

生物配列解析基礎

配列データベースとホモロジー検索

法政大学 生命科学部
応用植物科学科

大島 研郎



お問い合わせ

Google™ カスタム検索



生命科学部応用植物科学科
植物医科学専修

理工学研究科生命機能学専攻
植物医科学領域

HOME

応用植物科学科とは

4年間の学びの体系

授業内容

教員紹介

卒業生の声

交通アクセス

法政大学 生命科学部

応用植物科学科

2014年4月 開設

植物医科学専修



大島 研郎(おおしま けんろう)教授

専門(担当)分野

植物細菌学、植物メディカルゲノム学

経歴

東京大学アグリバイオインフォマティクス人材養成ユニット 特任助教、東京大学大学院農学生命科学研究科特任准教授

主な業績

植物病原細菌ファイトプラズマの全ゲノム解読、ファイトプラズマの病原性因子の解析など

本日の講義資料

[ホーム](#) > [教育プログラム](#) > [各講義のページ](#) > 1. 生物配列解析基礎



1. 生物配列解析基礎

授業の目標・概要

生命科学のためのデータベースの利用と基本的な解析手法について講義します。
配列データベースや機能データベースの使用法を紹介するとともに、ホモロジー検索、モチーフ解析、Perlプログラミング、系統解析などの基本的な手法について、実習形式で解説します。
バイオインフォマティクス関連の各種データベースにアクセスしたことのない人は、ぜひ本講義を受講して下さい。

担当教員

[清水謙多郎](#) (東大・農・応用生命工学専攻 / 教授)

[大島研郎](#) (法政大学生命科学部 / 教授)

お知らせ

ご自身のノートPCを利用される場合は[こちら](#)を参考にして必要なソフトウェアを予めインストールしておいてください。

講義日程

講師：大島研郎

- ▶ [2020_生物配列解析基礎_1回目_資料.pdf](#)
- ▶ [kiso1](#)
- ▶ [Mgenitalium.faa](#)
- ▶ [Mpneumoniae.faa](#)
- ▶ [parse-blast7.pl](#)
- ▶ [test1.seq](#)
- ▶ [test2.seq](#)
- ▶ [test3.seq](#)
- ▶ [Ureaplasma.faa](#)

本日の講義で使用する、Webページへのリンクが載せてあります。

デスクトップに「blast」フォルダを作成してください



test1.seq

test2.seq

test3.seq

Mgenitalium.faa

Mpneumoniae.faa

Ureaplasma.faa

parse-blast7.pl

の7つのファイルをダウンロードして

作成したblastフォルダに入れてください

講師：大島研郎

▶ 2020_生物配列解析基礎_1回目_資料.pdf

▶ kiso1

▶ Mgenitalium.faa

▶ Mpneumoniae.faa

▶ parse-blast7.pl

▶ test1.seq

▶ test2.seq













▶ test3.seq


▶ Ureaplasma.faa

BLAST (stand-alone BLAST) のインストール

- 矢印 1 のサイトにアクセスします。

<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>


	名前
	ChangeLog
	ncbi-blast-2.10.0+-4.src.rpm
	ncbi-blast-2.10.0+-4.src.rpm.md5
	ncbi-blast-2.10.0+-4.x86_64.rpm
	ncbi-blast-2.10.0+-4.x86_64.rpm.md5
	ncbi-blast-2.10.0+-src.tar.gz
	ncbi-blast-2.10.0+-src.tar.gz.md5
	ncbi-blast-2.10.0+-src.zip
	ncbi-blast-2.10.0+-src.zip.md5
	ncbi-blast-2.10.0+-win64.exe 
	ncbi-blast-2.10.0+-win64.exe.md5

- ▶ R (多くの科目で使用予定)
- ▶ Lhaplus (いくつかの科目で使
- ▶ RStudio (多くの科目で使用予
- ▶ Anaconda (フィールドインフ
- ▶ ActivePerl (生物配列解析基礎
- ▶ **BLAST** (生物配列解析基礎) 
- ▶ MEGA (生物配列解析基礎と生
- ▶ UCSF Chimera (構造バイオイ
- ▶ Modeller (構造バイオインフ
- ▶ ActivePython (構造バイオイ

Windowsの場合は、
このファイルをダウンロードします

ダウンロードしたファイルをダブルクリックしてインストールします

- コマンドプロンプトを立ち上げてください
(Mac OS の場合はターミナル)

 スタート → Windowsシステムツール → コマンドプロンプト

```
C:¥Users¥student>
```

- 以下, 省略して

```
>
```

と記述します

← スペースが入ります
↓

- 「blastp -help」と入力して, リターン

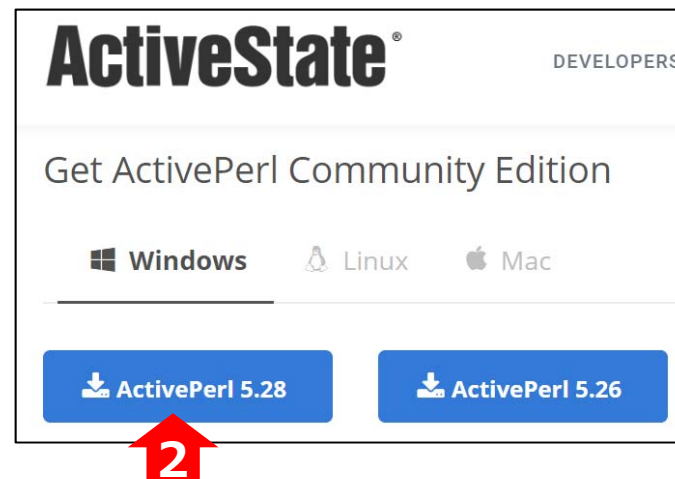
```
> blastp -help
```

BLASTについての説明が表示されれば, OKです

ActivePerl のインストール

- 矢印 1 のサイトにアクセスします。


<https://www.activestate.com/products/perl/downloads/>



ダウンロードしたファイルをダブル
クリックしてインストールします



- コマンドプロンプトを立ち上げてください

 スタート → Windowsシステムツール → コマンドプロンプト

- 「perl -v」と入力して, リターン
スペースが入ります

```
> perl -v
```

Perl についての説明が表示されれば, OKです

```
C:\Users\kenro>perl -v
```

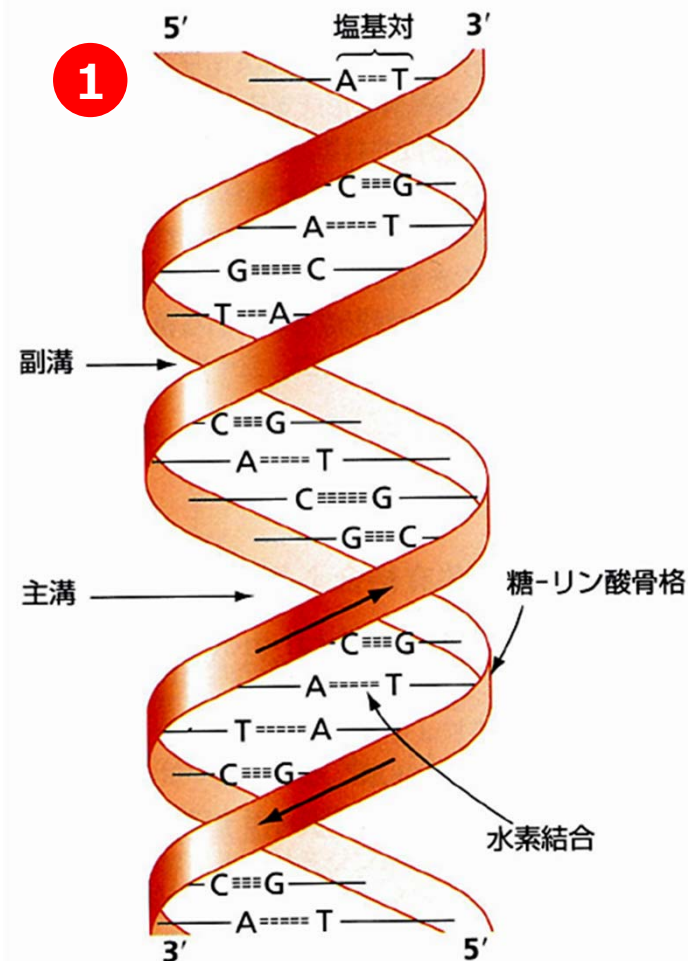
```
This is perl 5, version 28, subversion 1 (v5.28.1) built for MSWin32-x64-multi-thread  
(with 1 registered patch, see perl -V for more detail)
```

```
Copyright 1987-2018, Larry Wall
```


生物配列 = 塩基配列、およびアミノ酸配列

塩基配列 = DNAの塩基 (G A T C) の並び順

■ どのようにして、塩基配列を読むのか？



サンガー法



ABI377シーケンサー

3



ABI3100シーケンサー

次世代シーケンサー

4



Ion PGM

5



イルミナMiSeq

核酸配列データベース

GenBank (National Center for Biotechnology Information)

<http://www.ncbi.nlm.nih.gov/>

DDBJ (日本DNAデータバンク)

<http://www.ddbj.nig.ac.jp/>

EMBL (European Bioinformatics Institute)

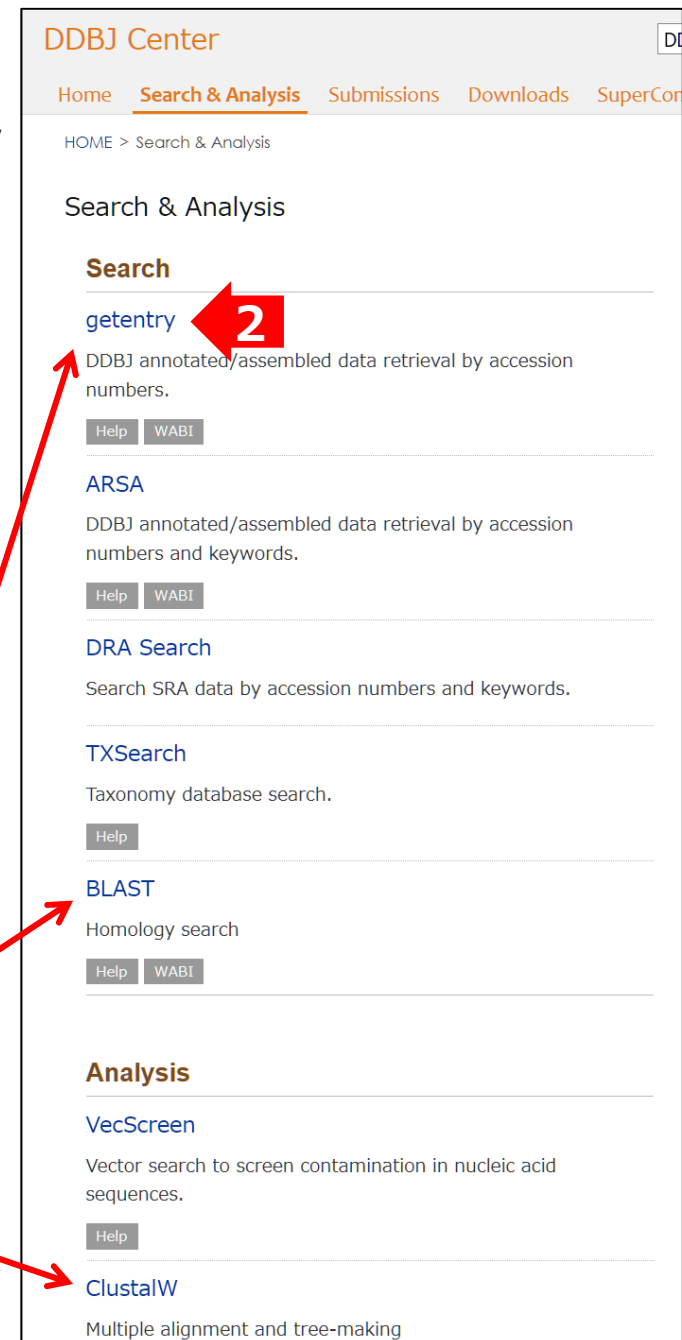
<http://www.ebi.ac.uk/embl/index.html>

- GenBank, DDBJ, EMBLのデータベースは、**3者が情報交換しながら連携**して、“国際データベース”として運営・維持されている
- データベースとは、関連性のある情報を集めて、**一定のフォーマット** (様式) に従って使いやすいように整理したもの。

