑪Gaps: マッチしない塩基数/全体の塩基数(その割合)

⑫Strand: 質問配列とデータベース配列の向き。

登録されているものと同じ: Plus、Complement: Minus

⑬アライメント: Query: 質問配列、Sbjct: データベースの配列

参考: [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)

# 検索結果



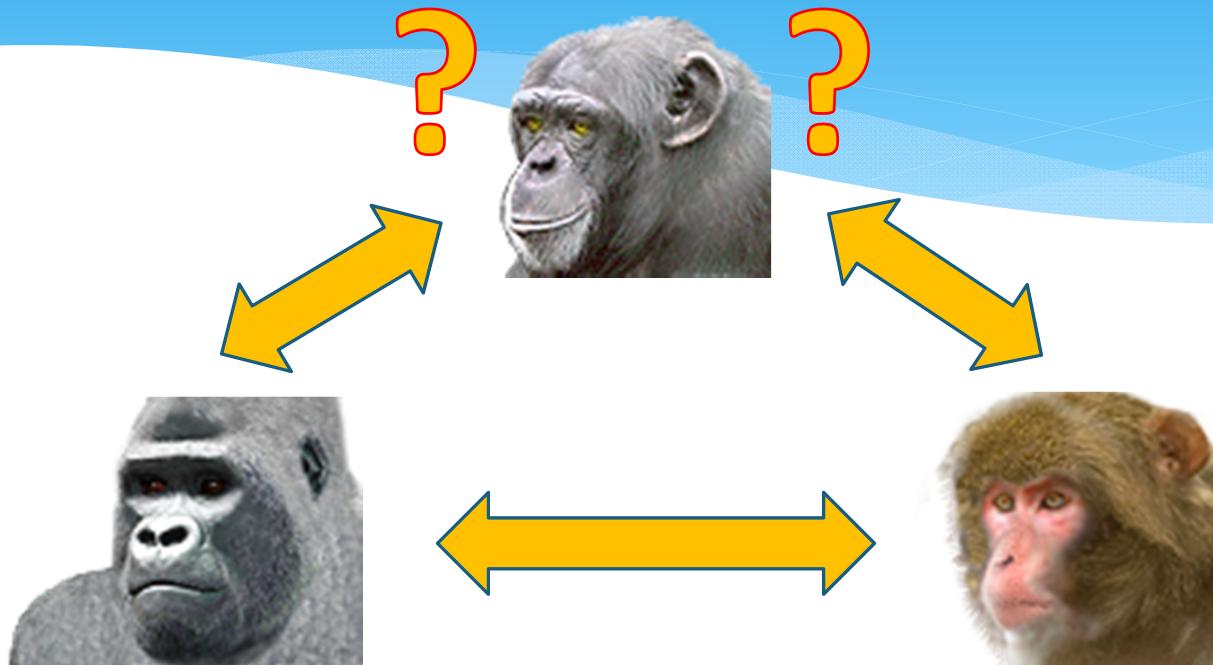
- SampleBはチンパンジーのミトコンドリアゲノムにコードされているcytochrome b遺伝子の配列



BLASTのデフォルト設定でも、高感度でデータベースから遺伝子1つ1つの情報を探すことができる。

※本資料付録のBLAST詳細設定を使えば、より高度な検索が可能です。

# 2つ以上の配列を比較



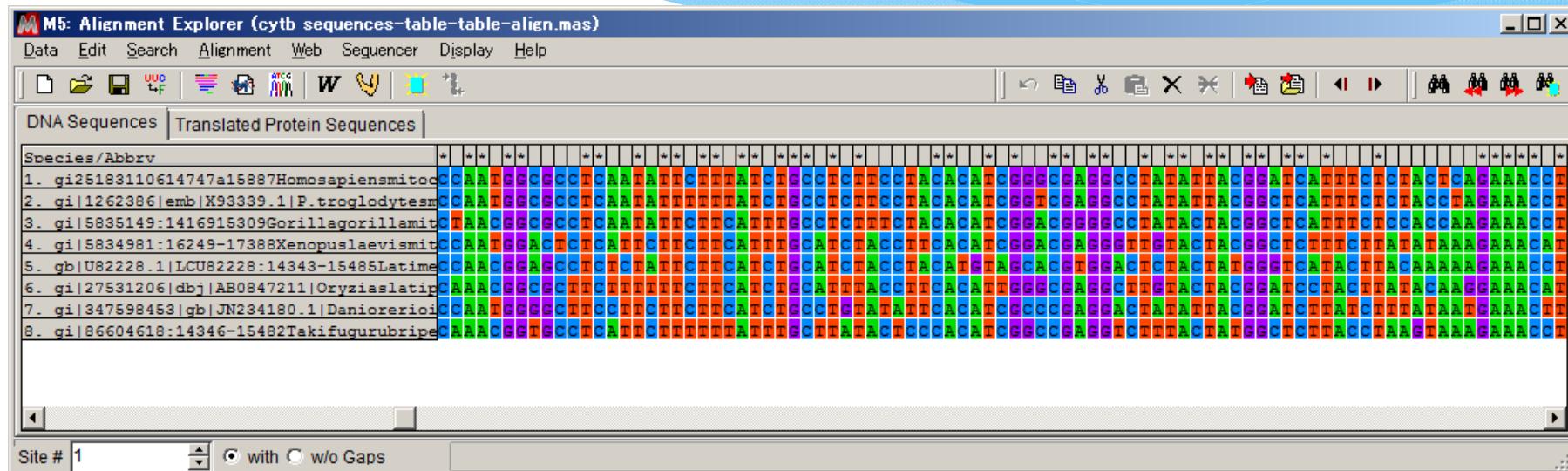
複数種のホモログを一度に比較したい場合は？

→ マルチプルアラインメント

生物アイコン © ライフサイエンス統合データベースセンター licensed under CC表示2.1 日本

## 4. マルチプルアラインメント ~MUSCLEとWebLogo

# マルチプルアラインメントとは



収集された複数のホモログ・オルソログについて  
ギャップを認識して桁揃えを行い、保存領域の検討や  
配列比較を行うための手法。

# マルチプルアラインメントの プログラム

- **2次元DP法**

2本の塩基配列のマルチプルアラインメントは、2本の漸化式を解くことで解が得られる。しかしN本になるとN本の漸化式を解く必要があり計算量が増大するため実用的ではない。

- **CLUSTALW**

アラインメントを1本の配列とみなし、2次元DP法によって配列を適当な順序で並置するプログレッシブアラインメント法(の中のツリーベース法)を利用したCLUSTALアルゴリズムによるプログラム。

- **MUSCLE**

PRRPやMAFFTのような、ペアワイズで基となるアラインメントを作成し、改良を加えていく手法を採用し、高速化を達成している。



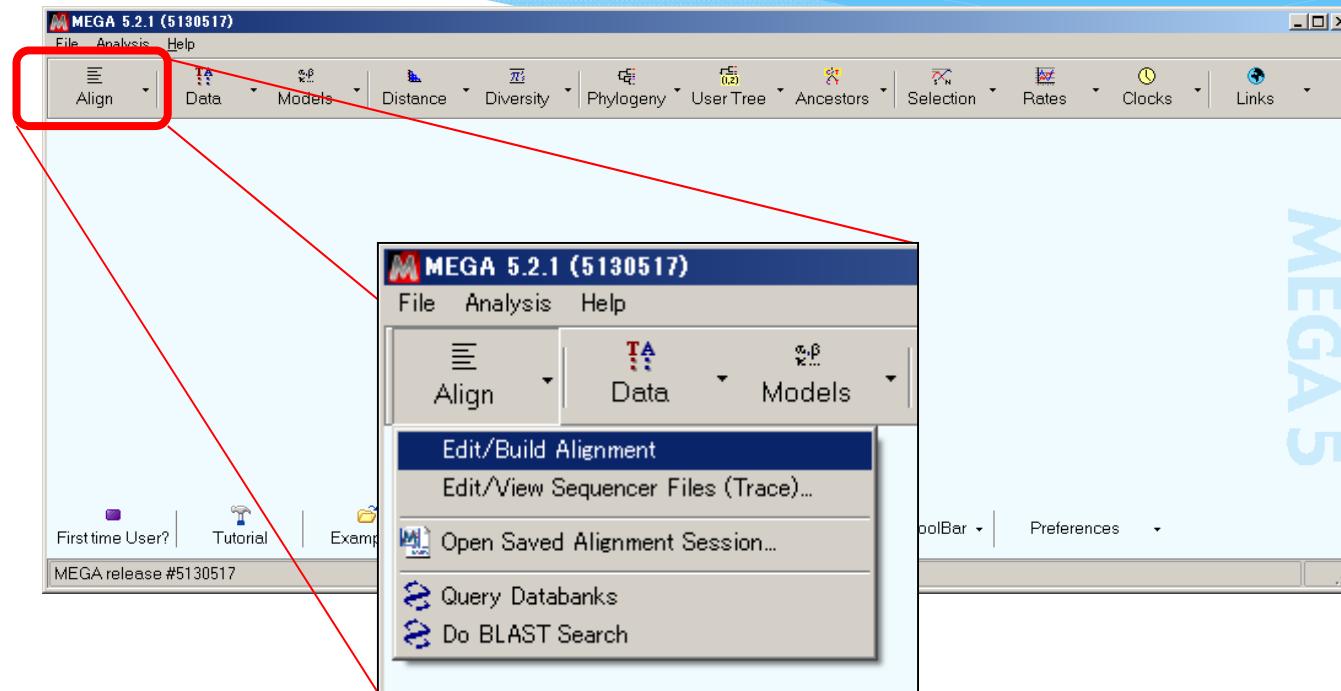
MEGAからMUSCLEを使って  
マルチプルアラインメントを行う。

# マルチプルアラインメント用 サンプル配列(T1R3)

## sampleC.fasta

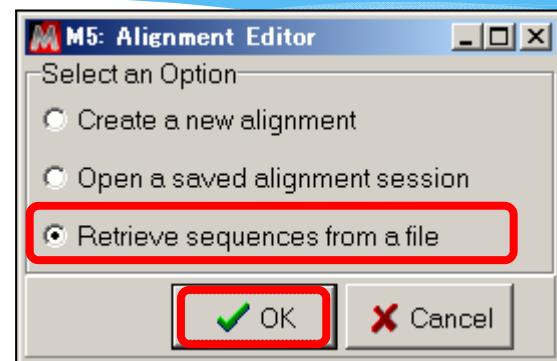
## 6種の哺乳類(ネコ、イヌ、ウシ、ヒト、マウス、ラット)の核ゲノムにコードされているT1R3遺伝子ホモログ一覧

# Alignment Explorerの起動



①MEGAメインウィンドウから  
「Align」ボタン→「Edit/Build Alignment」を選択

# Alignment Explorerの起動

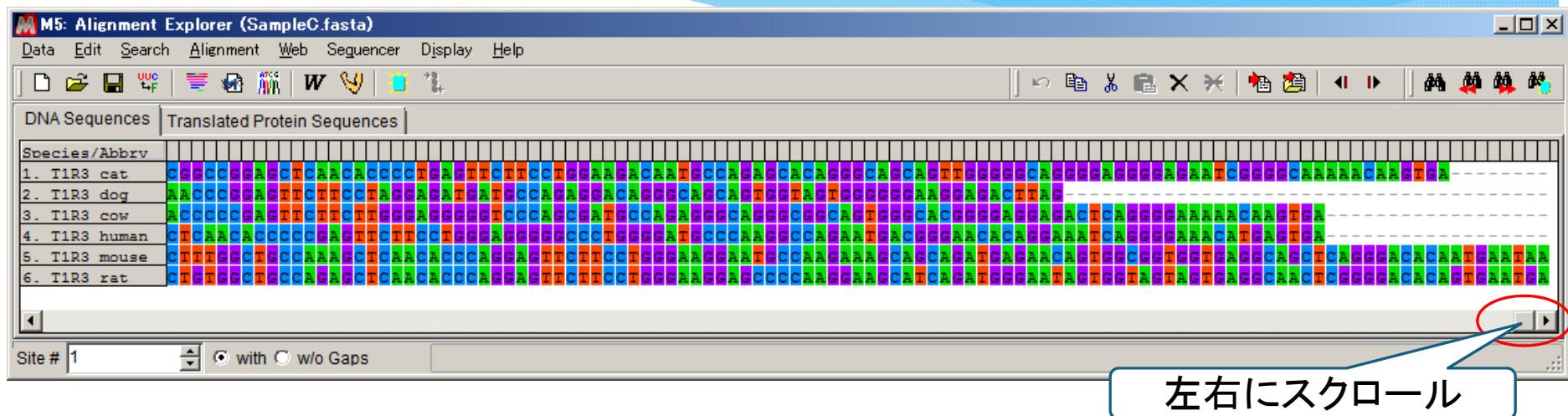


②「Retrieve sequences from a file」を選択し「OK」  
→sampleC.fasta  
ファイルを選択



Alignment Explorer起動

# SampleC:T1R3遺伝子のリスト



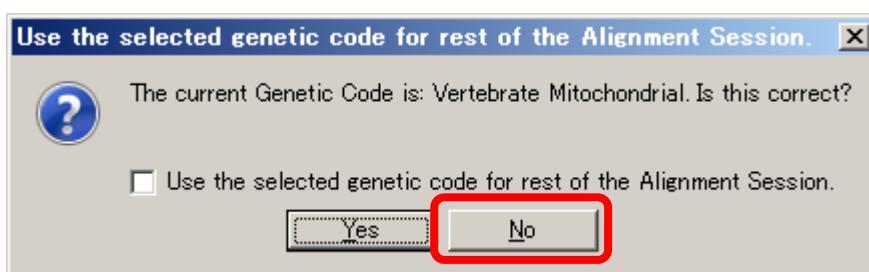
各生物種のT1R3遺伝子ホモログが  
左寄せで記述されています。  
(まだアラインメントされていません)

# アミノ酸に翻訳

※フレームシフトについては割愛

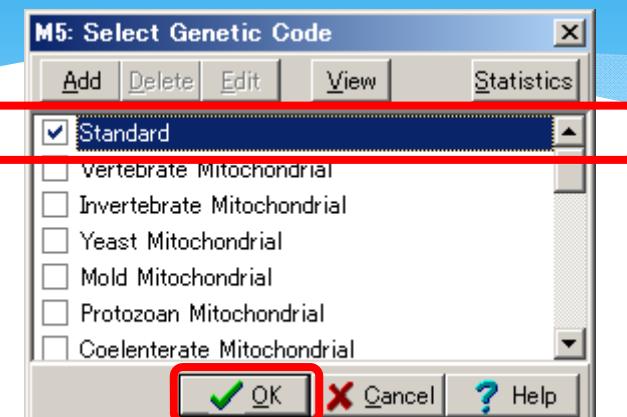


コドン表の確認



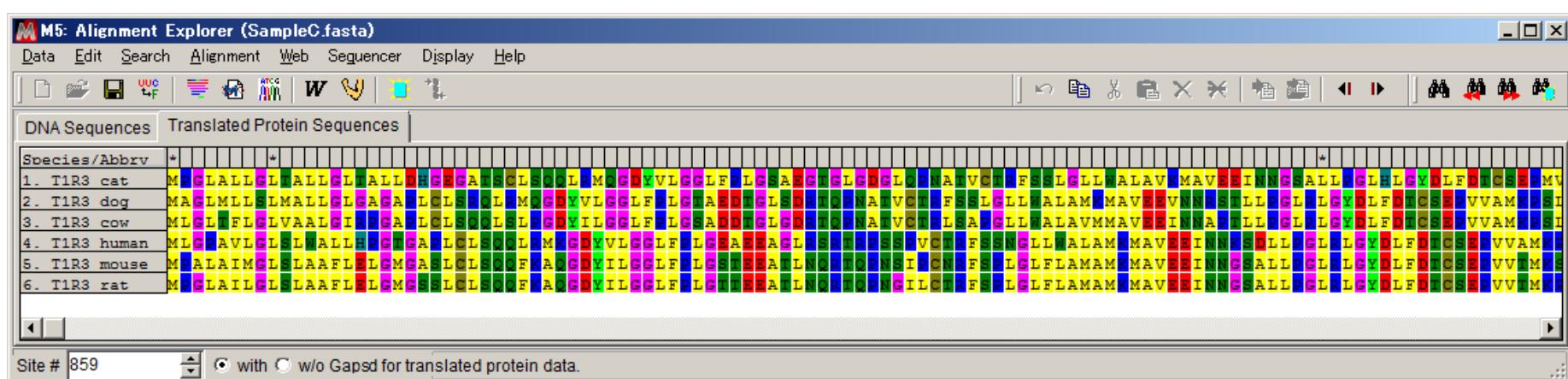
脊椎動物のミトコンドリア配列  
ではないので「No」

# アミノ酸に翻訳



脊椎動物の核ゲノム配列は  
「Standard」にチェックを入れ  
「OK」

# 变换



# アミノ酸に翻訳

The figure shows two windows from the BioEdit software. The left window, titled 'M5: Alignment Explorer (SampleC.fasta)', displays a DNA sequence alignment with six entries labeled 1 through 6. The sequences are color-coded by nucleotide (A, T, C, G). A red box highlights the first column, which starts with 'M' (methionine) for all entries, indicating it is a start codon. A blue box highlights the last column of each sequence, which ends with 'TAA' (stop codon). The right window shows the corresponding protein translation, where the amino acid sequence is aligned with the DNA. A red circle highlights a stop codon ('TAA') in the middle of a codon, while three blue circles highlight the start codons ('ATG') at the beginning of each protein sequence.

