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

法政大学

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

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、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）

講義日程

講師：大島研郎

- 2022_生物配列解析基礎_1回目_資料.pdf
- kiso1
- Mgenitalium.faa
- Mpneumoniae.faa
- parse-blast.py
- 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-blast.py

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

※ ユーザ名に日本語が使われているときなどに、デスクトップではうまく作動しない場合があります。その場合は、blastフォルダを C:ドライブの直下に移動させてみてください。

BLAST （stand-alone BLAST）のインストール

- 「講義で使用予定のソフトウェア」の中からBLASTにアクセスします.

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

Parent Directory
ChangeLog
ncbi-blast-2.11.0+-1.src.rpm
ncbi-blast-2.11.0+-1.src.rpm.md5
ncbi-blast-2.11.0+-1.x86_64.rpm
ncbi-blast-2.11.0+-1.x86_64.rpm.md5
ncbi-blast-2.11.0+-src.tar.gz
ncbi-blast-2.11.0+-src.tar.gz.md5
ncbi-blast-2.11.0+-src.zip
ncbi-blast-2.11.0+-src.zip.md5
ncbi-blast-2.11.0+-win64.exe
ncbi-blast-2.11.0+-win64.exe.md5
ncbi-blast-2.11.0+-x64-linux.tar.gz
ncbi-blast-2.11.0+-x64-linux.tar.gz.md5
ncbi-blast-2.11.0+-x64-macosx.tar.gz
ncbi-blast-2.11.0+-x64-macosx.tar.gz.md5
ncbi-blast-2.11.0+-x64-win64.tar.gz
ncbi-blast-2.11.0+-x64-win64.tar.gz.md5
ncbi-blast-2.11.0+.dmg
ncbi-blast-2.11.0+.dmg.md5

2

Windowsの場合は,
２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

※ PCのバージョンによっては古いソフトをインストールしないとうまく動かないことがあります

BLAST がインストールされているかどうかの確認方法

- コマンドプロンプト（またはWindows Terminal）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