DDBJ

日本DNAデータバンク。GenBankやEMBLと連携して国際塩基配列データベースを構築している。

http://www.ddbj.nig.ac.jp



データベース検索のページへ

ホモロジー検索のページへ

アラインメント、系統樹作成



Japanese

Google Custom Search

Search

About DDBJ

How to Use

Report/Statistics

FAQ

Contact Us

HOME > Search and Analysis > getentry

getentry

Help

Data retrieval by accession numbers etc

ID :

DNA Database : DDBJ / EMBL / GenBank / MGA

Protein Database : UniProt / PDB / DAD / Patent

Result :

Output Format :

Output Format :

Limit : Results

AP009356 と入力

Genbankフォーマット

```

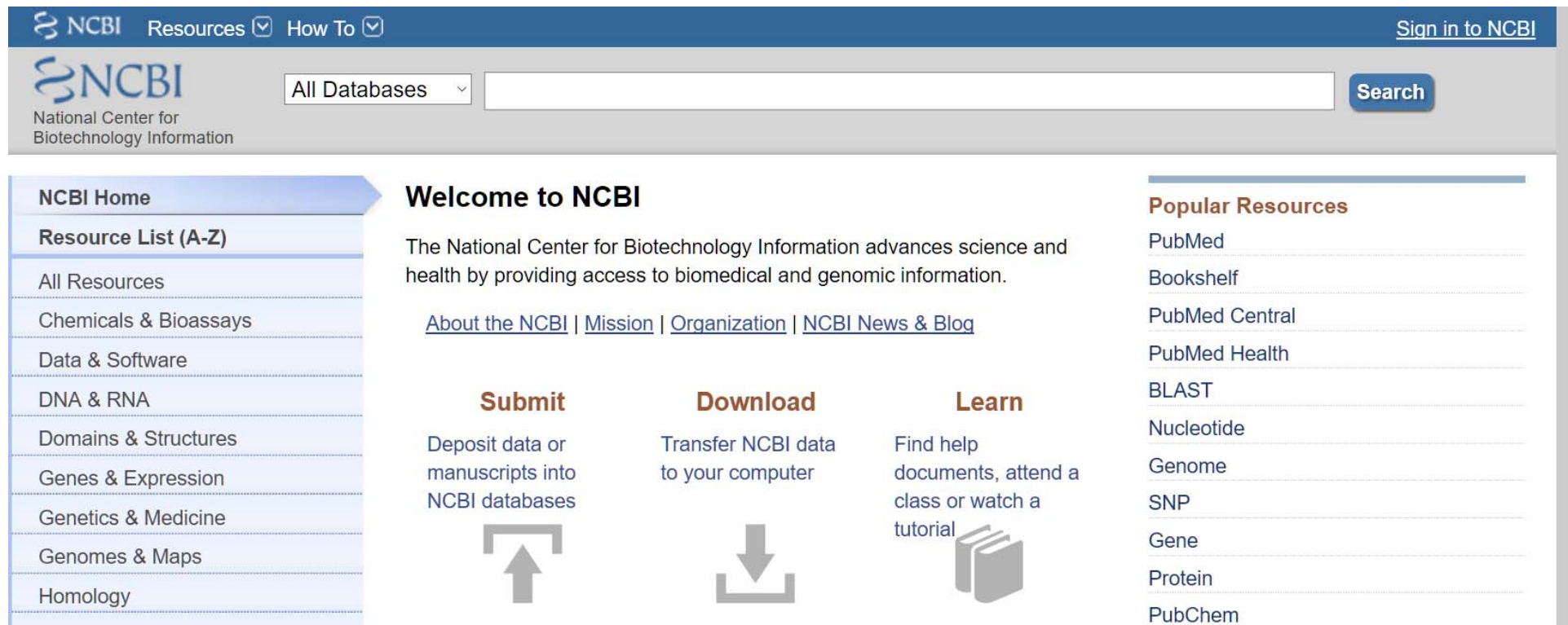
LOCUS       AP009356                80504 bp    DNA     linear   BCT 15-DEC-2007
DEFINITION  Onion yellows phytoplasma OY-W genomic DNA, partial sequeunce.
ACCESSION   AP009356
VERSION     AP009356.1
KEYWORDS    .
SOURCE      Onion yellows phytoplasma OY-W
  ORGANISM  Onion yellows phytoplasma OY-W
             Bacteria; Tenericutes; Mollicutes; Acholeplasmatales;
             Acholeplasmataceae; Candidatus Phytoplasma; Candidatus Phytoplasma
             asteris.
REFERENCE   1 (bases 1 to 80504)
AUTHORS     Oshima,K., Kakizawa,S., Arashida,R., Kagiwada,S. and Namba,S.
TITLE       Direct Submission
JOURNAL     Submitted (02-MAR-2007) to the DDBJ/EMBL/GenBank databases.
             Contact:Shigetou Namba
             The University of Tokyo. Graduate School of Agricultural and Life
    
```

National Center for Biotechnology Information

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

通称：NCBI

- 米国の国立衛生研究所 (NIH) の下の国立医学図書館 が運営するWebサイト
- GenbankやPubMed、BLASTなど、有用なデータベース・ツールがまとめられている

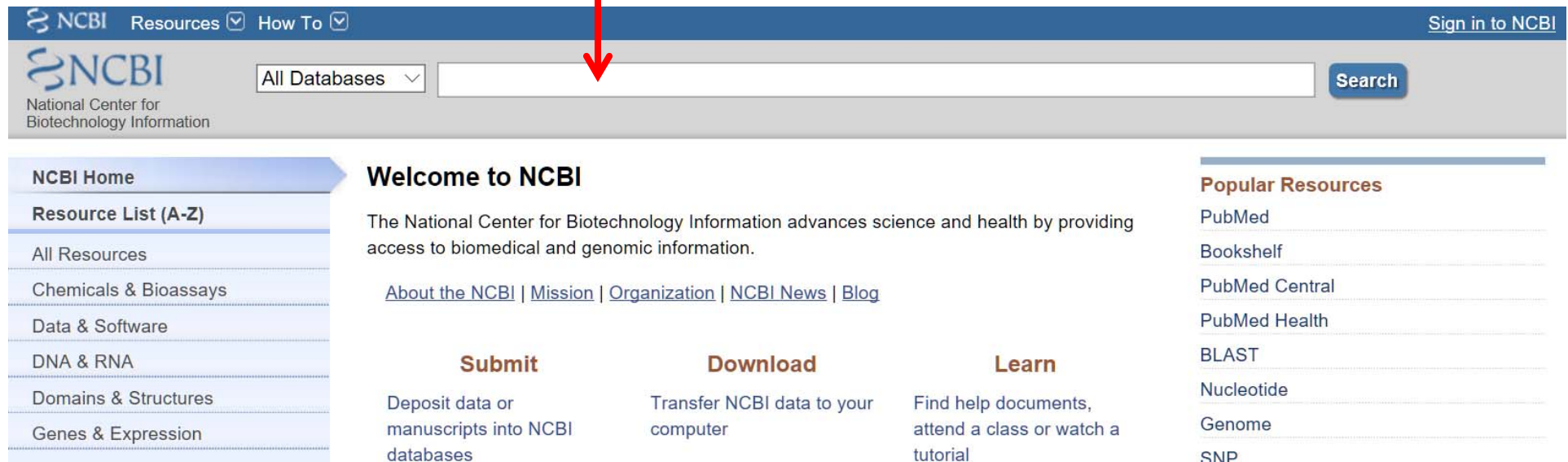


The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with the NCBI logo, "Resources" and "How To" dropdown menus, and a "Sign in to NCBI" link. Below the navigation bar is a search bar with a dropdown menu set to "All Databases" and a "Search" button. The main content area is divided into three columns. The left column contains a "Resource List (A-Z)" menu with items like "All Resources", "Chemicals & Bioassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", and "Homology". The middle column features a "Welcome to NCBI" message, a brief description of the center's mission, and links to "About the NCBI", "Mission", "Organization", and "NCBI News & Blog". Below this are three sections: "Submit" (Deposit data or manuscripts into NCBI databases), "Download" (Transfer NCBI data to your computer), and "Learn" (Find help documents, attend a class or watch a tutorial). The right column lists "Popular Resources" including PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem.

All Databases

- データベースの**統合検索システム**
- 主なデータベースは、PubMed・塩基配列データベース・アミノ酸配列データベース・ゲノムデータベース・3D高分子構造データベースなど
- All Databasesからは、これらのデータベースに対して**横断検索**ができる

例えば「replication protein phytoplasma」と入力してみる



The screenshot shows the NCBI All Databases search page. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' dropdown menus, and a 'Sign in to NCBI' link. Below this is the 'All Databases' search interface, featuring a search box with a dropdown menu set to 'All Databases' and a 'Search' button. A red arrow points to the search input field. The main content area is divided into three columns: 'NCBI Home' with a list of resources, 'Welcome to NCBI' with a description and links, and 'Popular Resources' with a list of frequently used tools and databases.

