

2014年1月23日

統合データベース講習会 AJACS肥後

配列比較解析の実際



科学技術振興機構
Japan Science and Technology Agency

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肥後」で検索

NBDC の広報サイト
バイオサイエンス ×DB=∞ 検索 Web events.biostciencedbc.jp/

ホーム シンポジウム 講習会 展示会 連載

統合データベース講習会: AJACS肥後

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。

講習会は2日間にわたりて開催されます。

バイオサイエンスデータベースセンターが開発している生命科学系データベースのカタログ・横断検索・アーリース、PubMedなどの文献関連のデータベース、タンパク質立体構造データベース PDBと立体構造を扱った Assemblなどのがんデータベース、BLASTやClustalWなどを用いた基本的な配列比較解析、次世代シーケンス解析ツールの使い方を中心に紹介します。参加者全員がハンズオンでコンピュータを使用を行います。

- 対象: 化学及血清療法研究所に所属する研究者
- 日時: 2014年1月22日(水) 13:00～17:30、1月23日(木) 8:50～16:00
- 会場: 化学及血清療法研究所 菊池研究所
- プログラム

講習資料はこちらのサイトをご覧ください。

1月22日(水)

12:30～13:00 受付

13:00～14:20 「NBDCの紹介とNBDCが提供するサービス」
／柳田達矢 (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)

14:20～14:25 休憩

14:25～15:55 「文献の検索とその整理方法・統合DBプロジェクトのサービスを中心に」
／川本祥子 (情報・システム研究機構 ライフサイエンス統合データベースセンター)

15:55～16:00 休憩

16:00～17:30 「立体構造データベース(仮)」
／川端 猛 (大阪大学蛋白質研究所)

1月23日(木)

8:50～10:20 「ゲノム情報を閲覧・取得し、活用する」
／坊農秀雅 (情報・システム研究機構 ライフサイエンス統合データベースセンター)

10:20～10:30 休憩

10:30～12:00 「配列比較解析の実際」
／大庭純一 (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)

12:00～13:00 昼食

13:00～16:00 「次世代シーケンサーを活用した研究事例と、それを支える公共ツール・データベース」
／大田達郎 (情報・システム研究機構 ライフサイエンス統合データベースセンター)

Index

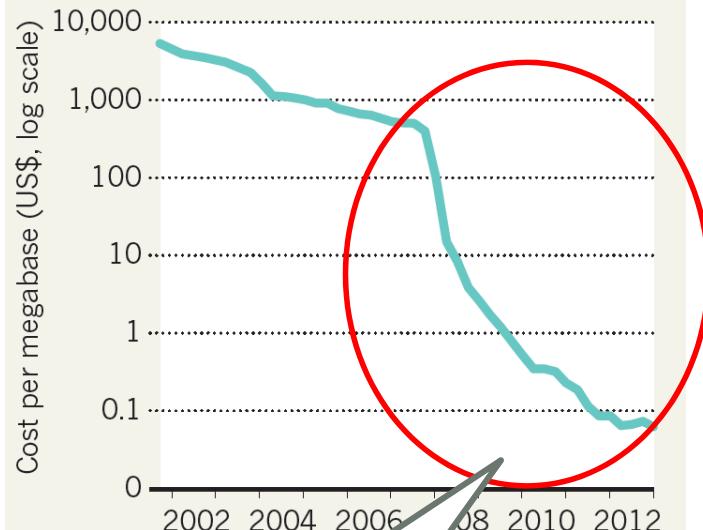
1. 配列比較解析の目的
2. 統合解析環境MEGA
3. 相同性検索 [ウイルス配列の調査]
4. マルチプルアラインメント [コウモリとハクビシン、ヒトから見つかったSARSウイルスの違いは？]
5. 系統樹構築 [SARSウイルスはどのように広まったか？]
6. 配列比較解析演習 [哺乳類p53遺伝子の比較解析]
7. 参考文献・リンク

1. 配列比較解析の目的

配列解析のコスト

PLUMMETING COSTS

Advances in sequencing technologies have driven a sharp drop in price.



4年でコスト
1/1000

Hayden EC. "Tepid showing for genomics X prize." Nature. 2013 May 30;497(7451):546-7. doi: 10.1038/497546a. <http://www.ncbi.nlm.nih.gov/pubmed/23719441>

BIOMEDICINE

NIH Seeks Better Database for Genetic Diagnosis

In 2011, Heidi Rehm, a molecular geneticist at the Harvard-affiliated Brigham and Women's Hospital in Boston, was asked to help physicians follow up on a pregnancy test result that showed a "mucosal translucency," a low-density area near the spinal cord of a fetus. It's viewed as a sign that the fetus might have a disfiguring condition called Noonan syndrome. So Rehm's lab did a DNA test, which came back positive for a gene variant listed in the lab's internal database as Noonan-related. The parents ended the pregnancy.

Many months later, however, the lab learned that the researcher who linked the variant to Noonan syndrome had concluded it was benign after all. But there was "no easy way to put that information in the public domain," Rehm says. So it had been set aside.

To make that less likely to happen in the future, the National Institutes of Health (NIH) last week awarded \$25 million for a new project called the Clinical Genome Resource, or ClinGen. The plan is to create a single reference point for data on medically important gene variants. John Hopkins University in Baltimore, Maryland, heads the donor list with 642 entries. Rehm's lab is second, with 116.

ClinGen is supposed to expand its collection and improve its quality.

experiment in which three clinical labs tested their prowess at making diagnoses from the same DNA sample. When they compared how they rated genes for medical significance, Rehm says, the labs disagreed on 20% of them.

According to Lissa Brooks, overseer of the project at NIH's National Human Genome Research Institute, there are now about 2000 databases on genes and disease worldwide.

Each lab concentrates mainly on its own work. ClinGen will scoop up clinically relevant information on gene variants from as many databases as will cooperate, review them, and share interpretations. The aim, Brooks says, is to create a "curated and annotated" collection of all medically relevant human gene variants.

The ambitious effort builds on a preexisting public database called ClinVar at NIH's National Center for Biotechnology Information. ClinVar

already contains more than 51,000 reported gene variants from 60 sign-off contributing

organizations. Rehm's lab is second, with 116 entries.

"Without a curated database of variants, we are not going to be able to move

into the whole-genome world," she says. "It just got to happen."

(is not curated now.) One funded effort—including Rehm's lab, Robert Nussbaum at the University of California, San Francisco, and others—will develop standards and solicit genetic data and associated medical records from patients and doctors. Keeping personal information private while sharing data online will be a big challenge, Rehm says.

A second group will classify gene variants according to medical significance and call on specialist panels to make calls about specific cases—for example, to decide whether a variant is pathogenic or not. The third group will work on computerized data classification and release. One ambitious goal is to devise algorithms that predict the medical relevance of variants.

Sherri Bale, managing director of GeneDx in Gaithersburg, Maryland, who took part in the DNA interpretation experiment with Rehm, says that DNA testing companies like hers recognize the value of the project. "Without a curated database of variants, we are not going to be able to move into the whole-genome world," she says. "It just got to happen."

—ELIOT MARSHALL

www.sciencemag.org SCIENCE VOL 342
Published by AAAS

Downloaded from www.sciencemag.org

27

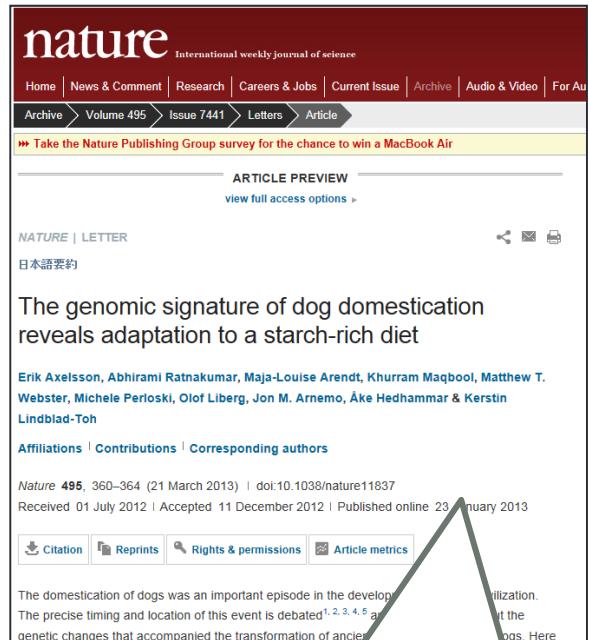
遺伝子や疾患関連
データベースは
世界に2000以上

Marshall E. "Biomedicine. NIH seeks better database for genetic diagnosis." Science. 2013 Oct 4;342(6154):27. doi: 10.1126/science.342.6154.27. <http://www.ncbi.nlm.nih.gov/pubmed/24092711>



配列解析の目的

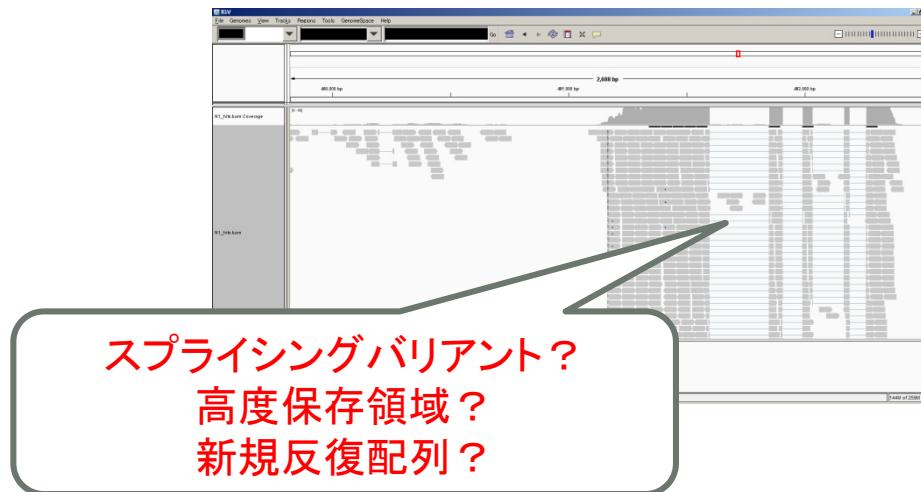
イヌとオオカミのゲノム再解析



Supplementaryとして
12個体分のSRA Data
198.3GB

Axelsson E. et. al. "The genomic signature of dog domestication reveals adaptation to a starch-rich diet"
Nature. 2013 Mar 21;495(7441):360-4. doi: 10.1038/nature11837.
<http://www.ncbi.nlm.nih.gov/pubmed/23354050>

- マクロな規模のデータは日に日に入手し易く。
- これらのデータから、ミクロな情報へズームアップしていく、疾患・医薬情報や生物学的意義のある情報を入手する。



分子生物学研究における配列解析 (想定されるワークフロー案)



- ・ある特定の遺伝子領域が、個体間や遺伝子間でどのように異なっているか？
- ・またその差は何故生じているのか？

検出と検討を行うのが
配列比較解析

主な配列比較解析手法

- 配列相同性検索
- マルチプルアラインメント
- 系統樹構築



全て統合解析環境
MEGAから実行可能

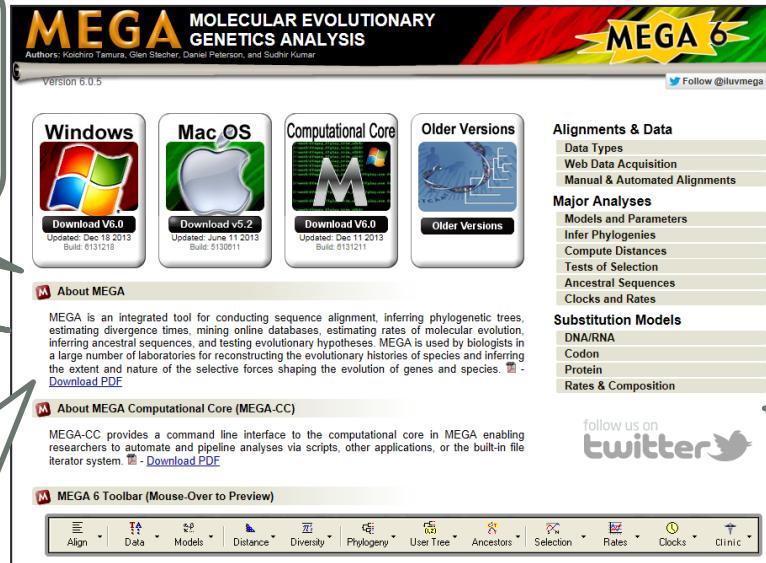
2. 統合解析環境MEGA

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)



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
身分と所属を入力)

1993年にVersion1が公開
されて以来、バージョンアップ
継続中。

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

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S (2013)
MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0.
Molecular Biology and Evolution 30: 2725-2729.

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

■ネアンデルタール人のゲノム解析

nature International weekly journal of science

出版者リスト ▾ 専門分野ゲートウェイ ▾ My Natureasia サイトマップ

Nature Japan » Nature » ハイライト » ネアンデルタール人女性のゲノム塩基配列

進化: ネアンデルタール人女性のゲノム塩基配列

Nature 505, 7481
2014年1月2日

近年、シベリア南部のアルタイ山脈にあるデニソワ洞窟の発掘で、おそらく25万年以上にわたって居住に利用されていた場所から多数のヒト族化石が見つかっている。今回、2010年にデニソワ洞窟の東の穴(east gallery)で見つかった約5万年前のあしゆび(趾)の骨(基節骨)から、高品質のゲノム塩基配列が得られた。この塩基配列はネアンデルタール人女性のものであり、彼女の両親は近親者同士で、おそらく異父母の兄弟姉妹か叔父と姪の関係であった。このような近親交配は、この女性の近い先祖においても普通に行われていた。他の旧人類や現代人のゲノムとの比較から、ネアンデルタール人とその近縁のデニソワ人、および初期現生人類の間で遺伝子流動が何度か起きたことが明らかになった。また、他の未知の旧人類集団からデニソワ人への遺伝子流動もあった可能性

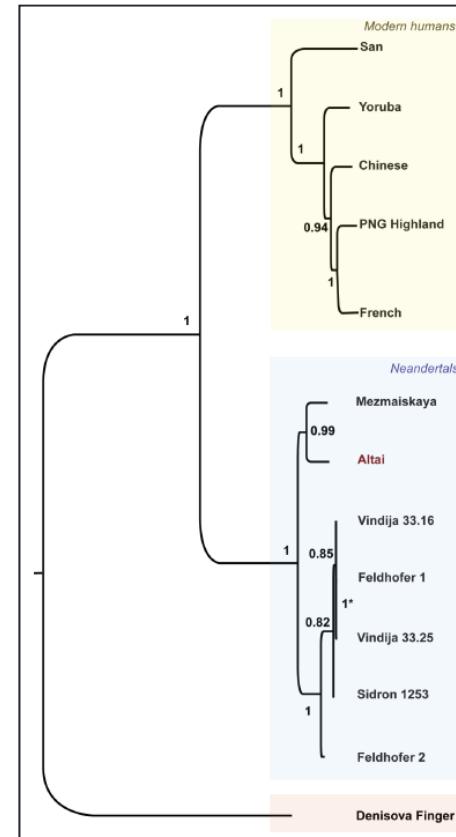


Credit: Bence Viola

<http://www.natureasia.com/ja-jp/nature/highlights/50475>

"The complete genome sequence of a Neanderthal from the Altai Mountains"
Kay Prüfer et. al.

Nature 505, 43–49 (02 January 2014) doi:10.1038/nature12886



between the seven Neandertals, five present-day
calculated using MEGA⁸.05 after the sequence

ansversions and tra
humans using MEGA⁹ (v5.1) on a mu
ndividual (FN673705), a Neandertal

MEGAの利用例

■ゾウキンザメのゲノム解析

nature International weekly journal of science

出版記リスト ▾ 専門分野ゲートウェイ ▾ My Natureasia サイトマップ

Nature Japan » Nature » ハイライト » ゆったり進化する生きもの:進化速度が最も遅い脊椎動物ゾウキンザメのゲノム塩基配列

Cover Story: ゆったり進化する生きもの:進化速度が最も遅い脊椎動物ゾウキンザメのゲノム塩基配列

Nature 505, 7482
2014年1月9日

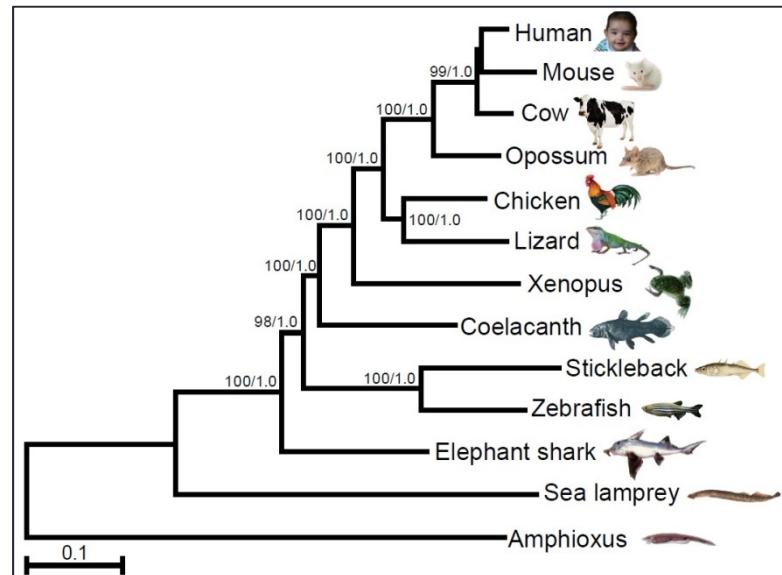
ゾウキンザメ(*Callorhinus milii*)は、オーストラリア南部およびニュージーランドの沖の温水海域だけに生息する軟骨魚類で、深さ200~500 m付近で暮らし、春には繁殖のために浅瀬に移動する。今回、ゾウキンザメのゲノム塩基配列が解読された。他の脊椎動物ゲノムとの比較により、このゲノムの進化速度が、シーラカンスを含めた既知の脊椎動物の中で最も遅いことが明らかになった。またゲノム解析によって、CD4受容体やそれに関連するサイトカイン類の一部を欠いた独特な適応免疫系の存在が明らかになり、軟骨魚類が始原的な顎口類適応免疫系を持つこと



ゲノム塩基配列が解読されたゾウキンザメ。
Credit: B. Venkatesh

<http://www.natureasia.com/ja-jp/nature/highlights/50711>

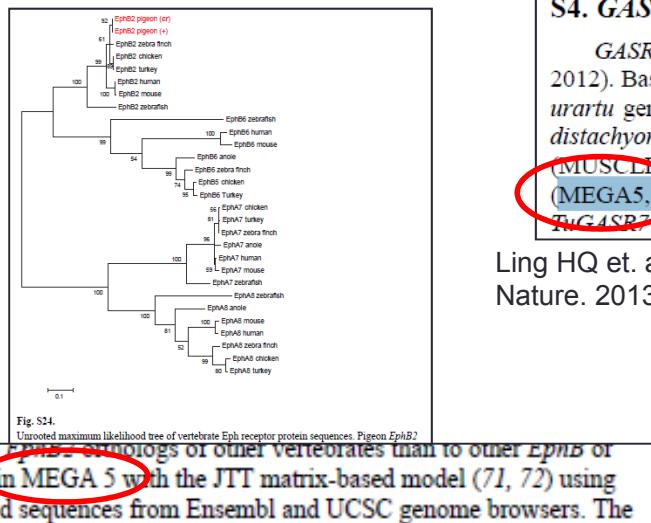
"Elephant shark genome provides unique insights into gnathostome evolution"
Byrappa Venkatesh et. al.
Nature 505, 174–179 (09 January 2014) doi:10.1038/nature12826



calculated by comparison with 100 random replicates from MEGA5. This analysis further shows that the elephant shark genome is slower-evolving than coelacanth, tetrapods, and birds.

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

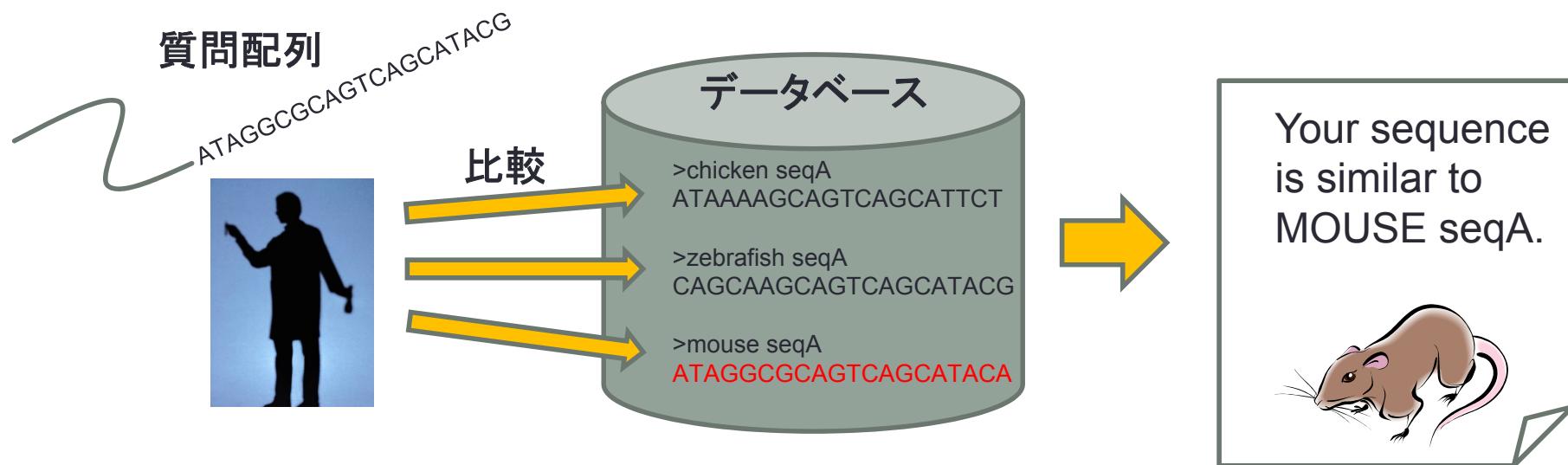
Ling HQ et. al., Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature. 2013 Apr 4;496(7443):87-90. doi: 10.1038/nature11997. Epub 2013 Mar 24.