6種で保存

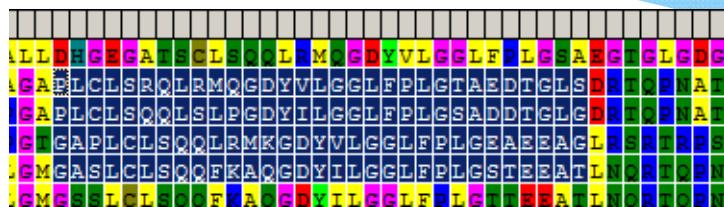
5'末端がM(メチオニン:開始コドン)

コドンの途中で途切れている

終始コドン

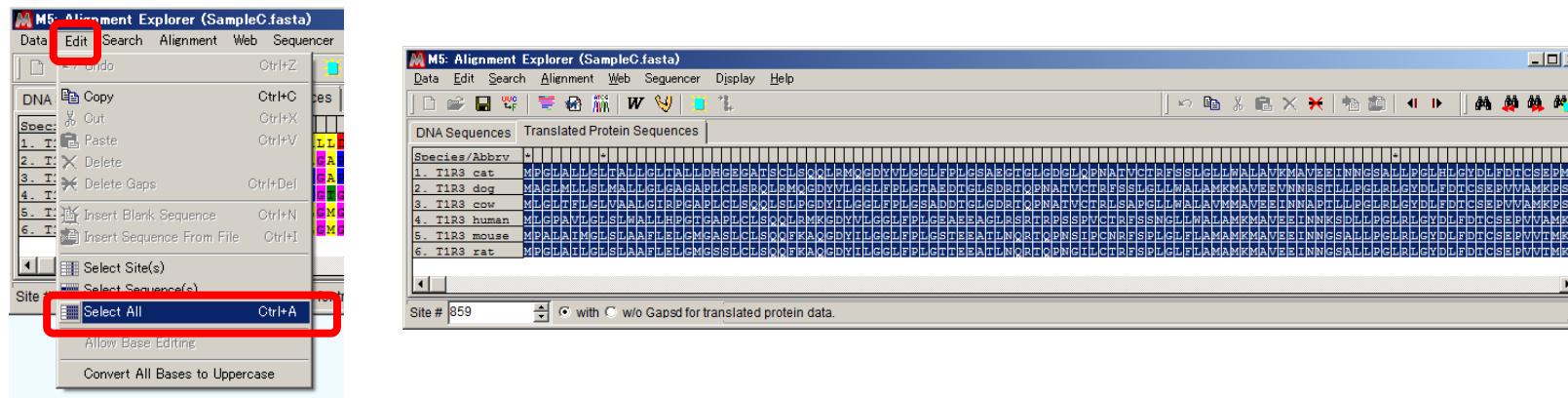
アミノ酸配列も左寄せで記述されていて、  
欠失や挿入が考慮されていません。  
約860bpの手作業によるアラインメントは時間がかかります。

# アミノ酸の自動アラインメント

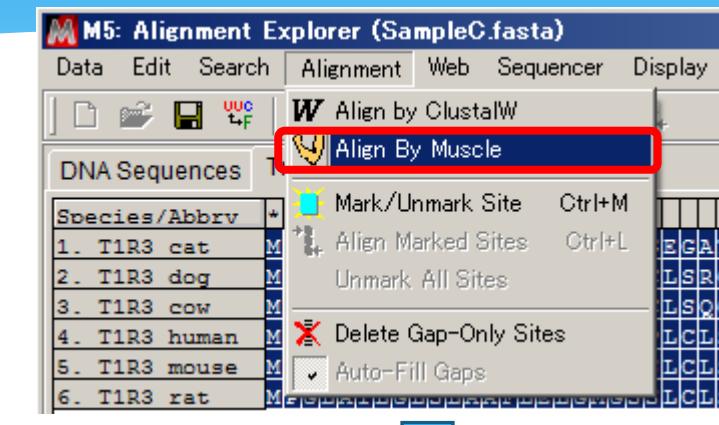


自動アラインメントは  
マウスで選択した  
「色が反転している範囲」だけ  
実施される。

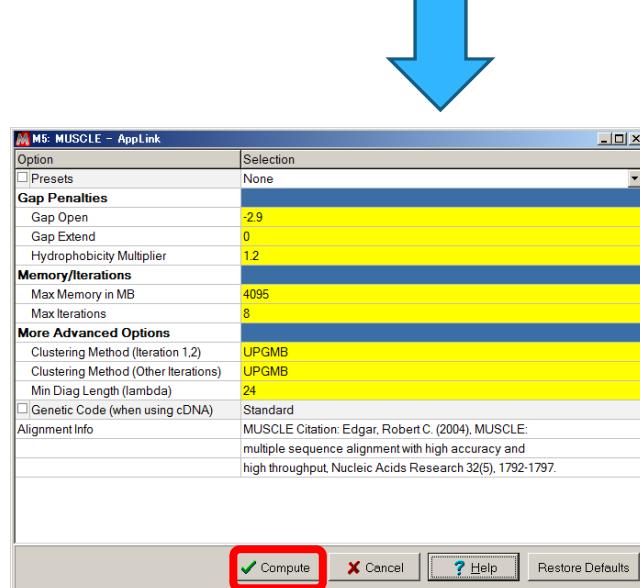
「Edit」→「Select All」で全体を選択



# アミノ酸の自動アラインメント

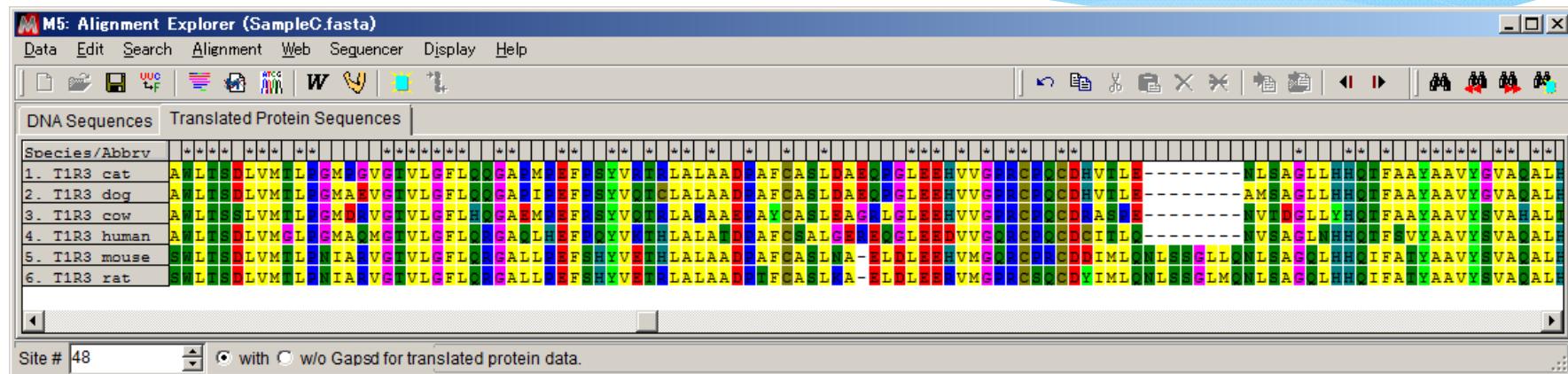


「Alignment」  
→「Align by Muscle」を選択



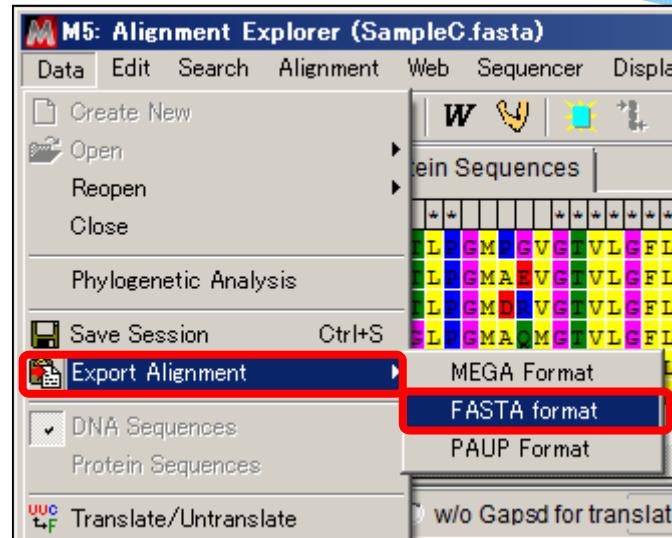
Muscleのアラインメントパラメータ設定  
→デフォルトのままで「Compute」をクリック

# アミノ酸の自動アラインメント



インデルが挿入されて、相同箇所が揃えられました。

# 自動アラインメント結果の保存



「Data」

→「Export Alignment」

→「FASTA format」を選択

→「T1R3amino」と名前を付けて保存

Fasta形式の **T1R3amino.fas** が保存されます。

※このファイルは後で系統樹構築に使用します。

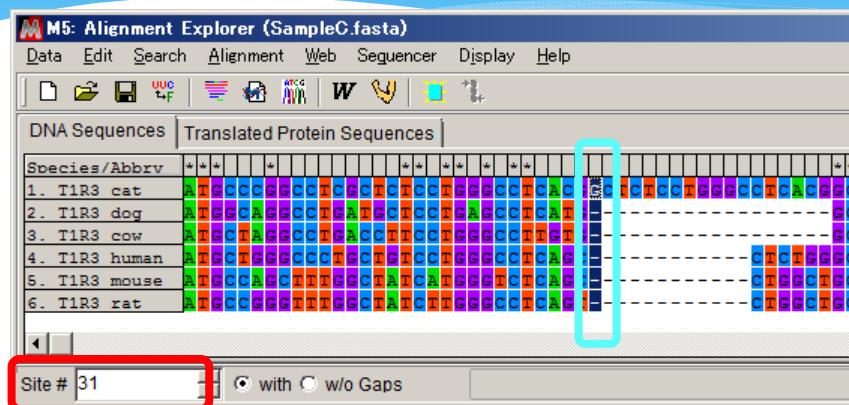
# 保存領域の図示

「DNA Sequences」タブをクリックすると、  
アラインメント済みの塩基配列が表示されます。

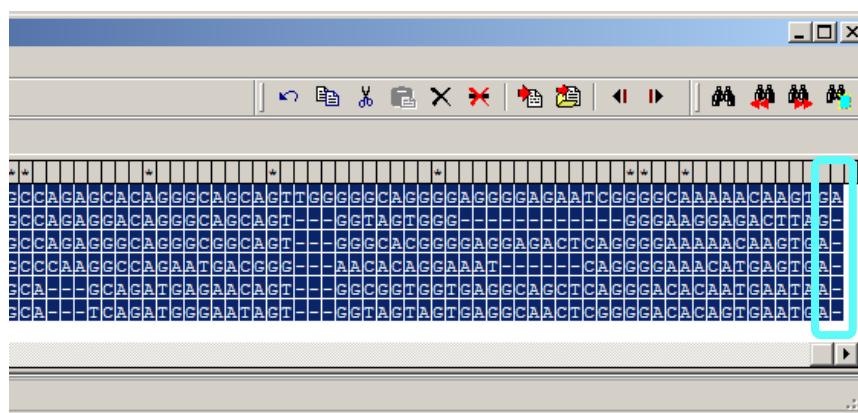


1～30bpの中で、その配列の6種での保存度を  
「WebLogo」サービスを利用して図示します。  
→そのためには1～30bpだけのFasta配列入手します。

# 不要配列の削除



①31番目の縦一列を選択



②「Shiftを押しながら」  
右端(2605番目)までドラッグ

# 不要配列の削除

③31番目以降が選択された状態

M5: Alignment Explorer (SampleC.fasta)

Data Edit Search Alignment Web Sequencer Display Help

DNA Sequences | Translated Protein Sequences

Species/Abbrv: 1. TLR3 cat, 2. TLR3 dog, 3. TLR3 cow, 4. TLR3 human, 5. TLR3 mouse, 6. TLR3 rat

Site # 2605 with w/o Gaps

Species/Abbrv	1. TLR3 cat	2. TLR3 dog	3. TLR3 cow	4. TLR3 human	5. TLR3 mouse	6. TLR3 rat
	A T G C C C S S C C T C G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S C C T C G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S C C T C G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S C C T C G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S C C T C G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S C C T C G C T C I C C T S S C C T C A C G G C T C T C T G G A C C
	A T G G C A G G C C T G A T G C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G G C A G G C C T G A T G C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G G C A G G C C T G A T G C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G G C A G G C C T G A T G C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G G C A G G C C T G A T G C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G G C A G G C C T G A T G C I C C T S S C C T C A C G G C T C T C T G G A C C
	A T G C T A G G C C T G A C C T I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T A G G C C T G A C C T I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T A G G C C T G A C C T I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T A G G C C T G A C C T I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T A G G C C T G A C C T I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T A G G C C T G A C C T I C C T S S C C T C A C G G C T C T C T G G A C C
	A T G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C
	A T G C C C A G G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C A G G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C A G G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C A G G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C A G G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C A G G C T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C
	A T G C C C S S G G T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S G G T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S G G T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S G G T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S G G T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C	A T G C C C S S G G T T G G G C C T G C T C I C C T S S C C T C A C G G C T C T C T G G A C C

④選択範囲を右クリック →Delete

Translated Protein Sequences

Copy Ctrl+C

Cut Ctrl+X

Paste Ctrl+V

**Delete**

Delete Gaps Ctrl+Del

Select Site(s)

Select Sequence(s)