NCBI Resources ▾ How To ▾ Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases ▾ Search

NCBI Home

- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP

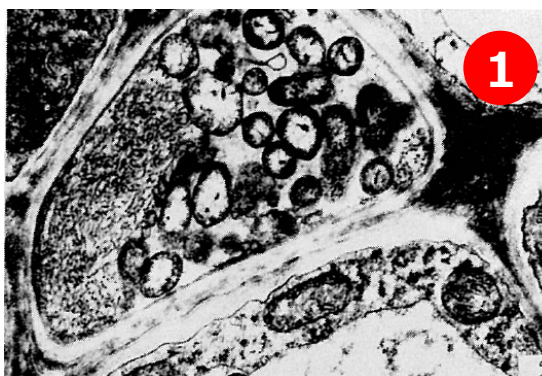
Literature

Bookshelf	1
MeSH	0
NLM Catalog	0
PubMed	20
PubMed Central	291

Genes

Gene	62
GEO DataSets	0
GEO Profiles	0
HomoloGene	0
PopSet	2

Name/Gene ID	Description
<input type="checkbox"/> pJHWp01 ID: 3206854	replication protein [<i>Aster yellows phytoplasma</i>]
<input checked="" type="checkbox"/> rep ID: 7439874	replication protein [<i>Onion yellows phytoplasma</i>]
<input type="checkbox"/> rep ID: 7439872	replication protein [<i>Onion yellows phytoplasma</i>]
<input type="checkbox"/> rep ID: 13915044	replication protein [<i>Onion yellows phytoplasma</i>]



■ ファイトプラズマ

Candidatus

Phytoplasma属細菌

- 植物の篩部細胞に寄生する植物病原細菌
- 感染植物では、がくや花弁が葉化する

rep replication protein [*Onion yellows phytoplasma*]

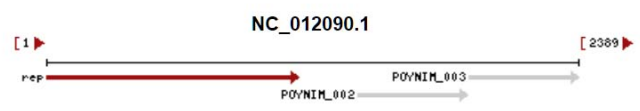
Gene ID: 7439874, updated on 6-Aug-2016

Summary

Gene symbol rep
Gene description replication protein
Locus tag POYNIM_001
Gene type protein coding
RefSeq status PROVISIONAL
Organism [Onion yellows phytoplasma \(strain: OY\)](#)
Lineage Bacteria; Tenericutes; Mollicutes; Acholeplasmatales; Acholeplasmataceae; Candidatus Phytoplasma; Candidatus Phytoplasma asteris

Genomic context

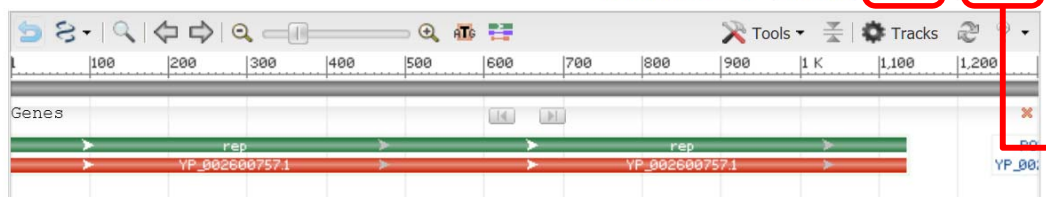
Location: plasmid: pOYNIM
Sequence: NC_012090.1 (1..1134)



Genomic regions, transcripts, and products

Genomic Sequence: NC_012090.1

Go to reference sequence details
 Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)



FASTAフォーマット

Onion yellows phytoplasma plasmid pOYNIM

NCBI Reference Sequence: NC_012090.1
[GenBank](#) [Graphics](#)
 >NC_012090.1:1-1134 Onion yellows phytoplasma plasmid pOYNIM, complete s
 ATGAAATTACGAATATGCGAACTTGTATTATAAAAAGCTTAATTACTAAAACAAAATAGAAAGCTATTT
 TAGAACTAAAAAAGAGCTATTCAAAATTATGCCTATATTTTGCATGATAAAGATATTTATCAAAATGA
 AAAAGAGGCTCAATTGAATGGTAAAAAAGTAGGAGATATAAAGCTCCTCATTGCGCATATATATGTAAGA
 TTTAATTATTCACAAGACTAACACATATCTCAATGGTTAATACCCAGAAAACCTTTGTATGTAATA
 TTAAGGTAGATTTAGTGATGCCTAATGATATGATTACGCTAATCGCTTAGATAAAGCATCAATATGA
 TGAAGAAGAGTAGTTAGTAATTTTATTGATTGAAAAGGCGAAGCTCAACAAGATATTTTAAAGAAAATAT
 AAAATGGATGCTCGCTTAAAAGATATTTACTTAAAATACATTCAGGAGAAAATTAAGAAATACAATAAA
 ATGATAATTTAATATCTAGAAAATAATATATGCTACTGCAATAGAAAAGCATTAAAGTTTGAAGAT
 TAGAGATTTAAAAAGAAAGTTAGACAAATGGAATGATTTTTATACTGGTTTAAAGTGGTTCAGGTA
 TCTACTTTAGCCAAAAAATAGCCGAAGATAAAAATATAGAGGCTTATATTTTCATCAGGTAGTAATGATA
 TTTTAGATGATTATGCGGTGAGGAATGATATTTTAGATGACCTACGTTCTAATTGTTTAGGTTTGTGTC
 TGATTTTAAAGATGTTAGATAATAAATCTGCCTCCAGTGTAAAGAGTCGTTATAAAAAATAAGTTTGA
 GAATGCAATTAATTATTACTACCGTTAAAAGTATTGATGATTTCTTTGAAGATATTTTAGAAAAAG
 ATGAAAGCATTATCAATTAACGCTGTTGTAATTAACATTTAAAATAGATTCAAATATATTTTATTA
 TAGTGTTTGGAATCCAATTGAAATGAAATATGATTTAATAGAAAAAACCTAATAATTTTAAATGAT
 TTTCAAATAAATCCTTATCAAAAAAGAAGCAAAGATTATATAAATCAATTTCTAATATAGATTTAG
 ATAAAGATTGTTAA

GenBankフォーマット

REFERENCE 5 (bases 1 to 1134)
 AUTHORS Namba, S., Oshima, K., Yamaji, Y., Kakizawa, S. and Ishii, Y.
 TITLE Direct Submission
 JOURNAL Submitted 02-FEB-2009 Contact: Shigetou Namba The University of
 Tokyo, Graduate School of Agricultural and Life Sciences, Yayoi
 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan
 COMMENT PROVISIONAL REFSEQ: This record has not yet been subject to final
 NCBI review. The reference sequence is identical to [AB079515](#).
 COMPLETENESS: full length

FEATURES
 Location/Qualifiers
 source 1..1134
 /organism="Onion yellows phytoplasma"
 /mol_type="genomic DNA"
 /strain="OY"
 /db_xref="taxon:100372"
 /plasmid="pOYNIM"
 /note="OY-NIM"
 gene 1..1134
 /gene="rep"
 /locus_tag="POYNIM_001"
 /db_xref="GeneID:7439874"
 CDS 1..1134
 /gene="rep"
 /locus_tag="POYNIM_001"
 /note="ORF5"
 /codon_start=1
 /transl_table=11
 /product="replication protein"
 /protein_id="YP_002600757.1"
 /db_xref="GeneID:7439874"
 translation: MWKSRGKELVINKTLITKTKIETLETKKKAQAYAYLLHKDKIY
 ONEKEAALNKKVVDIKAPRHHLYLRFNYSQTKHISQVNTGENFYSKIKGFSQAL
 NIMIHANRLDRHYDEKVEVSNFDWSEAGGDFPKRYKMDARLADILKIHGGEIKK
 YINIMNINLENNIYATAIEKAPKPRHKKKKVROMGCFITGSSGKSTLAKKIA
 EDKVEAYTSGGNQLDYOYGEQIILDGRKQLQLSLLKRLDKNTASVSRPK
 NWLEQQLIIITTVKSIDDFDFIKRDESIQLKRRKRLHKIKDSKYIYYSWNPHE
 MWYDLTEKPNLINDFQIKLSRKEANDYIKSISNIDLRKDC"

ORIGIN
 1 atgaattac gaattgca actgttat ataaactt taactaaa aactaaaat agaaagct atttt
 61 aaactatt tagaactaa aaaaactt atcaaat atccatat ttgcgat
 121 aaagattt atcaatga aaaaagcct cauttgat gaaaagct agaatata

データベースカタログ

http://integbio.jp/dbcatalog/?lang=ja

-生命科学系データベースを一覧から探す-
Integbio データベースカタログ

English

integbio.jp

全条件をリセット

一覧内を検索する

一覧を絞り込む

生物種

+ 動物 (658)

- 植物 (291)

シロイヌナズナ (72)

イネ (62)

ダイズ (20)

トマト (17)

ミヤコグサ (11)

コムギ (14)

オオムギ (11)

トウモロコシ (10)

クラミドモナス (6)

キャッサバ (6)

タバコ (6)

ヒメツリガネゴケ (5)

その他の植物 (152)

+ 原生生物 (60)

+ 菌類 (106)

+ 真正細菌 (159)

古細菌 (52)

ウイルス (54)

タグ<対象>

ゲノム/遺伝子 (7)

cDNA/EST (8)

データベースのレコード一覧 (全 1644件)

並べ替え: レコード公開順

生物種: **トマト** x



TFGD: Tomato Functional Genomics Database

運用機関: Cornell University

生物種: *Solanum lycopersicum* | *Solanum pennellii* | *Solanum habrochaites*

説明: トマトのファンクショナルゲノミクスのデータベースです。RNA-seqやマイクロアレイを用いた発現データ、代謝物データ、small RNAやmiRNAの情報が取られています。またトマトのESTおよびBAC... [詳細へ](#)



作物ゲノムリンク集

運用機関: 国立研究開発法人農業・食品産業技術総合研究機構

生物種: *Glycine max* | *Solanum lycopersicum* | *Raphanus sativus* | *Capsicum* | *Fragaria x ananassa* | ...

説明: 作物ゲノム育種研究センターによる作物ゲノム関連データベースのリンク集です。作物全体、ダイズ、ムギ類、果樹類、野菜類、飼料作物類、花き類に関するデータベースが収録（一部今後収録予... [詳細へ](#)



PODC: Plant Omics Data Center

運用機関: 明治大学

生物種: *Arabidopsis thaliana* | *Oryza sativa* | *Solanum lycopersicum* | *Sorghum bicolor* | *Vitis vinifera* | ...

説明: 高速シーケンサーにより得られた遺伝子発現プロファイルデータをもとに、文献情報に基づく知識情報を組み合わせた遺伝子発現ネットワーク情報を提供するデータベースです。種々の植物の遺伝... [詳細へ](#)

一括ダウンロード可



TOMATOMICS

運用機関: 明治大学 農学部

生物種: *Solanum lycopersicum*

説明: トマトの統合オミックスデータベースです。全てのトマトのEST配列、マイクロトムの完全長cDNA配列、ITAG2.4遺伝子モデルの配列を収録しています。各配列は予測されたゲノム上の位置に基づい... [詳細へ](#)

一括ダウンロード可



TOMATOMA (Tomato Mutants Archive)

運用機関: 筑波大学 遺伝子実験センター

生物種: *Solanum lycopersicum*

説明: 筑波大学 遺伝子実験センターではNBRPトマトの中核機関として、多数のトマト栽培種や近縁野生種の種子を保有しています。また、矮性トマト品種マイクロトムのEMSおよびガンマ線変異誘発系統... [詳細へ](#)

一括ダウンロード可

メニュー

- ホーム
- 本カタログについて
- 更新履歴
- データベース関係マップ
- ダウンロード
- お問い合わせ
- 類似サイトリンク集

新着情報

- 2018/02/06: 1件のレコードを追加しました。
- 2018/01/23: 3件のレコードを追加しました。
- 2018/01/16: 1件のレコードを追加しました。
- 2017/11/28: 1件のレコードを追加しました。
- 2017/11/28: 1件のレコードを追加しました。

本カタログの使い方



統合TVにて解説動画が公開されています (2017年10月04日版)

相同性検索（ホモロジー検索）

- 相同性検索は、配列の類似性から類縁の遺伝子・タンパク質を検索する方法で、進化・系統分類の解析、機能解析などを目的とした配列解析の最も基本的な手法の一つである。