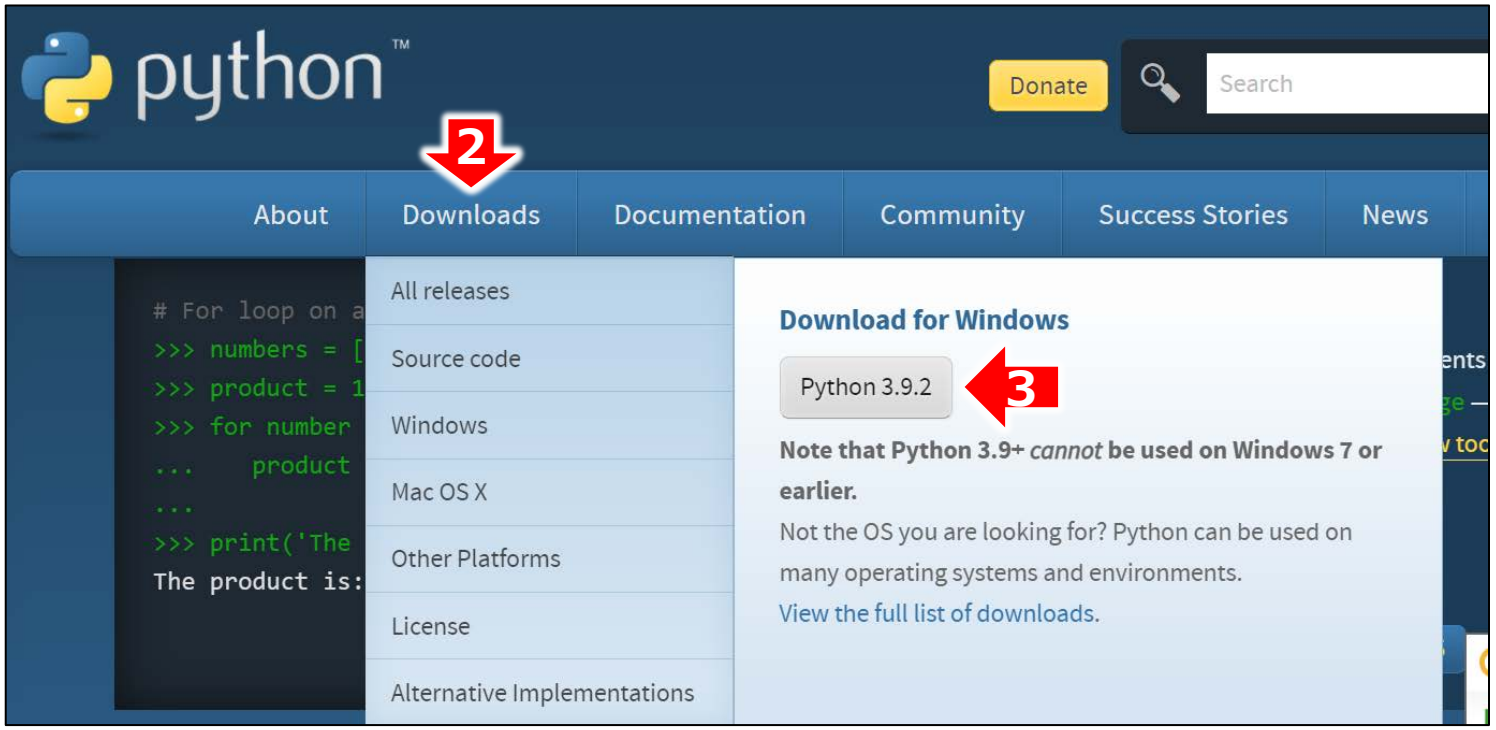
```
> blastp -help
```

BLASTについての説明が表示されれば，OKです

Python のインストール

- 「講義で使用予定のソフトウェア」の中からPythonにアクセスします.

https://www.python.org/



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

Python がインストールされているかどうかの確認方法

- コマンドプロンプト（またはterminal）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

スペースが入ります

- 「python -help」と入力して，リターン

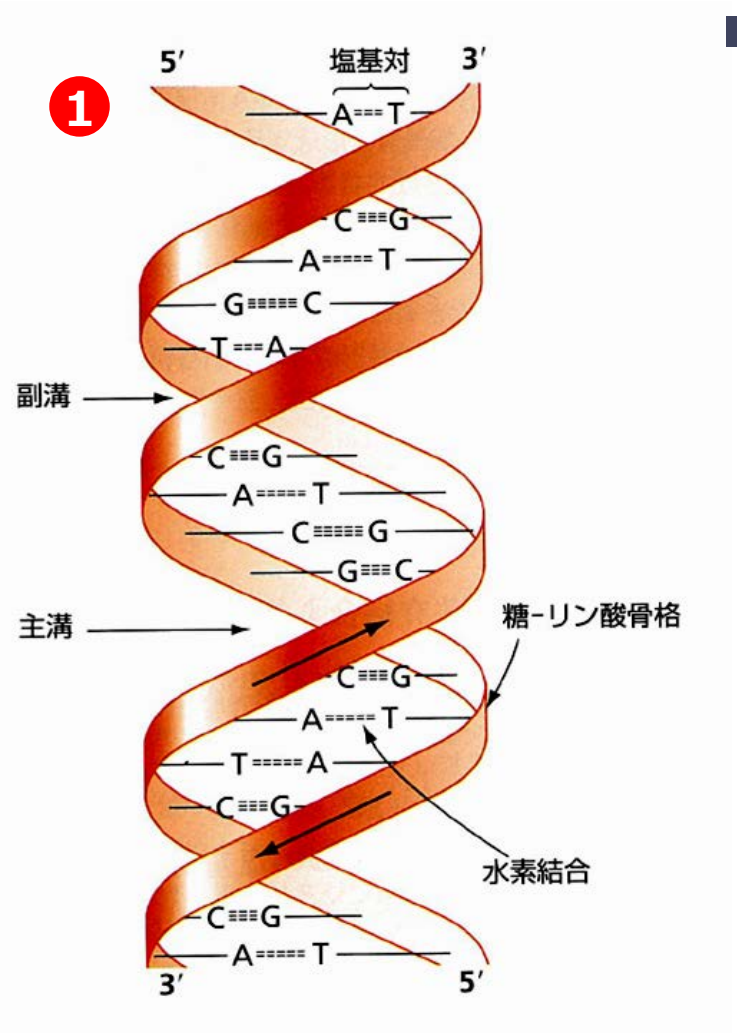
```
> python -help
```

Python についての説明が表示されれば，OKです

```
C:¥Users¥kenro>python -help
usage: python [option] ... [-c cmd | -m mod | file | -] [arg] ...
Options and arguments (and corresponding environment variables):
-b     : issue warnings about str(bytes_instance), str(bytearray_instance)
         and comparing bytes/bytearray with str. (-bb: issue errors)
```