今年・昨年のNature,Scienceの
様々な生物種のゲノム解析で利用実績あり。

3. 相同性検索

ウイルス配列の調査

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肥後ページ」の講義資料から
取得してください。

配列.zip

解凍

	p53.fas	2014/01/14 17:54	FAS File
	SampleA.txt	2014/01/09 18:04	テキスト ドキュメント
	SampleB.fas	2014/01/09 18:12	FAS File
	SampleC.fas	2014/01/13 14:37	FAS File
	Sprotein.fas	2014/01/09 19:12	FAS File

※もし今回取得できない場合でも、後日お試しいただけます。

相同性検索用サンプル配列

SampleA.txt

```
>SampleA
GATTACTCTGTGCTCTACAACACTAACATCTTTCAACCTTAAGTGCTATGGCGTTCTGCCACTAAGTTGAATGATCTTGCTT
CTCCAATGTCTATGCAGATTCTTTGATGCAAAGGAGATGATGTAAGACAAATAGCGCCAGGACAAACTGGTGTATTGCTGA
TTATAATTATAAATTGCCAGATGATTTCATGGGTTGTGCCTTGCTTGGAAACTAGGAACATTGATGCTACTTCAACTGGTAATT
ATAATTATAAATATAGGTATCTTAGACATGGCAAGCTTAGGCCCTTGAGAGAGACATATCTAATGTGCCTTCTCCCTGATGG
CAAACCTTGCACCCCCACCTGCTCTTAATTGTTATTGCCATTAAAAGATTATGGTTTACACCACTAGTGGCATTGGCTACCAA
CCTTACAGAGTTGAGTACTTTCTTGAACCTTTAAATGCACCGGCCACGGTTGTGGACCAAAATTATCCACTGACCTTATTA
AGAACCAAGTGTGTCATTAAATTAAATGGACTCACTGGTACTGGTGTAACTCCTCTTCAAAGAGATTCAACCATTCAA
CAATTGGCCGTGATGTTCTGATTTCACTGATTCCGTTGAGATCCTAAACATCTGAAATATTAGACATTCACCTGCTCTT
TGGGGGTGTAAGTGTATTACACCTGGAACAAATGCTTCATC
```

※参考:FASTA形式

- ・1行目に「大なり(greater-than)記号」と配列名(例:>sampleA)
- ・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ありません)
- ・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.faや.fas、.fasta等、アプリによって様々。
- ・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。

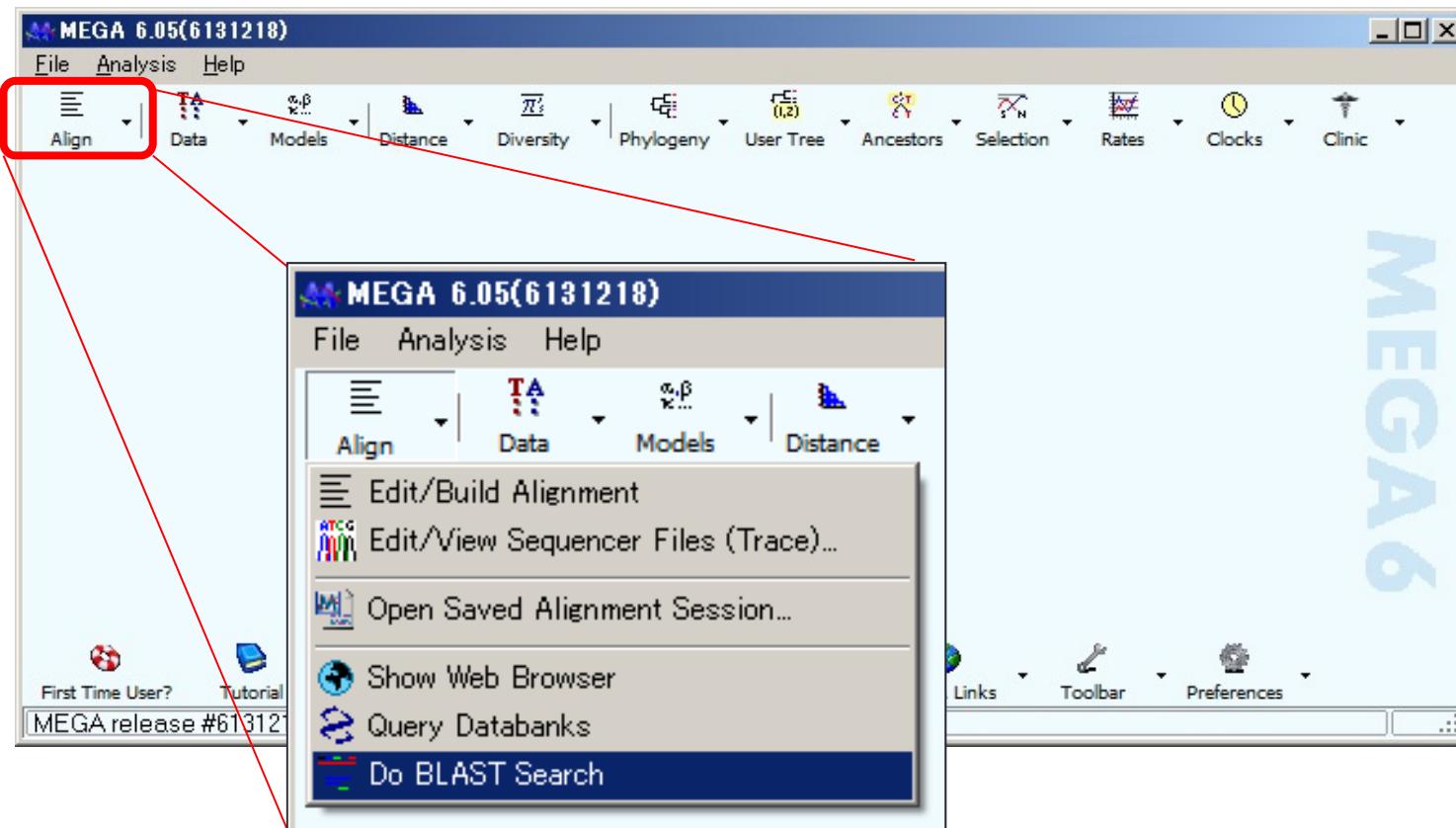
MEGA起動



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
ダウンロードを行ってください。

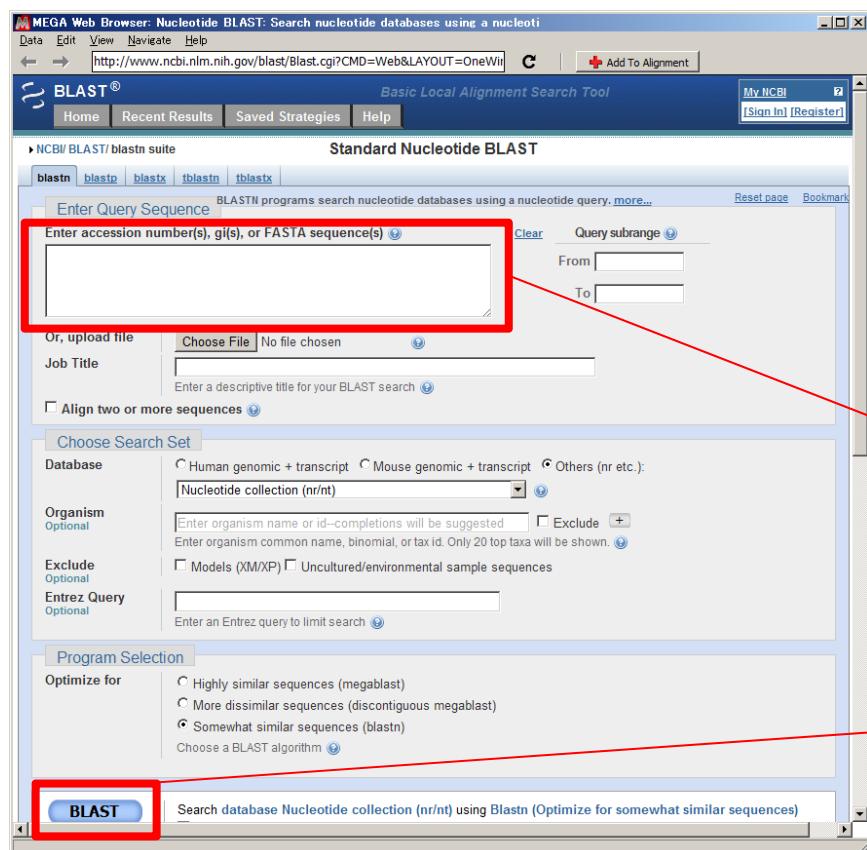
BLASTの実行



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「**SampleA.txt**」の配列をコピーして貼り付け

```
>SampleA
GATTACTCTGTGCTCTACAACTCAACATCTTTCAACCTTAAGTGCTATGGCGTTCTGC
CACTAAGTTGAATGATCTTGCTCTCCAATGTCTATGCAGATTCTTGTAGTCAAAGGAG
ATGATGTAAGACAAATAGCGCCAGGACAAACTGGTGTATTGCTGATTATAATTATAATTG
CCAGATGATTTCATGGGTGTGTCCTGCTTGAATACTAGGAACATTGATGCTACTTCAAC
```

②BLASTボタンをクリック

検索中…

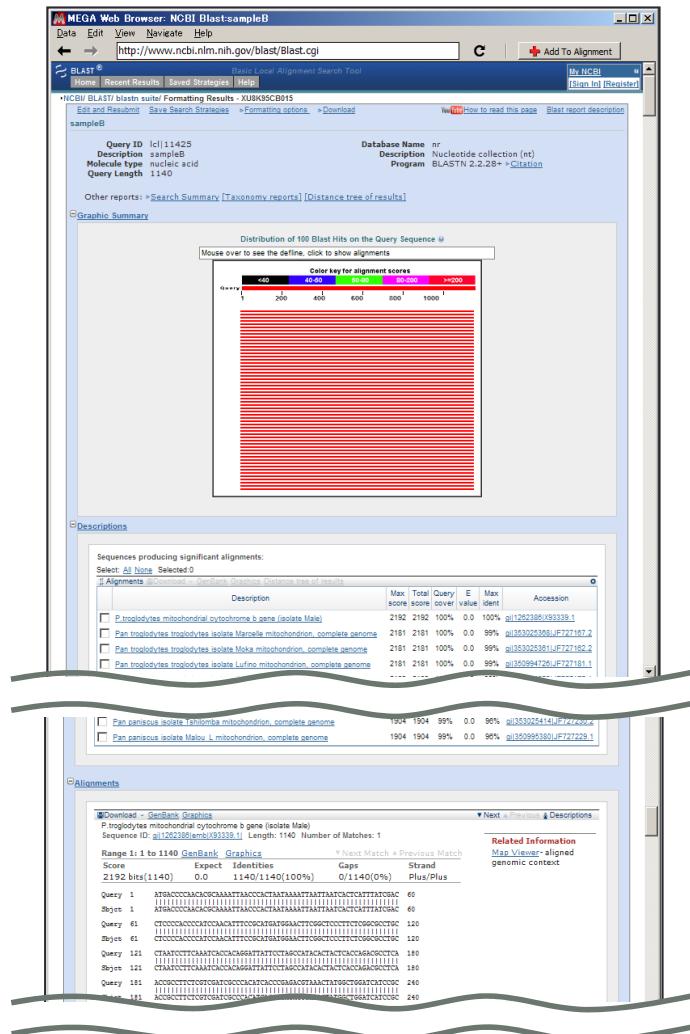
The screenshot shows a web browser window titled "MEGA Web Browser: NCBI Blast:SampleA". The address bar contains the URL "http://www.ncbi.nlm.nih.gov/blast/Blast.cgi". The main content is the NCBI BLAST interface, specifically the "Basic Local Alignment Search Tool". The job title is "SampleA". Below the job title, there is a table with the following information:

Request ID	CWKDDC62014
Status	Searching
Submitted at	Thu Jan 9 02:20:13 2014
Current time	Thu Jan 9 02:20:26 2014
Time since submission	00:00:12

Below the table, a message says "This page will be automatically updated in 2 seconds". At the bottom, there are links for "Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback" and the text "BLAST is a registered trademark of the National Library of Medicine. NCBI | NLM | NIH | DHHS".

※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。

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アクセッション番号

■ 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 search bar containing the URL <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>. Below the navigation bar is the BLAST logo and the text "Basic Local Alignment Search Tool". On the right, there are links for "My NCB", "Sign In", and "Register". The main content area displays search parameters and database details:

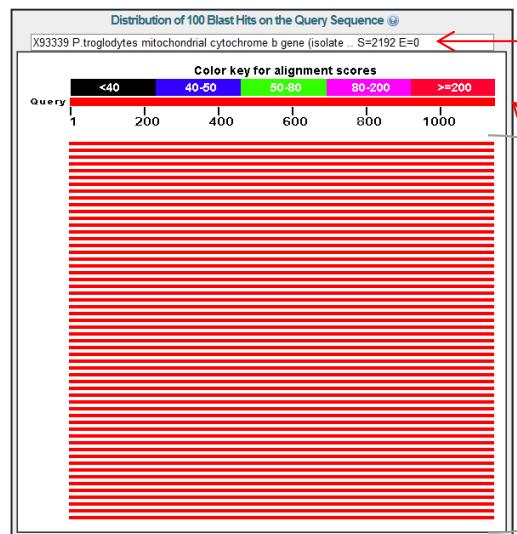
- sampleB**
- Query ID:** lcl|21223 (①)
- Description:** sampleB (②)
- Molecule type:** nucleic acid (③)
- Query Length:** 1140 (④)
- Database Name:** nr (⑤)
- Description:** Nucleotide collection (nt) (⑥)
- Program:** BLASTN 2.2.28+ (⑦)

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	③ Max score	④ Total score	⑤ Query cover	⑥ E value	⑦ Ident	⑧ Accession	⑨
<input type="checkbox"/>	SARS coronavirus SZ3, complete genome	1394	1394	100%	0.0	100%	gi 34482137 AY304486.1	
<input type="checkbox"/>	SARS coronavirus SZ16, complete genome	1383	1383	100%	0.0	99%	gi 34482139 AY304488.1	
<input type="checkbox"/>	SARS coronavirus SZ13, partial genome	1383	1383	100%	0.0	99%	gi 34482138 AY304487.1	
<input type="checkbox"/>	SARS coronavirus SZ1, partial genome	1377	1377	100%	0.0	99%	gi 34482140 AY304489.1	
<input type="checkbox"/>	SARS coronavirus isolate Tor2/FP1-10895, complete genome	1371	1371	100%	0.0	99%	gi 404325900 JX163928.1	
<input type="checkbox"/>	SARS coronavirus isolate Tor2/FP1-10851, complete genome	1371	1371	100%	0.0	99%	gi 404325840 JX163924.1	
<input type="checkbox"/>	SARS coronavirus HKU-39849 isolate recSARS-CoV HKU-39849, complete genome	1371	1371	100%	0.0	99%	gi 387912994 JN854286.1	

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

NCBI BLAST結果画面

■ Alignments

The screenshot shows the NCBI BLAST results page for a query against the SARS coronavirus SZ3 database. The top navigation bar includes 'Download' (①), 'GenBank' (②), and 'Graphics' (③) buttons. Below the search bar, it displays the target sequence information: 'SARS coronavirus SZ3, complete genome' (⑥), 'Sequence ID: gil34482137|gb|AY304486.1' (⑦), 'Length: 29741', and 'Number of Matches: 1' (⑧). The main content area shows the alignment details for Range 1: 22527 to 23251. It includes columns for Score (⑨), Expect (⑩), Identities (⑪), Gaps (⑫), Strand (⑬), Query sequence (⑭), Subject sequence (⑮), and Match length (⑯). A 'Related Information' section is also visible.

- ①Download: FASTA/Genbank形式で配列をダウンロード
- ②Genbank: Genbankのページへ。
- ③Graphics: Genbank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

NCBI BLAST結果画面

■ NCBI GenBankデータベース詳細

MEGA Web Browser: SARS coronavirus SZ3, complete genome - Nucleotide - NCBI

File Edit View Navigate Help

http://www.ncbi.nlm.nih.gov/nucleotide/34482137?report=genbank&log\$=nuclalign C Add To Alignment

NCBI Blast: SampleA SARS coronavirus SZ3, complete genome - Nucleotide - NCBI

Display Settings: GenBank

SARS coronavirus SZ3, complete genome

GenBank: AY304486.1

FASTA Graphics

Go to:

LOCUS AY304486 29741 bp **RNA** linear VRL 05-NOV-2003

DEFINITION SARS coronavirus SZ3, complete genome.

ACCESSION AY304486

VERSION AY304486.1 GI:34482137

KEYWORDS

SOURCE Civet SARS CoV SZ3/2003

ORGANISM Civet SARS CoV SZ3/2003

Viruses; ssRNA positive-strand viruses, no DNA stage; Nidovirales; Coronaviridae; Coronavirinae; Betacoronavirus.

REFERENCE 1 (bases 1 to 29741)

AUTHORS Guan,Y., Zheng,B.J., He,Y.Q., Liu,X.L., Zhuang,Z.X., Cheung,C.L., Luo,S.W., Li,P.H., Zhang,L.J., Guan,Y.J., Butt,K.M., Wong,K.L., Chan,K.W., Lim,W., Shortridge,K.F., Yuen,K.Y., Peiris,J.S. and Poon,L.L.

TITLE Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China

JOURNAL Science 302 (5643), 276-278 (2003)

PUBMED 12958366

REFERENCE 2 (bases 1 to 29741)

AUTHORS Guan,Y. and Zheng,B.J.

TITLE Direct Submission

JOURNAL Submitted (26-MAY-2003) Microbiology, The University of Hong Kong, University Pathology Building, Queen Mary Hospital, Pokfulam Road, Hong Kong, China

FEATURES

source Location/Qualifiers
1..29741
/organism="Civet SARS CoV SZ3/2003"
/mol_type="genomic RNA"
/isolate="SZ3"
/db_xref="taxon:231513"
/country="Hong Kong"

ORIGIN

```

1 ctacccgaga aaagccaaacc aacctcgatc tcttttagat ctgttctctta aacgaacttt
61 aaaatctgtg tagctgtgc tggctgtcat gccttagtca cttacgcgtt ataaacaata
121 ataaatttta ctgttgttgc caagaaacgt gtaactgttc ctttttttc agactgttta
...

```

SampleAは
SARSコロナウイルス
の部分配列

【参考】<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

【追加情報】SARSコロナウイルス

- 重症急性呼吸器症候群(じゅうしょうきゅうせいかいこうぐん、Severe Acute Respiratory Syndrome; SARS(サーズ))は、SARSコロナウイルスにより引き起こされる新種の感染症。新型肺炎(非典型肺炎、中国肺炎、Atypical Pneumonia)とも呼ばれる。
- 2002年11月(広州市呼吸病研究所は7月と発表)に中華人民共和国広東省で発生し、2003年7月に新型肺炎制圧宣言が出されるまでの間に8,069人が感染し、775人が死亡した。
- このウイルスの発生源はハクビシンが疑われていたが…

[wikipedia「SARS」(2014/01/21の記事)より抜粋]



SARSは終息。

しかし2013年5月、中東でMERSコロナウイルスのヒトへの感染が確認され、改めてSARSの起源や感染拡大方法について注目されている。

【追加情報】コーディング領域確認

The screenshot shows the NCBI BLAST search interface. At the top, there's a browser header for 'MEGA Web Browser: blastx: search protein database' with a URL 'http://www.ncbi.nlm.nih.gov/blast/Blast.cgi'. Below it is the main BLAST interface with tabs for 'Home', 'Recent Results', and 'Saved Strategies'. The current tab is 'blastx'. A sub-header says 'Translated BLAST: blastx'. There are four sub-tabs: 'blastn', 'blastp', 'blastx' (which is selected), and 'tblastn', 'tblastx'. A large text input field is labeled 'Enter Query Sequence' and contains a sequence labeled '>SampleA' followed by several lines of nucleotide sequence. To the right of this input field are 'Clear' and 'Query subrange' buttons, with 'From' and 'To' fields below them. Further down, there's a file upload section with a 'Choose File' button and a dropdown menu showing 'Standard (1)'. A text input field for a descriptive title is labeled 'SampleA'. A checkbox for 'Align two or more sequences' is present. At the bottom, there's a 'Choose Search Set' section with a 'Database' dropdown set to 'Non-redundant protein sequences (nr)'. Three numbered callouts point to specific elements: ① points to the 'blastx' tab, ② points to the 'SampleA' sequence in the query input, and ③ points to the 'blastx' button at the bottom.

①最初のページで、「blastx(質問配列をアミノ酸に変換
アミノ酸データベースを検索する)」を選択

②SampleAを入力

③下部blastボタンをクリック

【追加情報】コーディング領域確認

Sequences producing significant alignments:							
Select: All None Selected:0							
AT	Alignments	Download	GenPept	Graphics			
	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	spike protein, partial [Bat SARS-like coronavirus Rs4079] >qb AGZ48789.1 spike protein.	476	476	99%	2e-168	97%	AGZ48787.1
<input type="checkbox"/>	S1 protein [SARS coronavirus GD322]	493	493	99%	4e-168	98%	ABE77216.1
<input type="checkbox"/>	spike glycoprotein [SARS coronavirus GZ02]	498	498	99%	4e-164	99%	AAS00003.1
<input type="checkbox"/>	spike glycoprotein precursor [SARS coronavirus HKU-39849] >qb ADC35497.1 spike glyc	498	498	99%	4e-164	99%	ADC35483.1
<input type="checkbox"/>	spike glycoprotein [SARS coronavirus BJ302]	498	498	99%	4e-164	99%	AAR07629.1

SampleAはSARS コロナウイルスゲノムの
spikeタンパクコード領域

【追加情報】spikeタンパク質

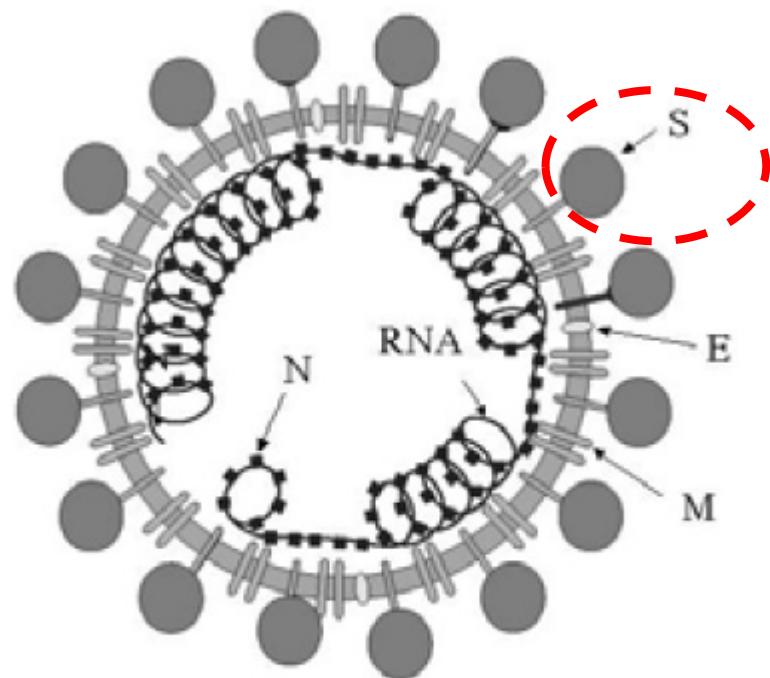


図1 コロナウイルス粒子の模式図

コロナウイルスのエンベロープには、S(spike)蛋白、M(membrane)蛋白、E(envelope)蛋白が存在し、その内部には約30kbの(+)鎖ゲノムRNAとそれに結合するN(nucleocapsid)蛋白が螺旋状の構造をなす。

"SARSコロナウイルス" 田口文広
ウイルス Vol. 53 (2003) No. 2 P 201-209

VI: SARS-CoV

ORF1a

ORF1b

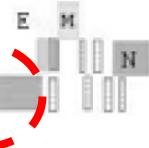


図2 コロナウイルス（グループI～IV）ゲノム構造の比較

5'末端からORF1a、1b、S、E、M、Nの遺伝子がマップされている。グループ2にはORF2とHE遺伝子がORF1bの下流にある。SARS-CoVのMとN遺伝子間に小さなORFが数個存在する。遺伝子名が明記されていないORF（□）は非構造蛋白をコードしていると考えられる。

S(spike)タンパク質コード領域は
ウイルスの持つ多くの生物活性
(受容体結合、細胞内侵入、
病原性等)に関わる重要な領域。

検索結果

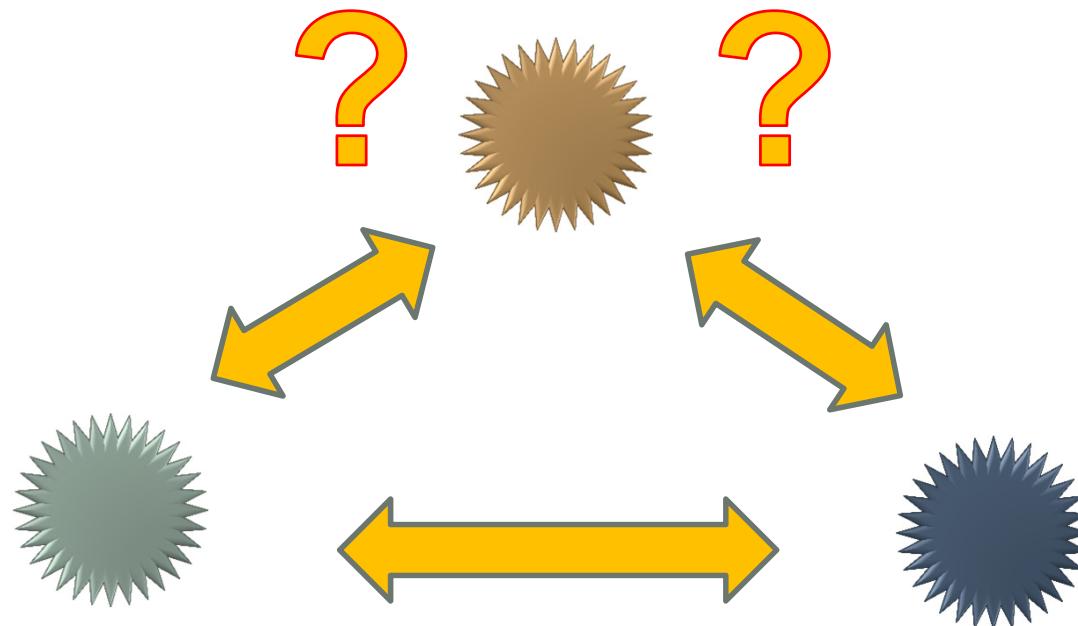
SampleAはハクビシンに感染していた
SARS コロナウイルスゲノムのspikeタンパクコード領域