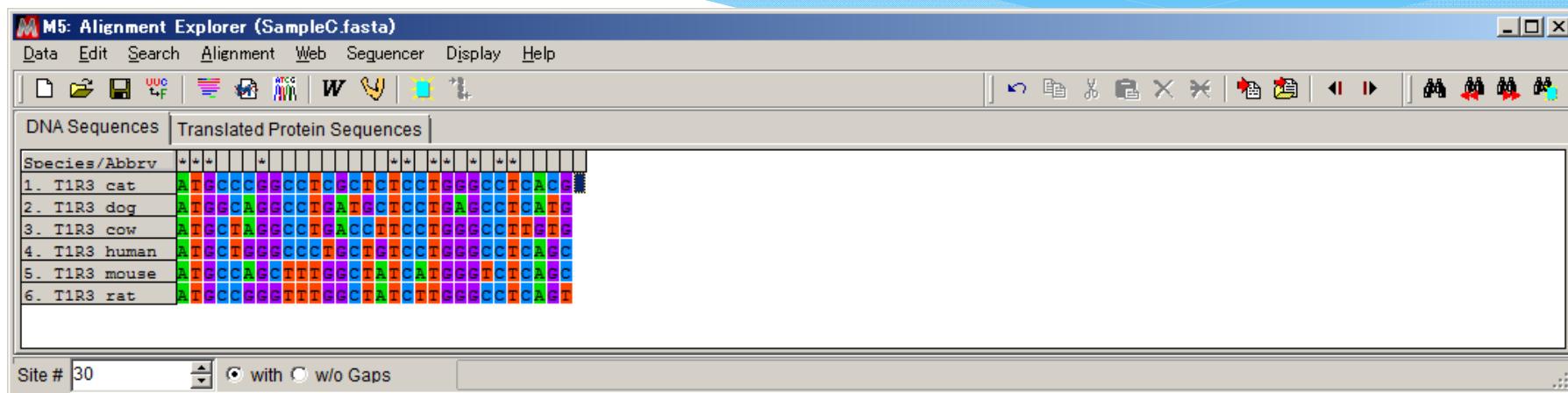
Select All Ctrl+A

Undo Ctrl+Z

with w/o Gaps

# 不要配列の削除

⑤1～30bpのみの配列の完成



⑥「Data」

→「Export Alignment」

→「FASTA format」を選択

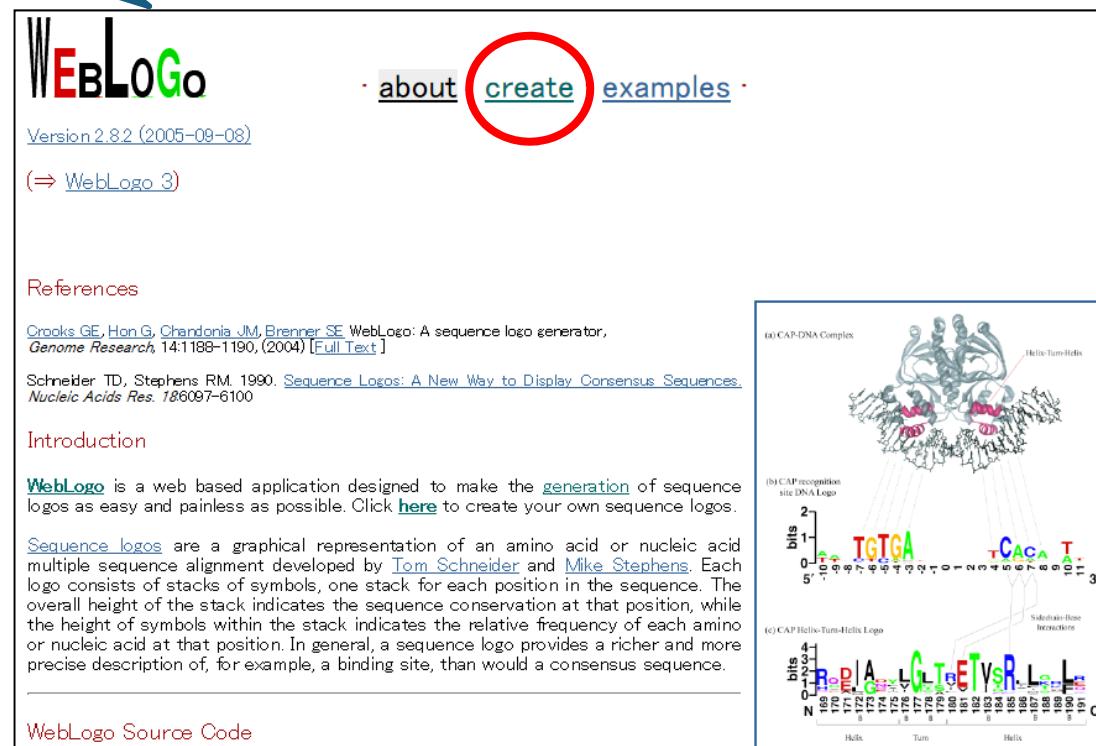
→「T1R3nuc30」と名前を付けて保存

Fasta形式の **T1R3nuc30.fas** が保存されます。

# WebLogo

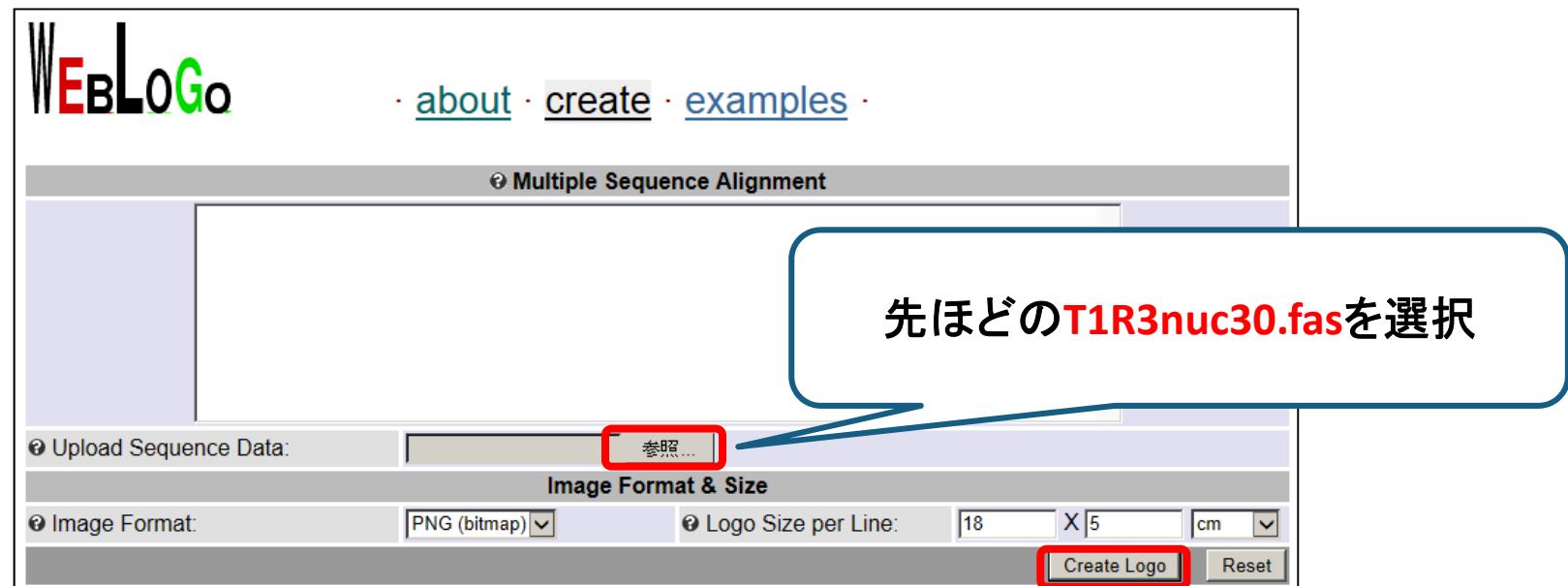
「WebLogo」で検索

<http://weblogo.berkeley.edu/>

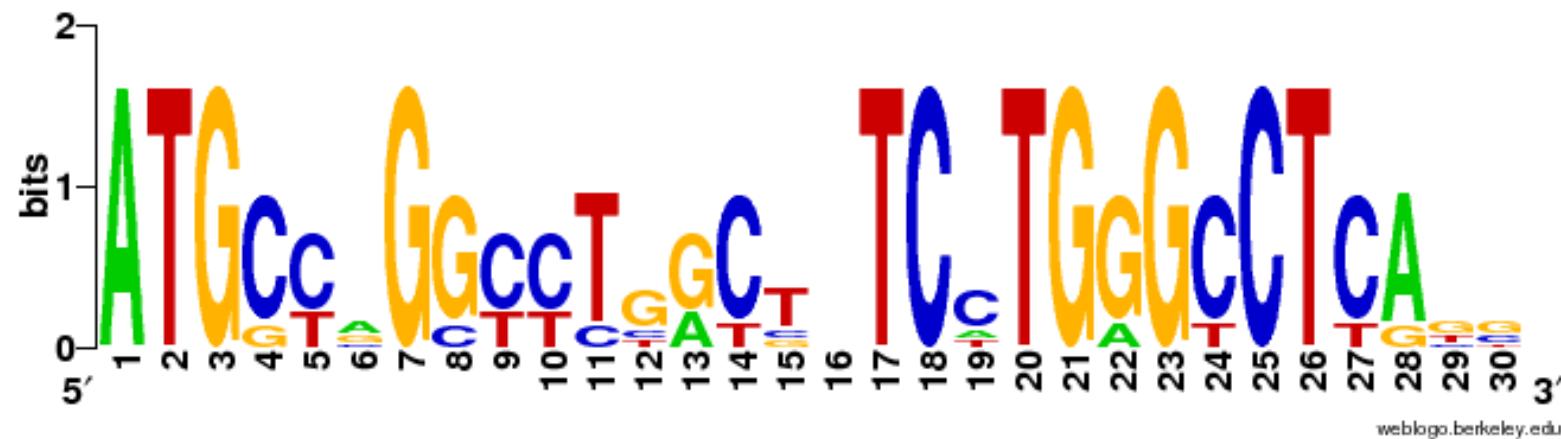


The screenshot shows the WebLogo homepage. At the top, there is a navigation bar with links for 'about', 'create' (which is circled in red), and 'examples'. Below the navigation bar, there is a section titled 'References' with two entries: one by Crooks et al. (2004) and another by Schneider and Stephens (1990). On the right side of the page, there are three small diagrams labeled (a), (b), and (c), which are examples of sequence logos. Diagram (a) shows a CAP-DNA complex with a helix-turn-helix motif. Diagram (b) shows a CAP recognition site DNA logo with a sequence TGTCA. Diagram (c) shows a CAP Helix-Turn-Helix logo with a sequence GATTC.

# 保存領域の図を作成する



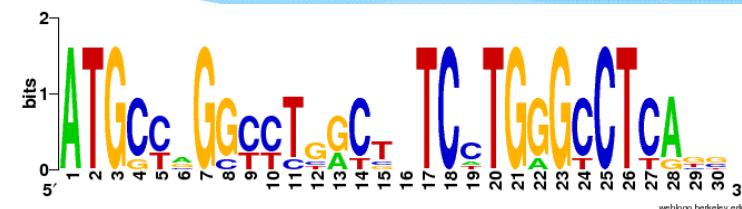
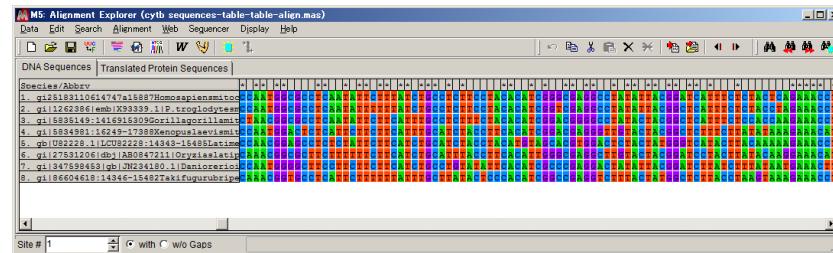
# WebLogoで保存領域を図示



T1R3配列の1~30bpにおける保存領域の図  
→サイトごとに縦に文字が長いほど保存されている

※ブラウザから右クリックで画像を保存できます。

# マルチプルアラインメントを元にした 配列比較のまとめ



- ・短い領域の配列比較や保存領域の確認は  
アラインメントやWebLogoの表示で可能

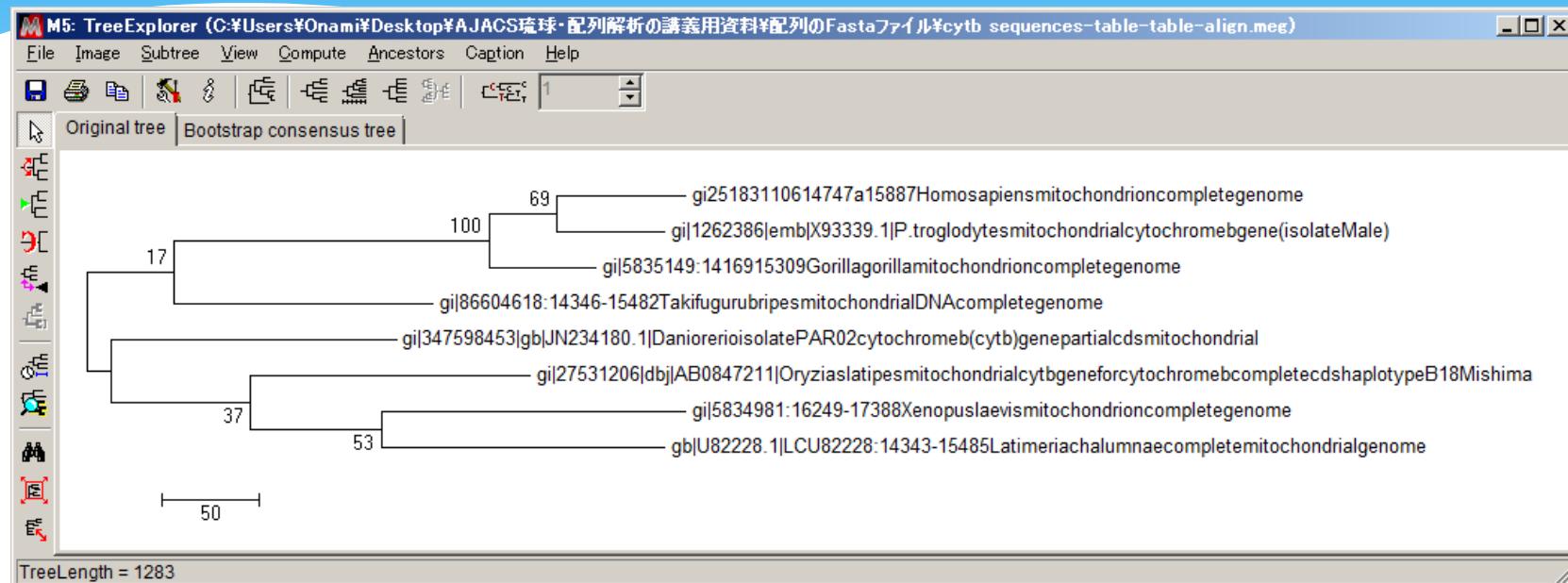
では、配列間の遺伝的距离や関係を示すには？

→分子系統樹

# 5. 系統樹構築

## ~近隣結合法と最尤法

# 系統樹



配列間の、遺伝的距离と統計的信頼性を  
記述する方法

# 系統樹推定の手法

- ・**距離行列法**

遺伝的距離が最小となる系統樹を候補とする。UPGMA法や近隣結合法を含む。  
進化速度が座位間で不均質な場合は向き。高速。

- ・**最大節約法**

データを説明するために系統樹全体で必要とする最小限の置換数が最も少なくて済むような系統樹を候補とする。高速。

- ・**最尤法**

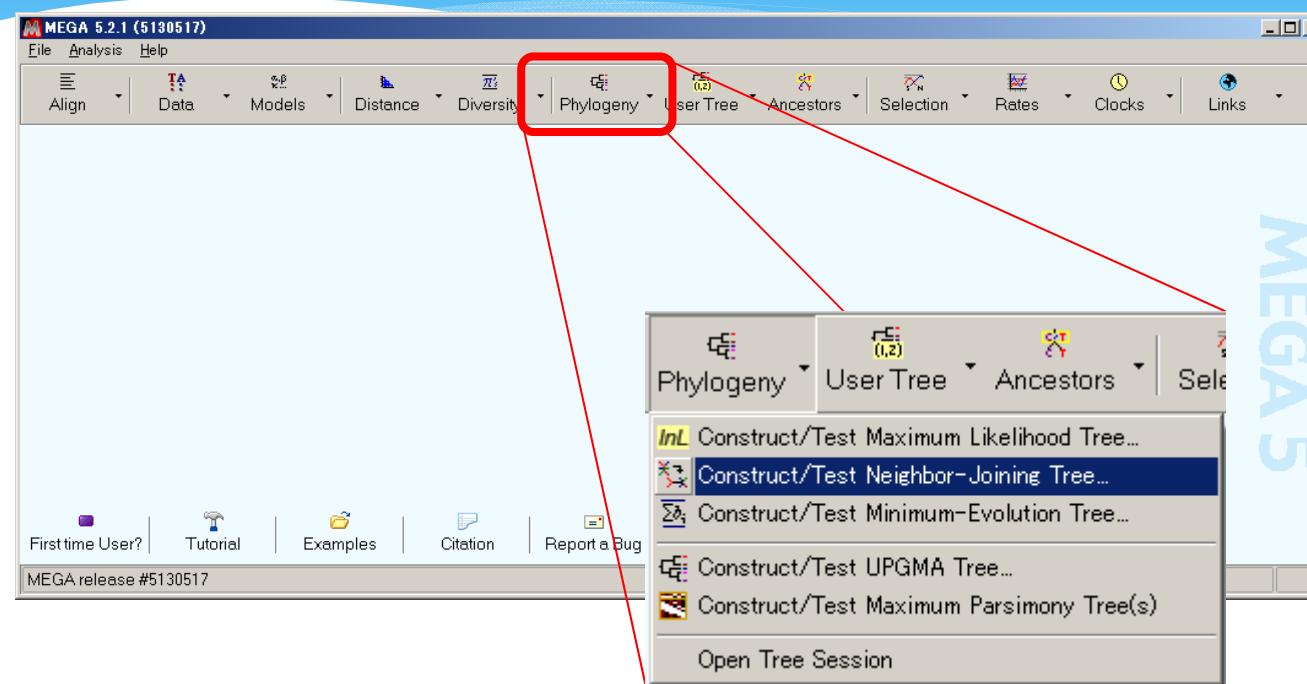
塩基・アミノ酸置換のモデルを明示した上で、そのモデルが実際の進化の過程の近似になっているかを評価基準とする。詳細な設定が可能で信頼度が高い。時間がかかる。

MEGAではこれら全ての手法が実施可能。生命科学の分野では近隣結合法(高速)や最尤法(低速、信頼度高)が頻繁に採用されている。

適切な手法・パラメータ選択と、現象の厳密な理解のため、  
ぜひ本スライド末の参考文献をご参照ください。

# T1R3アミノ酸配列の系統樹構築

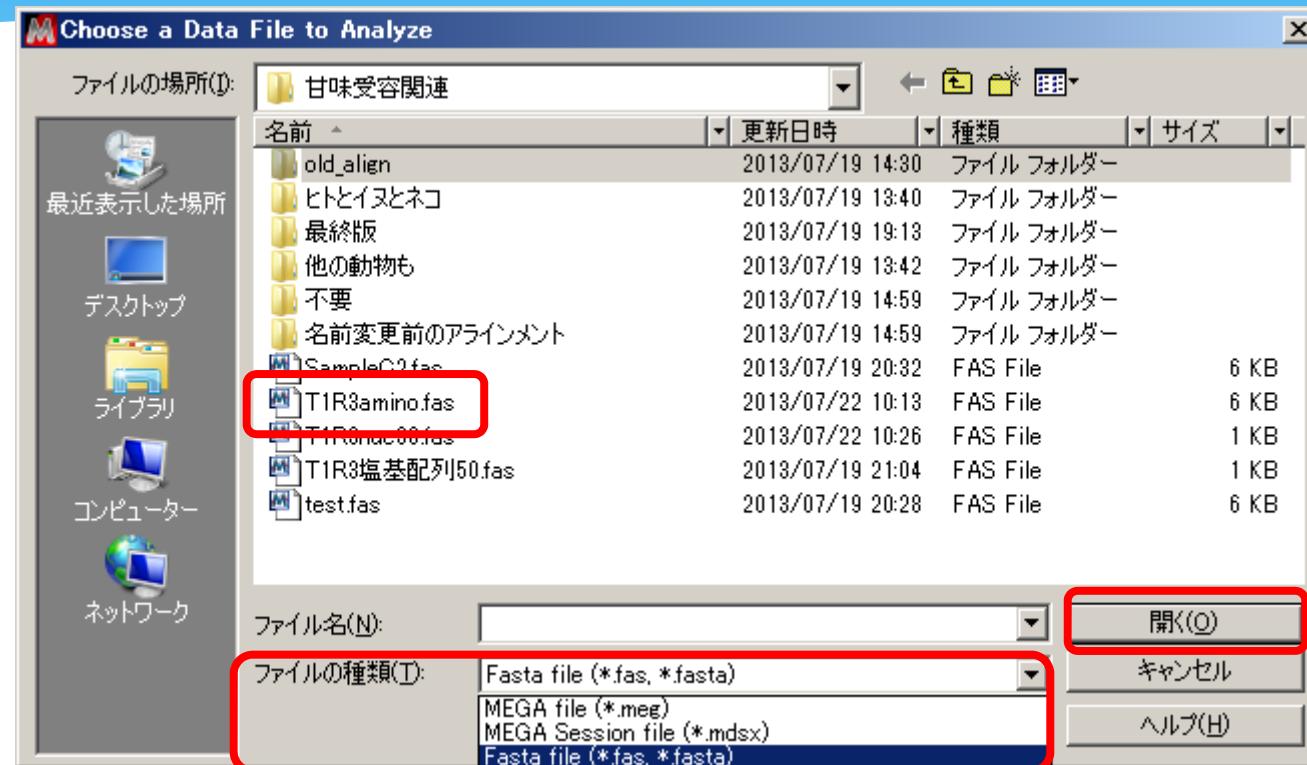
## 1. 近隣結合法



- ①MEGAメインウィンドウから「Phylogeny」ボタン  
→「Construct/Test Neighbor-Joining Tree」を選択

# T1R3アミノ酸配列の系統樹構築

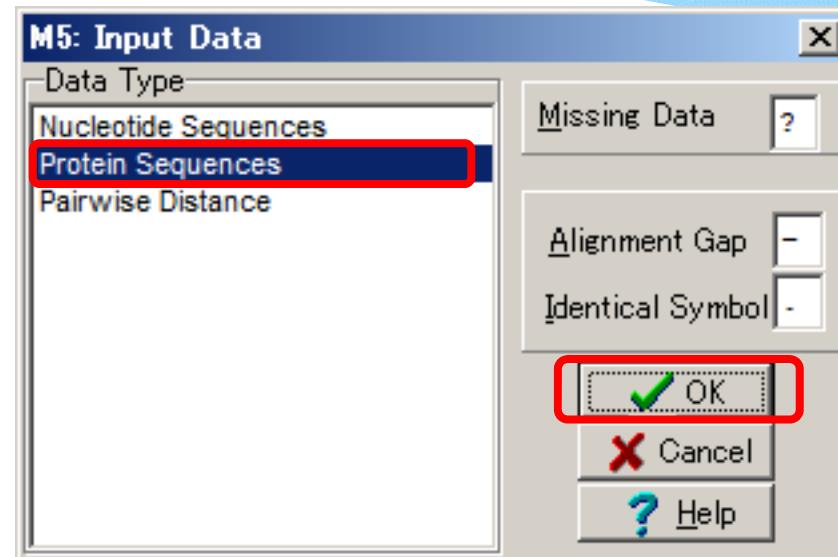
## 1. 近隣結合法



- ②ファイルの種類:「**Fasta file (\*.fas, \*.fasta)**」を選択し、先ほど保存した**T1R3amino.fas**を選択する。