生物配列　＝　塩基配列、およびアミノ酸配列

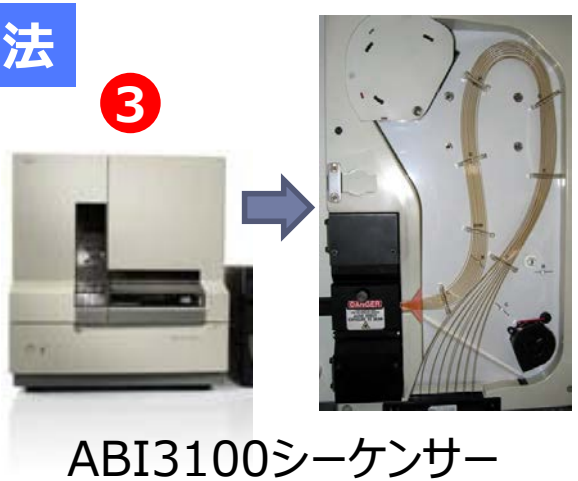
塩基配列　＝　DNAの塩基（ＧＡＴＣ）の並び順



■ 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



ABI3100シーケンサー

次世代シーケンサー

Ion PGM

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/

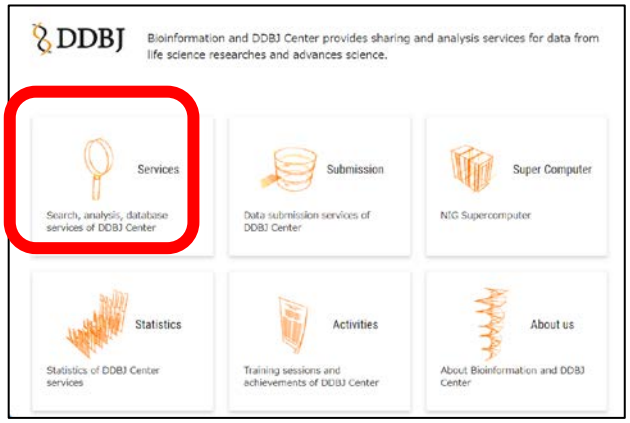
EMBL（European Bioinformatics Institute）
http://www.ebi.ac.uk/embl/index.html

- 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



AGD
database DDBJ
AGD is a controlled-access database for sharing individual-level genotype and phenotype information among specific researchers.

AOE
search DBCLS
Statistics and trends of gene expression data

ARSA
search DDBJ
DDBJ annotated/assembled data retrieval by accession numbers and keywords
HELP Web API

2 データベース検索

BioProject
database submission DDBJ
A BioProject is a collection of biological data related to a single project. A BioProject record provides links to the diverse data types generated for that project.

BioSample
database submission DDBJ
The BioSample database contains descriptions of biological source materials used in experimental assays.

CRISPRdirect
search DBCLS
Designing CRISPR/Cas9 guide RNA with reduced off-target sites

DBCLS SRA
search DBCLS
Statistics and trends of SRA data

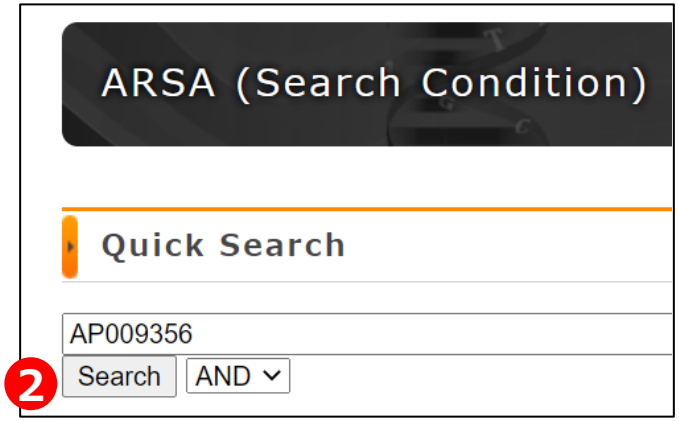
DDBJ 塩基配列の登録
database submission DDBJ
An annotated collection of genome, gene and transcript sequences.