- BLASTのデフォルト設定でも、高感度でデータベースから情報を探すことができる。
- アラインメントやe-valueなどで
2配列間の関係を表現可能。

※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。

3種以上の配列を一度に比較



複数種のホモログを一度に比較したい場合は？

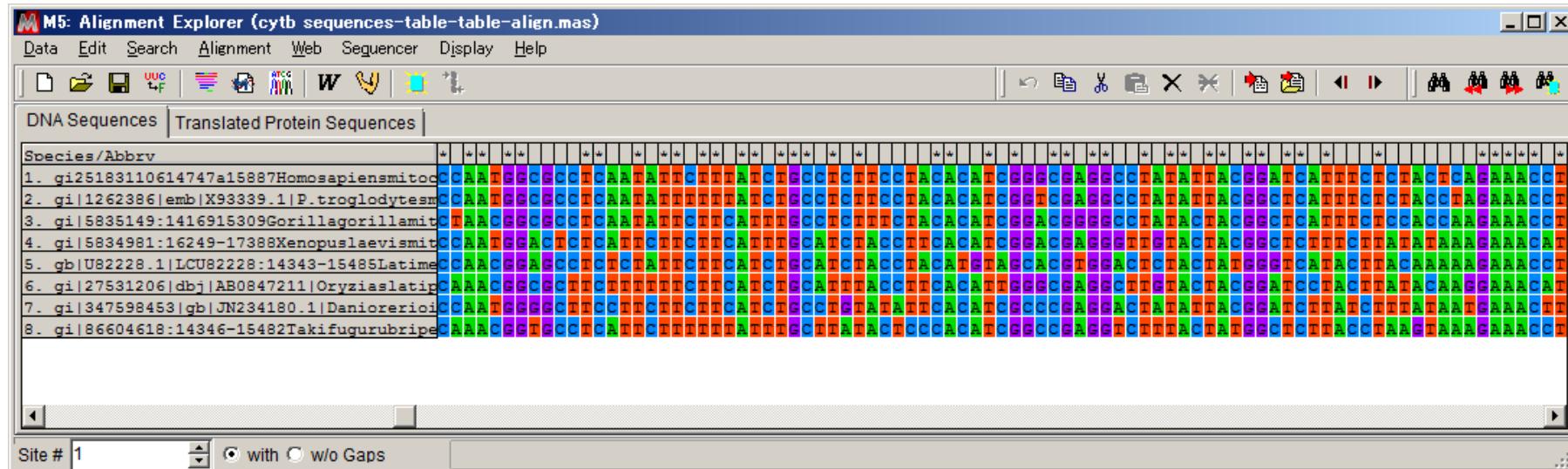


マルチプルアラインメント

4. マルチプルアラインメント

コウモリとハクビシン、ヒトから見つかったSARSウイルスの
違いは？

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 (SARSウイルスspikeタンパクコード領域)

sampleB.fas

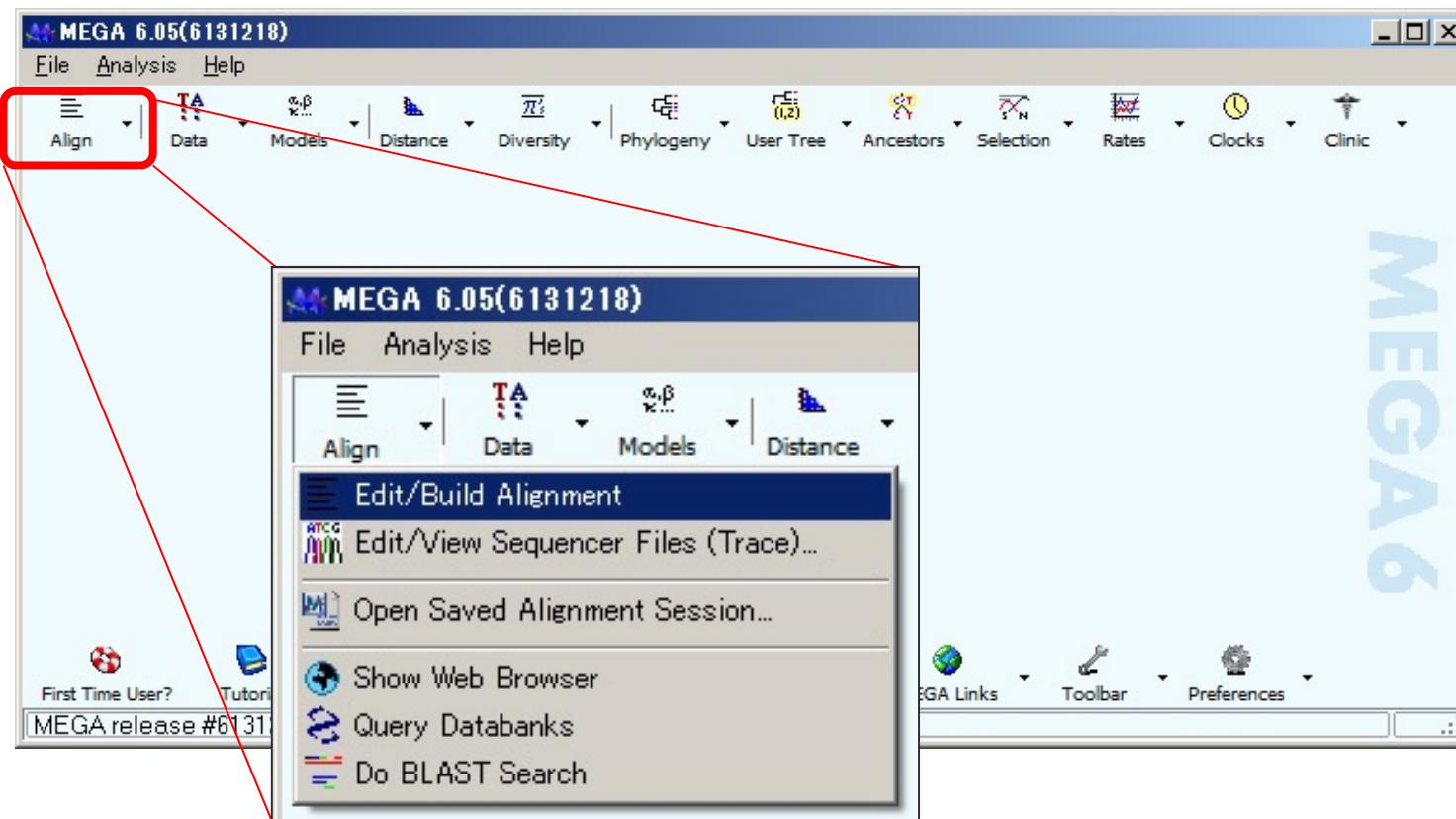
- Human_SARS_1～3
ヒトのSARS感染者から採取された3種のウイルス
- Civet_SARS_4
ハクビシン由来の1種のウイルス
- Bat_corona_5～17
コウモリから採取された13種類のコロナウイルス



```
>Human_SARS_1
GATTACTCTGTGCTCTACAACTCAACATTTTTAACCTTAAAGTGCTATGGCGTTCTGCCACTA
>Human_SARS_2
GATTACTCTGTGCTCTACAACTCAACATTTTTAACCTTAAAGTGCTATGGCGTTCTGCCACTA
>Human_SARS_3
GATTACTCTGTGCTCTACAACTCAACATTTTTAACCTTAAAGTGCTATGGCGTTCTGCCACTA
>Civet_SARS_4
GATTACTCTGTGCTCTACAACTCAACATCTTTAACCTTAAAGTGCTATGGCGTTCTGCCACTA
>Bat_corona_5
GATTACTCTGTACTCTACAACTCAACATCTTTAACCTTAAAGTGTACGGCGTTCTGCCACTA
>Bat_corona_6
GATTACTCTGTACTCTACAACTCAACATCTTTAACCTTAAAGTGTATGGCGTTCTGTCACTA
>Bat_corona_7
GATTACTCTGTACTCTACAACTCAACATCTTTAACCTTAAAGTGTATGGCGTTCTGCCACTA
>Bat_corona_8
GATTACACTGTTCTACAACTCAACTTCATTTAACCTTAAATGTTATGGAGTTCTCCCTCTA
>Bat_corona_9
GACTACACTGTTCTACAACTCAACCTCTTCGACTTTAACCTTAAATGTTATGGAGGTCTCCATCTA
>Bat_corona_10
GACTACACTGTTCTACAACTCAACTCTTCTCAACTTTAACCTTAAAGTGCTATGGAGTTCTCCCTCTA
>Bat_corona_11
GATTACACTGTTCTACAACTCACTCGACCTCTTCACACTTTAACCTTAAATGTTATGGAGTTCTCCCTCTA
>Bat_corona_12
GATTACACTGTTCTACAACTCAACTCTTTTGACTTTCAACTTCACTTAAATGTTATGGAGTTCTCCCTCTA
>Bat_corona_13
GATTACACTGTTCTACAACTCAACTTCATTTAACCTTAAATGTTATGGAGTTCTCCCTCTA
>Bat_corona_14
GATTACACTGTTCTACAACTCAACTTCATTTAACCTTAAATGTTATGGAGTTCTCCCTCTA
>Bat_corona_15
GACTATACGGCTTTTACAACTCAACCTTTCAACTTCAAAATGCTACGGAGTTCTCCCTCTA
>Bat_corona_16
GATTACACTGTTCTACAACTCAACTTCATTTAACCTTAAATGTTATGGAGTTCTCCCTCTA
>Bat_corona_17
GACTACTCAGTGCTTACAATTCTTGCCCTCTCAACATTCAAGTGTATGGCGTTACCTA
```

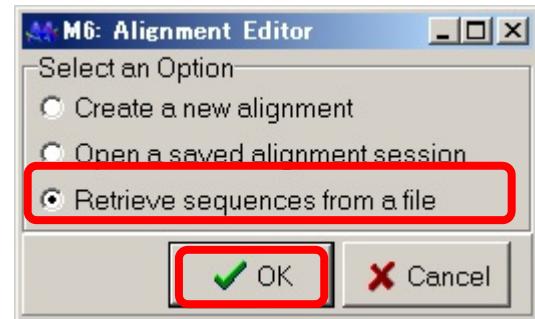
17種類のspikeタンパクコード領域の配列を比較してみましょう。

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

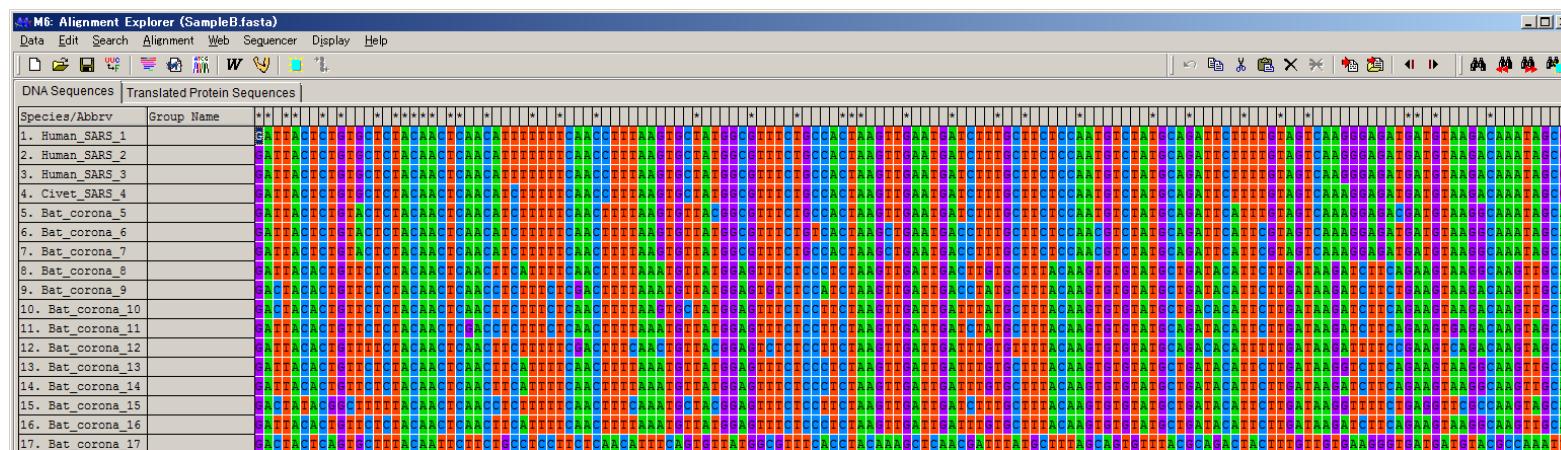
Alignment Explorerの起動



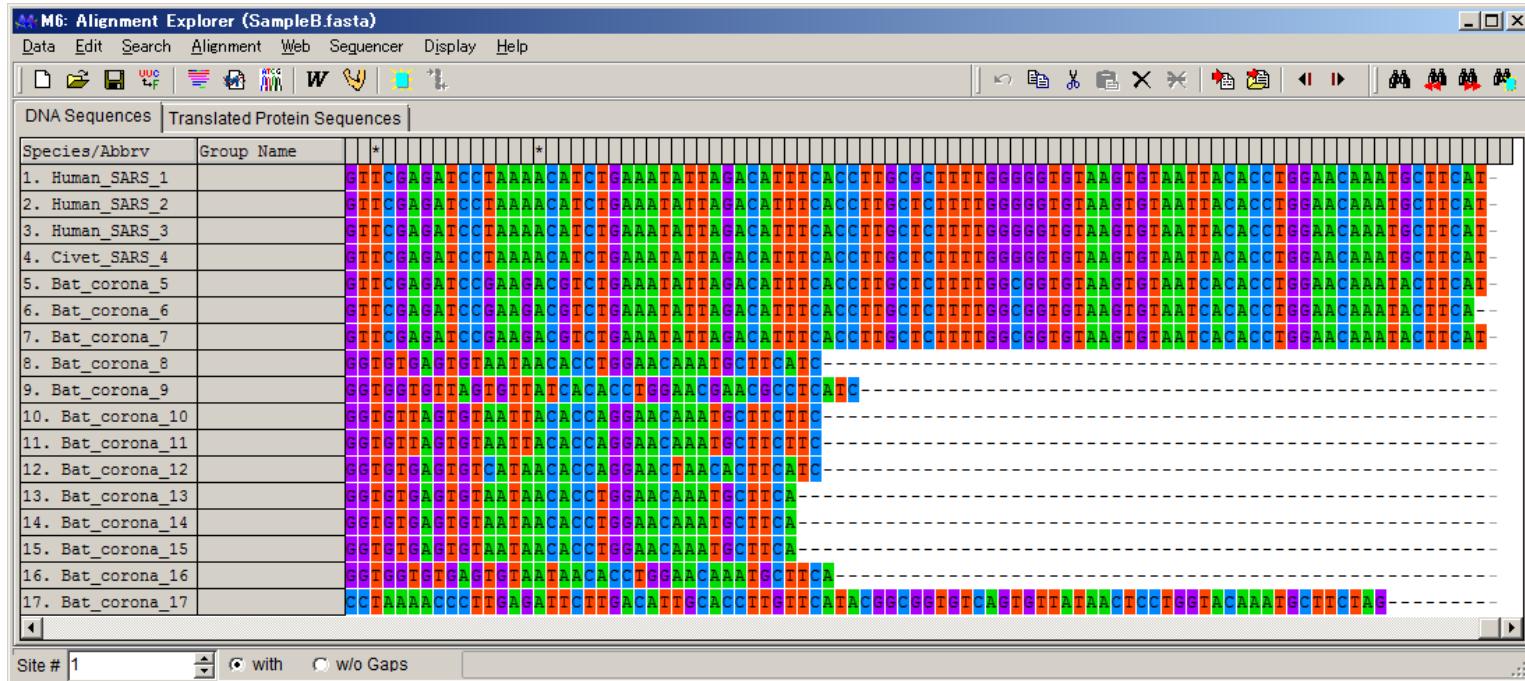
②「Retrieve sequences from a file」を選択し「OK」
→sampleB.fas
ファイルを選択



Alignment Explorer起動



SampleB:



約700bpの各ウイルスのspikeタンパクのホモログが
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)

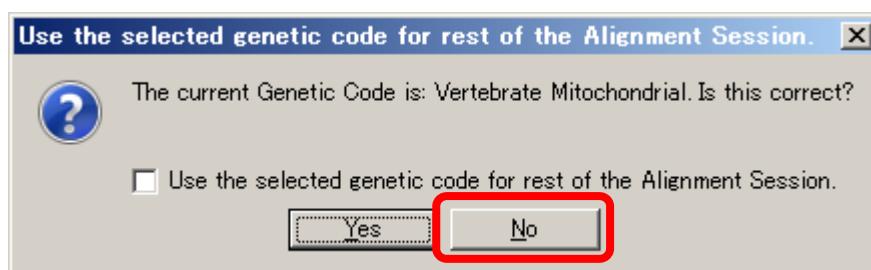
アミノ酸に翻訳

※フレームシフトについては割愛

The screenshot shows the M6 Alignment Explorer interface with 'SampleB.fasta' loaded. The 'Translated Protein Sequences' tab is highlighted with a red box and a callout bubble pointing to it with the text 'クリック' (Click). The main window displays two DNA sequences from 'Human_SARS_1' and 'Human_SARS_2'. The sequences are color-coded by codon: G (green), A (blue), T (orange), C (purple), T (orange), G (green), C (purple), T (orange), A (blue), C (purple), A (blue), A (blue). Above the sequences, there are several icons, including one labeled 'W' which is highlighted with a red box.

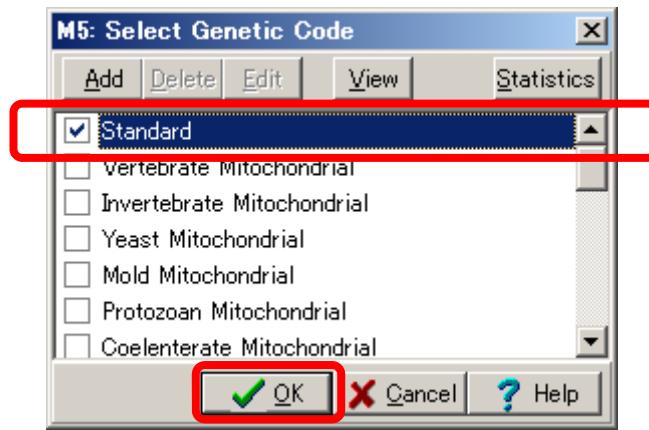


コドン表の確認



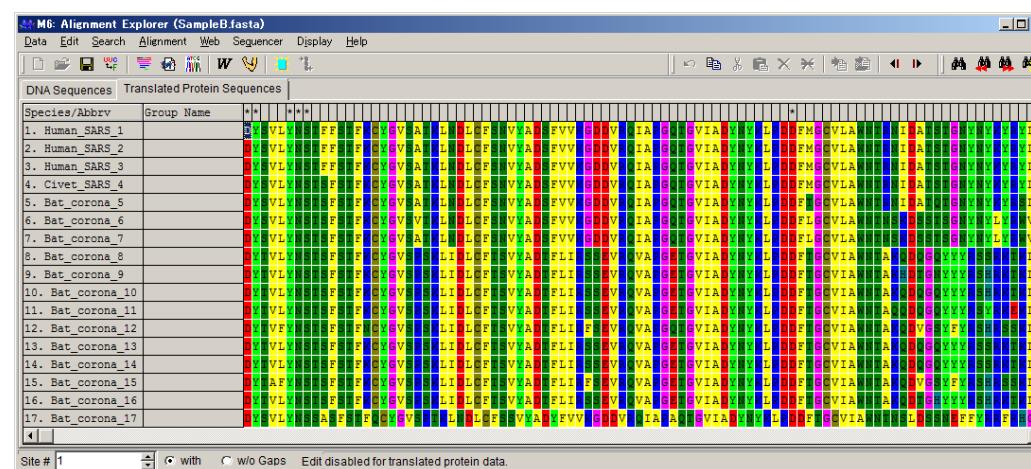
脊椎動物のミトコンドリア配列
ではないので「No」

アミノ酸に翻訳



「Standard」にチェックを入れ
「OK」

变换



アミノ酸に翻訳

6種で保存

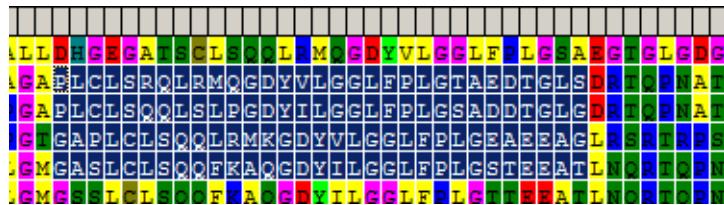
5'末端がM(メチオニン:開始コドン)
※今回の配列は部分配列のため
開始コドンになっていません。

コドンの途中で
途切れている

終始コドン

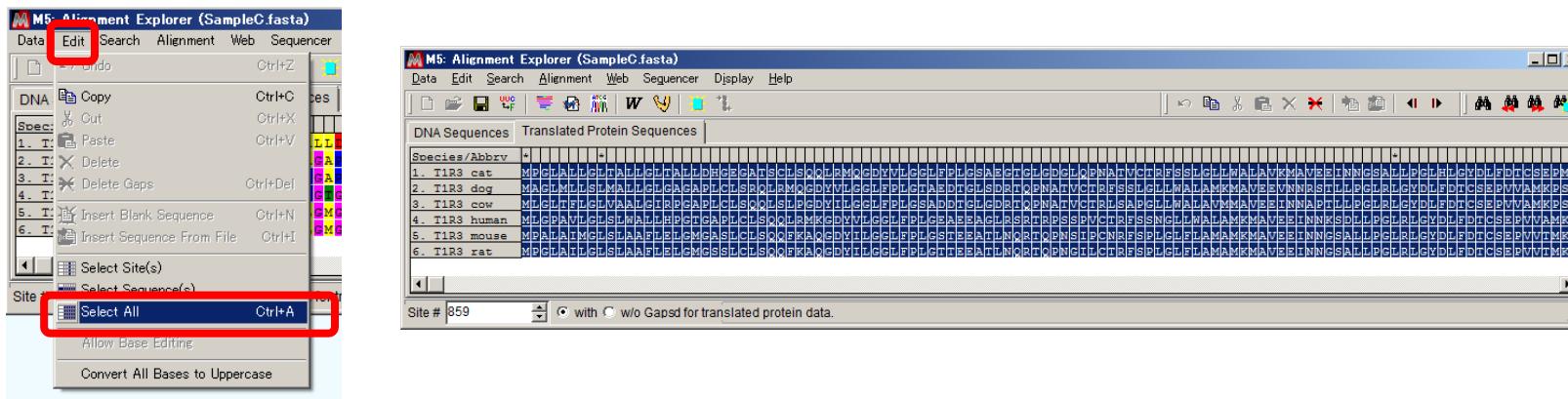
アミノ酸配列も左寄せで記述されていて、
欠失や挿入が考慮されていません。
手作業によるアラインメントは時間がかかります。

アミノ酸の自動アラインメント

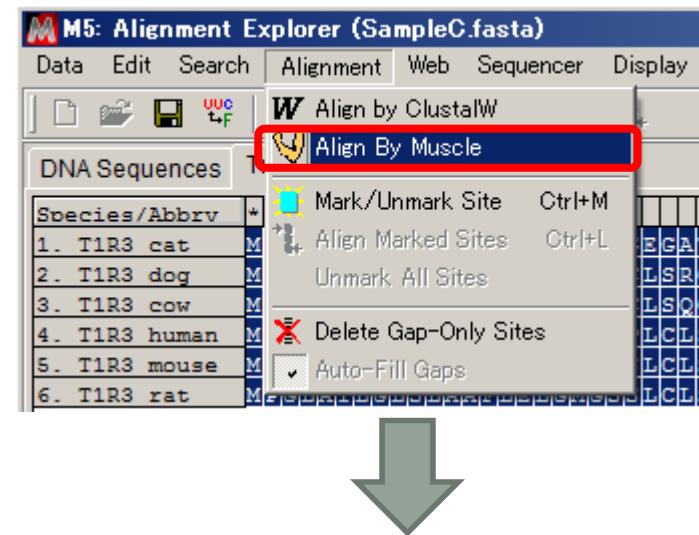


自動アラインメントは
マウスで選択した
「色が反転している範囲」だけ
実施される。

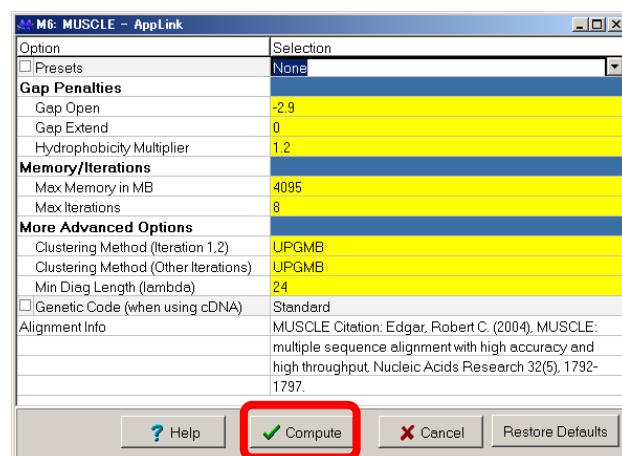
「Edit」→「Select All」で全体を選択



アミノ酸の自動アライメント

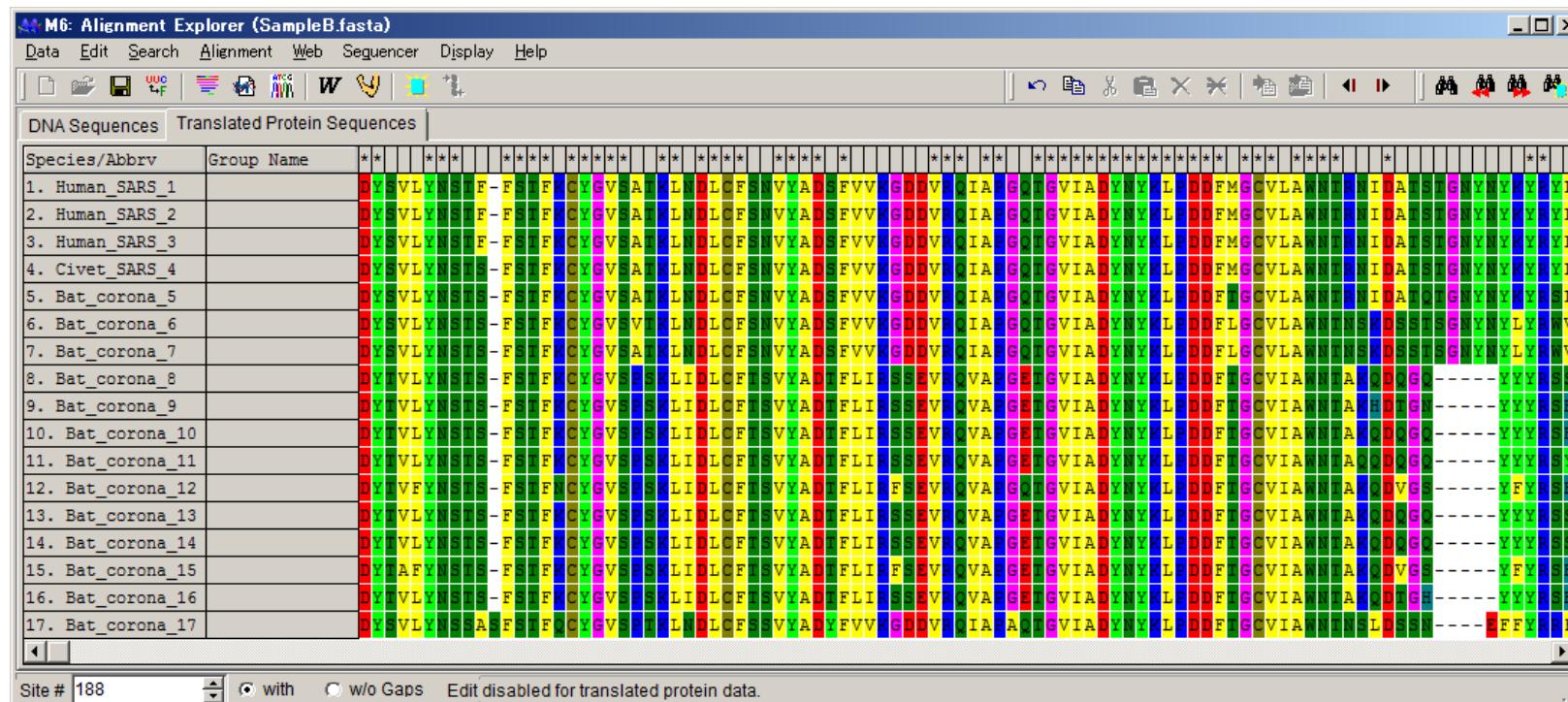


「Alignment」
→「Align by Muscle」を選択



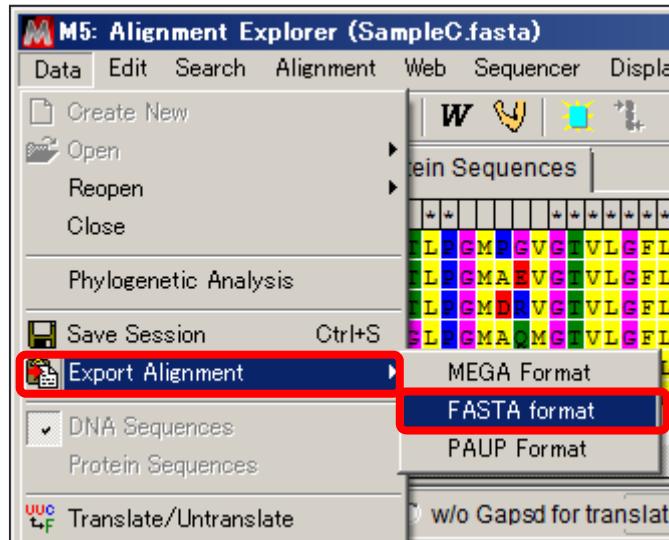
Muscleのアライメントパラメータ設定
→デフォルトのままで「Compute」をクリック