# T1R3アミノ酸配列の系統樹構築

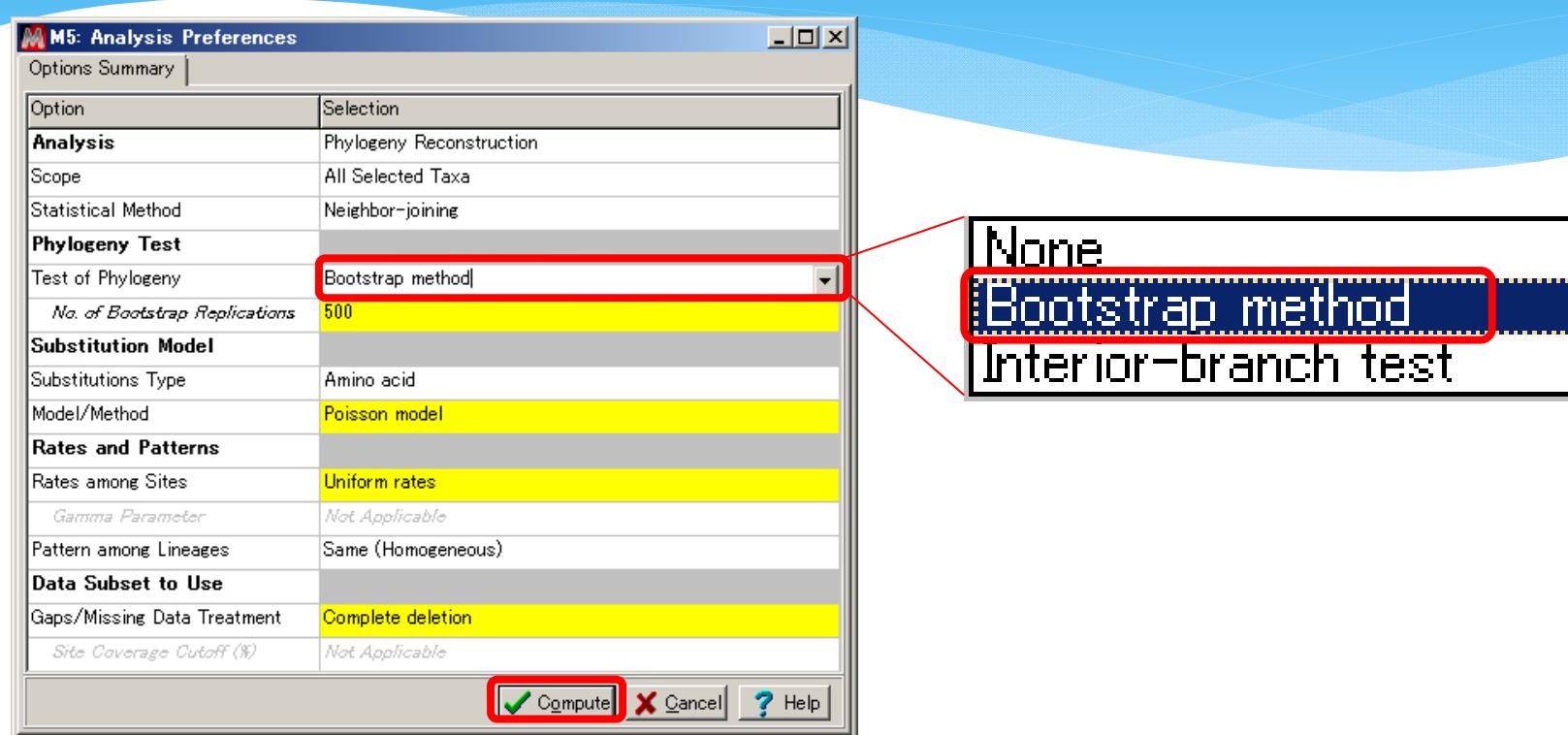
## 1. 近隣結合法



③Data Type: Protein Sequencesを選択し、「OK」

# T1R3アミノ酸配列の系統樹構築

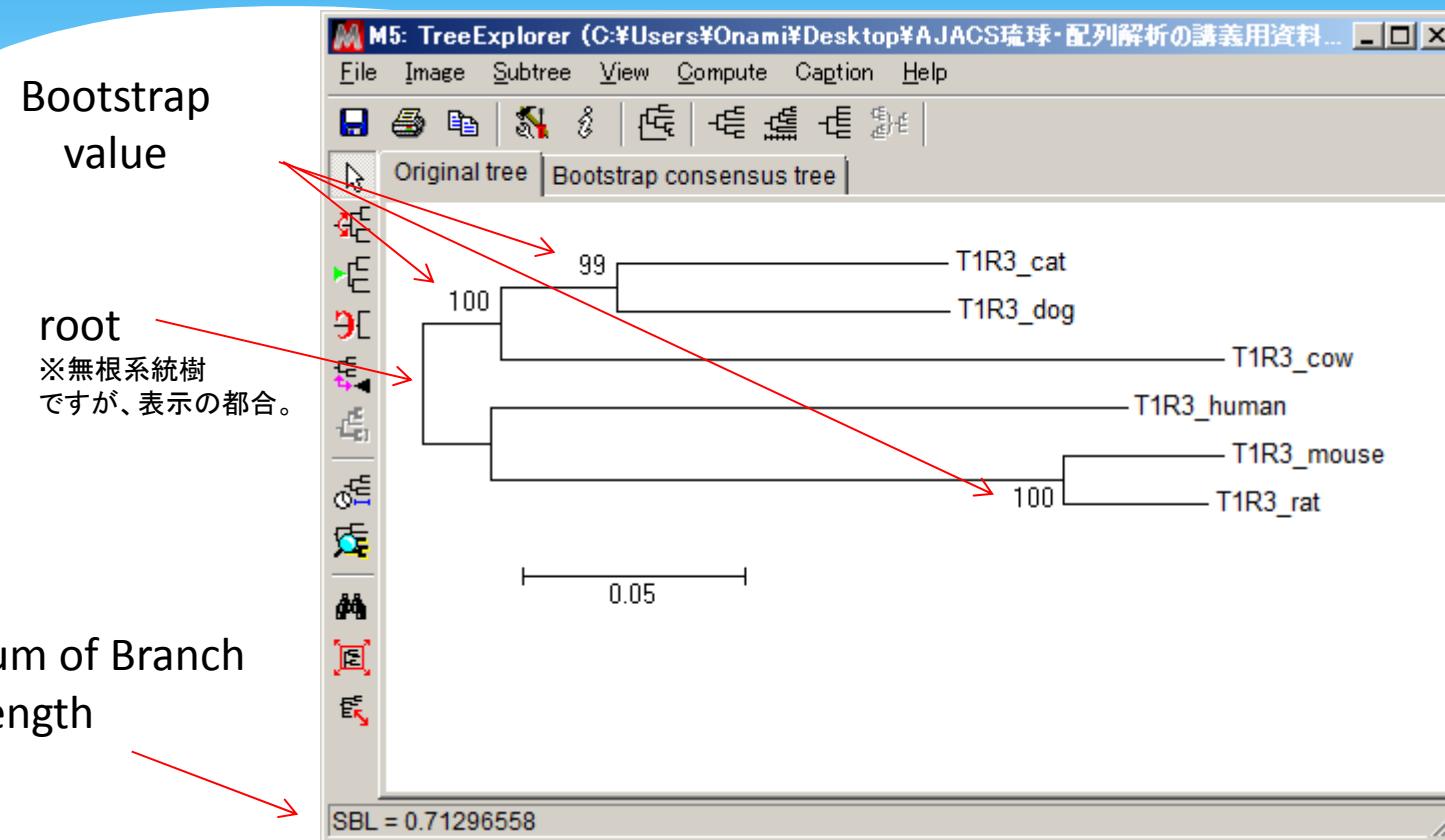
## 1. 近隣結合法



- ④ Test of Phylogeny: Bootstrap methodを選択し、「Compute」

# T1R3アミノ酸配列の系統樹構築

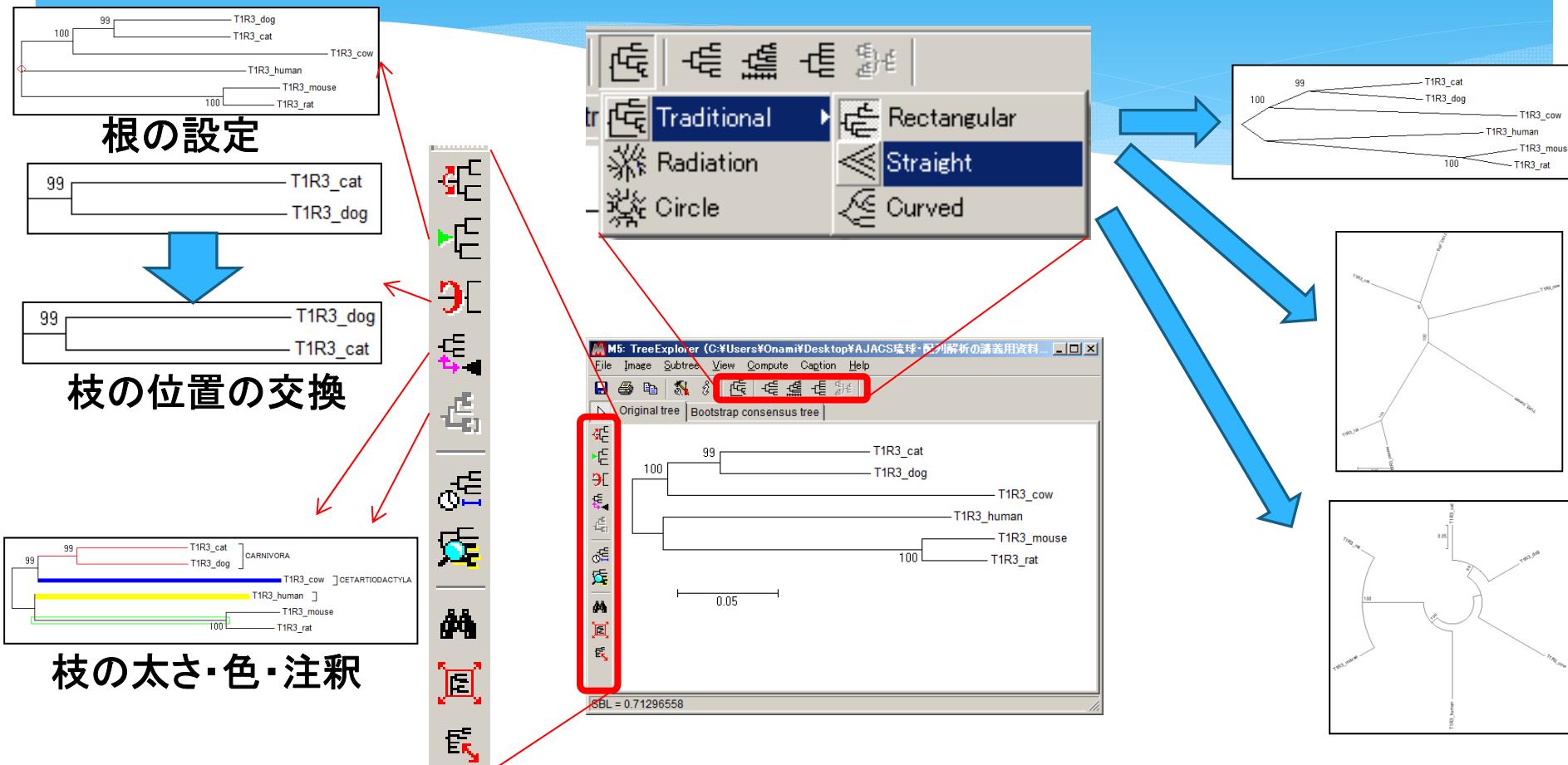
## 1. 近隣結合法



各枝の分岐が99%以上のBootstrap確率値で支持される  
近隣結合系統樹を構築することができました。

# T1R3アミノ酸配列の系統樹構築

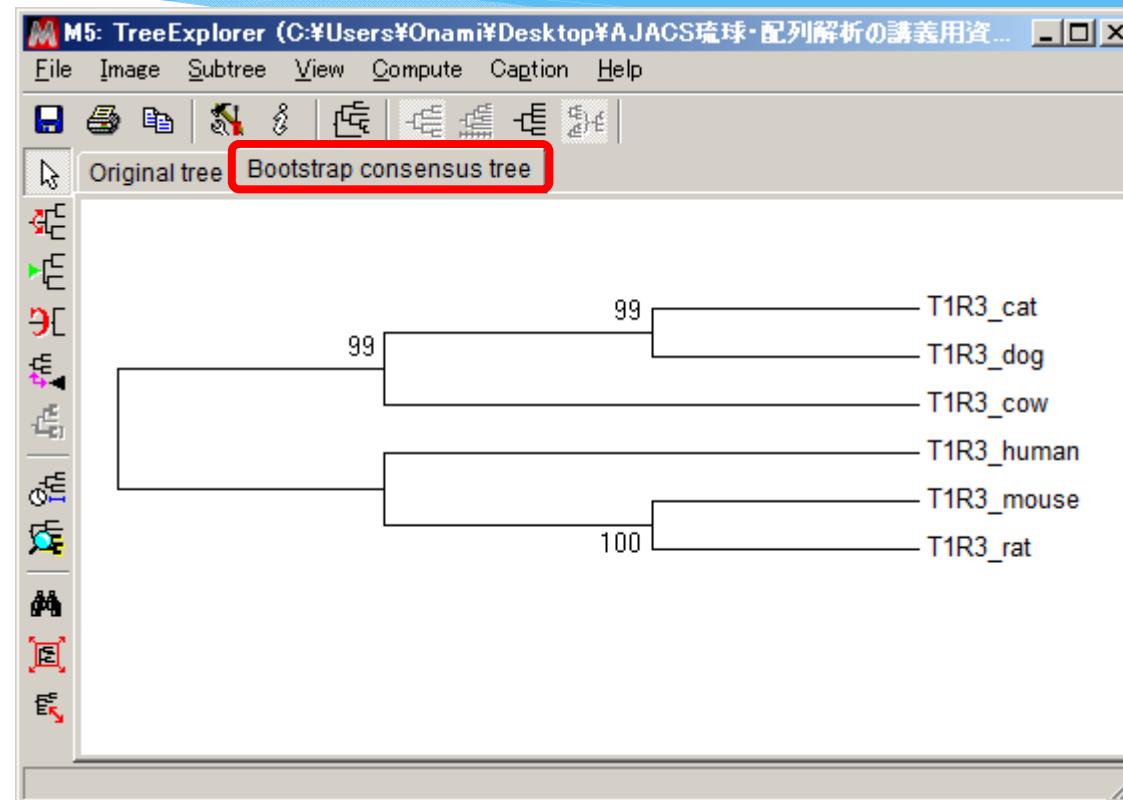
## 1. 近隣結合法



様々な表現の系統樹を作成することが可能

# T1R3アミノ酸配列の系統樹構築

## 1. 近隣結合法



「Bootstrap consensus tree」タブ  
→信頼性のある系統樹のトポロジー

# T1R3アミノ酸配列の系統樹構築

## 1. 近隣結合法

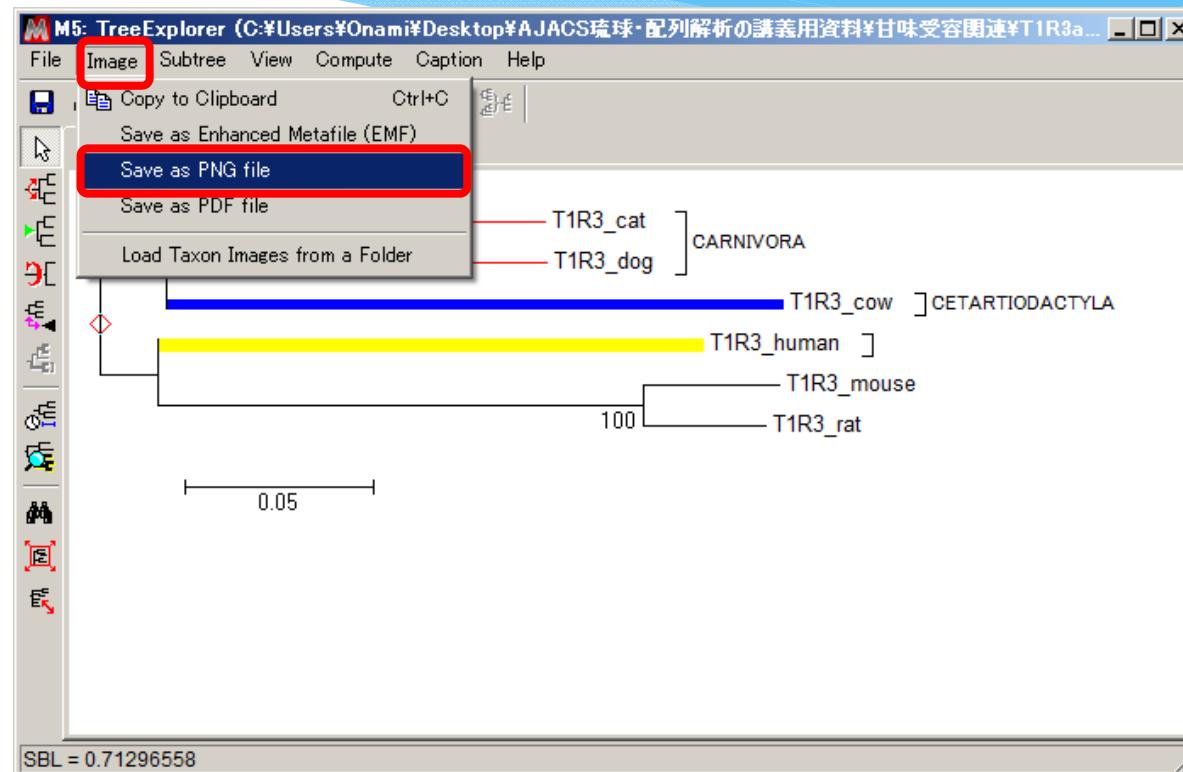


Image から「Save as PNG file」,「Save as PDF file」で  
系統樹をPNGやPDFファイルとして保存可能

# T1R3アミノ酸配列の系統樹構築

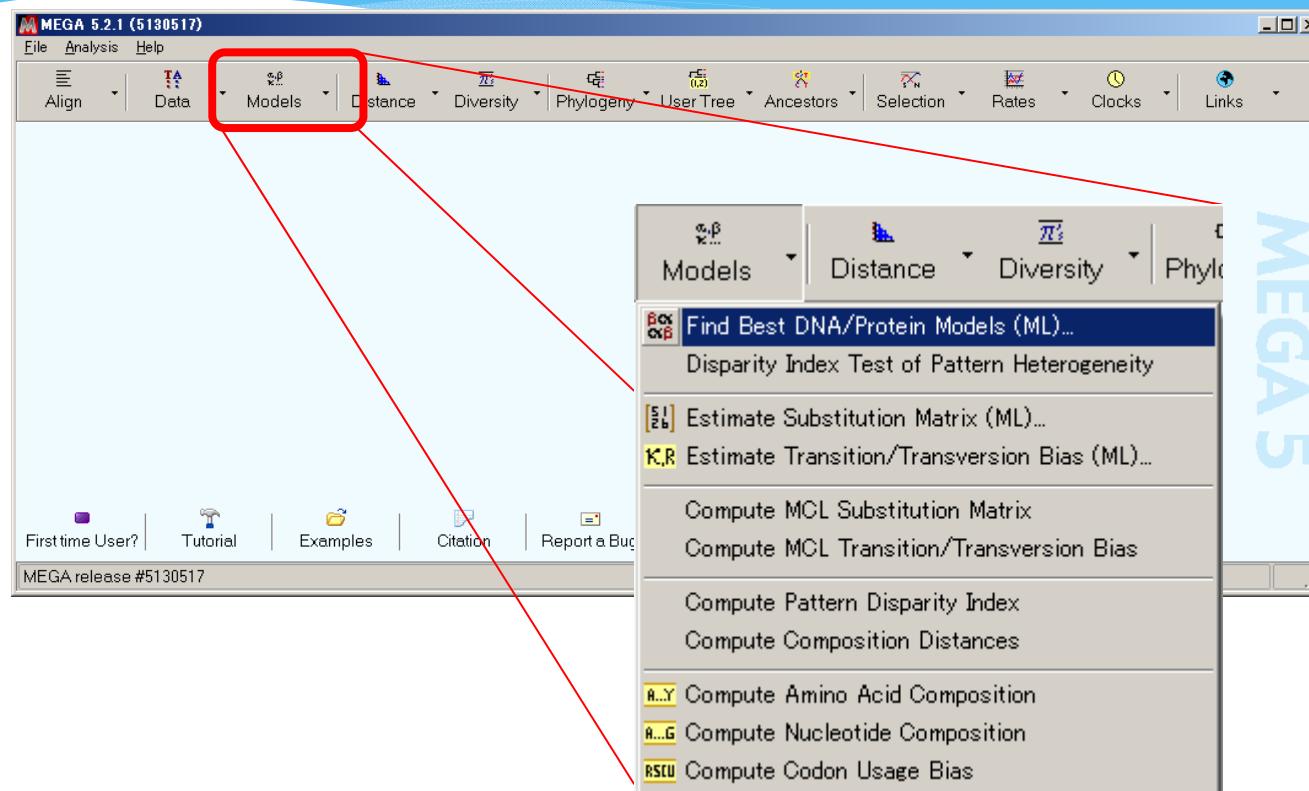
系統樹の信頼性をより高めるため、  
最尤法系統樹構築を行います(計算に時間がかかります)