DDBJ Search
search DDBJ
Search INSDC BioProject/BioSample/SRA and JGA data by accession numbers and keywords

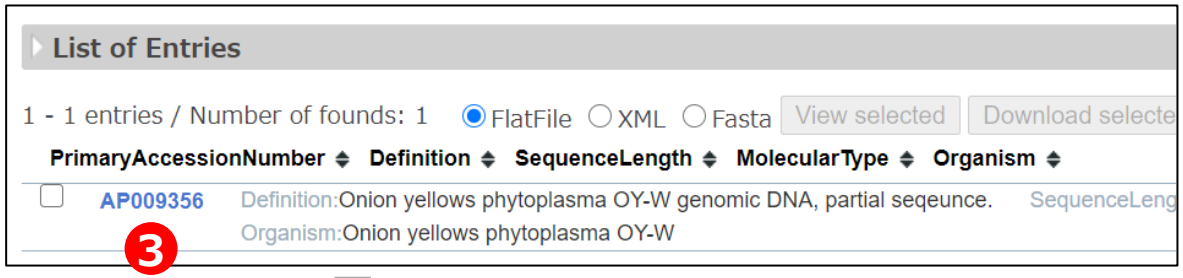
DDBJ-LD
database DDBJ
Linked data of DDBJ Center

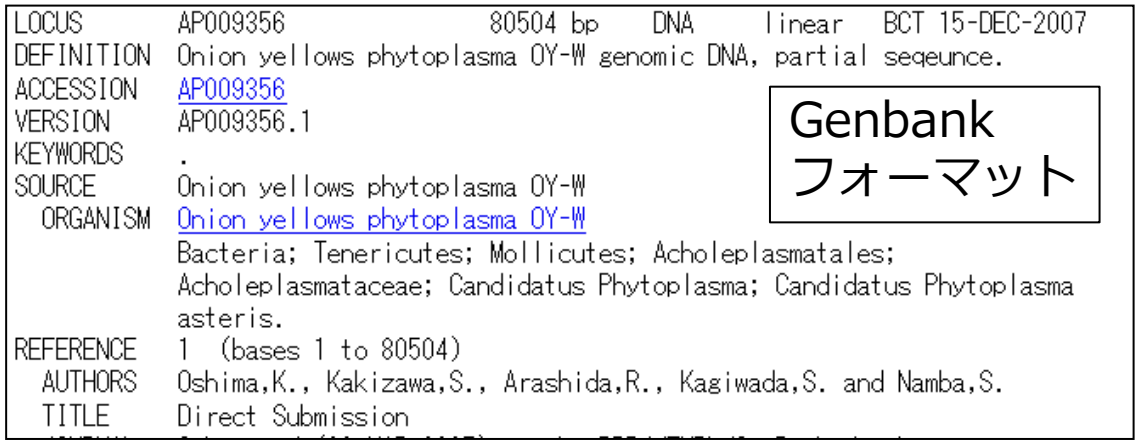
DFAST アノテーション
analysis submission annotation DDBJ
DFAST is an automatic annotation service for prokaryotic genomes. DFAST generates an annotation file submittable to DDBJ.
HELP

DRA
database submission DDBJ
DRA stores raw sequencing data and alignment information generated from high throughput sequencing platforms.



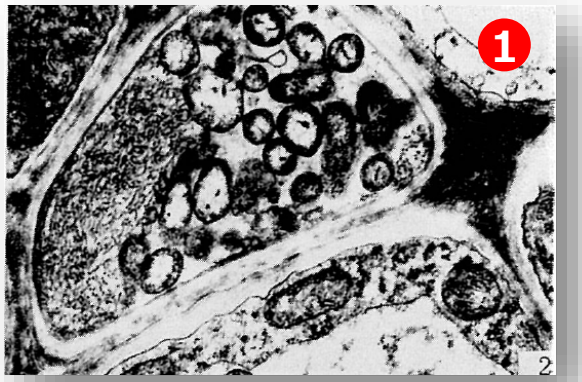
← AP009356 と入力





```
LOCUS       AP009356                80504 bp    DNA     linear   BCT 15-DEC-2007
DEFINITION  Onion yellows phytoplasma OY-W genomic DNA, partial seqeunce.
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
```