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

→「Srotein」と名前を付けて保存

Fasta形式の **Srotein.fas** が保存されます。

※このファイルは次の結果の可視化と、
系統樹構築に使用します。

短いアライメントの可視化

70番目から90番目までの配列の保存の程度を
グラフィカルに見てみたい。

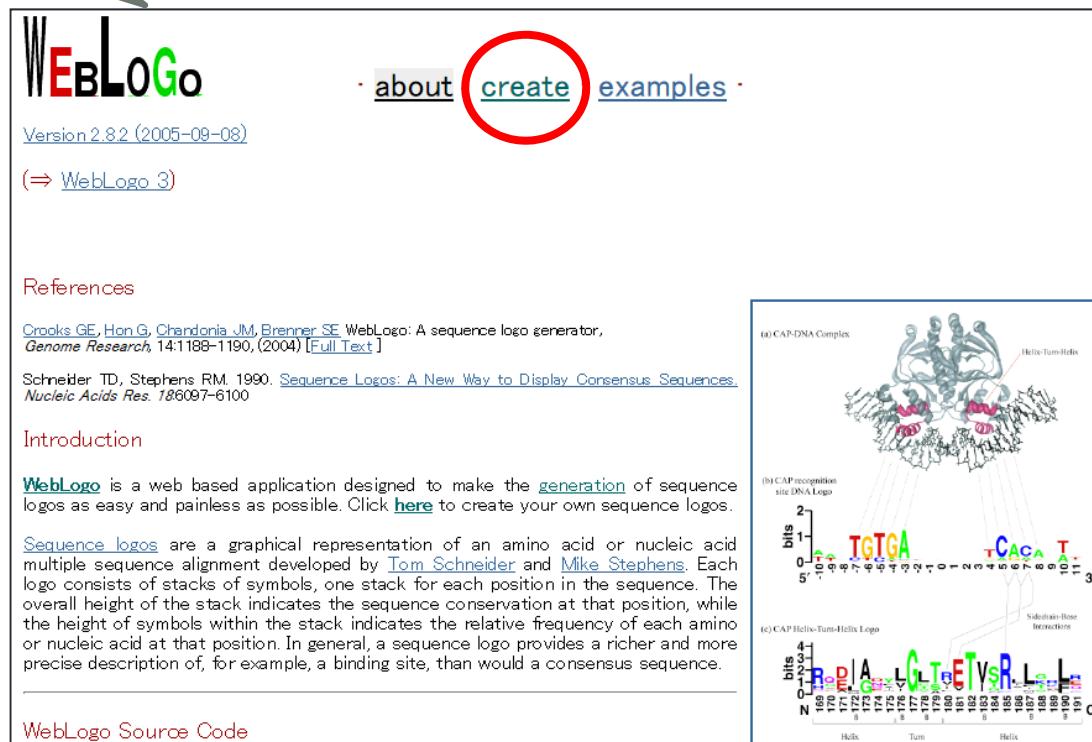


Webサービスとして提供されている
「**WebLogo**」を利用して図示します。

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. On the right side of the page, there are three small diagrams labeled (a), (b), and (c) illustrating different sequence logo applications.

Version 2.8.2 (2005-09-08)
 (⇒ [WebLogo 3](#))

References

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. *Genome Research*, 14:1188-1190, (2004) [Full Text]
 Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* 18:6097-6100

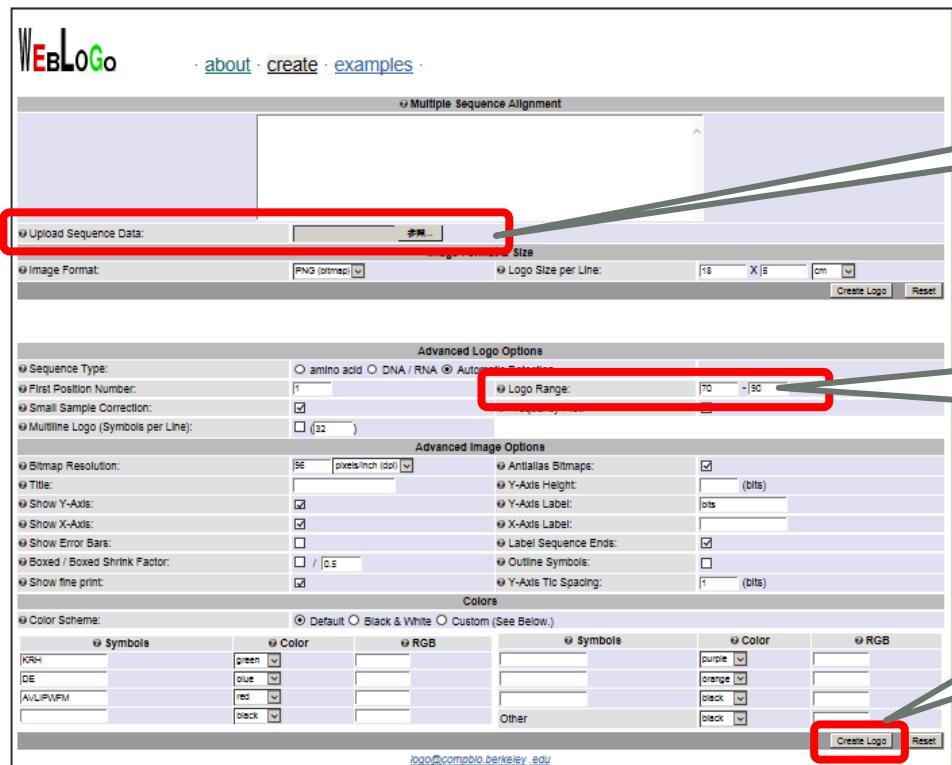
Introduction

[WebLogo](#) is a web based application designed to make the [generation](#) of sequence logos as easy and painless as possible. Click [here](#) to create your own sequence logos.

Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by [Tom Schneider](#) and [Mike Stephens](#). Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

[WebLogo Source Code](#)

WebLogoで保存領域の図を作成する



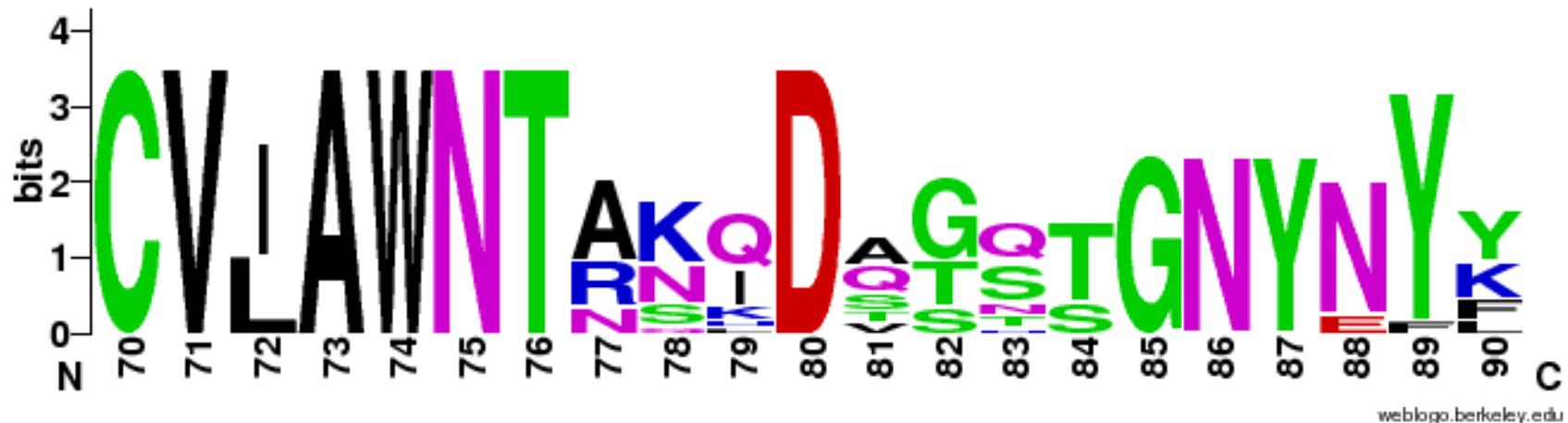
①先ほどのSprotein.fasを選択

②Logo Rangeに「70」～「90」と入力

③Create Logo ボタンをクリック

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

WebLogoで保存領域を図示



Sproteinの70～90番目のアミノ酸保存領域
→サイトごとに縦に文字が長いほど保存されている

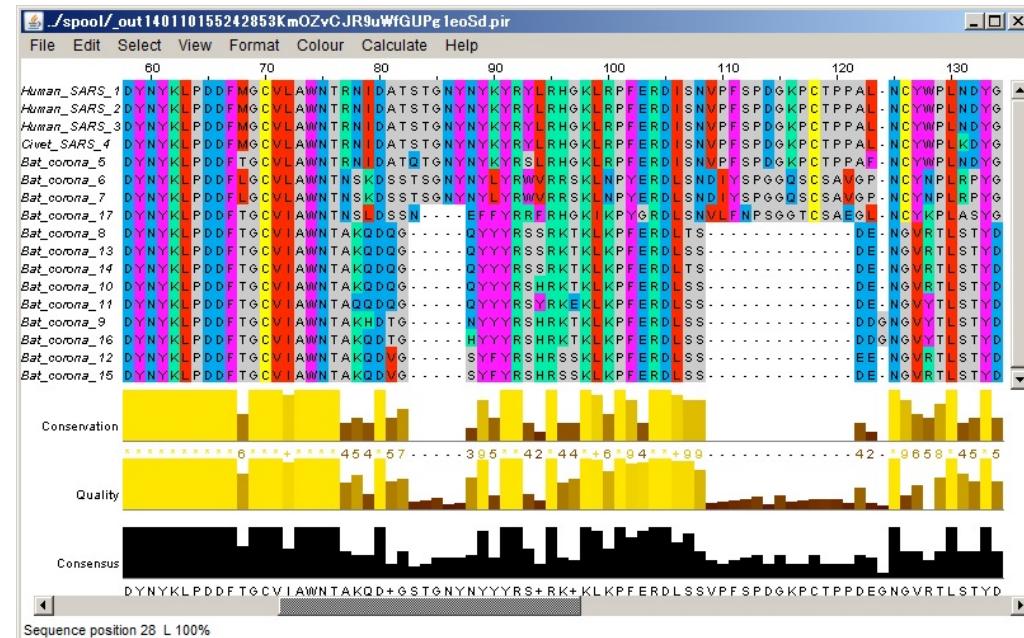
アラインメント全体の保存領域可視化は？ →Jalview

アラインメント全体を俯瞰的に可視化: Jalview

- OSを選ばず用いられる多機能なアラインメントエディタ
- スタンドアロンのみならず、ウェブブラウザに埋め込める
→ 今回はWeb上で既に公開されているサービスを利用

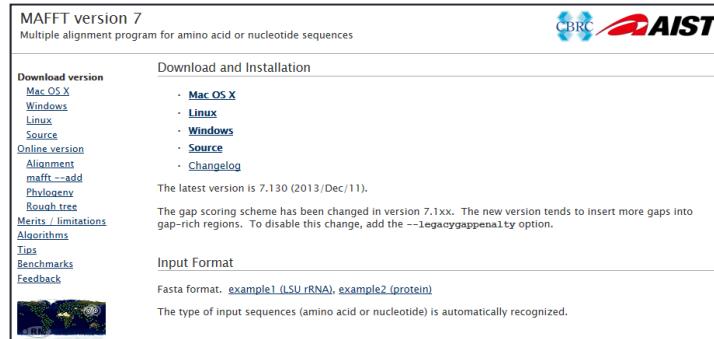
The screenshot shows the Jalview website homepage. It features a header with the Jalview logo and navigation links for Home, About, Help, Community, Development, Training, and Download. Below the header is a "Latest News" section with a vacancy for a Jalview Software Engineer posted on 01-01-2014. Another section mentions the "Last Day to register for the 4th Jalview and JABA residential workshop" posted on 04-12-2013. There are also sections for "Job vacancies for a Jalview developer and training/outreach officer in Dundee" posted on 22-03-2013 and a "View All news" link. On the right side, there's a "Analysis Services" section listing various tools like Jpred3, JABAWES, and Pfam, along with a "Biological Data Services" section. A "The Jalview Desktop" section shows a screenshot of the desktop application interface.

<http://www.jalview.org/>



MAFFTからJalviewへアクセス

- MAFFT = 多機能多重アラインメントツール。
→Jalviewの表示機能も搭載している、
産業技術総合研究所のMAFFT公開サービスへアクセス



<http://mafft.cbrc.jp/alignment/software/>

アラインメント・系統樹構築も可能な配列操作ツール。

→MAFFTの活用については、「AJACS蝦夷3」の理研CDB 原雄一郎先生の発表資料を参照

MAFFTのWebサービスへアクセス

①「MAFFT」で検索

②Alignmentをクリック

MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

Download and Installation

- Mac OS X
- Windows
- Linux
- Source
- Online version
- Alignment**
- mafft --add
- Phylogeny
- Rough tree
- Merits / limitations
- Algorithms
- Tips
- Benchmarks
- Feedback

The latest version is 7.130 (2013/Dec/11).

The gap scoring scheme has been changed in version 7.1xx. The new version tends to insert more gaps into gap-rich regions. To disable this change, add the --legacygappenalty option.

Input Format

Fasta format. [example1 \(LSU rRNA\)](#), [example2 \(protein\)](#)

The type of input sequences (amino acid or nucleotide) is automatically recognized.

CBRC AIST

<http://mafft.cbrc.jp/alignment/software/>

MAFFTへアラインメント結果をアップロード

MAFFT version 7
Multiple alignment program for amino acid or nucleotide sequences

Download version
Mac OS X
Windows
Linux
Source

Online version
Alignment
mafft→add
Phylogeny
Rough tree
Methods / Limitations
Algorithms
Tips
Benchmarks
Feedback

CBRC AIST

All jobs are reset at 4:00AM (JST) every Sunday.

Multiple sequence alignment and NJ / UPGMA phylogeny

MAFFT (2018 Sep 27)
We made a change in the scoring scheme in version 7.110.
For problems that require many gaps, alignment quality is (expected to be) improved.
For conserved dataset, the difference is small.

Scoring scheme:
 Now
 Conventional

Input:
Paste protein or DNA sequences in fasta format. [Example](#)

or upload a plain text file:

Use structural alignment(s)
 Allow unusual symbols (Selenocysteine "U", Inosine "I", non-alphabetical characters, etc) [Help](#)

UPPERCASE / lowercase:
 Same as input
 Amino acid → UPPERCASE / Nucleotide → lowercase

Direction of nucleotides sequences:
 Same as input
 Adjust direction according to the first sequence (accurate enough for most cases) [Beta](#)
 Adjust direction according to the first sequence (only for highly divergent data; extremely slow) [Beta](#)

Output order:
 Same as input
 Aligned

Notify when finished (optional; recommended when submitting large data):
Email address:

③先ほどのSprotein.fasを選択

④Submit をクリック

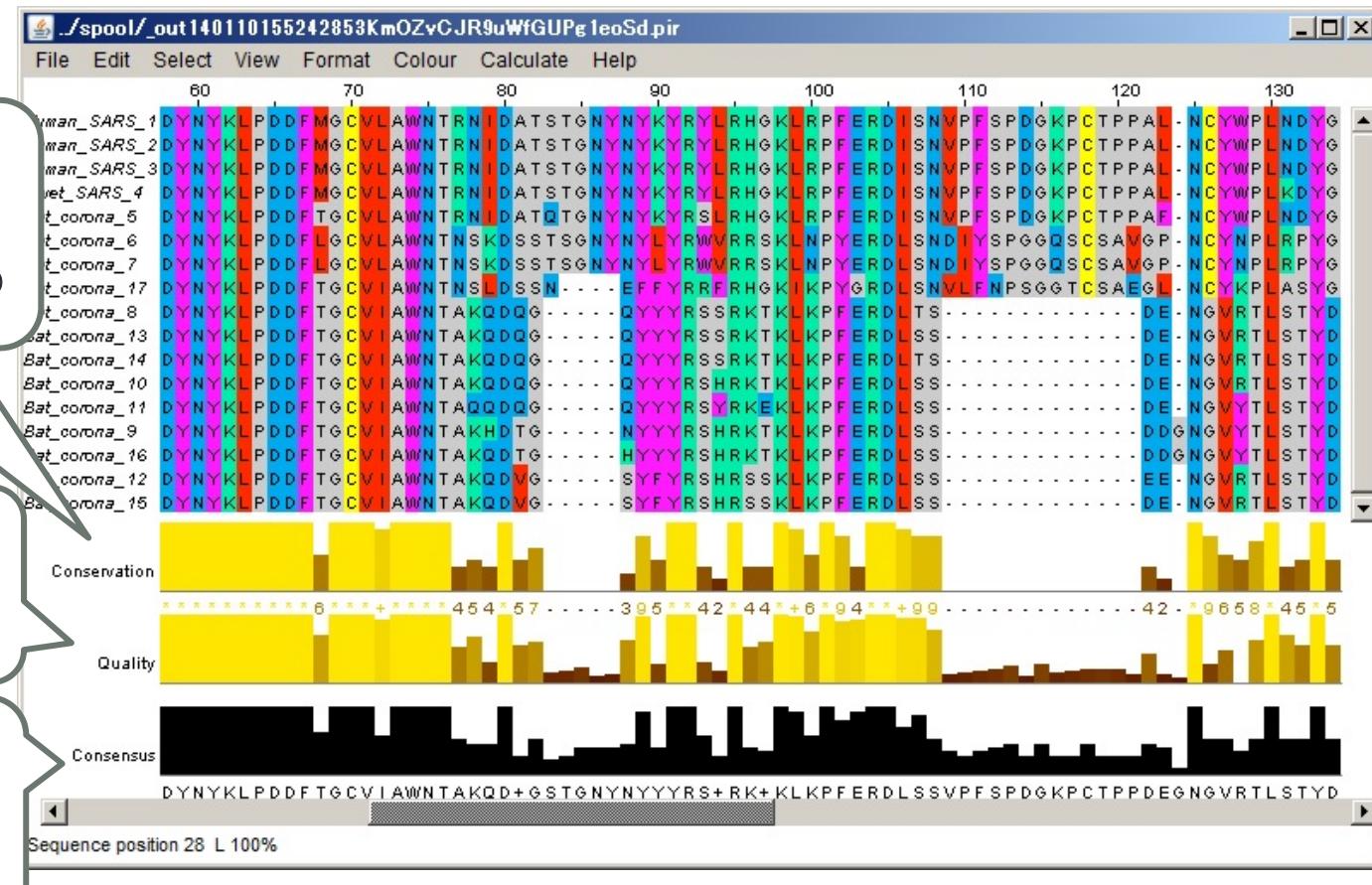
<http://mafft.cbrc.jp/alignment/server/>

Jalview アラインメントビューワ起動



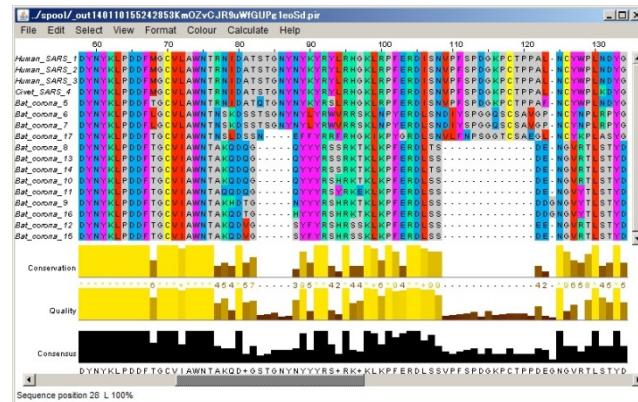
※動作がjava依存のため、
Webブラウザのバージョンにより
ボタンが表示されない場合があります。
別ブラウザを使用するか、
javaをバージョンアップすることで
改善する場合があります。

Jalview表示



"Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation." Livingstone C.D. and Barton G.J. (1993) CABIOS 9, 745-756

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、Jalview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

SARSウイルスはどのように広まったか？

系統樹



配列間の、遺伝的距离と関係性を
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**

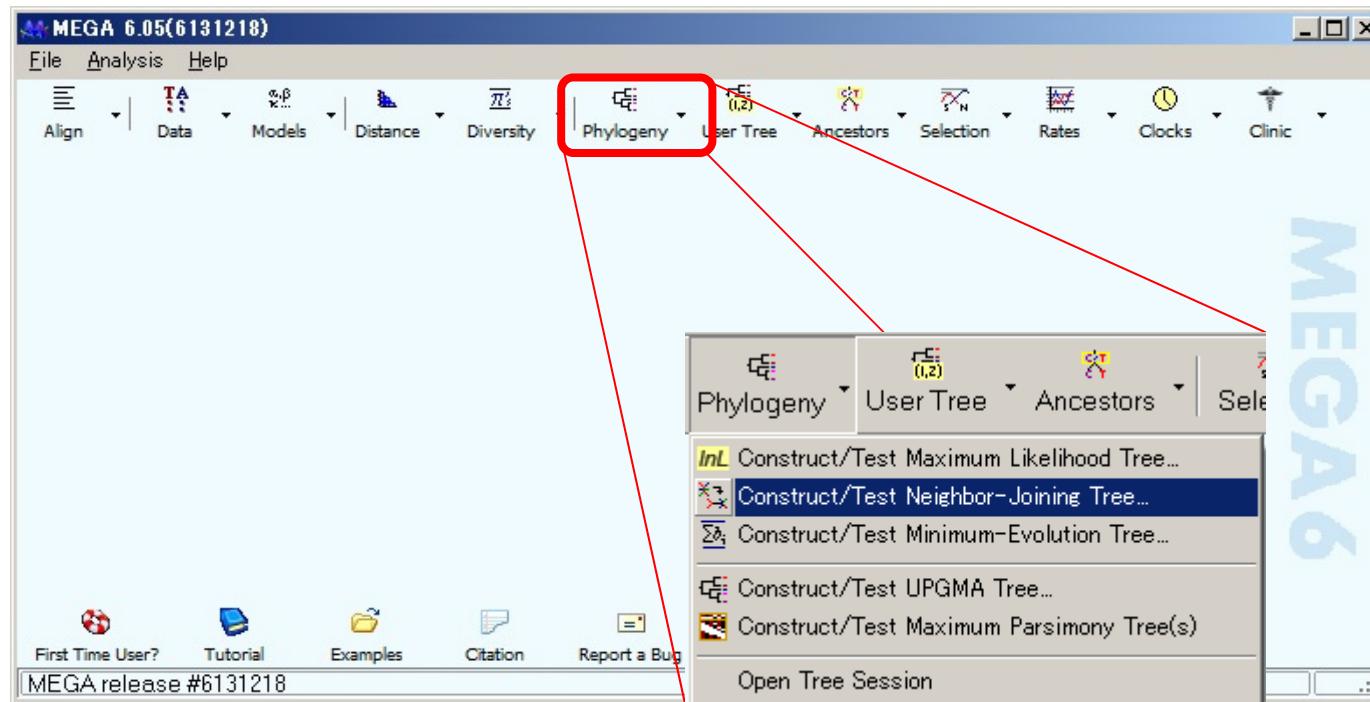
塩基・アミノ酸置換のモデルを明示した上で、そのモデルが実際の進化の過程の近似になっているかを評価基準とする。詳細な設定が可能で信頼度が高い。時間がかかる。