- 最尤法の流れ

1. 置換モデルの選択
2. 選択したモデルを利用した系統樹の推定

# T1R3アミノ酸配列の系統樹構築

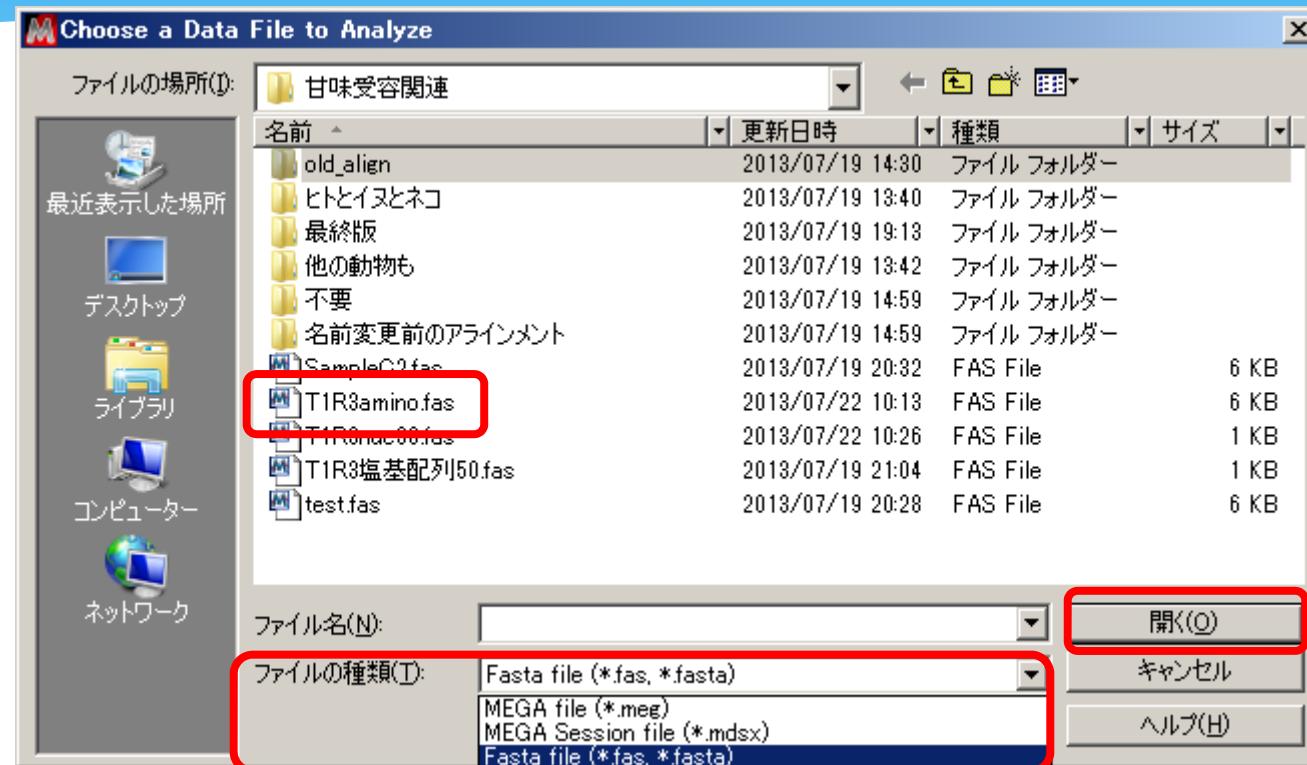
## 2. 最尤法(モデル選択)



- ①MEGAメインウインドウから「Models」ボタン  
→「Find Best DNA/Protein Model(ML)」を選択

# T1R3アミノ酸配列の系統樹構築

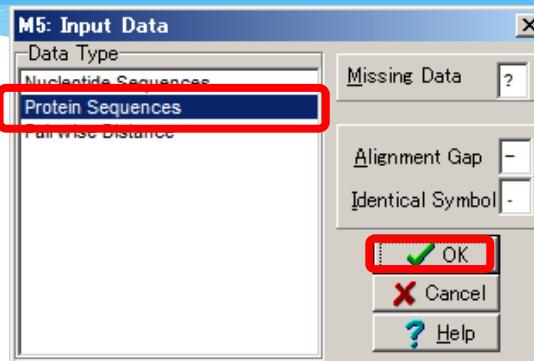
## 2. 最尤法(モデル選択)



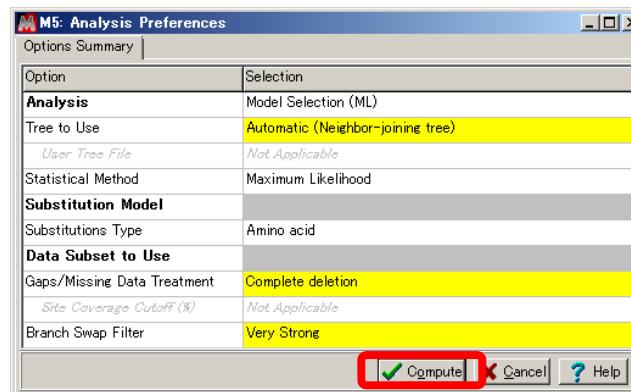
- ②ファイルの種類:「**Fasta file (\*.fas, \*.fasta)**」を選択し、先ほど保存した**T1R3amino.fas**を選択する。

# T1R3アミノ酸配列の系統樹構築

## 2. 最尤法(モデル選択)



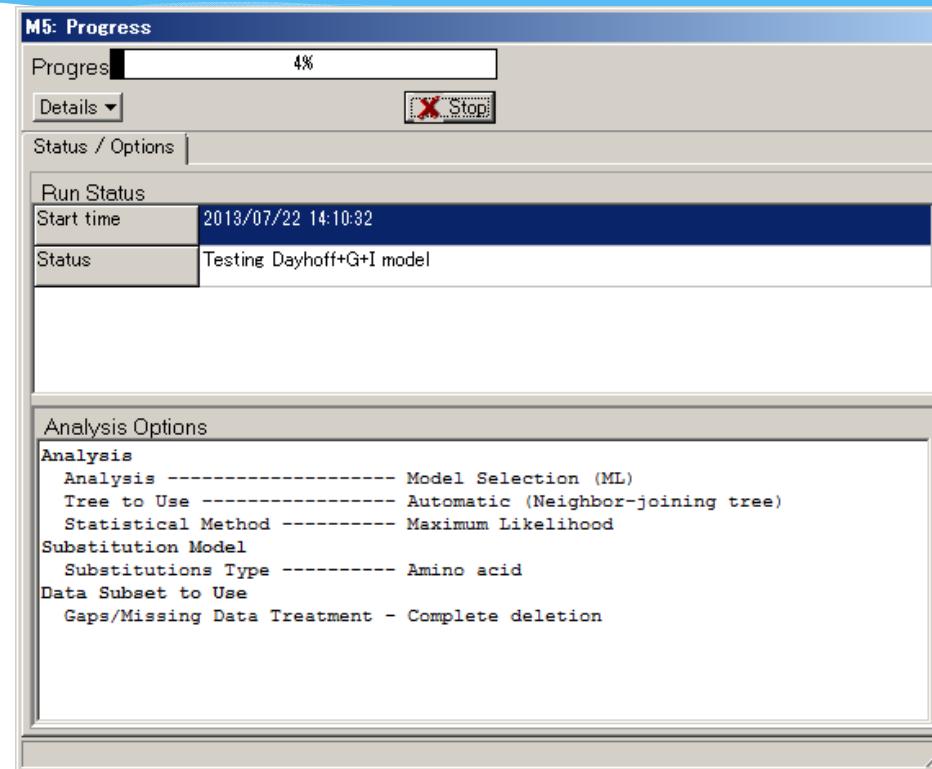
③Data Type: Protein Sequencesを選択し、「OK」



④そのまま「Compute」

# T1R3アミノ酸配列の系統樹構築

## 2. 最尤法(モデル選択)



⑤置換モデルの評価開始。  
(本件では2分程度かかります)

# T1R3アミノ酸配列の系統樹構築

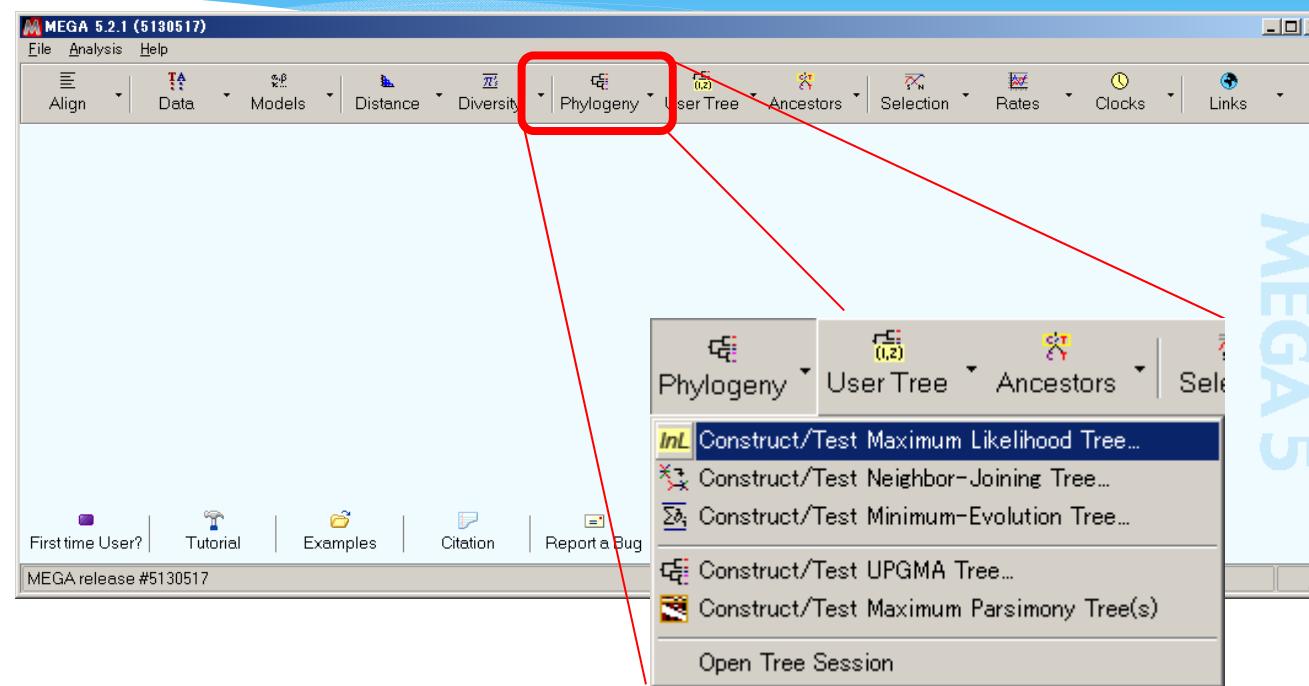
## 2. 最尤法(モデル選択)

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

Model	Parameters	BIC	AICc	<i>lnL</i>	(+/-)
JTT+G	10	11360.319	11295.183	-5637.569	n/a
JTT+G+I	11	11368.837	11297.191	-5637.569	0.0
JTT+I	10	11376.298	11311.162	-5645.559	0.3
JTT+G+F	29	11393.865	11205.193	-5573.422	n/a
JTT+G+I+F	30	11402.383	11207.217	-5573.422	0.0
JTT	9	11478.795	11420.169	-5701.067	n/a
WAG+G	10	11500.319	11295.183	-5637.569	n/a
WAG+G+I	11	11508.837	11297.191	-5637.569	0.0
WAG+I	10	11506.319	11411.063	-5710.509	n/a
WAG+O	10	11506.319	11319.059	-5630.355	n/a
nsREV+G+F	29	11507.771	11319.059	-5630.355	n/a
WAG+O+I+F	30	11513.839	11318.673	-5629.150	0.0
WAG+I+O	10	11514.319	11420.659	-5710.509	n/a
nsREV+G+I+F	29	11516.540	11321.083	-5630.355	n/a
WAG+I+O+F	30	11516.540	11321.083	-5630.355	n/a
WAG+I+F	29	11517.951	11319.754	-5641.714	n/a
JTT+I+F	30	11517.951	11319.754	-5641.714	n/a
WAG+I	10	11522.805	11457.727	-5718.541	0.9
Daetus+C+F	29	11522.805	11339.140	-5646.395	n/a
Daetus+C+I+F	30	11522.805	11340.647	-5646.395	n/a
cpREV+G+F	29	11524.843	11340.747	-5645.698	n/a
Daetus+G	29	11524.843	11478.771	-5729.054	n/a
Daetus+G+F	30	11524.843	11311.771	-5645.698	n/a
cpREV+G+I+F	30	11524.843	11311.771	-5645.698	n/a
Daetus+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+F	30	11524.843	11312.060	-5645.698	n/a
cpREV+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F	30	11524.843	11312.060	-5645.698	n/a
Daetus+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+G+I+F					

# T1R3アミノ酸配列の系統樹構築

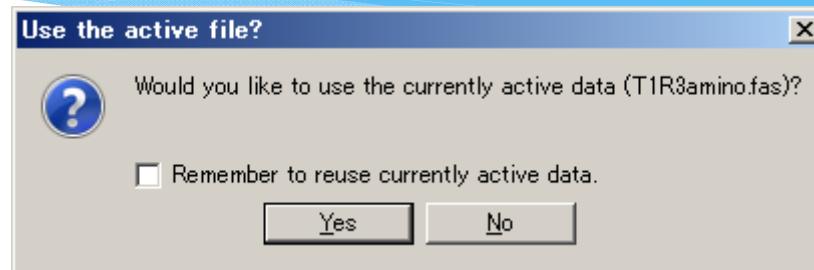
## 2. 最尤法(系統樹推定)



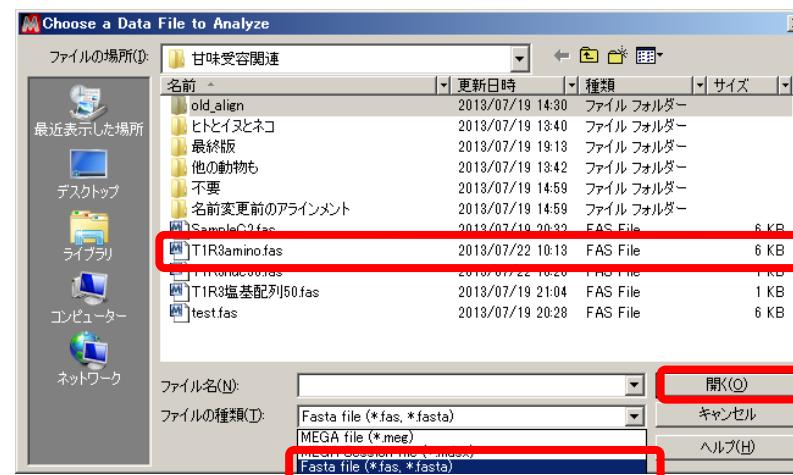
⑦置換モデルのウィンドウを消して、  
MEGAメインウィンドウから「Phylogeny」ボタン  
→「Construct/Test Maximum Likelihood Tree」を選択

# T1R3アミノ酸配列の系統樹構築

## 2. 最尤法(系統樹推定)



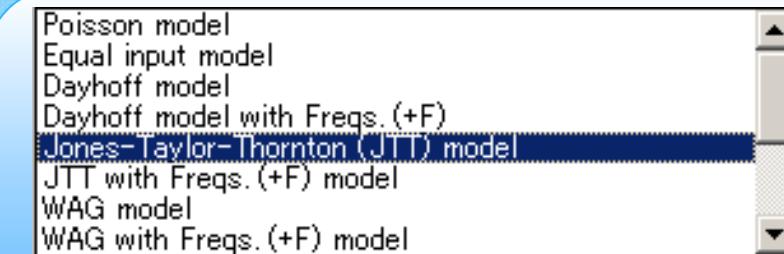
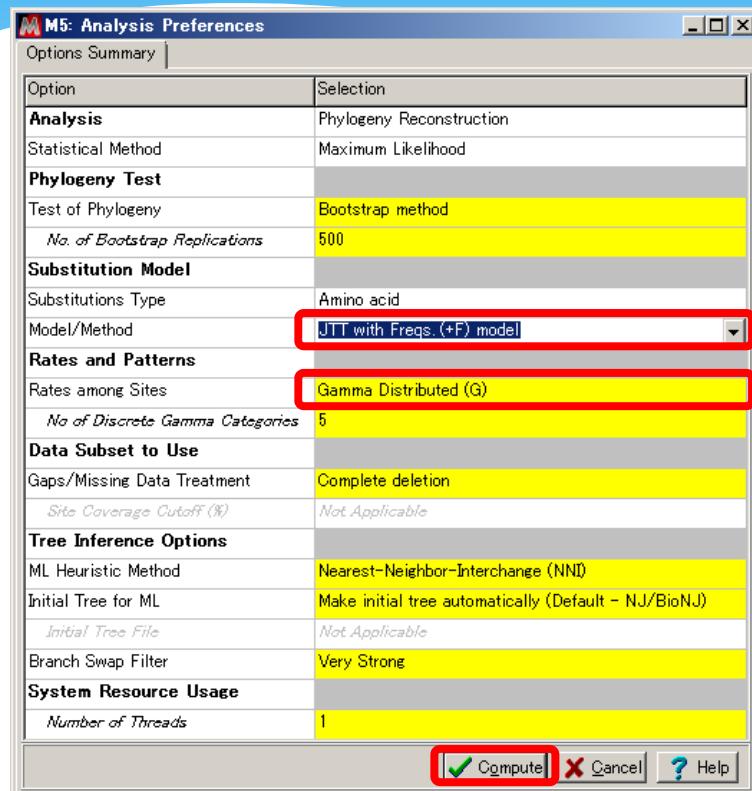
⑧現在開いているデータを利用するか聞かれた場合 →「Yes」