SSEARCH

FASTA

<http://fasta.genome.jp/>

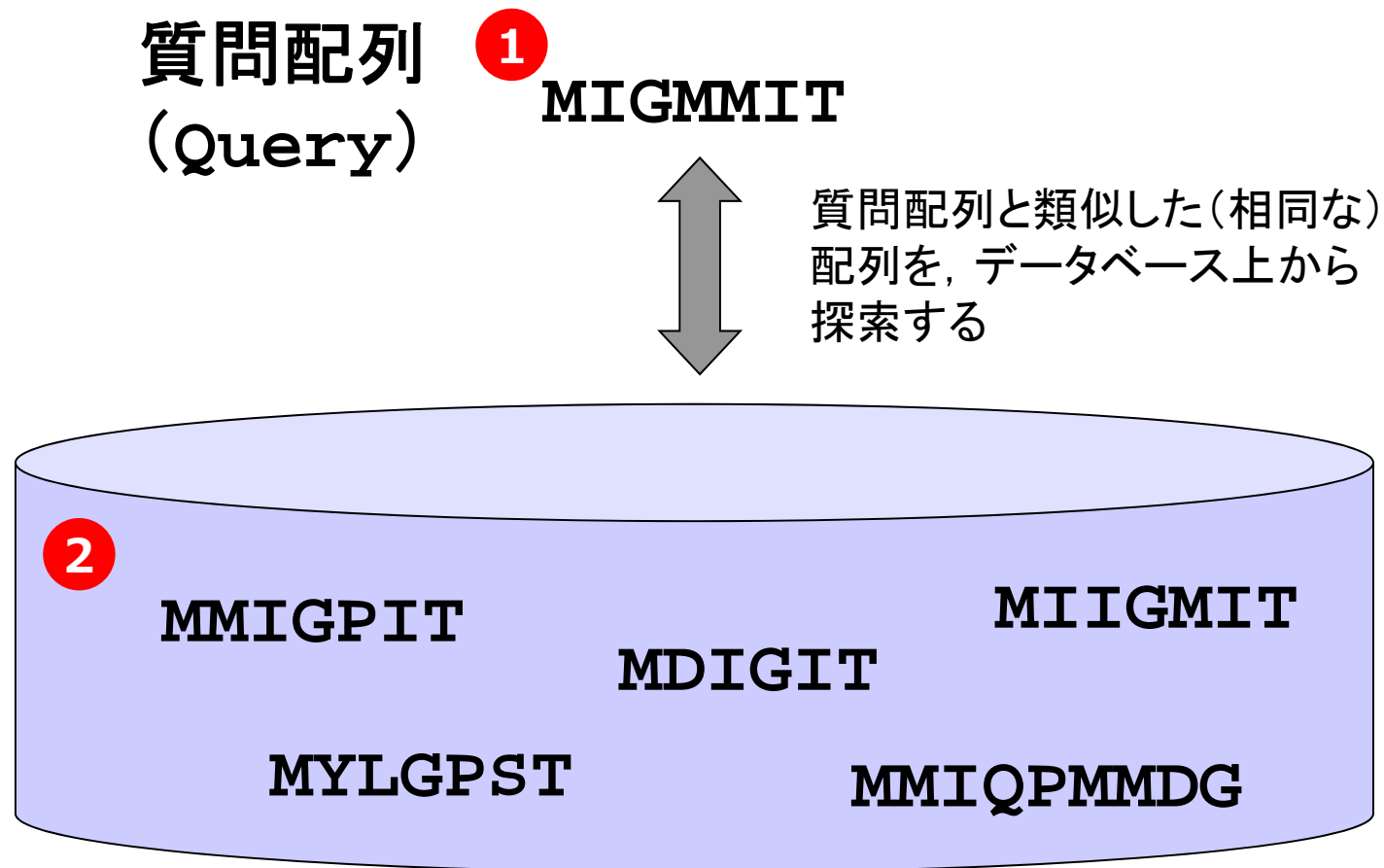
BLAST

<http://blast.genome.jp/>

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

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

相同性検索(ホモロジー検索)とは？



アラインメント

①

MIGMMIT
MMIGPIT

二つのアミノ酸配列を整列化させるにはどのように並べればよいか？

アラインメント(並置)

- ・2つの配列を要素ごとに対応づけて並べる操作
- ・進化の過程で生じ得る配列要素の挿入・欠失を ギャップ(-) に対応づける

②

グローバルアラインメント
– 配列全体の類似性を考慮

a = M-IGMMIT
b = MMIGP-IT

③

ローカルアラインメント
– 局所的な類似性を考慮

a = MIGMMIT---
b = ---MMIGPIT

アラインメントスコアの計算

- 配列の類似度 = アラインメントのスコア
- アラインメントのスコアの計算
 - ・ 対応する各要素の類似度スコアの和
 - ・ ギャップの挿入にはペナルティを与える

$$\begin{array}{l}
 \text{1} \quad \text{AFDC} \\
 \text{AEEC}
 \end{array}
 \quad
 \begin{array}{r}
 s(A, A) + s(F, E) + s(D, E) + s(C, C) = 8 \\
 3 \qquad \qquad -7 \qquad \qquad 3 \qquad \qquad 9
 \end{array}$$

$$\begin{array}{l}
 \text{2} \quad \text{AFDGC} \\
 \text{AEE-C}
 \end{array}
 \quad
 \begin{array}{r}
 s(A, A) + s(F, E) + s(D, E) + \text{gap} + s(C, C) = 0 \\
 3 \qquad \qquad -7 \qquad \qquad 3 \qquad \qquad -8 \qquad \qquad 9
 \end{array}$$

完全に一致するアミノ酸や、類似アミノ酸には高い点数を与えたい
 → 各アミノ酸の点数はどのように求めればよいか？

BLOSUMスコア (Henikoffらの方法)

BLOSUM: BLOcks amino acid Substitution Matrix

- 同一ファミリーのタンパク質のアライメントを用いて、アミノ酸の置換の頻度を調べて作成したスコア
- 良く似た配列の寄与が優勢になりすぎないように、例えば50%一致のパターンをひとまとめにして作成 → BLOSUM50

BLOSUM50マトリックス

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

Options for advanced blasting

Limit by [entrez query](#) or select from:

Compositional adjustments

Choose filter Low complexity Mask for lookup table only Mask lower case

Expect

Word Size

Matrix Gap Costs Existence: 11 Extension: 1

PSSM

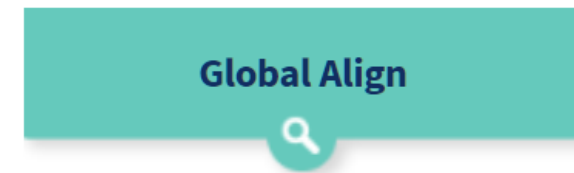
Other advanced

PHI pattern

アラインメントのアルゴリズム

■ Needleman-Wunschのアルゴリズム

- 2つの配列の最適なグローバルアラインメントを, ダイナミックプログラミング (動的計画法) により求める



Compare two sequences
across their entire span
(Needleman-Wunsch)

■ Smith-Watermanのアルゴリズム

- 2つの配列の部分配列間の一致を探索する
- 最も高いスコアをもつ一致箇所を示すアラインメントを求める
→ ダイナミックプログラミング (動的計画法)

FASTAとBLAST

- 動的計画法による検索方法 (SSERACH) は、 mn に比例した時間を要する (m, n は配列の長さ)
- 配列データベースに登録されている配列の数は膨大
→時間がかかりすぎてしまう

FASTA

- 最初に一致する配列断片を高速に検索して絞り込む
- Lipman and Pearson (1985)

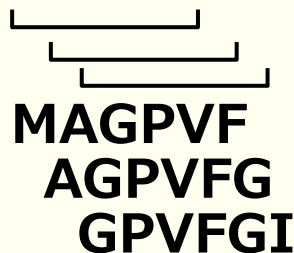
BLAST

- 最初に局所的に類似の部分配列を高速に検索して絞り込む
- Altschul (1990)

BLAST検索

- 他の方法に比べて高速であり，ホモロジー検索の方法として最もよく利用されている
 - 質問配列を固定長の断片（**ワード**）に区切る
 - まずは、**ワード単位**で類似する断片をデータベース上から検索
 - 類似度が**最大になるまで**両方向にアラインメントを伸ばす
 - 最後にこれらの**局所的**なアラインメントを結合する

1 MAGPVFGIPSCSF



MAGPVF
AGPVFG
GPVFGI

ワードの切り出し

Defaultの設定ではアミノ酸の場合は6文字，塩基配列は28文字.

↓ 一致する部分を検索

2 MSGPVFGIP...

←  →

一致したワードを中心にして両方向にアラインメントを伸ばしていく
(類似度が下がってきたらアラインメントを終了する)

NCBIのトップページの右にあるリンクからBLAST検索のページへ

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts into NCBI databases



Download

Transfer NCBI data to your computer



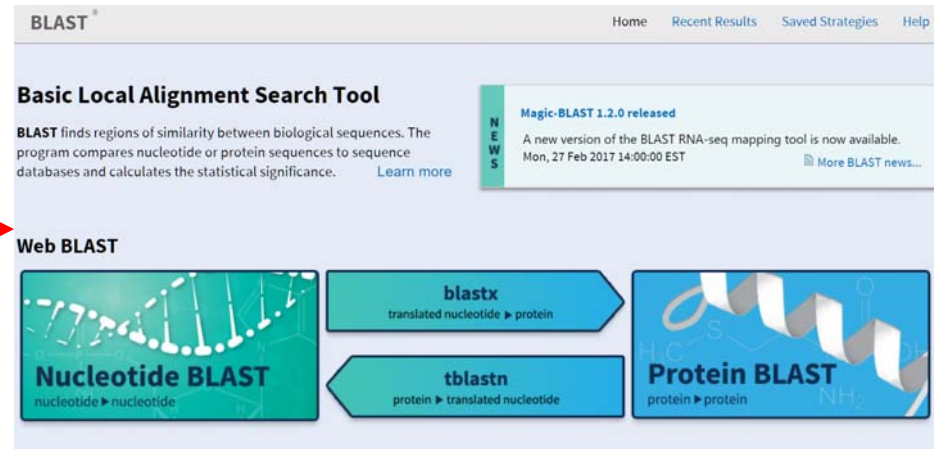
Learn

Find help documents, attend a class or watch a tutorial



Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST**
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem



プログラム	質問配列 (query)	検索対象
Protein BLAST	アミノ酸配列	アミノ酸配列データベース
blastx	塩基配列	アミノ酸配列データベース
Nucleotide BLAST	塩基配列	塩基配列データベース
tblastn	アミノ酸配列	塩基配列データベース
tblastx	塩基配列	塩基配列データベース

BLASTP検索 (protein blast)

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

The screenshot shows the BLASTP search interface. The 'Enter Query Sequence' section has a text input field with the red annotation '② 貼り付ける' (Paste) and a 'Clear' button. Below it is an 'Or, upload file' section with a '参照...' (Reference) button. The 'Choose Search Set' section has a 'Database' dropdown menu set to 'Non-redundant protein sequences (nr)' with a red arrow pointing to it and the annotation '③ データベースを選ぶ (nr)'. The 'Program Selection' section has radio buttons for 'blastp (protein-protein BLAST)', 'PSI-BLAST', and 'PHI-BLAST', with 'blastp' selected. At the bottom, there is a 'BLAST' button with a red arrow pointing to it and the annotation '④ 「BLAST」を押す'. There is also a checkbox for 'Show results in a new window'.

```
>sample1
MNRVFLFGKLSFTPNRLQTKNGTLGATFSMECLDS
SGFNNAKSFIRVTAWGKVASFIVAQNPGVMLFVEG
RLTTYKITNSENKNTYALQVTADKIFHPDEKTTNE
EPIKSTVVDSPFMNPKASVTEAEFEQAFPHQDET
D FNNITPIFENDVQLEEEESDD
```

① 配列をコピーする
(">"の行は入れても入れなくてもよい)

③ データベースを選ぶ
(nr)