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。

Sproteinの系統樹構築

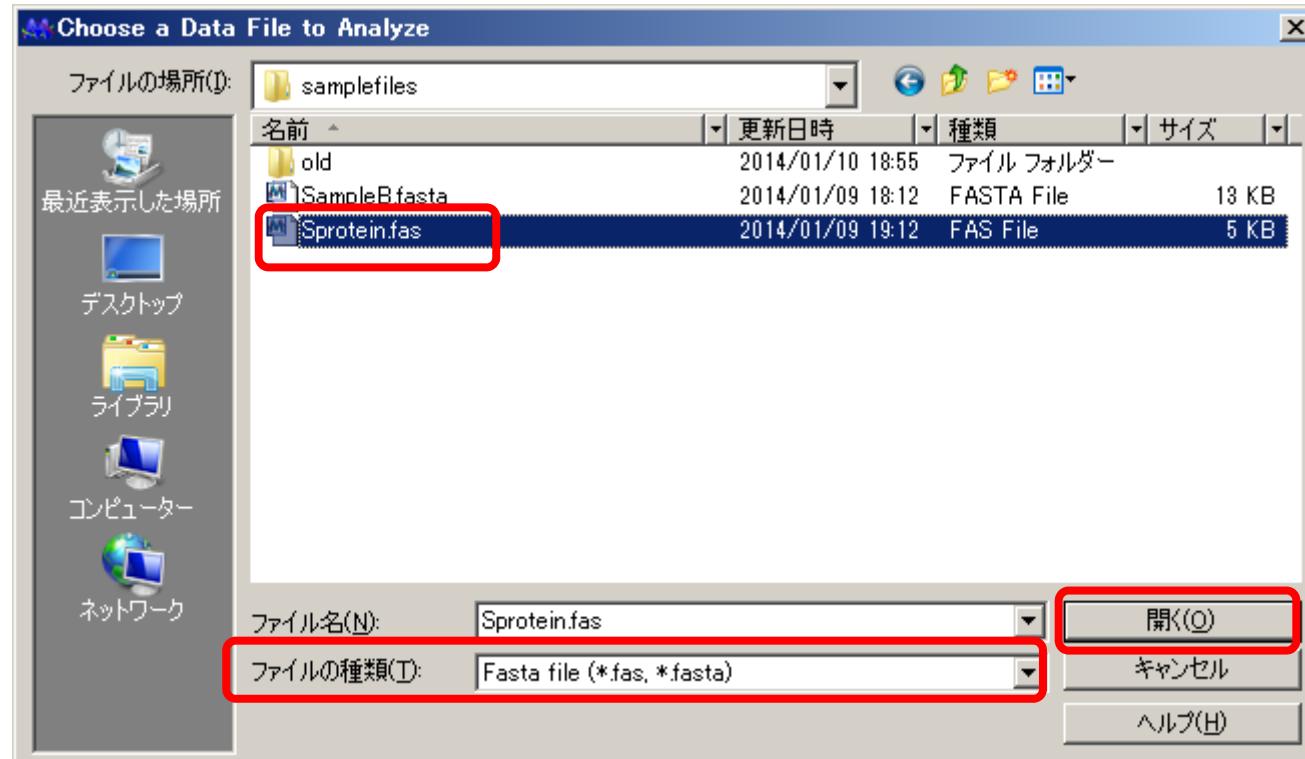
1. 近隣結合法



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

Sproteinの系統樹構築

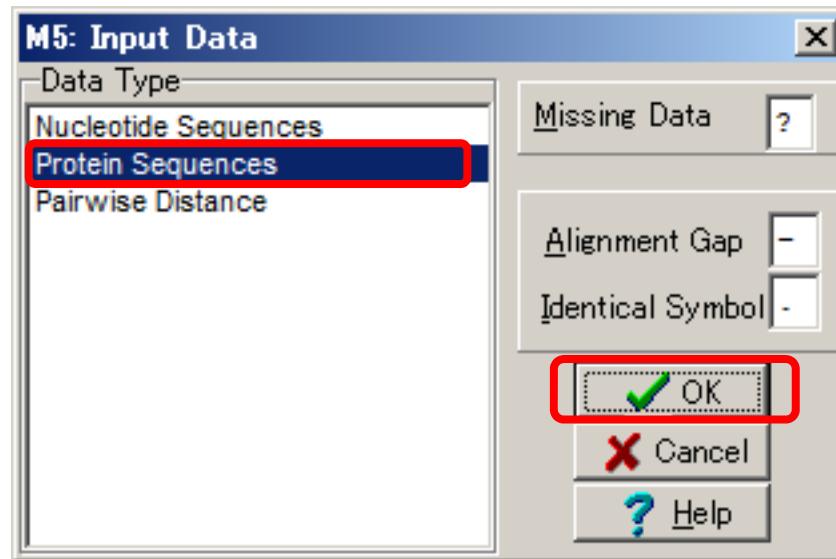
1. 近隣結合法



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

Sproteinの系統樹構築

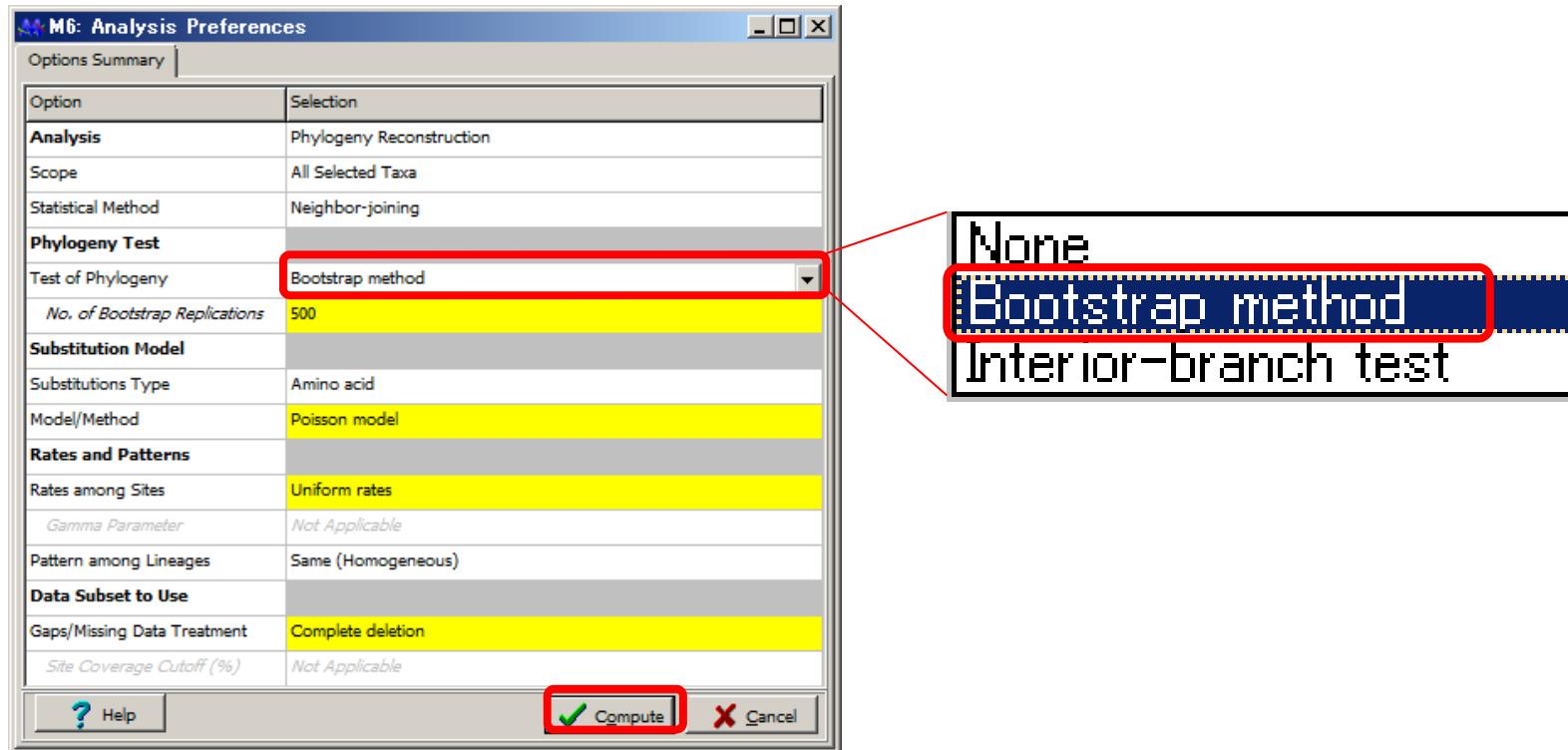
1. 近隣結合法



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

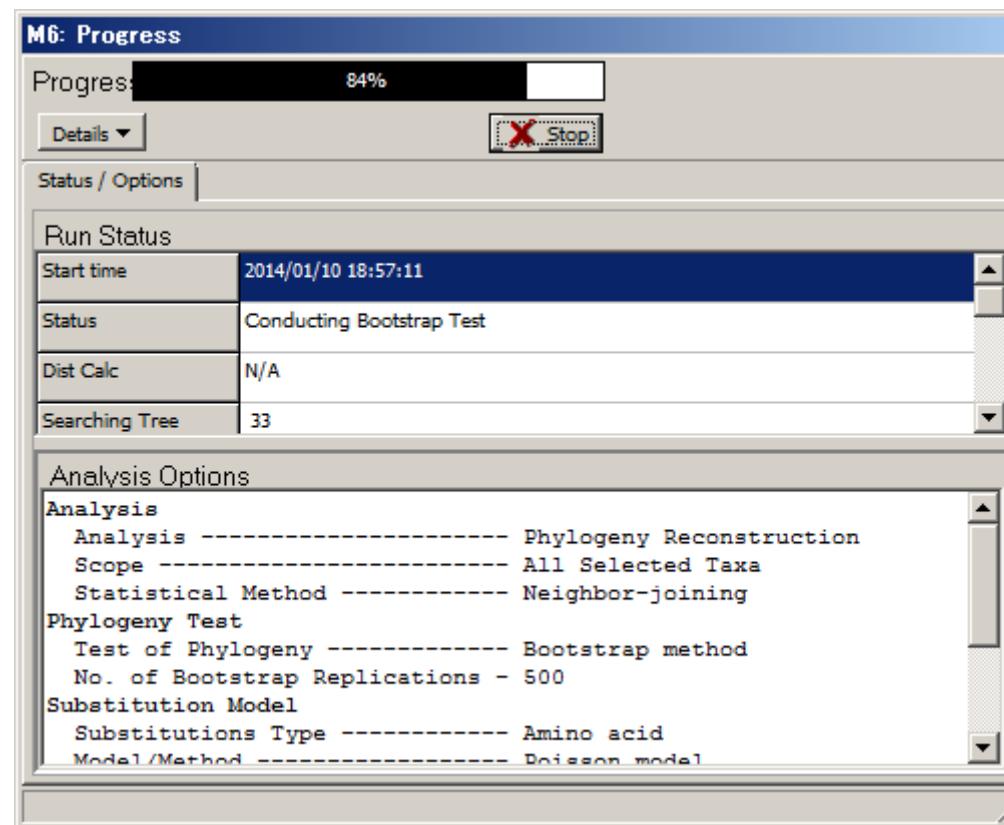
Sproteinの系統樹構築

1. 近隣結合法



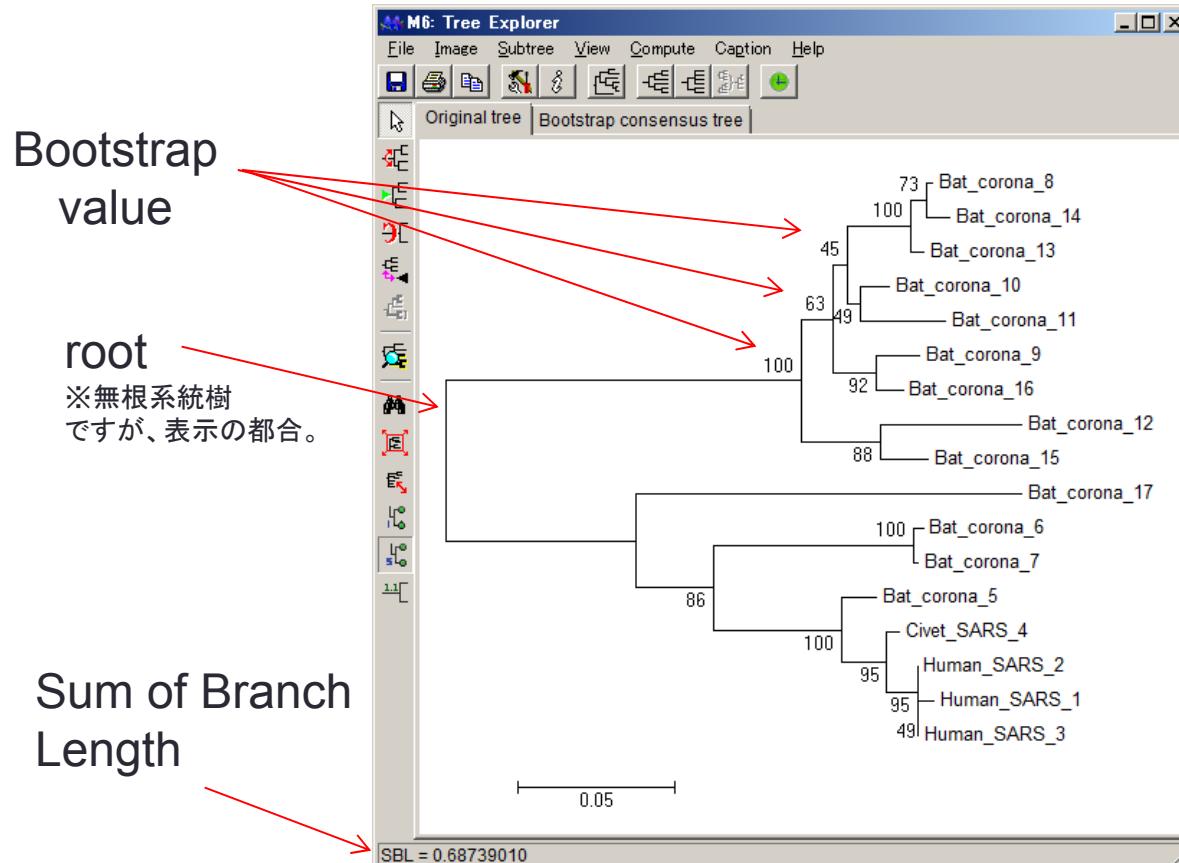
④Test of Phylogeny:Bootstrap methodを選択し、「Compute」