⑨ ⑧で聞かれなかった場合、ファイルの種類:「**Fasta file (\*.fas, \*.fasta)**」を選択し、先ほど保存した**T1R3amino.fas**を選択する。

# T1R3アミノ酸配列の系統樹構築

## 2. 最尤法(系統樹推定)



### ■ Model/Method

→「Jones-Taylor-Thornton(JTT) model」



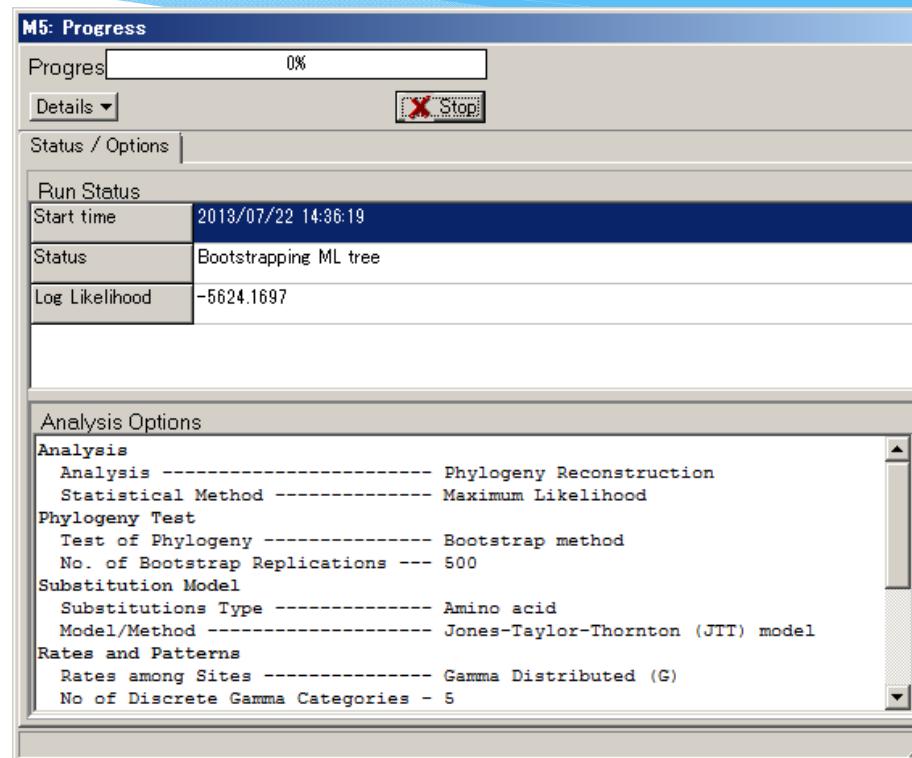
### ■ Rates among Sites

→「Gamma Distributed (G)」

- ⑩ 上記二か所のパラメータを「JTT+G」に合わせて設定し、「Compute」

# T1R3アミノ酸配列の系統樹構築

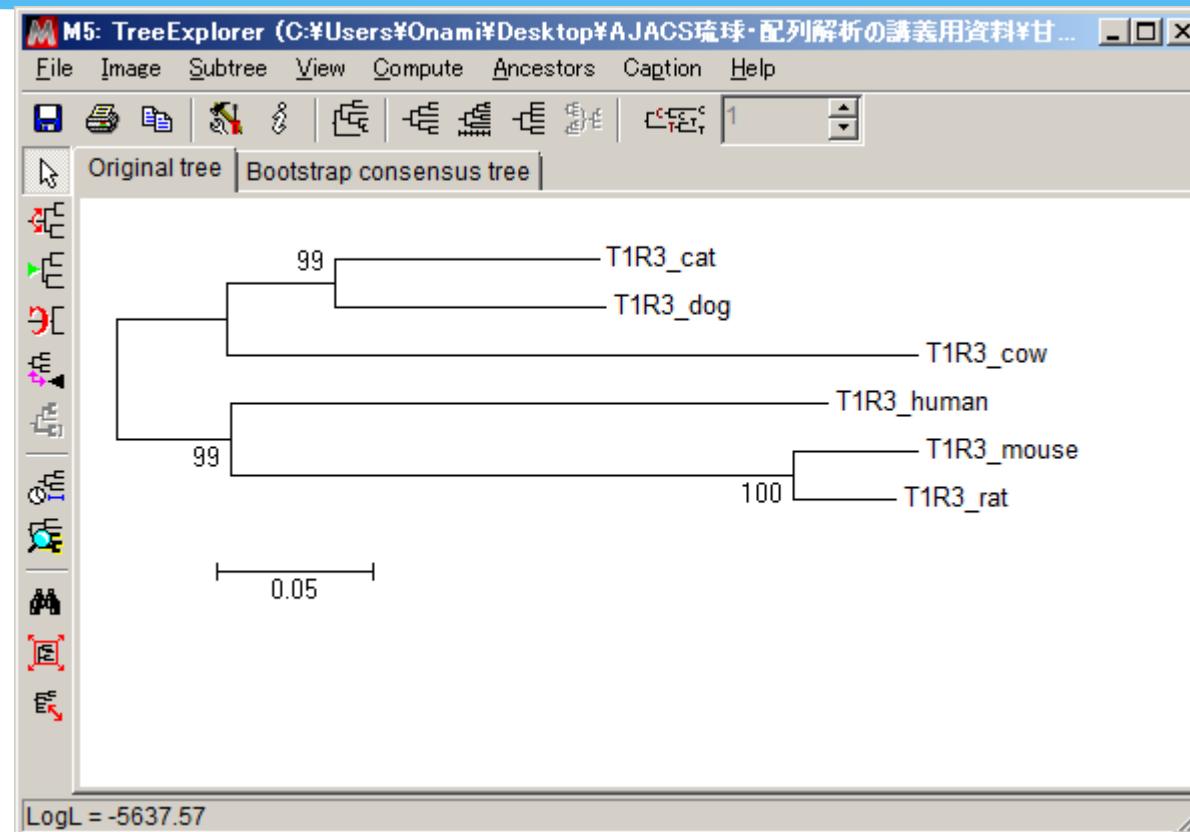
## 2. 最尤法(系統樹推定)



⑪系統樹推定開始。  
(本件では12分程度かかります)

# T1R3アミノ酸配列の系統樹構築

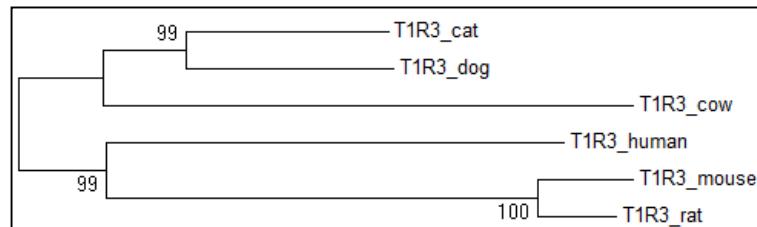
## 2. 最尤法(系統樹推定)



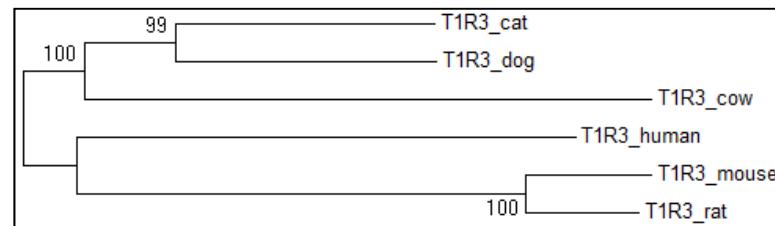
⑫各Bootstrap値が99以上の系統樹が推定できました。

# T1R3アミノ酸配列の系統樹構築

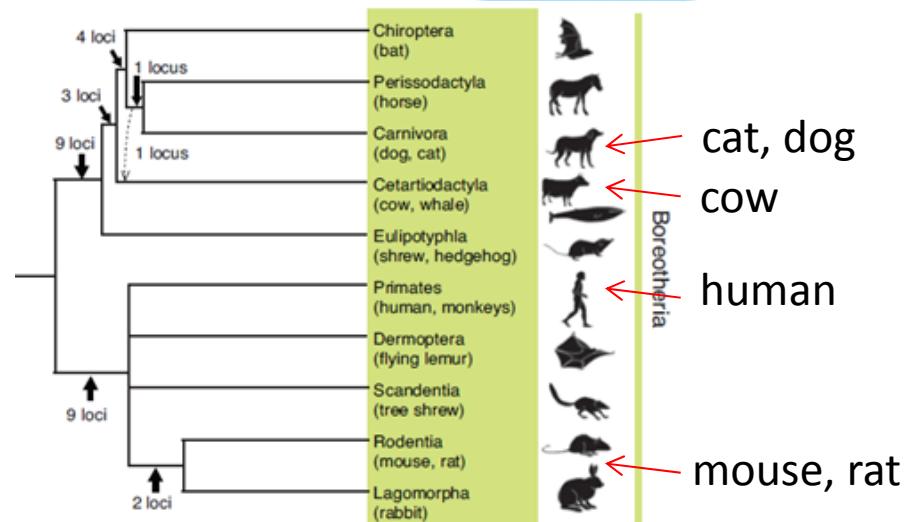
## 2. 最尤法(結果の検討)



最尤法で推定された系統樹



近隣結合法で推定された系統樹



# 系統樹構築のまとめ

配列のマルチプルアラインメント



近隣結合法・最尤法による系統樹推定

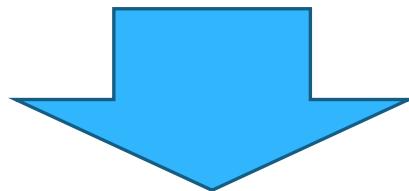
※ただし、手法やパラメータの選定、結果については、  
調べる対象や場合によって様々な考え方がある。  
(系統関係が樹状にならない場合も…)

→まずは一度MEGAのデフォルトの設定で結果を出力し、  
その情報として条件や対象の検討を行い、  
より確からしい方法に近づけていくのが良い。

# 6. 配列比較解析演習

# 【演習】配列比較解析

哺乳類では、先ほど確認したT1R3遺伝子と、  
T1R2遺伝子の転写産物が、  
二量体Gタンパク質共役型受容体を形成し、  
甘味受容に影響すると言われている。



ネコ、イヌ、ウシ、ヒト、マウス、ラットの  
T1R2遺伝子の配列解析を行い、  
T1R3とどのように違うか、確認してみましょう。

# 【演習】サンプル配列(T1R2)

## sampleD.fasta

6種の哺乳類(ネコ、イヌ、ウシ、ヒト、マウス、ラット)の核ゲノムにコードされているT1R2遺伝子ホモログ一覧

# 【演習】配列比較解析

ネコ、イヌ、ウシ、ヒト、マウス、ラットの  
T1R2遺伝子 (sampleD.fasta) の配列解析を行い、  
T1R3とどのように違うか確認してみましょう。(約10分)

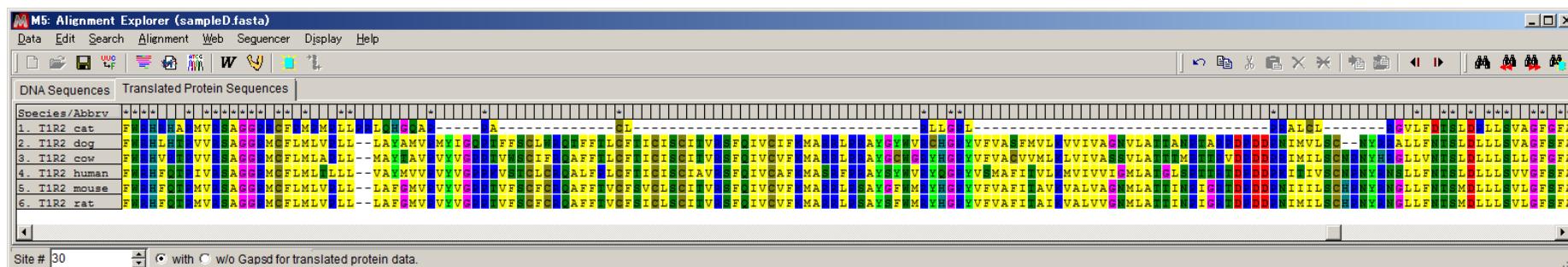
1. マルチプルアラインメントで配列をそろえてみましょう
2. 系統樹を構築してみましょう。(最尤法は時間がかかるため、近隣結合法のみ)
3. マルチプルアラインメントと系統樹の結果を検討してみましょう

【ヒント】ネコの配列に注目してください



# 【解答】1. T1R2のマルチプルアラインメント

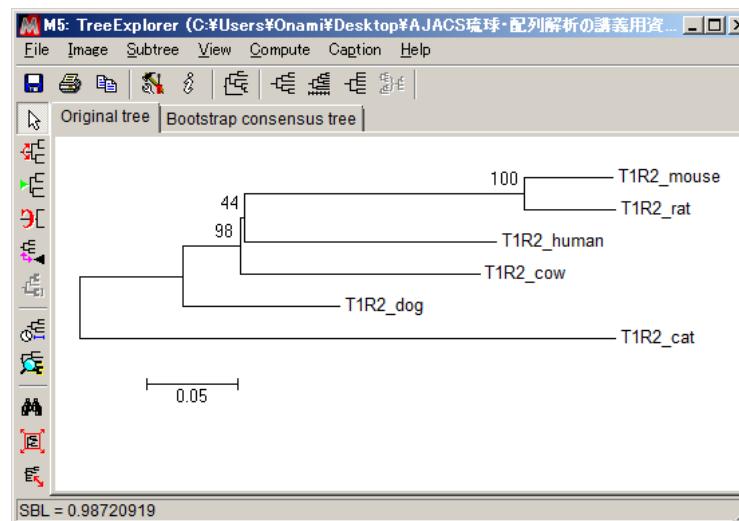
- ①MEGAを起動し、MEGAメインウィンドウから  
「Align」ボタン→「Edit/Build Alignment」を選択
- ②「Retrieve sequences from a file」を選択し「OK」 →sampleD.fasta選択
- ③Translated Protein Sequence タブをクリック、Standardを選択。
- ④「Edit」から「SelectAll」を選択。
- ⑤「Alignment」から→「Align by Muscle」を選択→「Compute」
- ⑥「Data」→「Export Alignment」→「FASTA format」を選択  
→「T1R2amino」と名前を付けて保存。Fasta形式の T1R2amino.fas が保存されます。



ネコのT1R2配列の中央部が、他の種のホモログと合っていない。

# 【解答】2. T1R2の系統樹構築

- ①MEGAメインウィンドウから「Phylogeny」ボタン  
→「Construct/Test Neighbor-Joining Tree」を選択
- ②ファイルの種類:「Fasta file (\*.fas, \*.fasta)」を選択し、T1R2amino.fasを選択する。
- ③Data Type: Protein Sequencesを選択し、「OK」
- ④Test of Phylogeny:Bootstrap methodを選択し、「Compute」

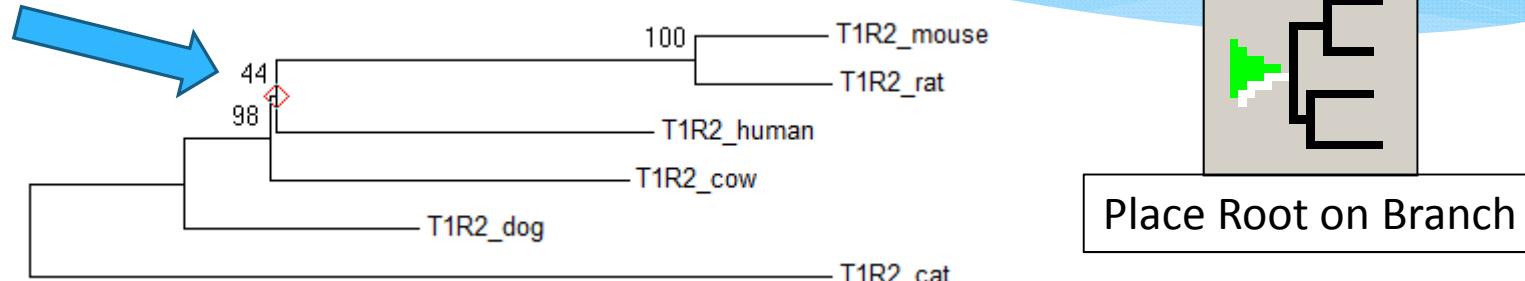


ネコの配列がアウトグループになっている? →根を設定していなかった

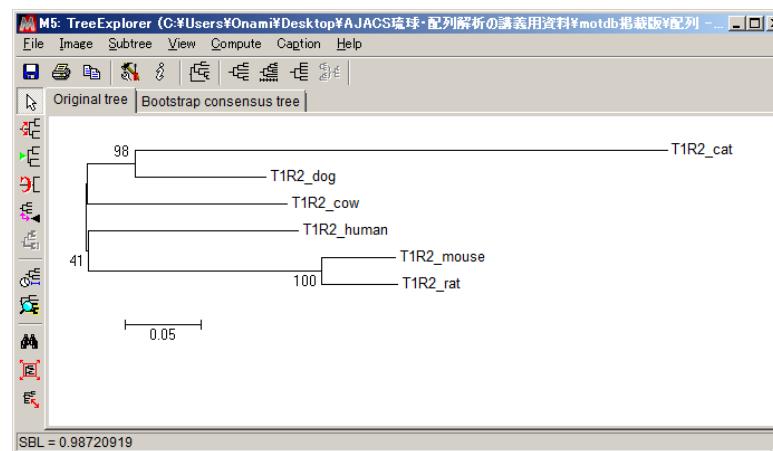
# 【解答】2. T1R2の系統樹構築・補足

## ■根の設定

- 通常の系統解析では明らかな外群配列を設定し、根を設定します。

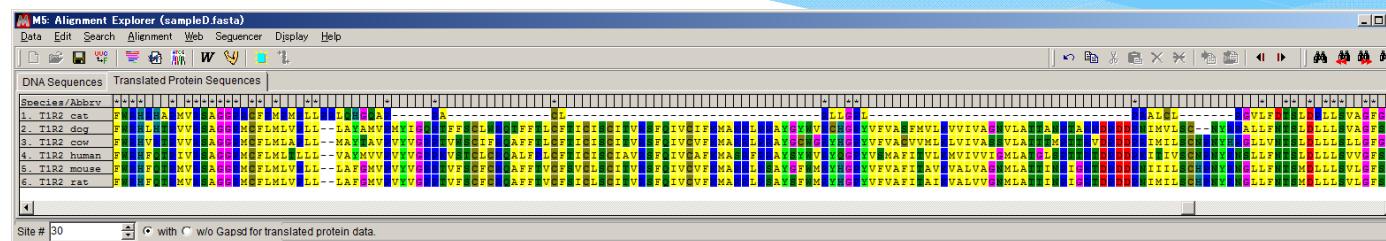


今回はmouse, rat, humanの根元を「根」とするため、上図の場所をクリックして「◇」印をつけ、右側の「Place Root on Branch」ボタンを押してください。