④ 「BLAST」を押す

nr : 冗長性をなくした (non-redundant) アミノ酸データベース

- 質問配列と類似した（**相同な**）アミノ酸配列のリストが表示される
- 一番上が最も**相同性の高い**アミノ酸配列（タンパク質）

5 Alignmentsのタブをクリック

Descriptions		Graphic Summary	Alignments	Taxonomy		
Sequences producing significant alignments		Download	Manage Columns	Show		
<input checked="" type="checkbox"/> select all 100 sequences selected		GenPept	Graphical 2	Distance 3 of ref 4 M		
1	Description	Max Score	Total Score	Query Cover	E value 3	Per. Ident
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Mycoplasma genitalium]	330	330	100%	3e-114	100.00%
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Mycoplasma pneumoniae]	202	202	100%	1e-63	58.18%
<input checked="" type="checkbox"/>	putative 19 kDa protein [Mycoplasma pneumoniae]	70.1	70.1	43%	9e-13	50.00%
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Brevibacillus borstelensis]	57.4	57.4	75%	4e-07	31.54%
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Firmicutes bacterium CAG:170]	57.0	57.0	83%	4e-07	25.36%
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Phorcysia thermohydrogeniphila]	56.2	56.2	68%	6e-07	32.17%
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Candidatus Colwellbacteria bacterium RIFCSPHIGHO2_02_FULL_43_15]	56.2	56.2	72%	1e-06	33.33%
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Veillonella seminalis]	54.7	54.7	61%	2e-06	33.98%
<input checked="" type="checkbox"/>	single-stranded DNA-binding protein [Candidatus Colwellbacteria bacterium RIFCSPLOWO2_12_FULL_43_11]	54.7	54.7	72%	3e-06	32.52%

1 ヒットした配列のGenBankフォーマットデータへのリンク

Download [GenPept](#) [Graphics](#)

single-stranded DNA-binding protein [Mycoplasma pneumoniae]

Sequence ID: [WP_010874586.1](#) Length: 166 Number of Matches: 1

See 16 more title(s) ▾

全長は166アミノ酸 6

Range 1: 1 to 165 [GenPept](#) [Graphics](#)

E-value 4

3 スコア

5 相同性 (identity) 相同性 (similarity) ギャップ

Score	Expect	Method	Identities	Positives	Gaps
202 bits(514)	1e-63	Compositional matrix adjust.	96/165(58%)	126/165(76%)	5/165(3%)

Query 1 MNRVFLFGKLSFTPNRLQTKNGTLGATFSMECLDSSGFNNAKSFIRVTAWGKVASFIVAQ 60

Sbjct 1 MNRVFLFGKLSF PN+LQT+ +GA+FS+ C+DSSGFN++KS+IR+TAWGKVASF++
MNRVFLFGKLSFDPNKLQTRTNNIGASFSLACIDSSGFNDSKSYIRITAWGKVASFVLT 60

Query 61 NPGVMLFVEGRLTTYKITNSEN----KNTYALQVTADKIFHPDEKTTNEEPI-KSTVVDS 115

Sbjct 61 PG +FVEGRL+TYK+ N + K TYALQV ADK++ PDE+ + E+P+ K+TV+DS
KPGDSVFVEGRLSTYKMNNRSDDPNSKATYALQVIADKVYRPDEENSLEQPVDKATVIDS 120

Query 116 PFMNPKASVTEAEFEQAFPHQDETDFNNITPIFENDVQLEEEESDD 160

Sbjct 121 PF+ K + TE E QAFP + + ++I PI ND QLEEEESDD
PFLAAKTNATENELAQAFPI SLDDDDDDINPILNNSQLEEEESDD 165 7

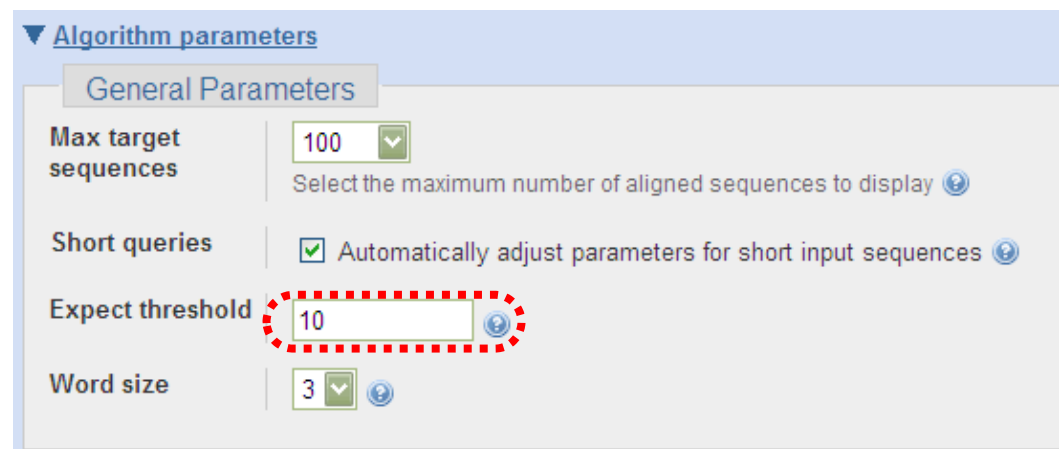
2 Query : 質問配列
中段 : 一致するアミノ酸、あるいは+ (類似アミノ酸)
Sbjct : Blast検索の結果, ヒットした配列

↑ 全長ではないので注意
(本当は166番目にEがある)

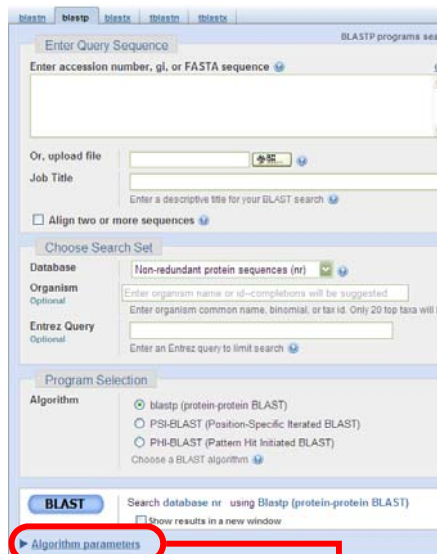
E-value

- BLAST検索では、**相同性の指標としてE-valueがよく用いられる**
- E-valueとは ⇒ ランダムな配列同士を比較したときに、今回の検索結果と同じスコアになる**配列数の期待値**
- E-valueが**小さい**ほど偶然には起こり得ない
= 「よく似ている」 ことを示している
- BLAST検索の際にE-valueの**しきい値**を設定することで、その値よりも小さいE-valueの検索結果しか表示されないようにすることもできる

1



The screenshot shows the 'Algorithm parameters' section of a BLAST search interface. Under the 'General Parameters' tab, the 'Expect threshold' is set to 10. A red dashed circle highlights the 'Expect threshold' input field, which is labeled with a red circle containing the number '1'. Other parameters shown include 'Max target sequences' (100), 'Short queries' (checked), and 'Word size' (3).



1



Algorithm parameters

General Parameters

Max target sequences | 100 | 検索結果の表示件数 2
 Select the maximum number of aligned sequences to display

Short queries | Automatically adjust parameters for short input sequences

Expect threshold | 10 | E-valueのしきい値 3

Word size | 3 | BLAST検索時のWordサイズ 4

Scoring Parameters

Matrix | BLOSUM62 | マトリックスの種類を選ぶ 5

Gap Costs | Existence: 11 Extension: 1 | ギャップのスコア設定 6

Compositional adjustments | Conditional compositional score matrix adjustment | E-value計算時の設定 7

Filters and Masking

Filter | Low complexity regions | 冗長配列を取り除く場合はチェック

Mask | Mask for lookup table only | 冗長配列を取り除く場合の設定 8
 Mask lower case letters | 小文字を無視する場合の設定

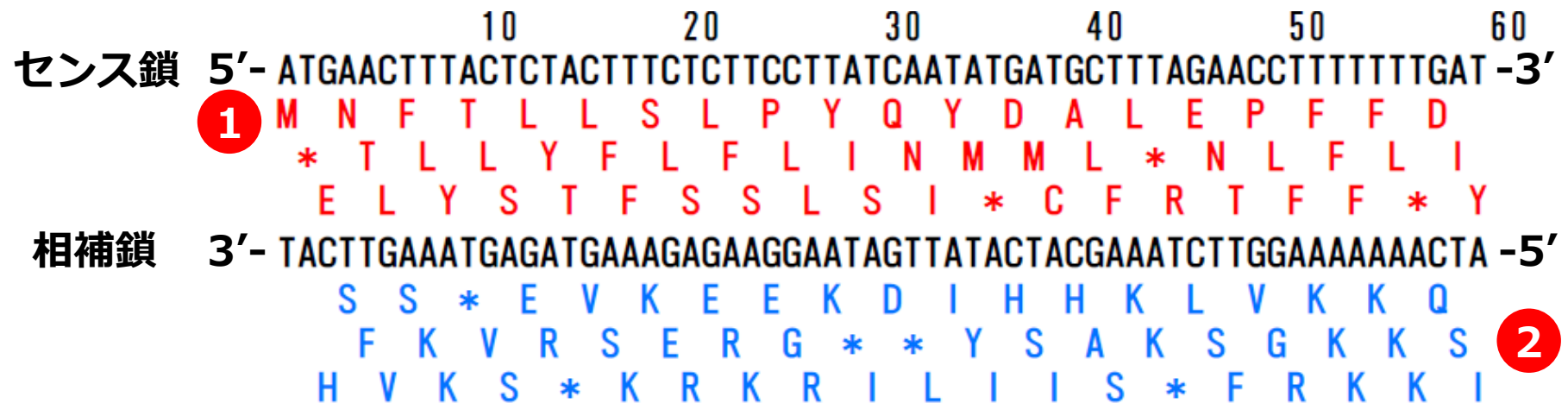
BLAST | Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

blastx

塩基配列を入力



6通りのreading frameのすべてについて翻訳し, アミノ酸配列データベースに対して検索してくれる



- ・塩基配列を決定したが, **どんなタンパク質コードされているかわからない**とき
- ・**non-coding領域**に, タンパク質がコードされていないかどうかを調べたいときなど

BLAST® » blastp suite » results for RID-8SS47UDA016



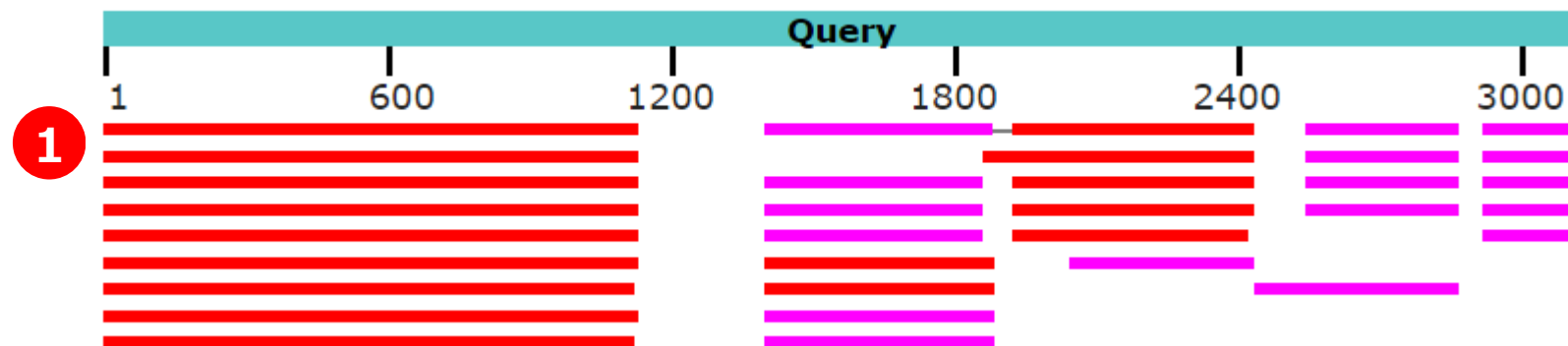
↑ 上部のBLASTをクリックし、blastx のページへ

- sample2の塩基配列を blastx検索にかける

>sample2

```
ATGAAATTAAGAATCTGCGAACTTGTTATTAATAAACTTTAATTACTAAAAC TAAA
ATAGAAACTATTTTAGAACTAAAAAAAAGCCATTCAA AATTATGCCTATATTTTG
CATGATAAAGATATTTATCAA AATGATAAAGAGGCTCAATTGAATGGTAAAAAAGTA
GGAGATATAAAAGCTCCTCATTGGCATATATTTTAAGATTTAA
```

- Graphic Summary のタブをクリック



- 5つのタンパク質がコードされていることがわかる

blastn (nucleotide blast)

上部のBLASTをクリックし、blastn のページへ

```
>sample3  
TTGAAGAGGACTTGGAACCTTCGAT
```

①配列をコピーする
(">"の行は入れても入れなくてもよい)

③データベースを選ぶ
(nr/nt)

④「BLAST」を押す

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nu

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear Query subrange From To

Or, upload file 参照...