計算中…



Sproteinの系統樹構築

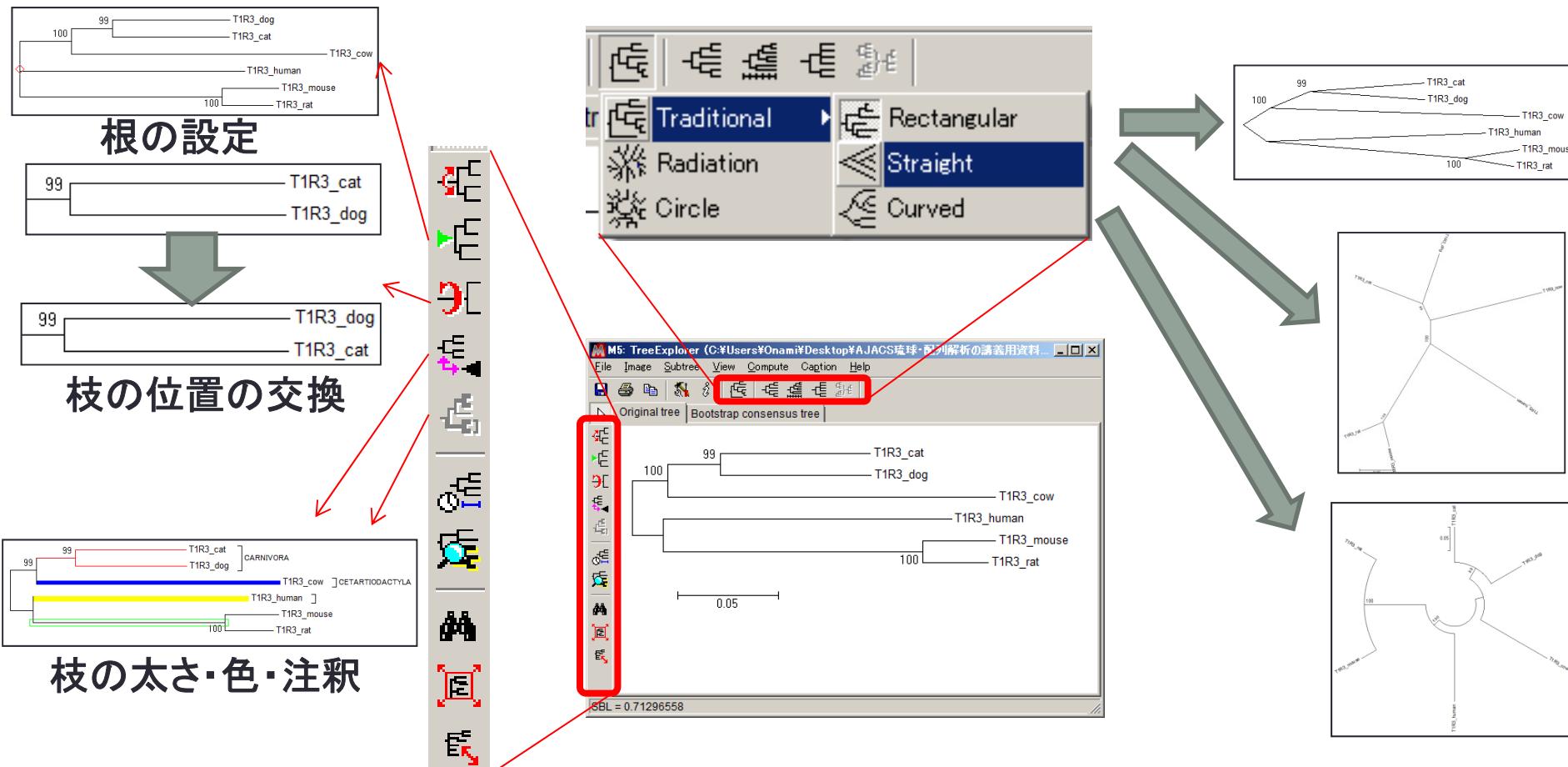
1. 近隣結合法



近隣結合系統樹を構築することができました。

Sproteinの系統樹構築

1. 近隣結合法

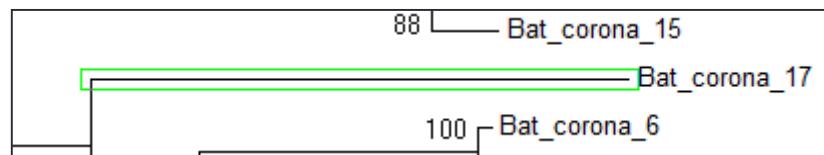


様々な表現の系統樹を作成することが可能

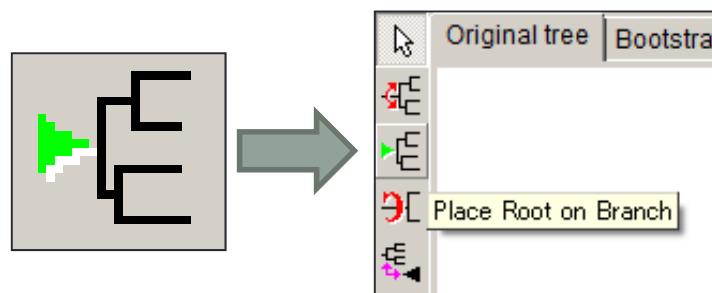
Sproteinの系統樹構築

1. 近隣結合法

Bat_corona_17は外群に位置することが知られている。
→Bat_corona_17の根元を根になるよう編集



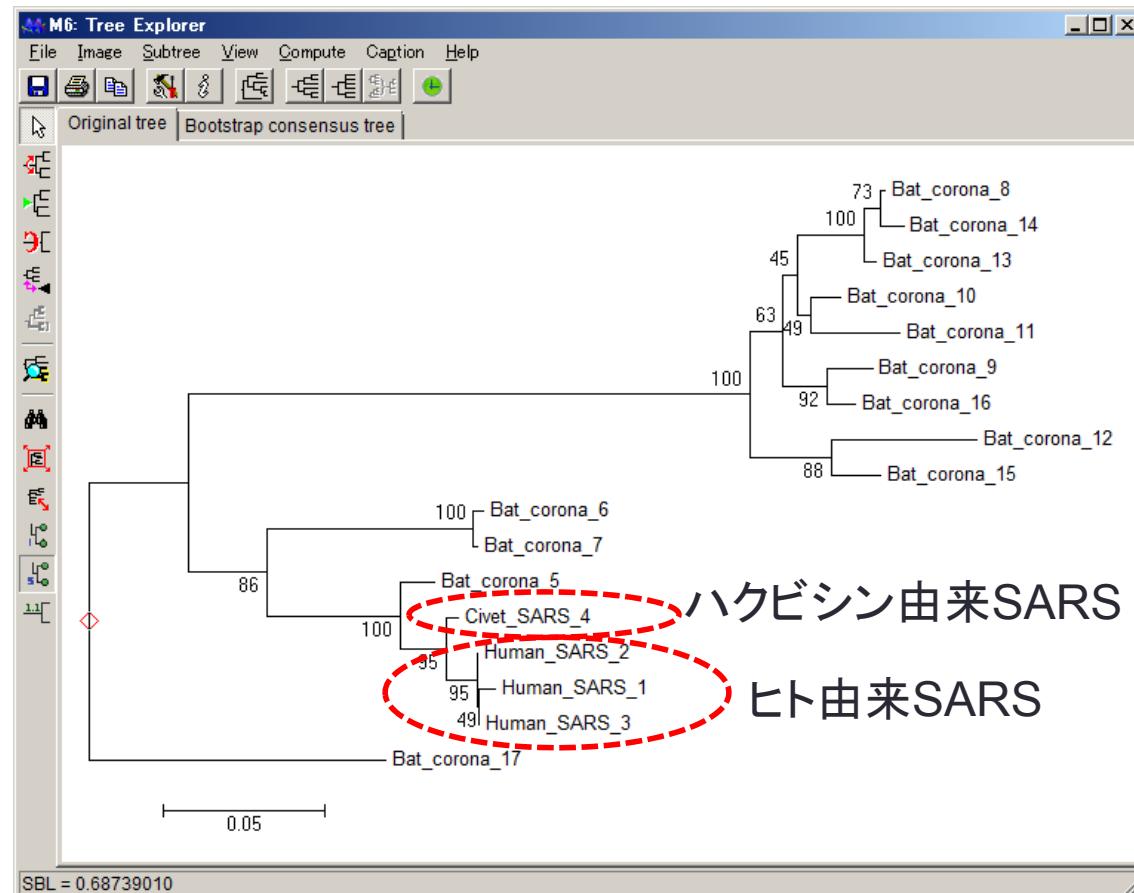
①Bat_corona_17の枝をクリックすると
緑色の枠で選択されます。



②緑色の右向き三角がついた
Place Root on Branch ボタンを
クリック。

Sproteinの系統樹構築

1. 近隣結合法



Bat_corona_17を外群とする系統樹ができました。

Sproteinの系統樹構築

1. 近隣結合法

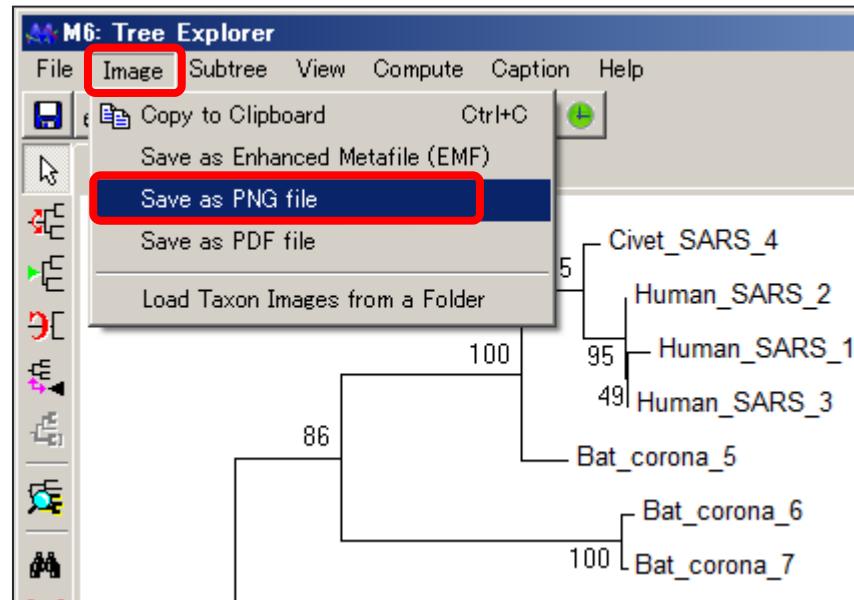


Image から「Save as PNG file」,「Save as PDF file」で
系統樹をPNGやPDFファイルとして保存可能

Sproteinの系統樹構築

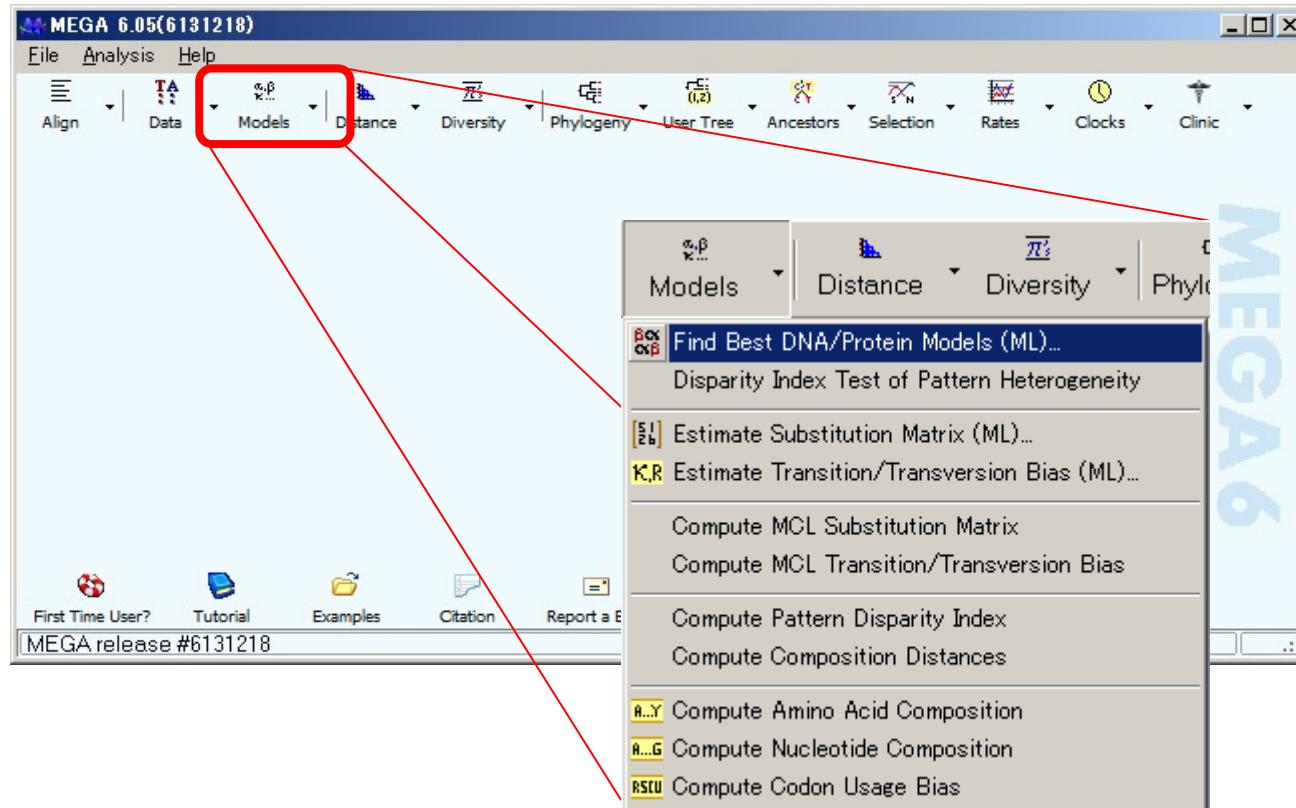
2. 最尤法

より信頼度の高い系統樹を得るために、
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

Sproteinの系統樹構築

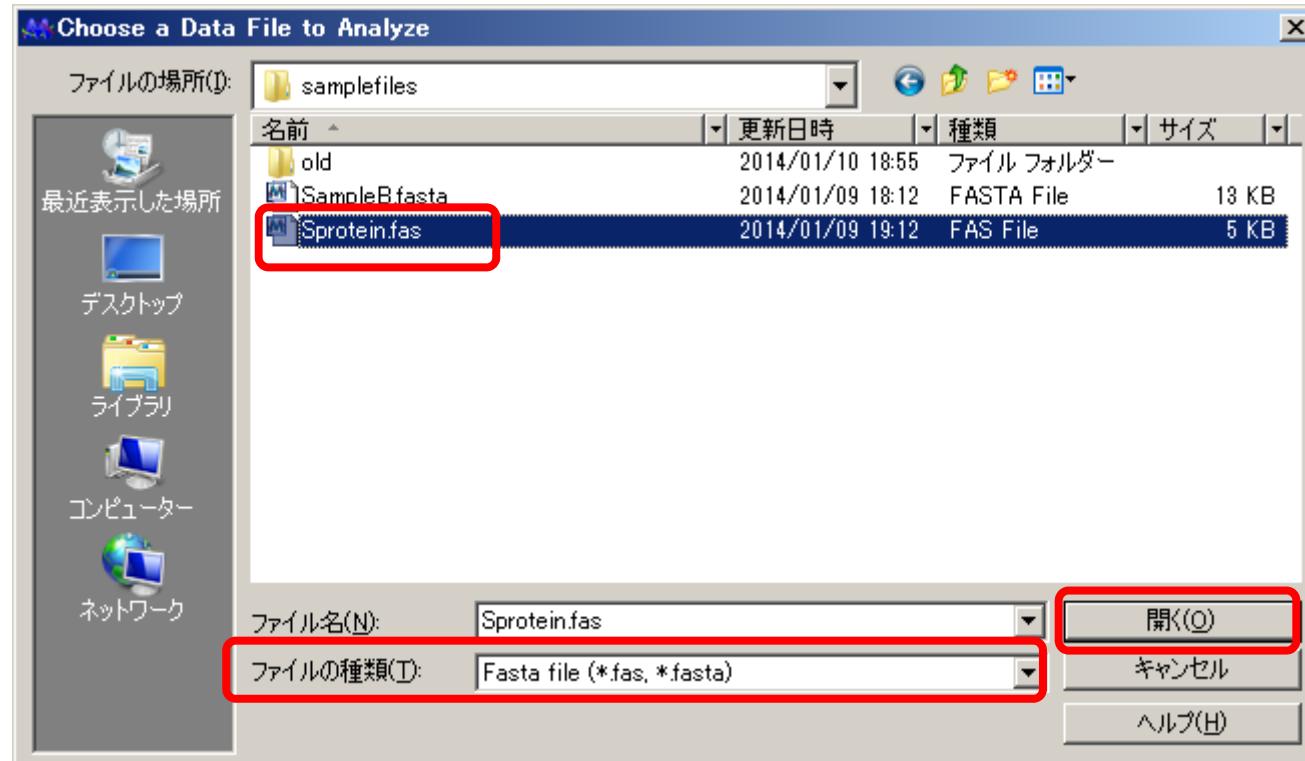
2. 最尤法(モデル選択)



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

Sproteinの系統樹構築

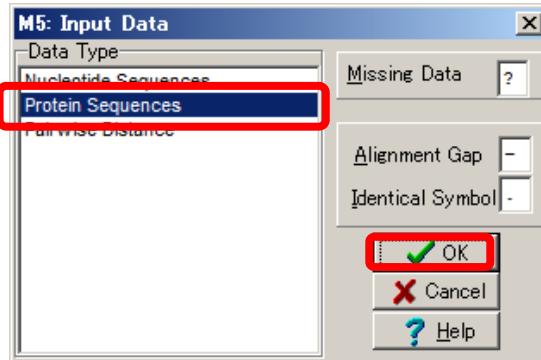
2. 最尤法(モデル選択)



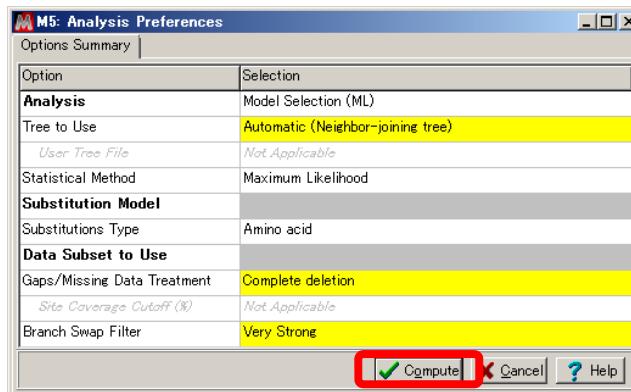
②ファイルの種類:「Fasta file (*.fas, *.fasta)」を選択し、先ほど保存したSprotein.fasを選択する。

Sproteinの系統樹構築

2. 最尤法(モデル選択)



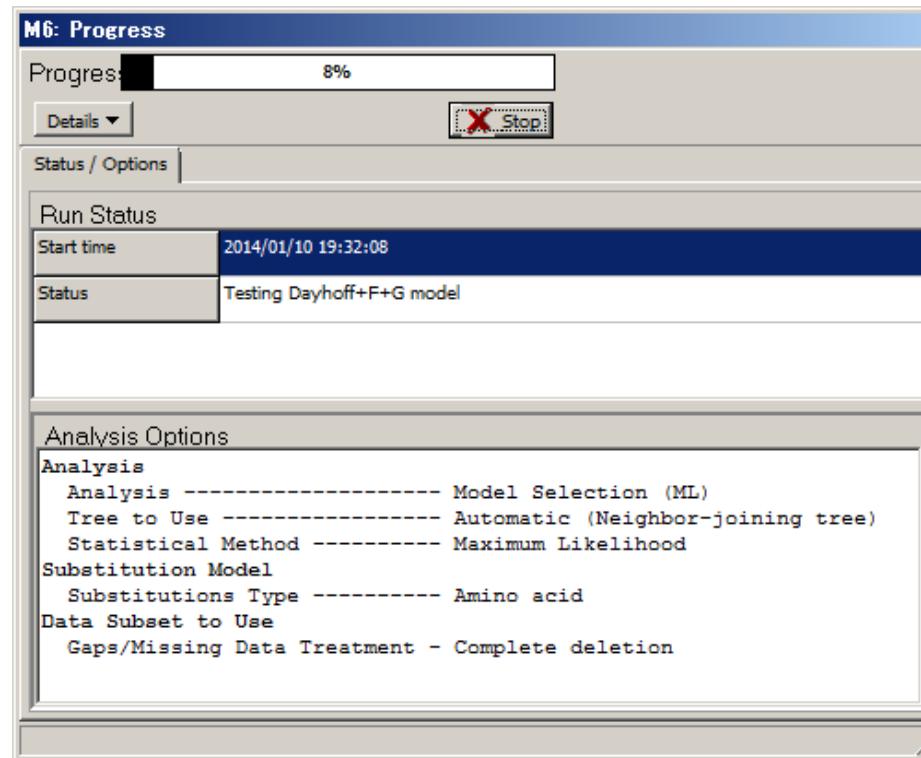
③Data Type: Protein Sequencesを選択し、「OK」



④そのまま「Compute」

Sprotein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8分程度かかります)

Sproteinの系統樹構築

2. 最尤法(モデル選択)

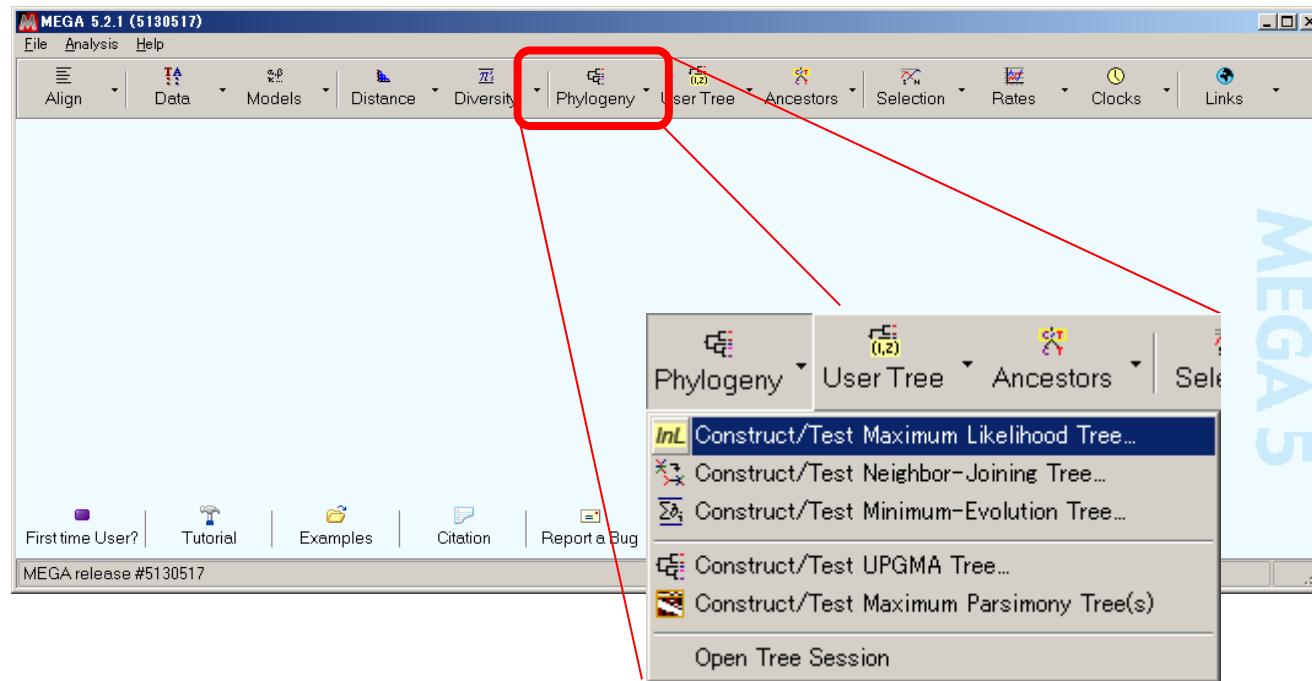
Table. Maximum Likelihood fits of 48 different amino acid substitution models

Model	Parameters	BIC	AICc	<i>InL</i>
WAG+G	32	3433.020	3233.889	-1584.664
WAG+G+I	33	3440.807	3235.472	-1584.431
JTT+G	32	3441.963	3242.833	-1589.135
JTT+G+I	33	3449.723	3244.388	-1588.895
LG+G	32	3450.494	3251.364	-1593.401
Dayhoff+G	32	3457.418	3258.288	-1596.863
LG+G+I	33	3458.285	3252.950	-1593.176
Dayhoff+G+I	33	3465.492	3260.157	-1596.780
JTT+I	32	3471.910	3272.779	-1604.109
rtREV+G	32	3477.877	3278.747	-1607.099
rtREV+G+I	33	3485.880	3280.545	-1606.974
Dayhoff+I	32	3488.149	3289.018	-1612.228
cpREV+G	32	3489.951	3290.821	-1613.120
cpREV+G+I	33	3496.214	3290.878	-1623.140
rtREV+I	32	3511.384	3312.253	-1623.846
cpREV+I	32	3522.866	3323.736	-1629.587
WAG+G+F	51	3536.804	3219.963	-1558.276
LG+G+F	51	3543.239	3226.397	-1561.489
WAG+G+I+F	52	3544.392	3221.366	-1557.946
JTT+G+F	51	3544.710	3227.869	-1562.226
LG+G+I+F	52	3550.702	3227.677	-1561.101
JTT+G+I+F	52	3552.254	3229.228	-1561.877
rtREV+G+F	51	3553.359	3236.518	-1566.559
Dayhoff+G+F	51	3558.236	3241.395	-1568.988

⑥アミノ酸置換モデルの内、BIC (Bayesian Information Criteria)が最も低いモデルは「WAG+G」

Sproteinの系統樹構築

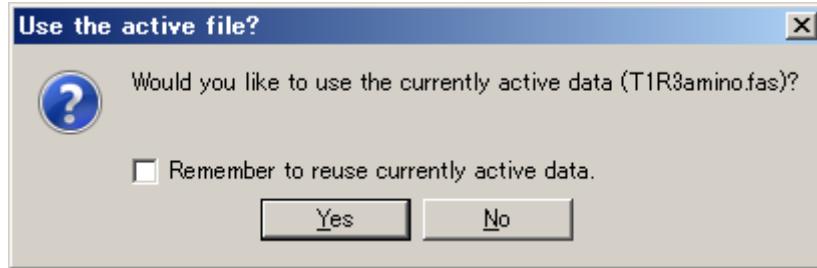
2. 最尤法(系統樹推定)



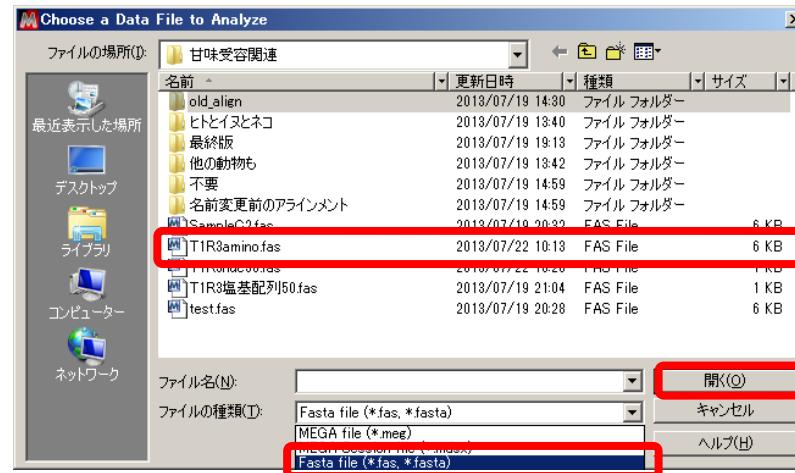
⑦置換モデルのウィンドウを消して、
MEGAメインウィンドウから「Phylogeny」ボタン
→「Construct/Test Maximum Likelihood Tree」を選択

Sproteinの系統樹構築

2. 最尤法(系統樹推定)



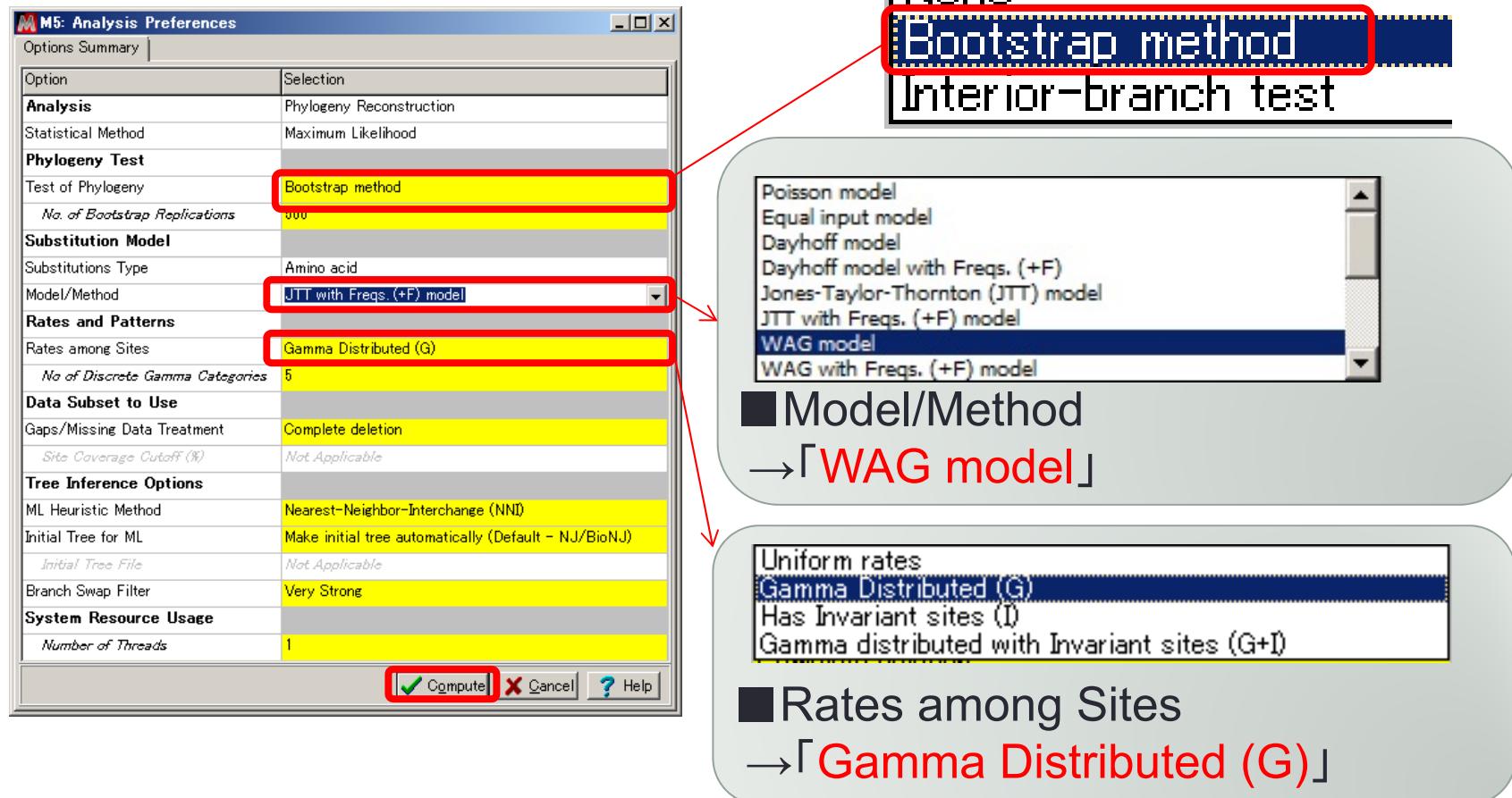
⑧現在開いているデータを利用するか聞かれた場合 →「Yes」



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、Sprotein.fasを選択する。

Sproteinの系統樹構築

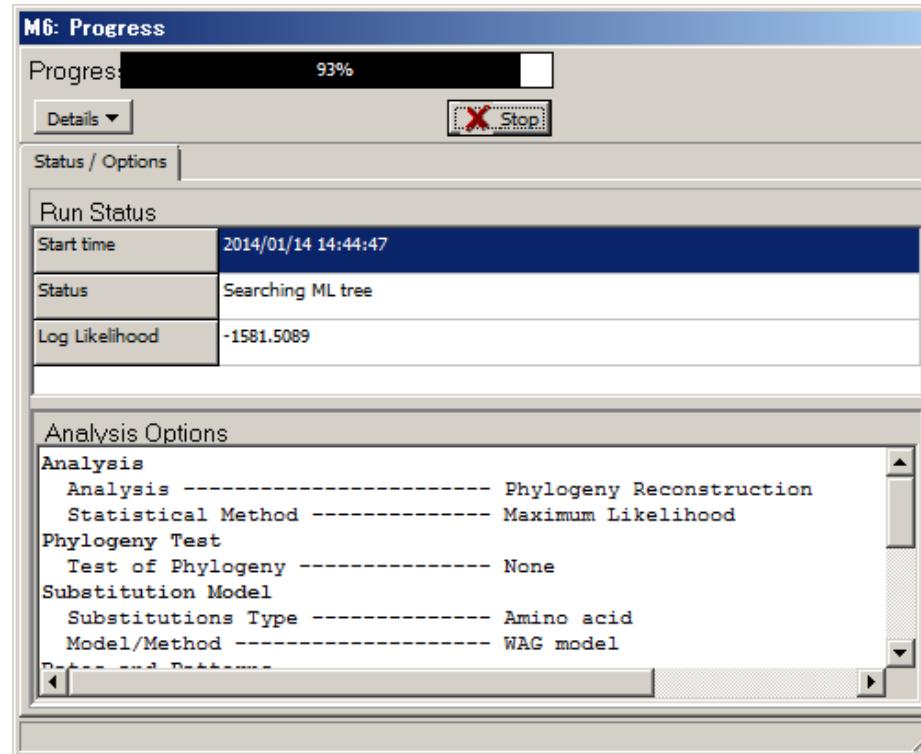
2. 最尤法(系統樹推定)



⑩ パラメータを「WAG+G」に合わせて設定し、「Compute」

Sproteinの系統樹構築

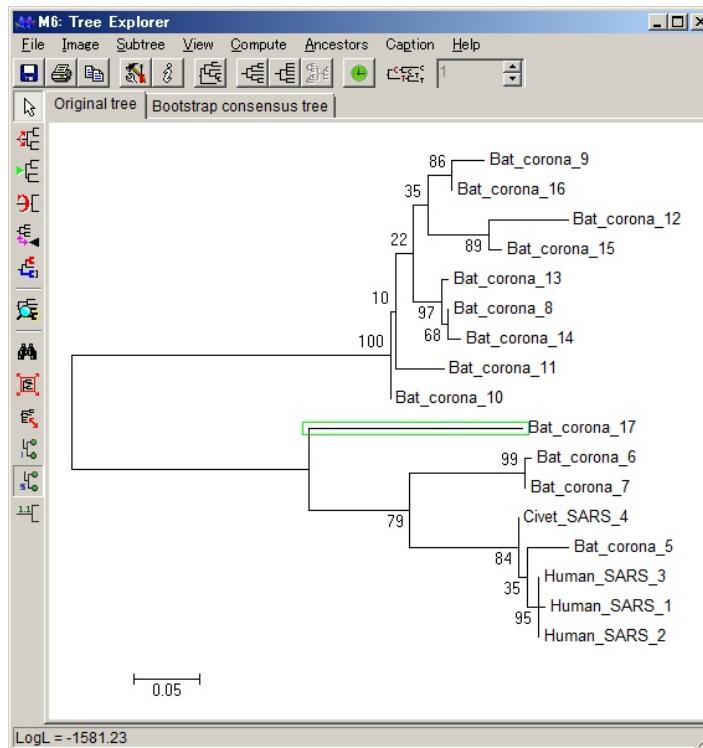
2. 最尤法(系統樹推定)



⑪ 系統樹推定開始。
(本件では約80分かかります)

Sproteinの系統樹構築

2. 最尤法(系統樹推定)