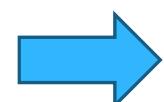


# 【解答】3. T1R2の椰討

## ■T1R2のアミノ酸配列のアラインメント



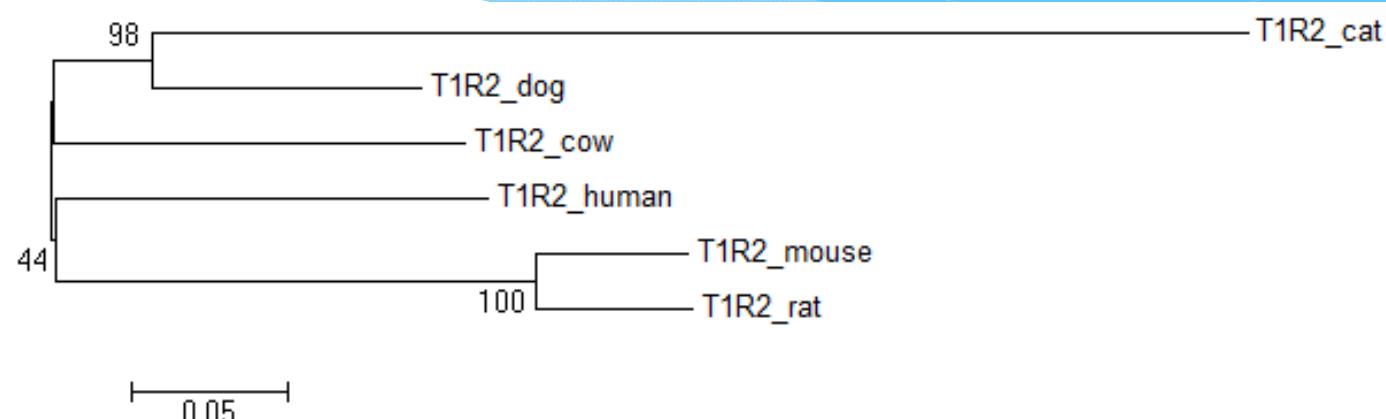
- ・ 中央部分の大規模な欠失をはじめ、ネコのT1R2には変異が多く含まれ他の種のホモログと大きく異なっている。
- ・ 中央部に終始コドンが含まれている。
- ・ ネコは甘いものにあまり興味を示さない。



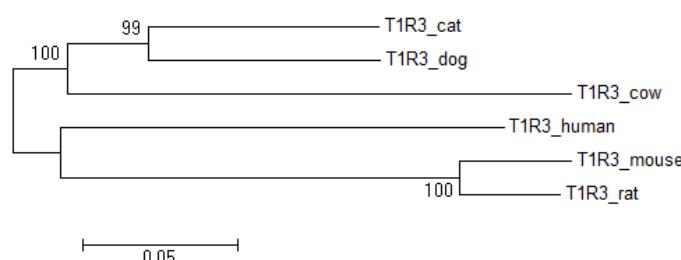
偽遺伝子化しているのでは？

# 【解答】3. T1R2の検討

## ■T1R2のアミノ酸配列の近隣結合法による系統樹



## ■T1R3のアミノ酸配列の近隣結合法による系統樹



ネコT1R2の進化速度が他と比べると異常に速い。

→ 偽遺伝子化している可能性が高い。

# 【解答】3. T1R2の検討

**PLOS GENETICS**

**OPEN ACCESS** Freely available online

**Pseudogenization of a Sweet-Receptor Gene Accounts for Cats' Indifference toward Sugar**

Xia Li<sup>1</sup>, Weihua Li<sup>1</sup>, Hong Wang<sup>1</sup>, Jie Cao<sup>1</sup>, Kenji Mochizuki<sup>1\*</sup>, Liquan Huang<sup>1</sup>, Alexander A. Bachmanov<sup>1</sup>, Danielle R. Reed<sup>1</sup>, Véronique Legrand-Defretin<sup>2</sup>, Gary K. Beauchamp<sup>1,3</sup>, Joseph G. Brand<sup>1,4,5\*</sup>

1 Morell Chemical Senses Center, Philadelphia, Pennsylvania, United States of America, 2 The WALTHAM Centre for Pet Nutrition, Merton Mowbray, Leicestershire, United Kingdom, 3 Department of Psychology, School of Arts and Sciences and Department of Anatomy, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, 4 Department of Biochemistry, School of Dental Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, 5 Veterans Affairs Affairs Medical Center, Philadelphia, Pennsylvania, United States of America

Although domestic cats (*Felis silvestris catus*) possess an otherwise functional sense of taste, they, unlike most mammals, do not prefer and may be unable to detect the sweetness of sugars. One possible explanation for this behavior is that cats lack the sensory system to taste sugars and therefore are indifferent to them. Drawing on work in mice, demonstrating that alleles of sweet-receptor genes predict low sugar intake, we examined the possibility that genes involved in the initial transduction of sweet perception might account for the indifference to sweet-tasting foods by cats. We determined that the sweet receptor genes of domestic cats act similarly to those of other feline species of the Felidae family, including obligate carnivores, tiger and cheetah. Because the feline sweet-receptor gene is formed by the dimerization of two proteins (T1R2 and T1R3; gene symbols *Tas1r2* and *Tas1r3*), we identified and sequenced both genes in the cat by screening a feline genomic BAC library and by performing PCR with degenerate primers on cat genomic DNA. Gene expression was assessed by RT-PCR of tissue of taste, *in situ* hybridization, and immunohistochemistry. The cat *Tas1r3* gene shows high sequence similarity with functional *Tas1r3* genes of other species. Message from *Tas1r3* was detected by RT-PCR of taste tissue. *In situ* hybridization and immunohistochemical studies demonstrate that *Tas1r3* is expressed, as expected, in taste buds. However, the cat *Tas1r2* gene shows a 247-base pair microdeletion in exon 3 and stop codons in exons 4 and 6. There was no evidence of detectable mRNA from *Tas1r2* by RT-PCR or in *in situ* hybridization, and no protein product expected by RT-PCR was detected by Western blotting in taste and *in situ*. All adult domestic cats also show the similar deletion and stop codons. We conclude that cat *Tas1r2* is an apparently functional and expressed receptor but that cat *Tas1r2* is an unexpressed pseudogene. A functional sweet-taste receptor heteromer cannot form, and thus the cat lacks the receptor likely necessary for detection of sweet stimuli. This molecular change was very likely an important event in the evolution of the cat's carnivorous behavior.

Citation: Li X, Li W, Wang H, Cao J, Mochizuki K, et al. (2005) Pseudogenization of a sweet-receptor gene accounts for cat's indifference toward sugar. *PLOS Genet* 1(1): e3. doi:10.1371/journal.pgen.0010003

**Introduction**

The domestic cat (*Felis silvestris catus*), of the family Felidae in the order Carnivora, is an obligate carnivore. Its sense of taste is distinguished by a lack of attraction to, or indifference toward, compounds that taste sweet to humans, such as sweet carbohydrates (sugars) and high-intensity sweeteners [1–3]. This behavior toward sweet stimuli is in marked contrast to the ability for sweets shown by most omnivores and herbivores, including humans, and many other mammals [4]. The indifference that cats display toward sweet-tasting compounds contrasts with their otherwise normal taste behavior toward stimuli of other taste modalities. For example, they show preference for selected amino acids [5] and generally avoid stimuli that to humans taste either bitter or very sour [1,5]. Congruent with these behavioral responses to taste stimuli, recordings from cat taste nerve fibers, and from units of the geniculate ganglion innervating taste cells, demonstrate that responses to bitter stimuli, as well as sour and umami acids are more intense, but cat do not show normal responses to sucrose and several other sugars [5–12]. The sense of taste in the cat, in general, is therefore similar to that of other mammals, with the exception of an inability to taste sweet stimuli.

The molecular basis for this sweet blindness in cats is not known. Because the taste blindness appears specific to this

single modality, we postulated that the defect in the cat (and likely in other obligate carnivores of Felidae) lies at the receptor step subenabling the sweet-taste modality. The possible defects at the molecular level that might cause this sweet blindness could range from a single to a few amino acid substitutions, such as it is found between near "taster" and "nontaster" strains of mice [13,14], to more radical mechanisms, such as an unexpressed pseudogene.

To distinguish among these possibilities, we identified the DNA sequence and examined the structures of the two known genes, *Tas1r2* and *Tas1r3*, that in other mammals encode the sweet-taste receptor heteromer, T1R2/T1R3. We compared these with the sequence and structure of the same genes in

Received February 16, 2005; Accepted March 26, 2005; Published July 25, 2005. DOI:10.1371/journal.pgen.0010003

Copyright © 2005 Li et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviations: bp, base pair; GPCRs, G protein-coupled receptors.

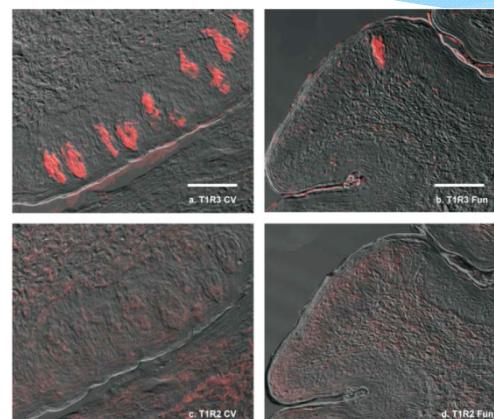
\*Editor: Jonathan Flint, University of Oxford, United Kingdom

†To whom correspondence should be addressed. Email: brand@vetscience.tufts.edu

‡Current address: Department of Remington Science, Tokyo University of Agriculture, Tokyo, Japan

PLOS Genetics | www.plosgenetics.org 0027 July 2005 | Volume 1 | Issue 1 | e3

2005年に、ネコでT1R2遺伝子が偽遺伝子化しており、味蕾で発現していないことが確かめられている。



T1R3蛋白質の発現

T1R2蛋白質の発現

チーターやトラも同様に発現していない。  
ネコ科の肉食性が甘味受容の喪失に影響した可能性があるとのこと。

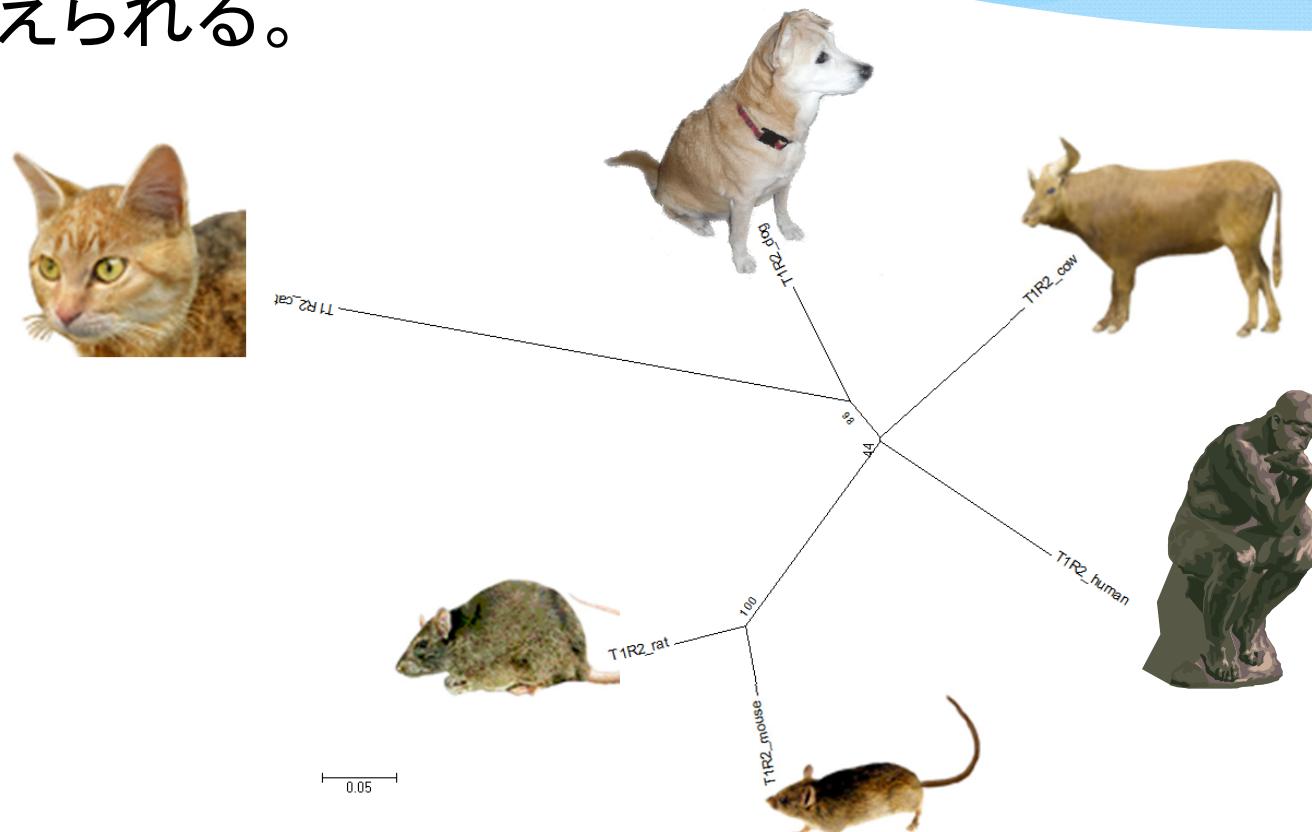
※今回はこの結果にウシの結果を加えて比較しました。

Xia Li et. al., PLoS Genet. 2005 July; 1(1): e3.

"Pseudogenization of a Sweet-Receptor Gene Accounts for Cats' Indifference toward Sugar"  
doi: 10.1371/journal.pgen.0010003 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1183522/

# 【解答】3. T1R2の検討

答. T1R2遺伝子はネコの系統で特異的に偽遺伝子化し、発現していないために、甘味受容の感覚を喪失したと考えられる。



生物アイコン © ライフサイエンス統合データベースセンター licensed under CC表示2.1 日本

# おまけ演習

小説「ロストワールド/ジュラシック・パーク2」に掲載されている恐竜の塩基配列が何の配列か確認し、そこに含まれるメッセージを推察しましょう。

## sampleA.txt

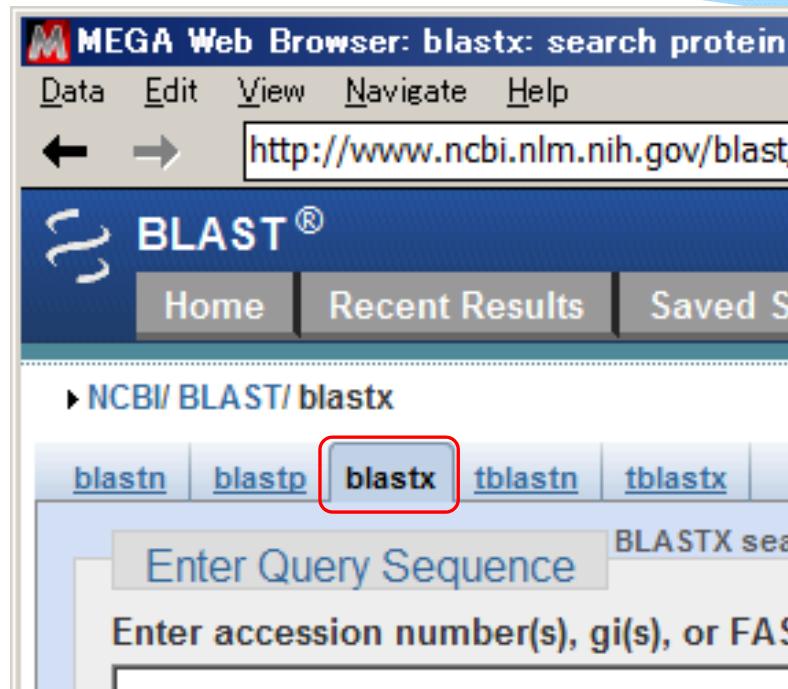
【ヒント】プログラムオプションで「blastx」を選択してください。



生物アイコン © ライフサイエンス統合データベースセンター licensed under CC表示2.1 日本

# 【おまけ演習】BLASTの設定

MEGAメインウィンドウから  
「Align」ボタン→「Do BLAST Search」を選択



配列を入力し、「BLAST」ボタンをクリック

# 【おまけ演習】解答

**Alignments**

Download ▾ GenPept Graphics erythroid transcription factor [Gallus gallus] Sequence ID: refNP\_990795.1 Length: 304 Number of Matches: 1 ▶ See 2 more title(s)

Range 1: 1 to 304 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
348 bits(893)	1e-113	Compositional matrix adjust.	304/318(96%)	304/318(95%)	14/318(4%)	+1
Query 121	MEFVALGGPDAGSPTPFPDeagaflglggerteaggllaSYPPSGRVSLVPWADTGTLG					300
Sbjct 1	MEFVALGGPDAGSPTPFPDEAGAFLGLGGGERTEAGGLLASYPPSGRVSLVPWADTGTLG					60
Query 301	TPQWVPPATQMEPPHYLEllqpprgspphpssgpllpplssgpppCEAREC	/MARKNCGAT				480
Sbjct 61	TPQWVPPATQMEPPHYLEllQPPRGSPHPSSGPLLPLSSGPPPCEARECV	NCGAT				116
Query 481	ATPLWRRDGTHYLICN <del>WASAC</del> GLYHRLNGQNRPLIRPKKRLLVSKRAGTVCSHERENCQT					660
Sbjct 117	ATPLWRRDGTHYLICN <del>A</del> GLYHRLNGQNRPLIRPKKRLLVSKRAGTVCS	NCQT				169
Query 661	STTTLWRRSPMDPVCONNI <del>H</del> GLYYKLHQVNRLPLTMRKDGIQTRNRKVsskgkkrrppg					840
Sbjct 170	STTTLWRRSPMDPV <del>N</del> --A <del>GLY</del> YKLHQVNRLPLTMRKDGIQTRNRKVSSKGKKRPPG					226
Query 841	ggnpusatagggapmggggdpsmooooooooaaappQSDALYALGPVVLSGHFLPfgnsggf					1020
Sbjct 227	GGNPSATAGGGAPMGGGDPSMooooooooaaappQSDALYALGPVVLSGHFLPFGNsggf					286
Query 1021	fgggaggYTAPPGLSPQI 1074					
Sbjct 287	FGGGAGGYTAPPGLSPQI 304					

ニワトリの赤血球タンパクの  
転写因子

~~MARK~~ ~~WAS~~ ~~HERE~~ ~~NIH~~  
 T P V Z I I I  
 T-----P V---Z I-----I I-----I



# 【おまけ演習】補足

## ■なぜ「MARK WAS HERE NIH」？

NIHの研究員だった Mark Boguski博士 が、  
Jurassic Parkの作者に空想の塩基配列を提供する際、  
BLASTで検索すると自分の名前が浮き出るよう、細工したこと。