Job Title ath-miR163 MIMAT0000184 Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.): Nucleotide collection (nr/nt)

Organism Optional Enter organism name or id--completions will be suggested

Entrez Query Optional Enter an Entrez query to limit search

Program Selection

Optimize for Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn) Choose a BLAST algorithm

BLAST Search database nr using Megablast (Optimize for highly similar sequences) Show results in a new window

ⓘ Your search parameters were adjusted to search for a short input sequence.

と表示され、短い配列用の設定で検索される

tblastn

アミノ酸配列を入力



データベース上の塩基配列を、6通りのreading frameのすべてについて翻訳し、このアミノ酸配列データに対して検索してくれる

- **EST配列**や**ドラフトゲノム**など、アノテーション情報が整備されていないデータから相同な配列を探したいときに便利

tblastx

塩基配列を入力



6通りのreading frameのすべてについて翻訳



データベース上の塩基配列も、6通りのreading frameのすべてについて翻訳し、このアミノ酸配列データに対して検索

- ・ 質問配列、データベースとも、**アノテーション情報が整備されていない**場合に便利

BLAST検索 (GenomeNet)

```
>sample5
MDENETQFNKLNQVKNKLLKIGVFGIGGAGNNIVDASLYHYPN
LASENIHFYAINSDLQHLAFKTNVKNKLLIQDHTNKGFAGG
DPAKGASLAISFQEQFNLTLDGYDFCILVAGFGKGTGTGATP
VFSKILKTKKILNVAIVTYPVSLNEGLTVRNKATKGLEILNKA
TDSYMLFCNEKCTNGIYQLANTEIVSAIKNLIELITIPLOQN
IDFEDVRAFFQTKKTNQDQQLFTVTHPFSFSFDSKDSIEQFA
KQFKNFEKVSYFDHSIVGAKKVVLLKANINQKIVKLNFKQIQD
IIWTKIDNYQLEIRLGVDVFTTIPNIQIFILSEHKNPVSLPI
DNKSTENNQNKLKLLDELKELGMKYVKHQNQIY
```

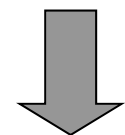
①配列をコピーする
(">"の行は入れても入れなくてもよい)

③Favorite organisms を選択

④「mge mpn uur」と入力

mge: *Mycoplasma genitalium*
mpn: *Mycoplasma pneumoniae*
uur: *Ureaplasma parvum*

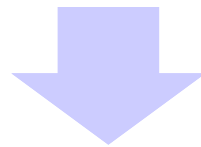
⑤「Compute」を押す



Entry	bits	E-val
Top 10 <input type="button" value="Clear"/> Select operation <input type="button" value="Exec"/>		
<input checked="" type="checkbox"/> mge:MG_224 ftsZ; cell division protein FtsZ ; K03531 cell divisi...	679	0.0
<input checked="" type="checkbox"/> mpn:MPN317 ftsZ, F10_orf380; cell division protein FtsZ ; K03531...	358	e-100
<input checked="" type="checkbox"/> uur:UU317 hypothetical protein	28	0.53
<input checked="" type="checkbox"/> mpn:MPN257 galE, A65_orf338; UDP-glucose 4-epimerase	28	0.68

} *Ureaplasma*は、ftsZを持っていないことがわかる

- **大量の質問配列**についてBLAST検索を行いたい
- 自分の持っている**未公開のデータ**に対して検索したい
- ホモロジー検索を用いて**比較ゲノム解析**を行いたい

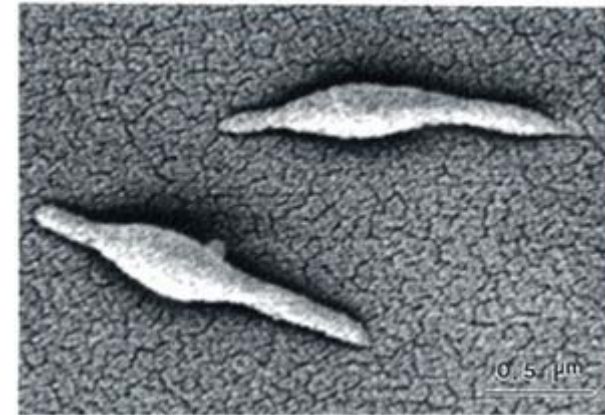


Stand-alone BLASTを利用する

(ローカルなコンピュータで動くBLASTのプログラム)

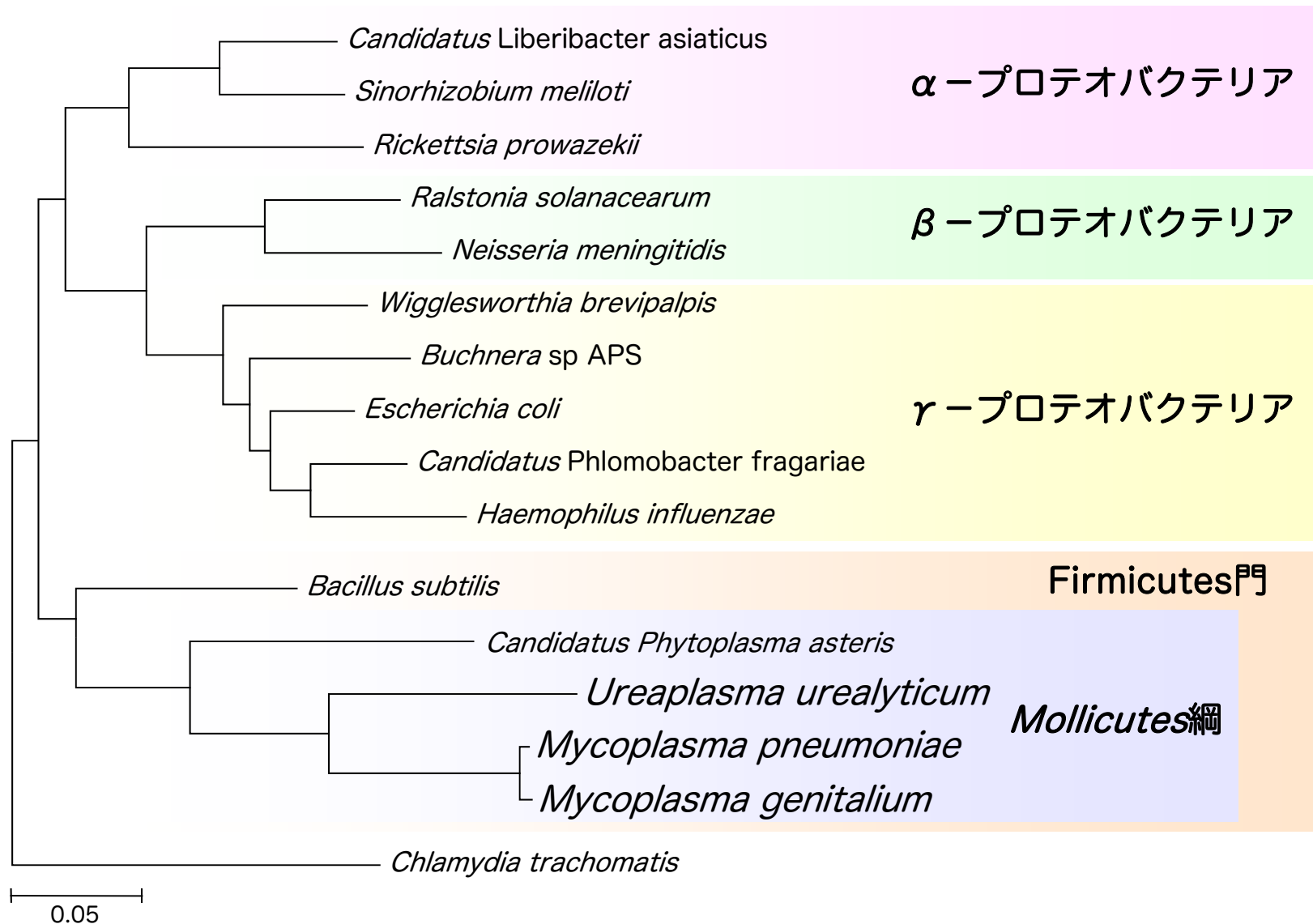
細菌の全ゲノム解読の歴史

生物種	ゲノムサイズ (Mbp)	全ゲノム解読 された年
<i>Haemophilus influenzae</i>	1.83	1995
★ <i>Mycoplasma genitalium</i>	0.58	1995
★ <i>Mycoplasma pneumoniae</i>	0.82	1996
・		
・		
・		
<i>Bacillus subtilis</i>	4.21	1997
<i>Escherichia coli</i>	4.67	1997
・		
★ <i>Ureaplasma parvum</i>	0.75	2000
・		
・		



- ◆ マイコプラズマ類は、ゲノムサイズが小さいため、ゲノムプロジェクトで取り上げられることが多かった

マイコプラズマの系統学的位置



■ blastフォルダに移動します

```
> cd C:\Users\iu\Desktop\blast
```

「cd (スペース)」を打ち込む → blastフォルダをドラッグして
→ コマンドプロンプトの上にドロップ → リターンを押します

以下のように表示されます

```
C:\Users\iu\Desktop\blast>
```

■ blastフォルダ内のファイルを表示します

```
> dir
```

```
2009/03/11  19:52    <DIR>          .
2009/03/11  19:52    <DIR>          ..
2005/04/21  23:34    222,447 Mgenitalium.faa
2005/04/21  23:33    307,006 Mpneumoniae.faa
.
.
```


データベースの準備

- 練習用に *Mycoplasma genitalium* のゲノムデータを用います
- blastフォルダの中に **Mgenitalium.faa** という **Multi-FASTAフォーマット形式** のファイルが置いてあります
- 中身を見てみましょう

```
> more Mgenitalium.faa
```

moreコマンドについて

指定したファイルの内容を表示します。次ページを見るには [Space]キー、1行ずつ見るには[Enter]キー、終了するには[Q]キー押します。

blastフォルダ内のファイルを、メモ帳等で開いてもOKです

データベースの準備

- stand-alone BLASTはMulti-FASTAフォーマットのままでは、データベースとして使うことができません
- **BLAST用のデータベースへ変換する**ために以下のコマンドを実行します

```
> makeblastdb -in Mgenitalium.faa -dbtype prot
```