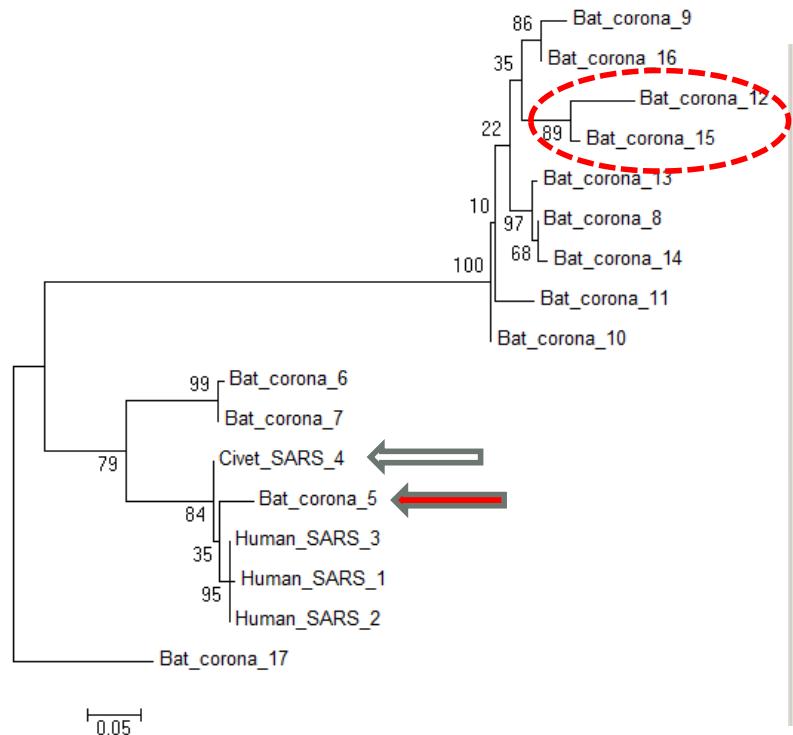


⑫最尤法で系統樹が推定できました。
Bat_corona_17を外群として、先ほどの
近隣結合法の系統樹と比較してみましょう。

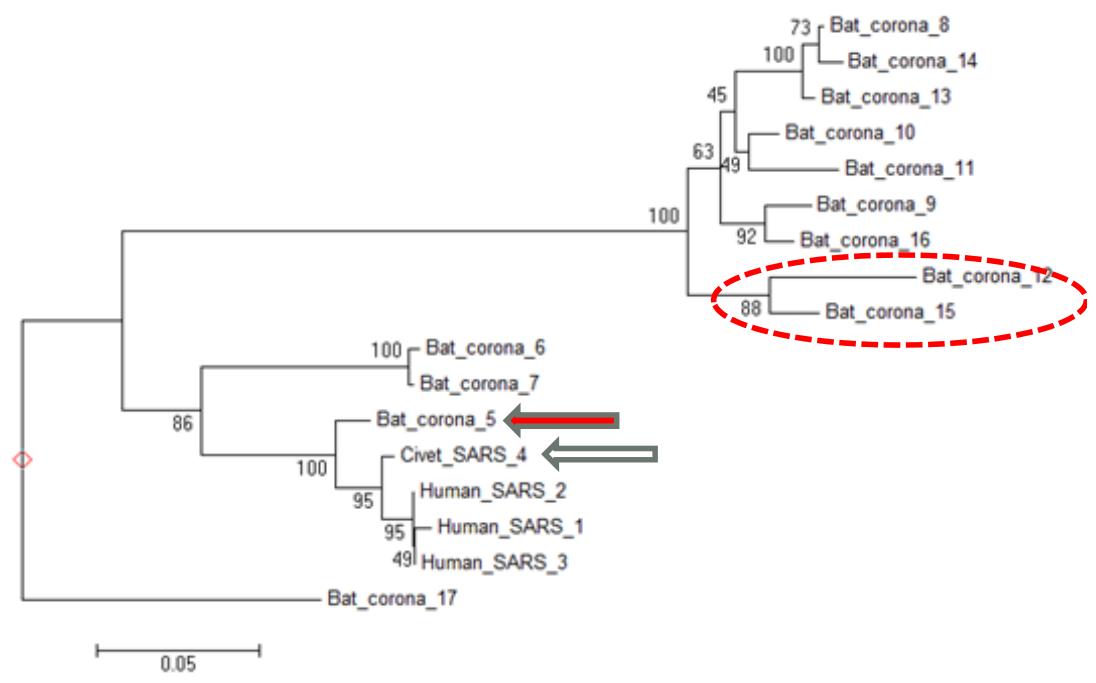
Sproteinの系統樹構築

2. 最尤法(結果の検討)

■最尤法で構築した系統樹



■近隣結合法で構築した系統樹



系統樹のトポロジーが異なっている。

Sproteinの系統樹構築

2. 最尤法(結果の検討)

複数の系統関係が候補として示された場合の、
次の一手は？

A. 情報の追加

- ・別の置換モデルを使って再度最尤法実施
- ・別の遺伝子や、ゲノム全体を使って系統樹構築を試行
- ・アミノ酸ではなく塩基で系統関係を推定
- ・分子情報だけでなく形態や性質についてもパラメータを定め、系統関係を検討してみる。
- ・歴史記録や別の研究で得られた情報をCalibration pointとして設定し、再度系統樹構築
- ・イントロンやエキソン、non-coding配列を分けて系統樹構築に利用

B. 唯一の系統樹を確定できない理由の検討

- ・ゲノム融合・組換えや別種交雑の可能性
- ・一部の種のSproteinが、特殊な選択圧を受けていた可能性
- ・短期間に3種以上の系統へ、急速な分化が起こった可能性
- ・アラインメントが間違っている/不可能な可能性

→対象の生物種や遺伝子により、
様々なケースが考えられる。

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、
改め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。

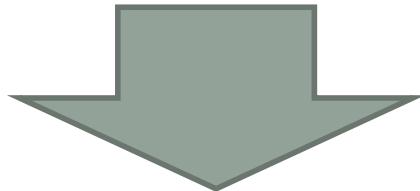
6. 配列比較解析演習

哺乳類p53遺伝子の比較解析

【演習】配列比較解析

・背景

SARS コロナウイルスは、コロナウイルスの中で
コウモリ野生集団がリザーバとして保持していた系統の中から
出現したことが、Sprotein(spikeタンパクコード領域)の
系統解析から示唆されました(*1)。



多様なコロナウイルスを長期に渡り保持してきた
コウモリのゲノムの特徴は？

*1 Ge XY et. al. "Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor." Nature. 2013 Nov 28;503(7477):535-8. doi: 10.1038/nature12711. Epub 2013 Oct 30.

【演習】配列比較解析

・背景

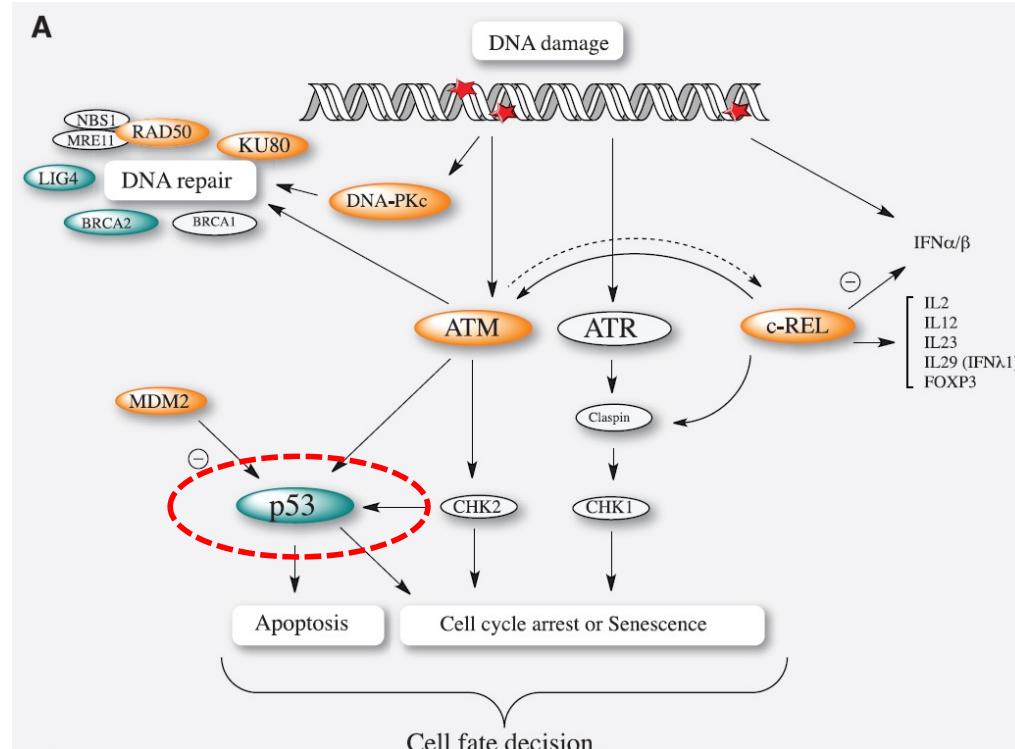


・2種のコウモリゲノム解読

Zhang G et. al.

Comparative analysis of bat genomes provides insight into the evolution of flight and immunity.

Science. 2013 Jan 25;339(6118):456-60.



哺乳類の系統の中でもコウモリのゲノムは多くの遺伝子修復関連の遺伝子が正の選択を受けている(図内: 橙色・青色の遺伝子)。
→p53について、アラインメント・系統樹を作成し、その状態を確認してみましょう。

【演習】サンプル配列(p53)

sampleC.fas

- Bat1_p53
- Bat2_p53
- Horse_p53
- Cat_p53
- Dog_p53
- Cow_p53
- Pig_p53
- Human_p53
- Mouse_p53
- Rat_p53

```
>Bat1_p53
MDDQPLEVVDPPLSQETFSDLWNSATAWGPSPLTRFPHAWNPEDSQAVGGLDWMPFSLEDCLDNGPNEASSMATT PAPAAA V PAPATSTLSSVSPSPKTYPGSYGFRFLKGTAKS V TCTYSPVLNKFCOM
AKSCPVQLVWSSPPPLHSRVRAMAIYKKSEHMTEVRRCPHHERCSEYS DGLAPP H LIRVEGNLRAEYLDDMNTFRHSV/VPYEPPEVGSEYATHYNFMCNSSCMGGMNRRLPILTITLEDSGNLLGRNSFEVRI
CACPGDRRTEEEENFRKGEPSPKQPGPSKQPGPSKTRALPTDSSPPP KKADEEYFTLQIRGRERFETFRKLNEALELQDDVLAGKDPGSKTHSHLKPKKGQSTS RHKRMLFKREGPDSD
>Bat2_p53
MDIPOSELNMEPPLSQETFSDLWKL LPQNVLPPDILSPNEFLPSL VNWLD E QNESPRVPA AAT PAPATSWPLSSFVPSQKTYPGSYDFRLGFLNSGTAKS V CTYSPTLNKFCQLAKTCPVQLWSSPPPLG
TRVRAMAIYKKSEYMTEVRRCPHHERCS D YDGLAPP QHLIRVEGNLRAEYLDDKHTFRHSV/VPYEPPEVGSDCTIHYNFMCNSSCMGGMNRRLPILTITLEDSGNLLGRNSFEVRCACPGRDRRTEEE
KKGEPCPKKPGSTKRALPTDTSSSPSKMPLDEEYFTLQIRGRKNEI RNEALELKDAQGEPRGSRAHSHLKSKGOSTSCHKLTKREGPDSD-----
>Horse_p53
MEETQTELGIEPPLSQETFSDLWKL LPQNVLPPDILSPAVNNLSPD VNWLD E QN E PRM P A A P P A T S W P L S S F V P S Q K T Y P G C Y G F R L G F L N S G T A K S V T C T Y S P T L N K F C Q L A K T C P V Q L V W S S P P
PPGTRVRAMAIYKKSEFMTEVRRCPHHERCS D SDG LAPP QHLIRVEGNLRAEYL D E R N T F R H S V / V P Y E P P E V G S D C T I H Y N F M C N S S C M G G M N R R P L I T I T L E D S G N L L G R N S F E V R V C A C P G R D R R T E E
NFRKKECPPEPPRSTKRVLSSNTSSSPQKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQTGEPGGSKAHSSHLSKSKKGQSTS RHKLMLFKREGPDSD-----
>Dog_p53
MQEPQSELNIDPPLSQETFSELWNLLPENNVLSSELCPA V D E L L P E S V V N W L D E Q D S D A P R M P A T S A T P A G P A P S W P L S S V P S Q K T Y P G C Y G F R L G F L N S G T A K S V T W T Y S P L N K F C Q L A K T C P V Q L V W S S P
PPPPGTRVRAMAIYKKSEFMTEVRRCPHHERCS D SDG LAPP QHLIRVEGNLRAEYLDDRNTFRHSV/VPYEPPEVGSDCTIHYNFMCNSSCMGGMNRRLPILTITLEDSGNLLGRNSFEVRCACPGRDRRTEEE
ENFHKGEGCPEPPGSKTRALPTSSSPQKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQGEPRGSRAHSHLKA KKGQSTS RHKLMLFKREGPDSD-----
>Cat_p53
MQEPPLELTIEPPLSQETFSELWNLLPENNVLSSELSAMNELPLSEDVANWLDEAPDASGMSA V P A P A P A P A T P A P A I S W P L S S F V P S Q K T Y P G A Y G F H L G F L O S G T A K S V T C T Y S P P L N K F C Q L A K T C P V Q L
WVRSPPPGTCTVRAMAIYKKSEFMTEVRRCPHHERCPDSSDGLAPPQHLIRVEGNLHAKYLD DRNTFRHSV/VPYEPPEVGSDCTIHYNFMCNSSCMGGMNRRLPILTITLEDNGKLLGRNSFEVRCACPGRD
RRTEENFRKKGECPPEPPGSKTRALPTSTPPQKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQSGKEPGGSRASHSHLKA KKGQSTS RHKPMLKREGPDSD-----
>Cow_p53
MEEQSQAEVNVEPPLSQETFSDLWNLNPENNLLSSELSAPVDDLLPYQPTDVWLD ECPNEAPQMPE S A P A P P A T P A P A T S W P L S S F V P S Q K T Y P G N Y G F R L G F L Q S G T A K S V T C T Y S P S L N K F C Q L A K T C P V Q L
WVDSPPPGTCTVRAMAIYKKLEHMTEVRRCPHHERCS D SDYDGLAPPQHLIRVEGNLRAEYLDDRNTFRHSV/VPYEPPEVGSDCTIHYNFMCNSSCMGGMNRRLPILTITLEDSCGNLLGRNSFEVRCACPGRD
RRTEENLRKKGQSCPPEPPRSTKRALPTNTSSSPQKKPLDGEYFTLQIRGFKRYEMFRELNALELKDALDGREPGE SRAHSSHLKSKKKPSRSHKKPMFKREGPDSD-----
>Pig_p53
MEESQSELGVEPPLSQETFSDLWKL LPQNVLPPDILSSLA AVNDLLSPVNTWL DENPDSRV P A P A T P A P A P A T S W P L S S F V P S Q K T Y P G S Y D F R L G F L H S G T A K S V T C T Y S P A L N K F C Q L A K T C P V Q L
WVSSPPPGTCTVRAMAIYKKSEYMTEVRRCPHHERCS D SDYDGLAPPQHLIRVEGNLRAEYLDDRNTFRHSV/VPYEPPEVGSDCTIHYNFMCNSSCMGGMNRRLPILTITLEDASGNLLGRNSFEVRCACPGRD
RRTEENFLKKGQSCPPEPPGSKTRALPTSTSSPVQKKPLDGEYFTLQIRGRERFEMFRELNALELKDAQTA RESGENRAHSHLKS KGGQSPSRSHKKPMFKREGPDSD-----
>Human_p53
MEEPQSDPSVEPPLSQETFSDLWKL LPQNVLPPDILSSLA AVNDLLSPVNTWL DENPDSRV P A P A T P A P A P A T S W P L S S F V P S Q K T Y Q G S Y G F R L G F L H S G T A K S V T C T Y S P A L N K F C Q L A K T C P V Q L
QLAKTCPVWVWDSTPPGTRVRAMAIYKKQSHMTEVRRCPHHERCS D SDG LAPPQHLIRVEGNLREYLD DRNTFRHSV/VPYEPPEVGSDCTIHYNFMCNSSCMGGMNRRLPILTITLEDSSGNLLGRNSFEV
HVCACPGRDRRTEEENLRKKGEPHELPPGSKTRALSNNTSSSPQKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPGGSRAHSSHLKS KGGQSTS RHKLMLFKTEGPDSD-----
>Mouse_p53
MEESQSDISIPLSQETFSDLWKL LPQNVLPPDILSSLA AVNDLLSPVNTWL DENPDSRV P A P A T P A P A P A T S W P L S S F V P S Q K T Y Q G N Y G F H L G F L Q S G T A K S V M C T Y S P P L N K F C Q L A K T C P V Q L
QLWVSPAGSRVRAMAIYKKQSHMTEVRRCPHHERCS D SDG LAPPQHLIRVEGNLYPEQDRTFRHSV/VPYEPPEAGSEYTTIHYKYM CNSSCMGGMNRRLPILTITLEDSSGNLLGRDSFEVRCACPGRD
RRDRTEENFRKKEVLCPELPPGSKAKRALTCTSASPPQKKPLDGEYFTLQIRGRERFEMFRELNALELKDAHATEEGSDSRAHSSLQPRFAQLIKEESPNC-----
>Rat_p53
MEDQS QDMSIPLSQETFSDLWKL LPQNVLPPDILSSLA AVNDLLSPVNTWL DENPDSRV P A P A T P A P A P A T S W P L S S F V P S Q K T Y Q G N Y G F H L G F L Q S G T A K S V M C T Y S I S L N K F C Q L A K T C P V Q L
CPVQWVSTTPPGTRVRAMAIYKKQSHMTEVRRCPHHERCS D SDG LAPPQHLIRVEGNLYPEA EYLDDRQTRHSV/VPYEPPEVGSDCTIHYKYM CNSSCMGGMNRRLPILTITLEDSSGNLLGRDSFEVRCACPGRD
CPGRDRRTEEENFRKKEEHCPELPPGSKAKRALTSTS SSSPQKKPLDGEYFTLQIRGRERFEMFRELNALELKDAAREAESGDSRAHSSYPTKKGQSTS RHKPMLKVGFDSD-----
```

2013年に解読されたコウモリゲノム(2種)に含まれていた
p53配列と、哺乳類8種のp53アミノ酸配列

【演習】配列比較解析

コウモリ2種、ウマ、ネコ、イヌ、ウシ、ブタ、ヒト、マウス、ラットのp53遺伝子(sampleC.fas)の配列解析を行い、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】

- ・アミノ酸309～342番目付近はp53のNLS(核局在シグナル)関連領域として知られています。
- ・コウモリのMDM2(p53抑制遺伝子)のNES(核外搬出シグナル)は、哺乳類の中でコウモリだけが特徴的な変異を含んでいることが知られています。

【参考】p53の構造

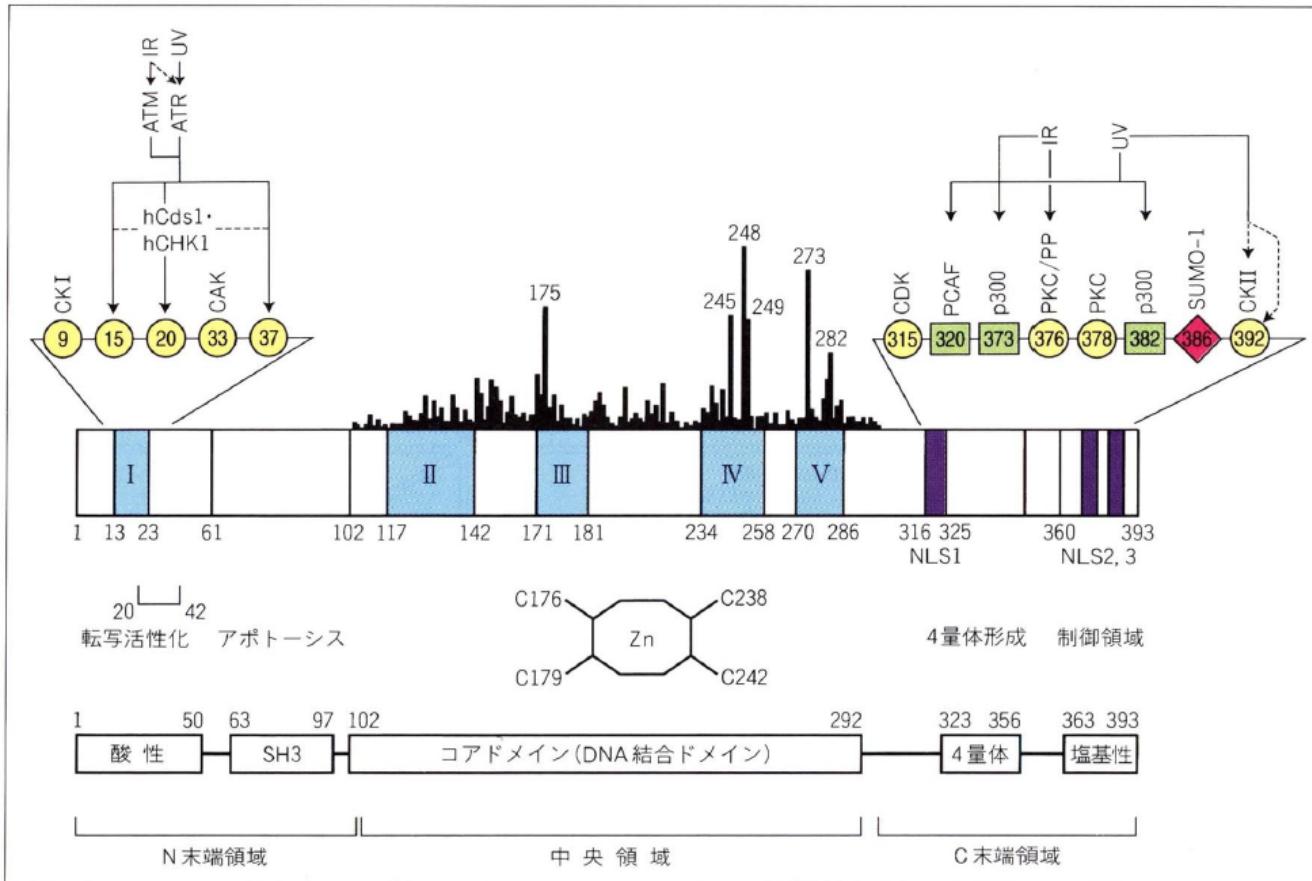


図 1 p53 の変異、機能ドメインと修飾

図の上部にコアドメインにおける各アミノ酸残基の変異頻度とホットスポット部位のコドン番号を示す。p53 ドメイン構造の下には、各領域、ドメインの最初と最後のアミノ酸番号を示した。点線：不明。I～V：保存領域、NLS：核移行シグナル、○：プロテインキナーゼ [CKI, CKII : カゼインキナーゼ I, II, ATM : ATM キナーゼ, ATR : ATR キナーゼ, CAK : サイクリン活性化キナーゼ, CDK : サイクリン依存キナーゼ, PKC : プロテインキナーゼ C, PP : ホスファターゼ]。□ : ヒストンアセチラーゼ [PCAF : p300/CBP-associated factor, p300]。◇ : SUMO-1 による修飾。

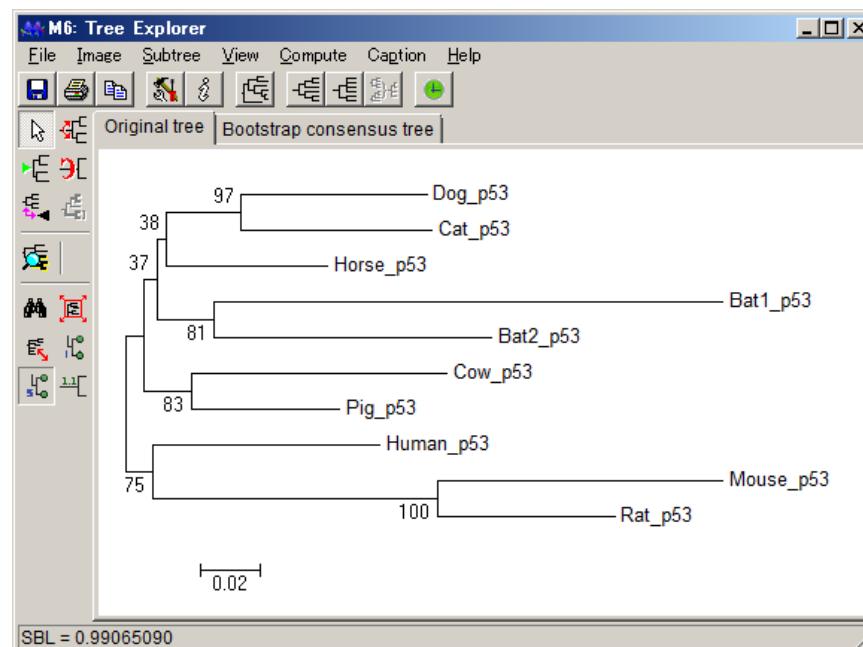
【解答】p53のマルチプルアラインメント

- 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」 →sampleC.fas選択
- ③Translated Protein Sequence タブをクリック、Standardを選択。
- ④「Edit」から「SelectAll」選択。
- ⑤「Alignment」から→「Align by Muscle」を選択→「Compute」
- ⑥「Data」→「Export Alignment」→「FASTA format」を選択
→「p53」と名前を付けて保存。Fasta形式の p53.fas が保存されます。

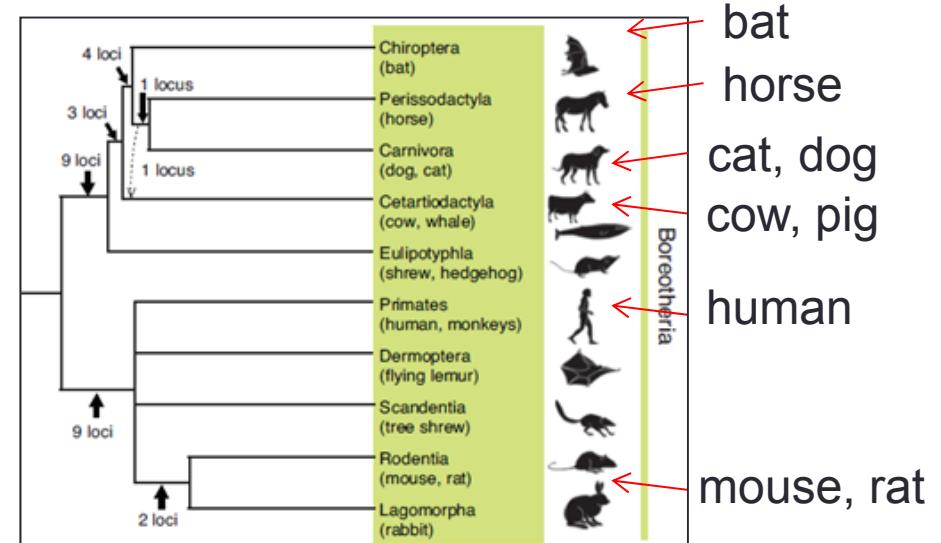
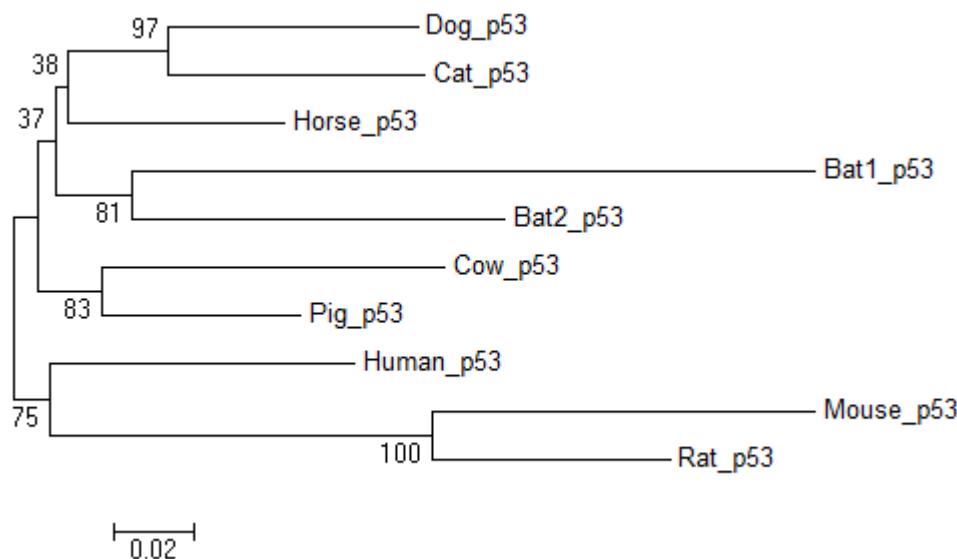


【解答】p53の系統樹構築

- 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、p53.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」



【解答】p53配列比較結果の検討



参考:既知の哺乳類の系統関係

【検討例】

- 既知の哺乳類の系統関係と、トポロジーは一致。
- Bat1の枝が長い。進化速度が速いため？
- Dog, Cat, Horse, Batの根元のBootstrap値が低いのは
短期間に急速な分化が起こったため？

【解答】p53のアライメント検討 p53のNLS

重複

【検討例】

- ・Bat1の309～342番目付近:NLS関連領域の一部が重複している。

→コウモリゲノムの論文では、p53の核内局在機構を独自に進化させたことで、独特な免疫機構・飛行のための形態形成に繋がったのではと推論を行っている。

*2 Zhang G et. al. "Comparative analysis of bat genomes provides insight into the evolution of flight and immunity." Science. 2013 Jan 25;339(6118):456-60. doi: 10.1126/science.1230835. Epub 2012 Dec 20.