- NCBI Problems (このような"汚染"は問題である、という意見)  
[http://www.ncbi.nlm.nih.gov/Class/FieldGuide/problem\\_set.html](http://www.ncbi.nlm.nih.gov/Class/FieldGuide/problem_set.html)
- NCBI Introductory course(トレーニングのネタとして利用)  
[http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/BLAST/q\\_jurassicparkDNA.html](http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/BLAST/q_jurassicparkDNA.html)

# 最後に…

## ■配列解析の基礎をお送りしました。

- ・各ソフト・プログラムのパラメータや詳細の説明は除外させていただきました。  
→まずは習うより慣れた方がよいという考えです。

## ■CUI(Linux等でコマンドで実行するインターフェース)とはどう違う？

- ・多数のサンプルについて繰り返して自動実行したり、  
詳細なパラメータや出力形式を設定する場合はCUIが優秀。  
※コマンドを覚えたりマニュアルを読み込む必要があります。

ご清聴ありがとうございました。

## 7. 参考

### ~文献情報とNCBI BLASTの詳細

# 参考文献(Web)

**■BLASTの評価**

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>

**■BLASTの結果**

[ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)

**■WebLogo**

<http://weblogo.berkeley.edu/>

**■恐竜の遺伝子(空想)**

<ftp://ftp.ncbi.nih.gov/pub/FieldGuide/lostworld.txt>

**■解析に利用した配列のGenBank Accession番号**

※AY819787についてはイントロン予想領域を削除し使用しました。

	T1R3コード領域	T1R2コード領域
cat	AY819786	AY819787
dog	AY916759	AY916758
cow	XM_588865.6	NM_001206529.1
human	BK000152	NM_152232
mouse	BC153121.1	NM_031873.1
rat	AF456324.1	NM_001271266.1

# 参考文献(書籍)

## ■全般

- ・バイオインフォマティクス ゲノム配列から機能解析へ 第2版  
マウント デービッド W. (著), 岡崎 康司 (著), 坊農 秀雅 (著) メディカル・サイエンス・インターナショナル (2005)
- ・生物配列の統計 岸野洋久 浅井潔 (2003)

## ■恐竜の遺伝子(空想)

- ・*The Lost World*, by Michael Crichton (Ballantine Books, 1996)

## ■MEGA5関連

- ・実験医学増刊 使えるデータベース・ウェブツール 有田正規・編  
第4章 5. 最尤法による分子系統解析の実際(田村浩一郎) 羊土社 2011

## ■相同性検索/マルチプルアラインメント

- ・あなたにも役立つ バイオインフォマティクス 菅原英明・編 共立出版(2002)
- ・ゲノムネットのデータベース利用法 金久実 共立出版(2002)
- ・バイオインフォマティクスの実際 村上 康文(編集), 古谷 利夫(編集) 講談社(2002)
- ・NCBI BLAST HELP [http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\\_TYPE=BlastDocs](http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs)

## ■系統樹構築

- ・分子系統学 長谷川政美・岸野洋久 岩波書店(1996)
- ・分子進化学入門 木村資生・編 (1984)

# 参考文献(論文)

## ■MEGA5

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S

“MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.” *Molecular Biology and Evolution* 28: 2731–2739. 2011

## ■シーラカンスのゲノム

Woolston C. “Living fossil” genome unlocked.” *Nature*. 2013 Apr 18;496(7445):283. doi: 10.1038/496283a.

Amemiya CT et. al., “The African coelacanth genome provides insights into tetrapod evolution.” *Nature*. 2013 Apr 18;496(7445):311–6. doi: 10.1038/nature12027.

## ■カワラバトのゲノム

Shapiro MD et. al., “Genomic diversity and evolution of the head crest in the rock pigeon.” *Science*. 2013 Mar 1;339(6123):1063–7. doi: 10.1126/science.1230422. Epub 2013 Jan 31.

## ■ウラルツコムギのゲノム解析

Ling HQ et. al., “Draft genome of the wheat A-genome progenitor *Triticum urartu*.” *Nature*. 2013 Apr 4;496(7443):87–90. doi: 10.1038/nature11997. Epub 2013 Mar 24.

## ■BLAST

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. “Basic local alignment search tool.”, *J Mol Biol*. 1990 Oct 5;215(3):403–10.

## ■FASTA

Pearson WR, Lipman DJ. “Improved tools for biological sequence comparison.” *Proc Natl Acad Sci U S A*. 1988 Apr;85(8):2444–8.

## ■2次元DP法

Lipman DJ, Altschul SF, Kececioglu JD., “A tool for multiple sequence alignment.” *Proc Natl Acad Sci U S A*. 1989 Jun;86(12):4412–5.

## ■Clustal

Higgins DG, Sharp PM. “CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.” *Gene*. 1988 Dec 15;73(1):237–44.

## ■ClustalW

Thompson JD, Higgins DG, Gibson TJ. “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.” *Nucleic Acids Res*. 1994 Nov 11;22(22):4673–80.

## ■MUSCLE

Edgar RC. “MUSCLE: multiple sequence alignment with high accuracy and high throughput.” *Nucleic Acids Res*. 2004 Mar 19;32(5):1792–7. Print 2004.

## ■T1R2,T1R3 甘味受容遺伝子の偽遺伝子化と、ネコの砂糖への無関心

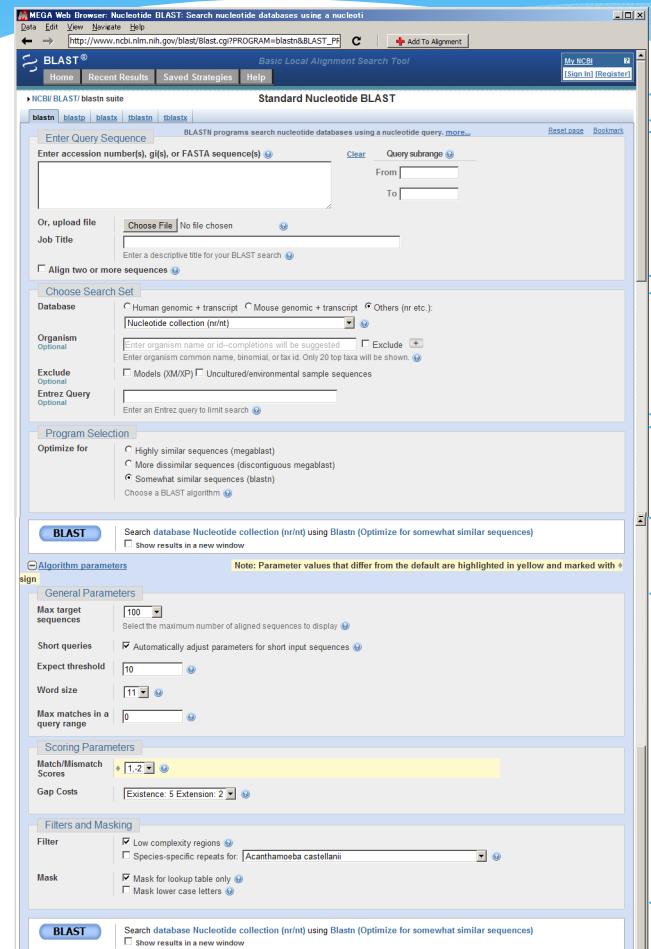
Xia Li et. al., *PLoS Genet*. 2005 July; 1(1): e3. “Pseudogenization of a Sweet-Receptor Gene Accounts for Cats’ Indifference toward Sugar” doi: 10.1371/journal.pgen.0010003 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1183522/

## ■哺乳類の系統関係

Nishihara H, Maruyama S, Okada N. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals.

*Proc Natl Acad Sci U S A*. 2009 Mar 31;106(13):5235–40. doi: 10.1073/pnas.0809297106. Epub 2009 Mar 13.

# NCBI BLASTの詳細設定



## ■BLASTプログラムの選択タブ

blastn、blastp、blastx、tblastn、tblastxから選択

## ■Enter Query Sequence

質問配列の設定

## ■Choose Search Set

検索するデータベースや対象生物種を設定

## ■Program Selection

blastn及びblastp専用の設定項目。  
BLASTアルゴリズムを選択

## ■Algorithm Parameters

blastn専用の設定項目。普段は隠されている。  
結果表示やスコアリング、マスキング関連の設定

# NCBI BLASTの詳細設定①

## ■ BLASTプログラムの選択

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

プログラム	質問配列	データベース	備考
blastn	塩基	塩基	
blastp	アミノ酸	アミノ酸	
blastx	塩基 <small>(アミノ酸に翻訳して検索)</small>	アミノ酸	
tblastn	アミノ酸	塩基 <small>(アミノ酸に翻訳して検索)</small>	
tblastx	塩基 <small>(アミノ酸に翻訳して検索)</small>	塩基 <small>(アミノ酸に翻訳して検索)</small>	

※プログラム内で塩基→アミノ酸に翻訳される場合、6通りの読み枠で変換される。  
この場合、検索する配列の種類が多くなり、計算に時間がかかることがある。

参考 : [http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=FAQ#expect](http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect)

# NCBI BLASTの詳細設定②

## ■ Enter Query Sequence

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

Or, upload file [?](#) Choose File No file chosen [?](#)

Job Title [?](#)

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

①

②

③

④

⑤

①質問配列入力: accession番号、gi番号、FASTA配列(タイトル無でも可)

②質問配列の幅: 質問配列の何番目から何番目を使用するか

③ファイル選択: 配列ファイルを質問配列とする場合

④Job Title: 検索セッションのタイトル

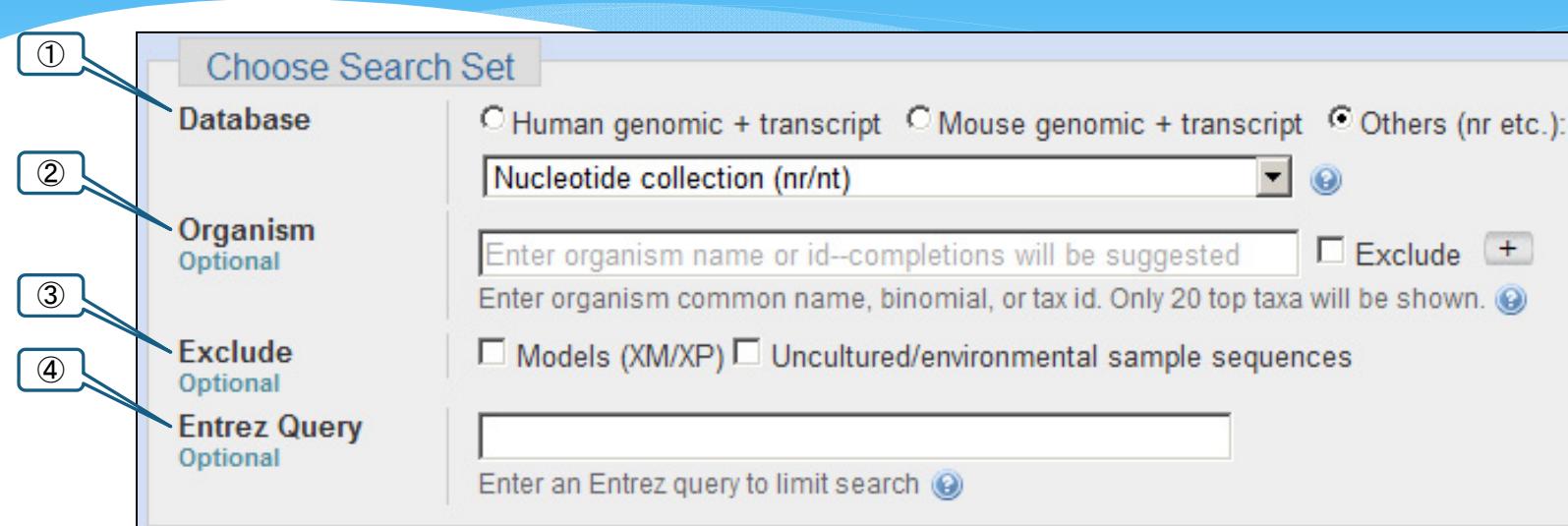
⑤複数アライメント: 対応する配列をこの下に表示される

ボックスに入力し、アライメント実行。

参考: [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)

# NCBI BLASTの詳細設定③

## ■ Choose Search Set



①データベース:検索対象のNCBIデータベースを選択

②生物種:データベース内の生物種で区切る場合

③Exclude: Modelやメタゲノムを除外対象とするか。

④Entrez query:Entrezデータベースに限定する場合その番号

参考: [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)

# NCBI BLASTの詳細設定④

## ■ Program Selection

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm 

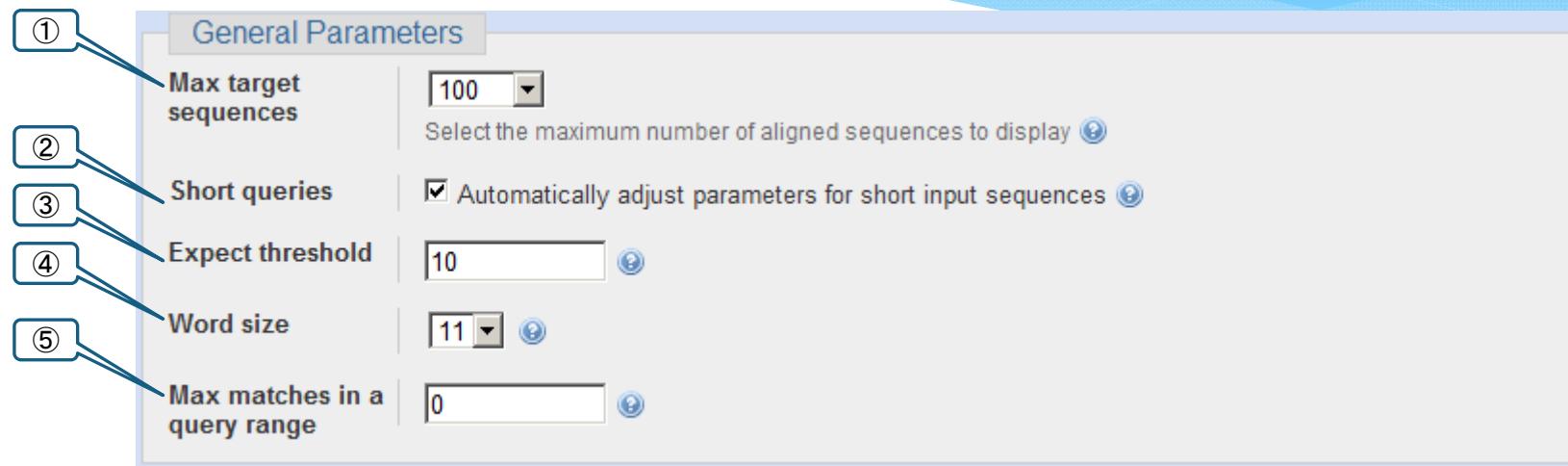
### BLASTアルゴリズムの選択

・塩基配列、アミノ酸配列、両配列の変換を含む場合で異なる。各アルゴリズムの詳細は下記参照。

参考: <http://www.ncbi.nlm.nih.gov/blast/html/BLASThomehelp.html>

# NCBI BLASTの詳細設定⑤

## ■Algorithm Parameters ~General Parameters



①Max target sequences: アラインメント表示される配列の最大数

②Short queries: クエリが短い場合にパラメータを最適化する

③Expect threshold: 統計的有意性の閾値。小さいほど厳密な検索を行うがヒット確率が低下する。

④Word size: 単語をヒットさせる際の読み枠の大きさ。

⑤Max matches in a query range: クエリ幅内でマッチする数の制限

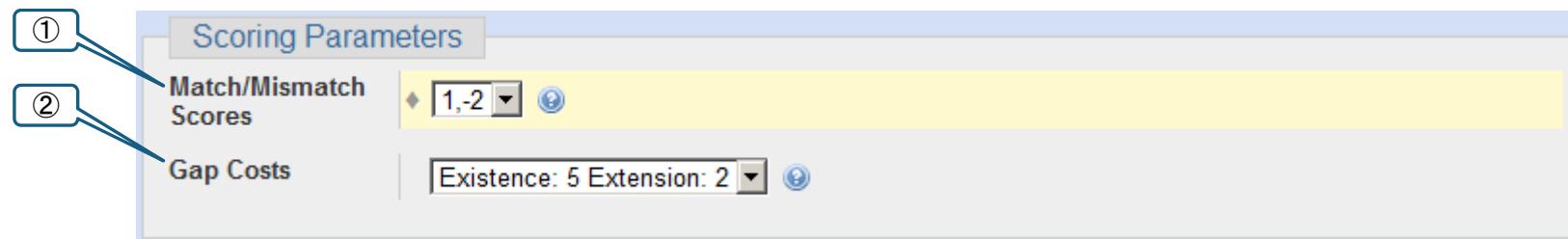
参考: [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)

<http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#expect>

<http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#wordsize>

# NCBI BLASTの詳細設定⑥

## ■ Algorithm Parameters ~Scoring Parameters

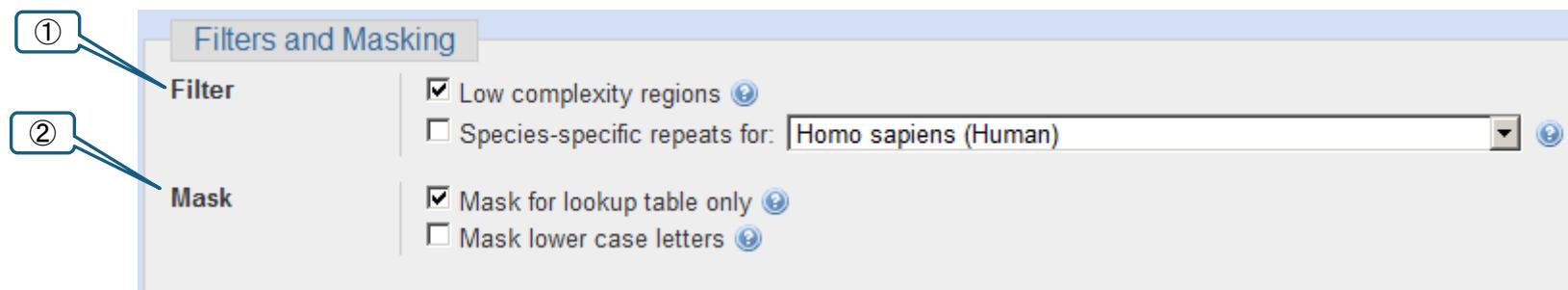


- ①Match/Mismatch Scores: マッチやミスマッチの場合のペナルティスコア  
②Gap Costs: ギャップ開始やギャップ伸長のペナルティスコア。megablastの時はlinearで自動で設定される。

参考: [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)

# NCBI BLASTの詳細設定⑦

## ■ Algorithm Parameters ~Filters and Masking



- ①Filter: 同じ配列の連続や繰り返しが続く(Low complexity)場合や種特異的な反復配列を含む場合、フィルタをかけるか  
②Mask: lookup tableの検索を除外してLow complexity領域を検索する、質問配列の小文字(lower case)を検索範囲から除外する。

参考: [ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo\\_NewBLAST.pdf](ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf)  
<http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#filter>