1

2

-in オプション：データベース指定

-dbtype オプション：データがアミノ酸配列の場合は prot

データがアミノ酸配列の場合は nucl

stand-alone BLASTの実行

- Query (質問配列) にはtest1.seqを用います

```
> more test1.seq
```

```
>gi|16130505|ref|NP_417075.1| uracil-DNA-glycosylase  
[Escherichia coli str. K-12 substr. MG1655]  
MANELTWHDVLAEEKQQPYFLNTLQTVASERQSGVTIYPPQKDVFNFRFTELG  
DVKVVILGQDPYHGPQAHGLAFSVRPGIAIPPSLLNMYKELENTIPGFTRPNH  
GYLESWARQGVLLLNTVLTVRAGQAHSHASLGWETFDTKVISLINQHREGVVFL  
LWGSHAQKKGAIIDKQRHHVLKAPHPSPLSAHRGFFGCNHFVLANQWLEQRGET  
PIDWMPVLP AESE
```

楽にコマンドを入力するコツ

ファイル名 (例えば test1.seq) を入力するときに, 「t」や「te」など
最初の数文字を入力した後, **Tabを押す**ことで, その文字から始まるファイル名
を自動的に表示させることができます

stand-alone BLASTの実行

- test1.seqを質問配列として使い, Mgenitalium.faaデータベースに対してblastp検索を行うには, 以下のコマンドを実行します

```
> blastp -db Mgenitalium.faa -query test1.seq
```

1

2

-db : データベースを指定

-query : 質問配列 (query) を指定

検索結果をテキストファイルとして出力する

- 検索結果をファイルとして出力するには, `-out` オプションを用います

```
> blastp -db Mgenitalium.faa -query test1.seq  
-out result1.txt  
> more result1.txt
```

`-out` : 出力ファイル指定

楽にコマンドを入力するコツ

↑ (上矢印) を押すと, 過去に入力したコマンドが出てきます

- リダイレクトという機能を使って出力することもできます

```
> blastp -db Mgenitalium.faa -query test1.seq  
> result1.txt
```

■ メモ帳やワードパッドを使って result1.txt を開いてください

検索対象として用いた
データベース →

質問配列の名前 →

アラインメント →

```

BLASTP 2.9.0+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference for composition-based statistics: Alejandro A. Schaffer,
L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri
I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001),
"Improving the accuracy of PSI-BLAST protein database searches with
composition-based statistics and other refinements", Nucleic Acids
Res. 29:2994-3005.

Database: Mgenitalium.faa
         484 sequences; 175,929 total letters

Query= gi|16130505|ref|NP_417075.1| uracil-DNA-glycosylase [Escherichia
coli str. K-12 substr. MG1655]

Length=229

Sequences producing significant alignments:

gi|12044949|ref|NP_072759.1| uracil DNA glycosylase (ung) [Mycopl... 108      6e-31
gi|12045134|ref|NP_072945.1| guanosine-3',5'-bis(diphosphate) 3'-... 23.1    2.8
gi|12044874|ref|NP_072684.1| GTP-binding protein, putative [Mycop... 22.3    5.1
gi|12045072|ref|NP_072883.1| cytadherence accessory protein (hmw2... 21.6    8.8

>gi|12044949|ref|NP_072759.1| uracil DNA glycosylase (ung) [Mycoplasma
genitalium G-37]
Length=245

Score = 108 bits (271), Expect = 6e-31, Method: Compositional matrix adjust.
Identities = 72/226 (32%), Positives = 106/226 (47%), Gaps = 14/226 (6%)

Query  6      TWHDVLAEEKQPPYFLNLTQTVASERQSGVTIYPPQKDVFNAFRFTLGDVKVVILGQDP 65
      +W  + EE ++PYF  L+ + + +   TI P  + +F  F F +  D KV+I GQDP
Sbjct 17      SWRAFIDEEVKPPYFQALLEKALK---ATIIPKPELIFRVFSFFKPIDTKVIIFGQDP 73

Query  66      YHGPQQAHLAFSVRPGIAIPPSLLNMYKELENTIPGFTRPN---HGYLESWARQGVLLL 122
      Y  P  A GLAF+      P SL  +  LE  P  + +      +L +WA QGVLLL
Sbjct  74      YPSPNDACGLAFASNNS-KTPASLKRRIILRLEKEYPSLQESSWQQNFLLNWAEQGVLLL 132
    
```

スコア →
E value →

E value設定

- E-valueが配列同士の相同性が高いことを示しています
- BLAST検索の際にE valueのしきい値を設定することで、その値よりも小さいE valueの検索結果しか出力されなくなります
- しきい値を設定するには、-evalueオプションを用います

```
> blastp -db Mgenitalium.faa -query test1.seq  
-out result1.txt -evalue 1e-10  
  
> more result1.txt
```

1 (いち) と 1 (エル) の違いに注意してください

BLASTX

- 次にblastX検索を行ってみましょう
- test2.seqには塩基配列データが入っています

```
> more test2.seq  
> blastx -db Mgenitalium.faa -query test2.seq  
-evaluate 1e-10 -out result2.txt  
> more result2.txt
```

- メモ帳やワードパッドを使って result2.txt を開いてください

大量Queryのホモロジー検索法

- stand-alone BLASTは、**Multi-FASTA形式の質問配列**にも対応しています。
- 例えば、下のような**複数の配列を含むファイル**を質問配列として用いると、それぞれをBLAST検索した結果が**つながった一つ**のファイルとして出力されます。

```
>gi|49176138|ref|NP_416237.3| 6-phosphofructokinase II [Escherichia coli K12]
MVRIYTLTLAPSLDSATITPQIYPEGKLRCTAPVFEPGGGGINVARAIAHLGGSATAIFPAGGATGEHLV
SLLADENVPVATVEAKDWTRQNLHVHVEASGEQYRFVMPGAALNEDEFRQLEEQVLEIESGAILVISGSL
PPGVKLEKLTQLISAAQKQGIRCIIVDSSGEALSAALAIGNIELVKPNQKELSALVNRELTQPDDVRKAAQ
EIVNSGKAKRVVSLGPGQALGVDSENCIQVPPPVKVSQSTVGAGDSMVGAMTLKLAENASLEEMVRFV
AAGSAATLNQGTRLCSHDDTQKIYAYLSR

>gi|16132212|ref|NP_418812.1| phosphoglyceromutase 2 [Escherichia coli K12]
MLQVYLVRHGETQWNAERRIQGQSDSPLTAKGEQQAMQVATRAKELGITHI ISSDLGRTRRTAEIIAQAC
GCDIIFDSRLRELNMGVLEKRRHIDSLTEEEENWRRQLVNGTVDGRIPEGESMQELSDRVNAALESCRDLP
QGSRPLLVSHGIALGCLVSTILGLPAWAERLRLRNCISISRVVDYQESLWLASGWVETAGDISHLDPAL
DELQR

>gi|16131851|ref|NP_418449.1| glucosephosphate isomerase [Escherichia coli K12]
MKNINPTQTAAWQALQKHFDEMKDVTIADLFAKDGDRFSKFSATFDDQMLVDYSKNRITEETLAKLQDLA
KECDLAGAIKSMFSGEKINRTENRAVLHVALRNRSNTPIILVDGKDVMPVNAVLEKMKTFSEAIISGEWK
GYTGKAITDVVNIIGGSDLGPYMVTEALRPYKNNLNMHFVSNVDGTHIAEVLKKNVPETTLFLVASKTF
TTQETMTNAHSARDWFLKAAGDEKHAVAKHFAALSTNAKAVGEFGIDTANMFEFWDWVGGRYSLWSAIGLS
IVLSIGFDNFVELLSGAHAMDKHFSTTPAEKNLPVLLALIGIWINNFFGAETEAAILPYDQYMHRFAAYFQ
QGNMESNGKYVDRNGNVVDYQTGPIIWGEPGTNGQHAFYQLIHQGTKMVPCDFIAPAIITHNPLSDHHQKL
LSNFFAQTEALAFGKSREVVEQEYRDQGKDPATLDYVVPFKVFEGNRPTNSILLREITPFSLGALIALYE
HKIFTQGVILNIFTFDQWGVELGKQLANRILPELKDDEKISSHDSSTNGLINRYKAWRG
```

大量Queryのホモロジー検索法

- test3.seqには、100個分のアミノ酸配列がMulti-FASTAフォーマットで記述してあります

```
> more test3.seq
```

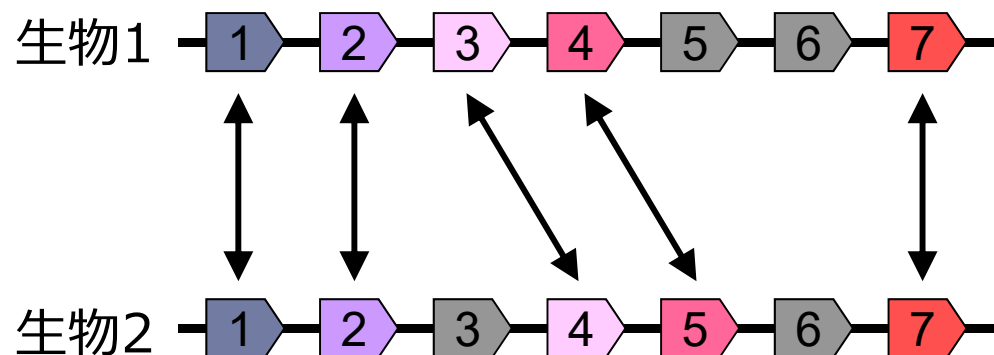
- これらと相同なアミノ酸配列がMgenitalium.faa内にあるかどうかを調べるために、以下のコマンドを実行してください

```
> blastp -db Mgenitalium.faa -query test3.seq  
-evalue 1e-10 -out result3.txt
```

- メモ帳やワードパッドを使って result3.txt を開いて、結果を確認してください

ホモロジー検索を用いた比較ゲノム解析

- アミノ酸配列が類似したタンパク質は、**機能も似ている**ことが推測されます
- 類似性が高く、おそらく共通の祖先タンパク質から派生したと考えられるタンパク質のことを、「**オーソログ**」と呼びます
- 片方の生物種の**すべてのタンパク質**を質問配列として用いて、相手の**すべてのタンパク質**に対してホモロジー検索を行うことで、オーソログ遺伝子を網羅的に同定できます



ホモロジー検索を用いた比較ゲノム解析

- Mpneumoniae.faaには, *Mycoplasma pneumoniae* のゲノムにコードされる全アミノ酸配列がMulti-FASTAフォーマットで記述してあります

```
> more Mpneumoniae.faa
```

- これらと相同なアミノ酸配列を *M. genitalium* が持っているかどうかを調べるために, 以下のコマンドを実行してください

```
> blastp -db Mgenitalium.faa -query Mpneumoniae.faa  
-evaluate 1e-10 -out result4.txt
```

- メモ帳やワードパッドを使って result4.txt を開いて、結果を確認してください

perlを用いたデータ処理

- 大量の質問配列を使ってBLAST検索を行うと、**結果が羅列**した形で出力されます
- **Perl**などのプログラミング言語を用いることで、この中から、必要な情報だけを取り出すことができます
- 質問配列のアクセッション番号や、検索の結果ヒットしたタンパク質の情報などのリストを作成してみましょう