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

参考文献(論文・Web)

■配列解析のコストのグラフ

Hayden EC. "Tepid showing for genomics X prize." *Nature*. 2013 May 30;497(7451):546-7. doi: 10.1038/497546a.
<http://www.ncbi.nlm.nih.gov/pubmed/23719441>

■データベースの数について

Marshall E. "Biomedicine. NIH seeks better database for genetic diagnosis." *Science*. 2013 Oct 4;342(6154):27. doi: 10.1126/science.342.6154.27.
<http://www.ncbi.nlm.nih.gov/pubmed/24092711>

■イヌとオオカミのゲノム再解析

Axelsson E. et. al. "The genomic signature of dog domestication reveals adaptation to a starch-rich diet"
Nature. 2013 Mar 21;495(7441):360-4. doi: 10.1038/nature11837. Epub 2013 Jan 23
<http://www.ncbi.nlm.nih.gov/pubmed/23354050>

■統合解析環境MEGA

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S (2013) "MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0." *Molecular Biology and Evolution* 30: 2725-2729.

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

■ネアンデルタール人のゲノム

Kay Prüfer et. al. "The complete genome sequence of a Neanderthal from the Altai Mountains" *Nature* 505, 43–49 (02 January 2014) doi:10.1038/nature12886
<http://www.natureasia.com/ja-jp/nature/highlights/50475>

■ゾウギンザメのゲノム

Byrappa Venkatesh et. al. "Elephant shark genome provides unique insights into gnathostome evolution"
Nature 505, 174–179 (09 January 2014) doi:10.1038/nature12826
<http://www.natureasia.com/ja-jp/nature/highlights/50711>

■カララバトのゲノム

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

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.

■SARSコロナウイルスの構造

田口文広"SARSコロナウイルス" ウイルス Vol. 53 (2003) No. 2 P 201-209

■SARSコロナウイルスの比較

Ge XY et. al. "Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor."
Nature. 2013 Nov 28;503(7477):535-8. doi: 10.1038/nature12711. Epub 2013 Oct 30.

■p53の構造

土田信夫ら "DNAダメージによるp53蛋白質の癌抑制シグナルとヒト癌における遺伝子変異" 蛋白質核酸酵素 2000年7月号 1742-1751

参考文献(論文・Web)

■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.

■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

■WebLogo

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188–1190, (2004)

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

■コウモリのゲノム

Zhang G et. al. "Comparative analysis of bat genomes provides insight into the evolution of flight and immunity."

Science. 2013 Jan 25;339(6118):456–60. doi: 10.1126/science.1230835. Epub 2012 Dec 20.

■Jalview

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G. J. (2009) "Jalview Version 2 – a multiple sequence alignment editor and analysis workbench"

*Bioinformatics*25 (9) 1189–1191 doi: 10.1093/bioinformatics/btp033

<http://www.jalview.org/>

■MAFFT

Katoh, Standley 2013 (*Molecular Biology and Evolution* 30:772–780) MAFFT multiple sequence alignment software version 7: improvements in performance and usability.

<http://mafft.cbrc.jp/alignment/software/>

■アミノ酸の保存度の計算

"Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation." Livingstone C.D.

and Barton G.J. (1993) *CABIOS* 9, 745–756

■p53の構造

土田信夫・西垣玲子・高崇峰・土田幸子・中島琢磨 DNAダメージによるp53蛋白質の癌抑制シグナルとヒト癌における遺伝子変異 蛋白質核酸酵素 2000年7月号 1742–1751

■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.

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)

■MEGA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

■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)

参考配列

■ SampleA、SampleB

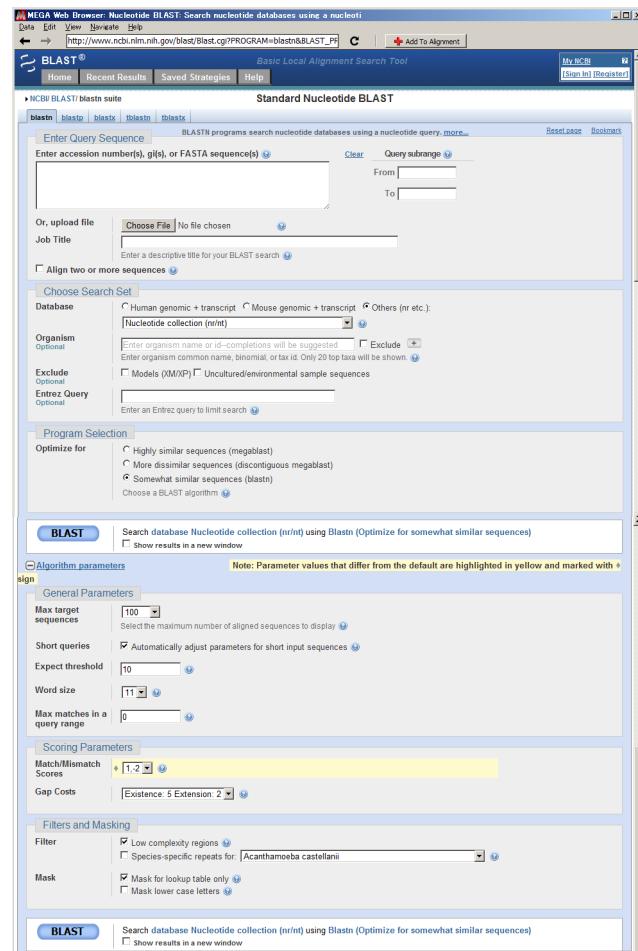
配列名	NCBI nucleotide title
SampleA / Civet_SARS_4	gi 34482137 gb AY304486.1 _SARS_coronavirus_SZ3_complete_genome
Human_SARS_1	gi 30248028 gb AY274119.3 _SARS_coronavirus_TOR2_complete_genome
Human_SARS_2	gi 41323719 gb AY390556.1 _SARS_coronavirus_GZ02_complete_genome
Human_SARS_3	gi 30275666 gb AY278488.2 _SARS_coronavirus_BJ01_complete_genome
Bat_corona_5	gi 556015127 gb KC881006.1 _Bat_SARS-like冠状病毒_Rs3367_complete_genome
Bat_corona_6	gi 556015076 gb KC880986.1 _Bat_SARS-like冠状病毒_Rs3369_spike_protein_(S)_gene_partial_cds
Bat_corona_7	gi 556015114 gb KC881005.1 _Bat_SARS-like冠状病毒_RsSHC014_complete_genome
Bat_corona_8	gi 255733149 gb FJ588686.1 _Bat_SARS-CoV_Rs672/2006_complete_genome
Bat_corona_9	gi 76160337 gb DQ022305.2 _Bat_SARS冠状病毒_HKU3-1_complete_genome
Bat_corona_10	gi 72256267 gb DQ071615.1 _Bat_SARS冠状病毒_Rp3_complete_genome
Bat_corona_11	gi 89514824 gb DQ412043.1 _Bat_SARS冠状病毒_Rm1_complete_genome
Bat_corona_12	gi 89514809 gb DQ412042.1 _Bat_SARS冠状病毒_Rf1_complete_genome
Bat_corona_13	gi 556015106 gb KC881001.1 _Bat_SARS-like冠状病毒_Rs4108_spike_protein_(S)_gene_partial_cds
Bat_corona_14	gi 556015102 gb KC880999.1 _Bat_SARS-like冠状病毒_Rs4081_spike_protein_(S)_gene_partial_cds
Bat_corona_15	gi 556015090 gb KC880993.1 _Bat_SARS-like冠状病毒_Rs4075_spike_protein_(S)_gene_partial_cds
Bat_corona_16	gi 556015088 gb KC880992.1 _Bat_SARS-like冠状病毒_Rs4085_spike_protein_(S)_gene_partial_cds
Bat_corona_17	gi 301298998 gb GU190215.1 _Bat冠状病毒_BM48-31/BGR/2008_complete_genome

参考配列

■ SampleC

配列名	NCBI Protein title
Bat1_p53	gi 431894020 gb ELK03826.1 Cellular tumor antigen p53 [Pteropus alecto]
Bat2_p53	gi 432105617 gb ELK31811.1 Cellular tumor antigen p53 [Myotis davidii]
Horse_p53	gi 320202967 ref NP_001189334.1 cellular tumor antigen p53 [Equus caballus]
Cat_p53	gi 538225 dbj BAA05653.1 p53 [Felis catus]
Dog_p53	gi 4996230 dbj BAA78379.1 P53 [Canis lupus familiaris]
Human_p53	gi 23491729 dbj BAC16799.1 P53 [Homo sapiens]
Mouse_p53	gi 200203 gb AAA39883.1 p53 [Mus musculus]
Rat_p53	gi 189083686 ref NP_112251.2 cellular tumor antigen p53 [Rattus norvegicus]

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

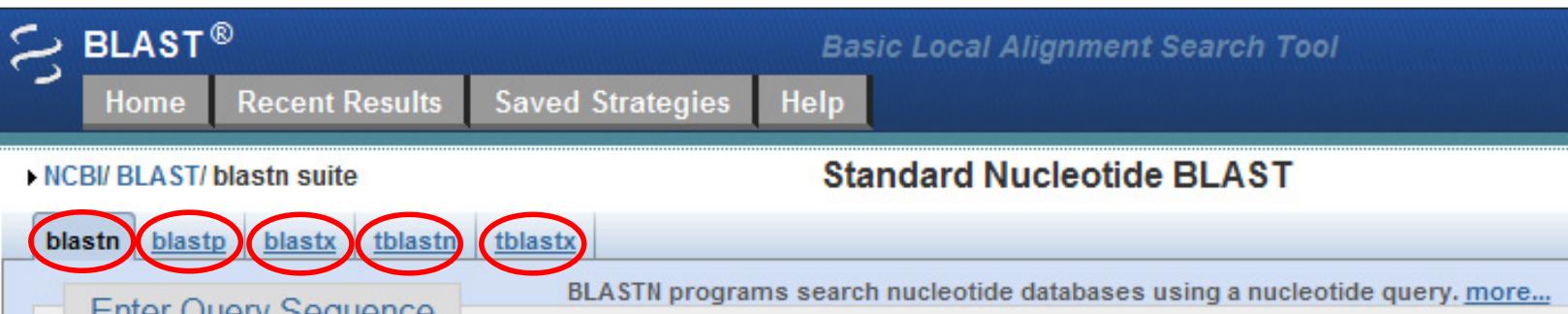
blastn及びblastp専用の設定項目。
BLASTアルゴリズムを選択

■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



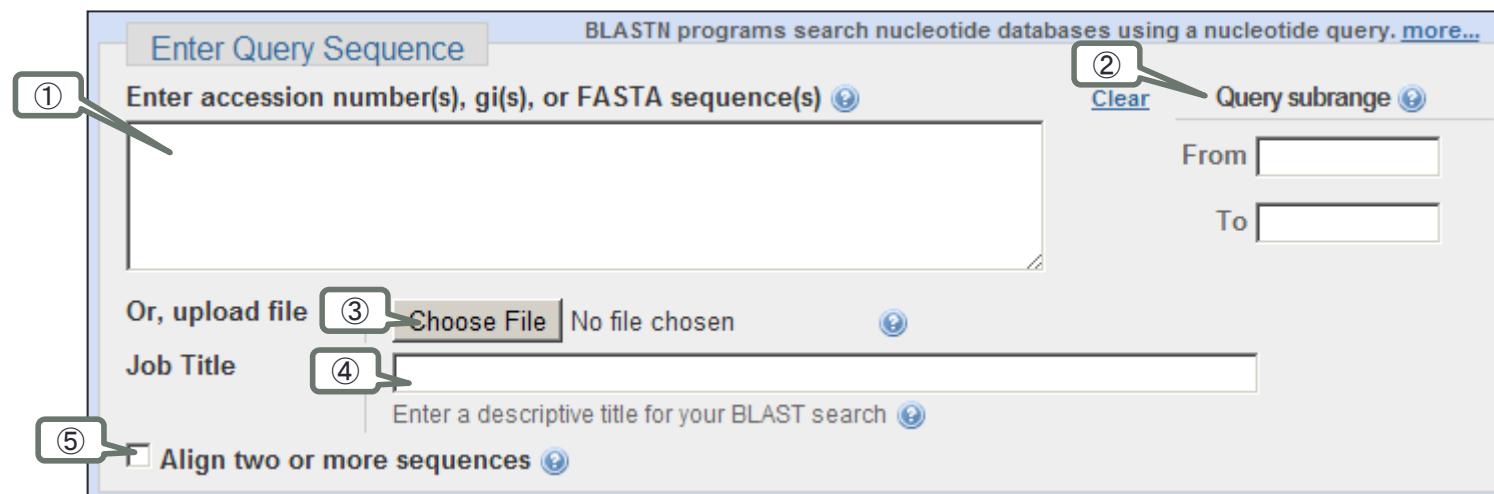
The screenshot shows the NCBI BLAST search interface. At the top, there's a navigation bar with links for Home, Recent Results, Saved Strategies, and Help. Below that, a breadcrumb trail shows 'NCBI/ BLAST/ blastn suite'. The main title is 'Standard Nucleotide BLAST'. Below the title, five programs are listed: 'blastn', 'blastp', 'blastx', 'tblastn', and 'tblastx'. The first four are circled in red. A text box below says 'BLASTN programs search nucleotide databases using a nucleotide query.' followed by a link 'more...'. There's also a button 'Enter Query Sequence'.

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

NCBI BLASTの詳細設定②

■Enter Query Sequence



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

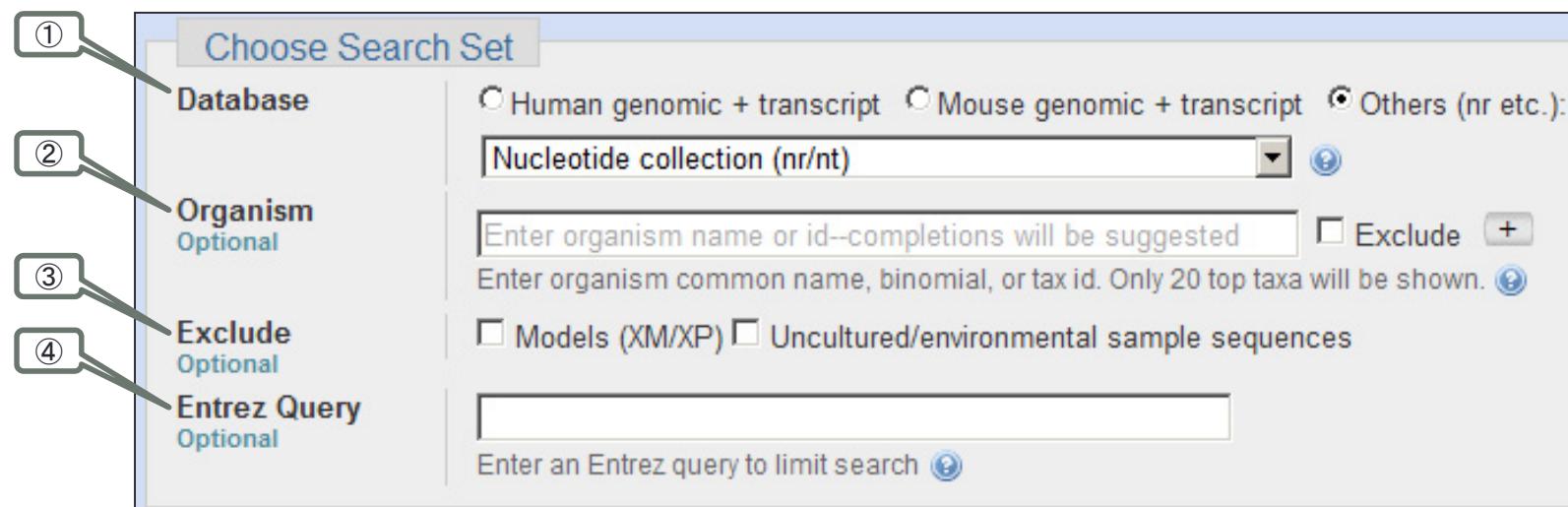
⑤複数アライメント: 対応する配列をこの下に表示される

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

NCBI BLASTの詳細設定③

■Choose Search Set

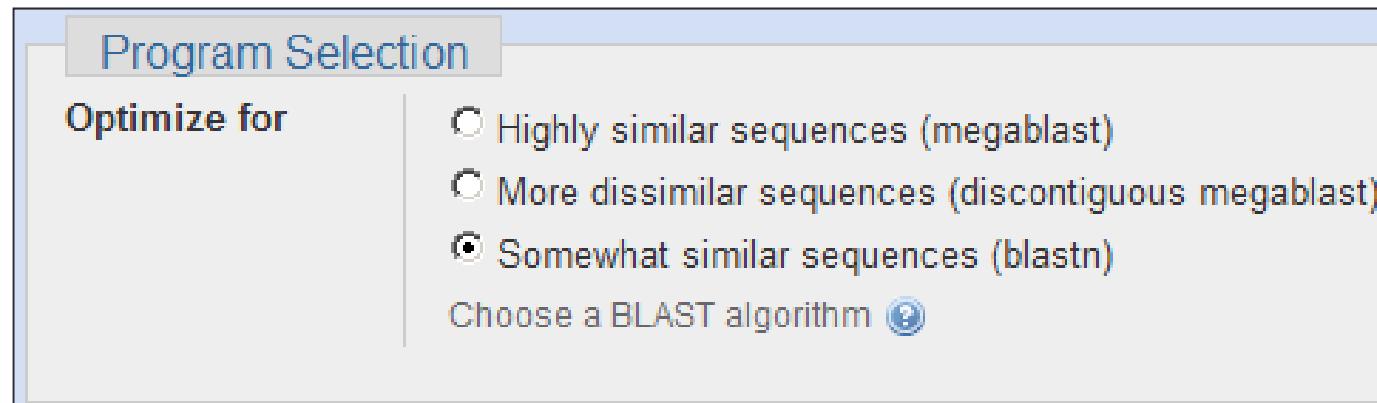


- ①データベース:検索対象のNCBIデータベースを選択
②生物種:データベース内の生物種で区切る場合
③Exclude: Modelやメタゲノムを除外対象とするか。
④Entrez query:Entrezデータベースに限定する場合その番号

参考: ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf

NCBI BLASTの詳細設定④

■Program Selection



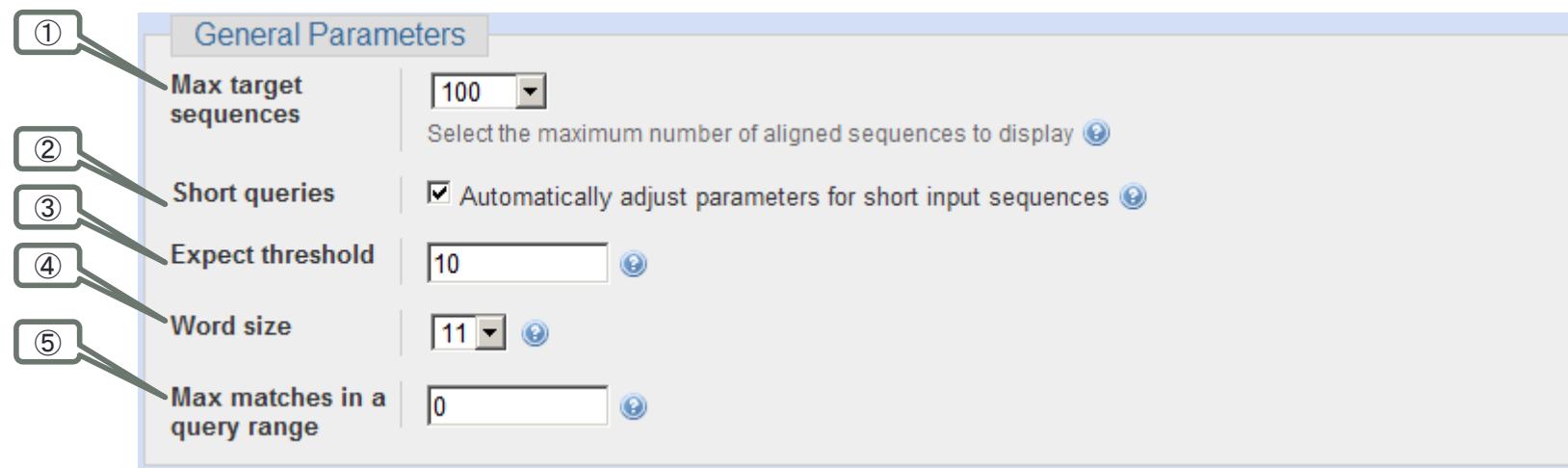
BLASTアルゴリズムの選択

・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

<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

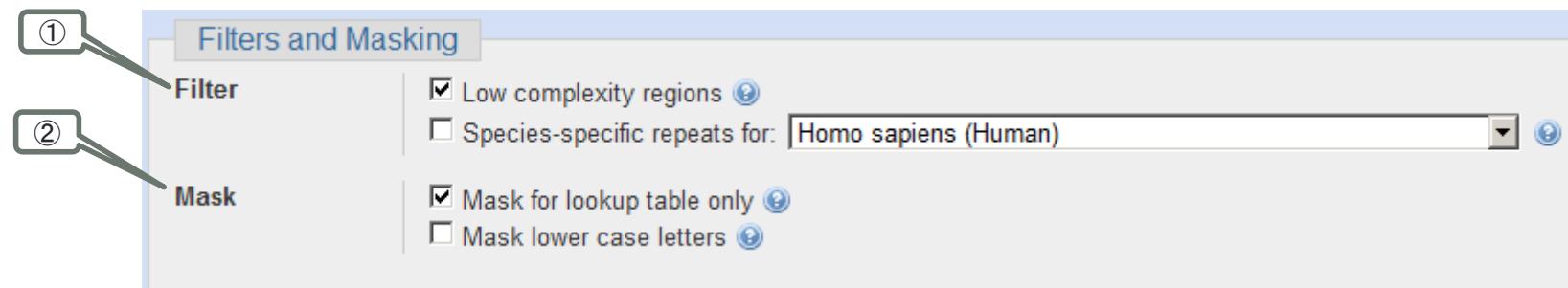


- ①Match/Mismatch Scores: マッチやミスマッチの場合のペナルティスコア
②Gap Costs: ギャップ開始やギャップ伸長のペナルティスコア。megablastの時はlinearで自動で設定される。

参考: ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf

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

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