Query GI	ref No.	Function	Length	Score	E-value	Identity
16132212	NP_014926.1	Yor283wp	230	62.8	4.00E-11	48%
16131851	NP_009755.1	Glucose-6-phosphate isomerase; Pgi1 p	554	641	0	73%
16131757	NP_010335.1	triosephosphate isomerase; Tpi1 p	248	192	4.00E-50	60%
16131754	NP_011756.1	phosphofructokinase alpha subunit; Pfk1 p	987	184	2.00E-47	51%
16131018	NP_009362.1	Pyruvate kinase; Cdc19p	500	40.8	2.00E-04	50%
16130827	NP_009938.1	3-phosphoglycerate kinase; Pgc1 p	416	255	7.00E-69	57%
16130826	NP_012863.1	aldolase; Fba1 p	359	352	4.00E-98	68%
16130686	NP_011770.1	enolase I; Eno1 p	437	359	1.00E-100	62%
16130106	NP_009965.1	ribokinase; Rbk1 p	333	35.4	0.012	59%
16129807	NP_009362.1	Pyruvate kinase; Cdc19p	500	247	3.00E-66	49%
16129733	NP_012483.1	Glyceraldehyde-3-phosphate dehydrogenase	332	427	1.00E-120	77%

Database: Mgenitalium.faa
484 sequences; 175,929 total letters

1 Query= gi|13507740|ref|NP_109689.1| DNA polymerase III beta subunit
[Mycoplasma pneumoniae M129]

Length=380

Sequences producing significant alignments:	Score (Bits)	E Value
gi 12044851 ref NP_072661.1 DNA polymerase III, subunit beta (dn...	525	0.0

2 >gi|12044851|ref|NP_072661.1| DNA polymerase III, subunit beta
(dnaN) [Mycoplasma genitalium G-37]
Length=364

3

Score = 525 bits (1352), **Expect = 0.0**, Method: Compositional matrix adjust.
Identities = 257/364 (71%), Positives = 315/364 (87%), Gaps = 0/364 (0%)

Query	17	LNNVIVSNNKMKPYHSYLLIEATEKEINFYANNEYFSAKCTLAENIDVLEEGERVIVKGI	76
		+NNVI+SNNK+KP+HSY LIEA EKEINFYANNEYFS KC L +NID+LE+G +IVKGI	
Sbjct	1	MNVIIISNNKIKPHHSYFLIEAKEKEINFYANNEYFSVKCNLNKNIDILEQGSLIVKGI	60
Query	77	FSELINGIKEDIITIQEKDQTLVKTCKTNINLNTIDKKEFPRIKFNQNVLDKEFDELKI	136
		F++LINGIKE+IITIQEKDQTLVKTCKT+INLNTI+ EFPRIRFN+ DL EF++ KI	
Sbjct	61	FNDLINGIKEEIIITIQEKDQTLVKTCKTSINLNTINVNEFPRIKFNKNDLSEFNQFKI	120
Query	137	QHSLLTGKGLKIAHAVSTFRESTRKFNQVNFNGSNGKQIFLEASDSYKLSVYEIKQKTD	196
		+SLL KG+KKI H+VS RE + KFNQVNFNGSNGK+IFLEASD+YKLSV+EIKQ+T+P	
Sbjct	121	NYSLLVKGIKKIFHSVSNNREISSKFNQVNFNGSNGKEIFLEASDQYKLSVFEIKQETEP	180
Query	197	FNFIVETNLLSFINSFNPEGDLISIFFRKEHKDDLSTELLIKLDNFLINYTSINESFPR	256
		F+FI+E+NLLSFINSFNPE I ++RK++KD STE+LI +DNF+I+YTS+NE FP	
Sbjct	181	FDFILESNNLLSFINSFNPEEDKSIVFYRKNKDSFSTEMLISMDNFMISYTSVNEKFPE	240

- "Query=" で始まる行に質問配列の情報が書かれており, ">" で始まる行にヒットした遺伝子の情報が書かれています.
- これらの情報を抜き出して表示するプログラム **parse-blast7.pl** を用意しておきました.

```
> more parse-blast7.pl
```

parse-blast.pl

```
#!/usr/local/bin/perl

use strict;
use warnings;
use Getopt::Std;

my $mode = 0;
my $name = "";
    .
    .
```

- Perlのプログラミングについては, 次回の講義で扱います.

- 以下のコマンドを入力し, result4.txtを処理します
→ list1.txt というファイルが新たに出来上がります

```
> perl parse-blast7.pl -i result4.txt -o list1.txt
```

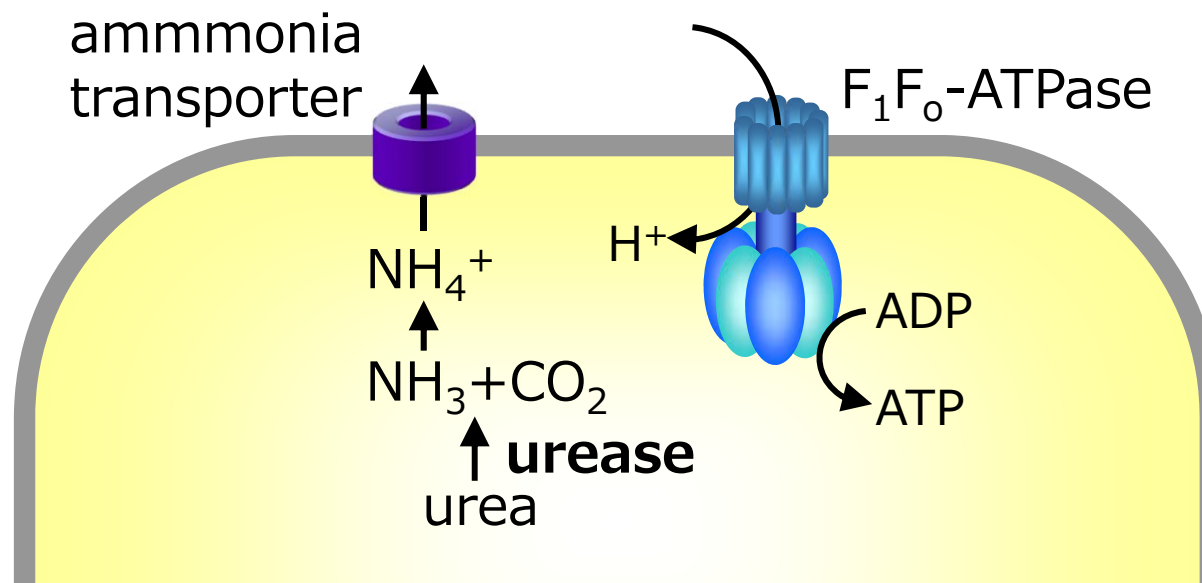
- Excel を開きます (空白のブック)
- list1.txt を Excel上にドラッグ&ドロップしてください

1 質問配列の情報		2 BLAST検索でヒットした配列の情報 (ヒットしなかった場合は空欄)			3 スコア, E-value, Identity		
Query GI	Query	Hit_ref No.	Hit_Function	Hit_Length	Score	E-value	Identity
gi 13507740	DNA polymerase III beta subu	NP_072661.1	DNA polymerase III, subunit k	364	516	1.00E-148	70%
gi 13507741	similar to J-domain of DnaJ[M	NP_072662.1	dnaJ-like protein [Mycoplasm	310	437	1.00E-125	83%
gi 13507742	DNA gyrase subunit B [Myco	NP_072663.1	DNA gyrase subunit B (gyrB)	650	1184	0	86%
gi 13507743	DNA gyrase subunit A [Myco	NP_072664.1	DNA gyrase subunit A (gyrA)	836	1330	0	84%
gi 13507744	seryl-tRNA synthetase [Mycc	NP_072665.1	seryl-tRNA synthetase (serS	417	669	0	76%
gi 13507745	thymidylate kinase [Mycoplas	NP_072666.1	thymidylate kinase (tmk) [Myc	210	280	1.00E-77	62%
gi 13507746	similar to DNA-polymerase su	NP_072667.1	hypothetical protein MG007 [254	281	4.00E-78	72%
gi 13507747	thiophene and furan oxidator	NP_072668.1	thiophene and furan oxidator	442	573	1.00E-166	63%
gi 13507748	hydrolase [Mycoplasma pneur	NP_072669.1	hypothetical protein MG009 [262	365	1.00E-103	64%
gi 13507749	hypothetical protein MPN010						
gi 13507750	hypothetical protein MPN011						
gi 13507751	hypothetical protein MPN012						
gi 13507752	hypothetical protein MPN013						
gi 13507753	hypothetical protein MPN014	NP_072670.1	hypothetical protein MG010 [218	230	9.00E-63	70%
gi 13507754	hypothetical protein MPN015	NP_072671.1	hypothetical protein MG011 [287	325	3.00E-91	82%
gi 13507755	similar to ribosomal S6modific	NP_072672.1	hypothetical protein MG012 [287	368	1.00E-104	62%

4

M. genitaliumゲノム上には、これらと相同なタンパク質がコードされていない

- ◆ *Ureaplasma* はウレアーゼを用いて尿素を分解し，その結果生じたプロトン濃度勾配を利用してATPを合成する



Query GI	Query	Hit_ref No.	Hit_Function
gi 1335799	urease complex component[Ureaplasma parvum		
gi 1335799	urease complex component[Ureaplasma parvum		
gi 1335799	urease complex component[Ureaplasma parvum		
gi 1335799	urease complex component[Ureaplasma parvum		
gi 1335799	urease subunit alpha [Ureaplasmaparvum serovar		
gi 1335799	urease complex component[Ureaplasma parvum		
gi 1335799	urease complex component[Ureaplasma parvum		
gi 1335799	ferrichrome transport ATP-bindingprotein [Ureap	NP_109882.1	cobalt transport ATP-binding protein [Myc
gi 1335799	hemolysin [Ureaplasma parvumserovar 3 str. ATC		
gi 1335800	hypothetical protein UU437[Ureaplasma parvum	NP_110226.1	UV protection protein MucB [Mycoplasma
gi 1335800	holliday junction DNA helicase(fragment) [Ureapl	NP_110225.1	Holliday junction DNA helicase RuvB [Myc

ウレアーゼは，*Ureaplasma*ゲノムにだけコードされていることがわかる

本日の課題

- Ureaplasma.faa には, *Ureaplasma parvum*のゲノムにコードされる全タンパク質がMulti-FASTAフォーマットで記述してあります
- 「Mpneumoniae.faa」をデータベース, 「Ureaplasma.faa」を質問配列にしてBLAST検索を行い, *Ureaplasma*のタンパク質と相同なものが*M. pneumoniae*ゲノム上にもあるかどうか調べてください (E-valueのしきい値は, $1e-3$ に設定してください)
- parse-blast7.plを使って, ヒットしたアミノ酸配列のリストを作成してください
- 作成したエクセルファイルを提出してください

- 作成した**エクセルファイル**を、メールに添付して提出してください
- 送付先は「kenro@hosei.ac.jp」です
- メールのはじめの件名は「**BLAST課題**」にしてください
- メール本文に、以下のように「氏名」「所属」「学生証番号」「今日の講義の感想」を記載してください

氏名：○○ ○○

所属：××××専攻 △△△△研究室

学生証番号：□□□□□

講義の感想：