Genbank
フォーマット



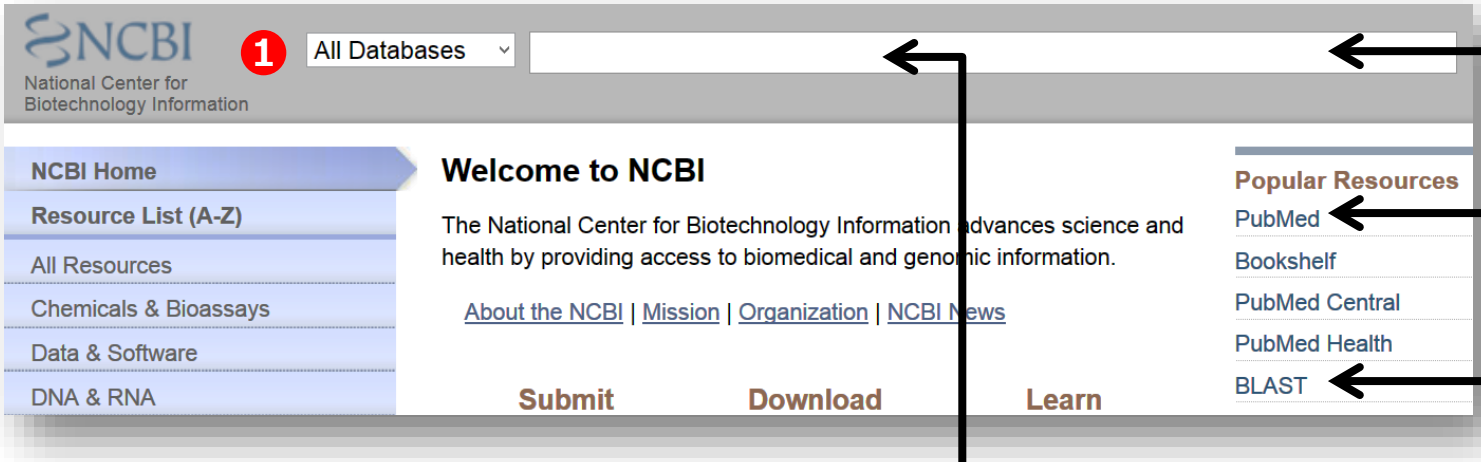
ファイトプラズマ

- 植物の師部細胞に寄生する**植物病原細菌**
- 感染植物では、がくや花弁が**葉化**する

National Center for Biotechnology Information　通称：NCBI

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

- 米国の国立衛生研究所 (**NIH**) 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がまとめられている



All Databases
横断検索

PubMed
文献検索 4

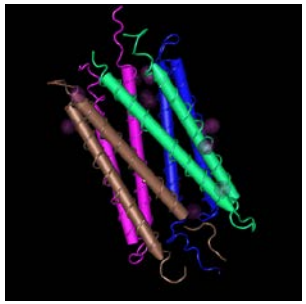
BLAST
相同性検索 5

試しに「phytoplasma」と入力してみる

各データベースの該当件数が表示される

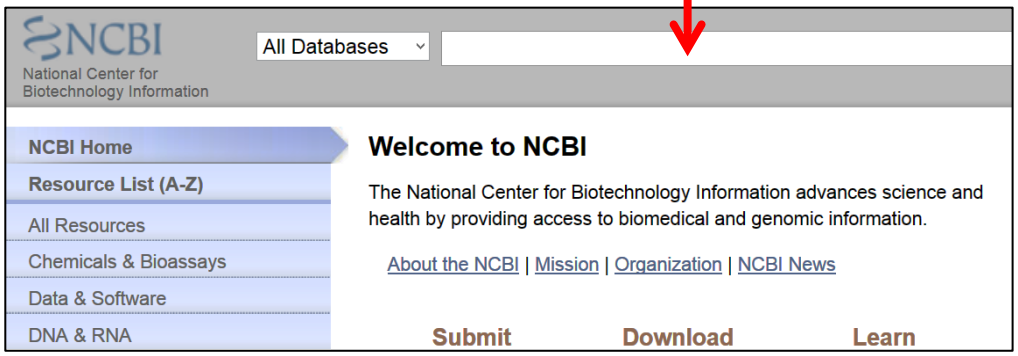
Protein	43,816
Protein Clusters	0
Sparcle	4
Structure 2	4

3
PHYL1の立体構造



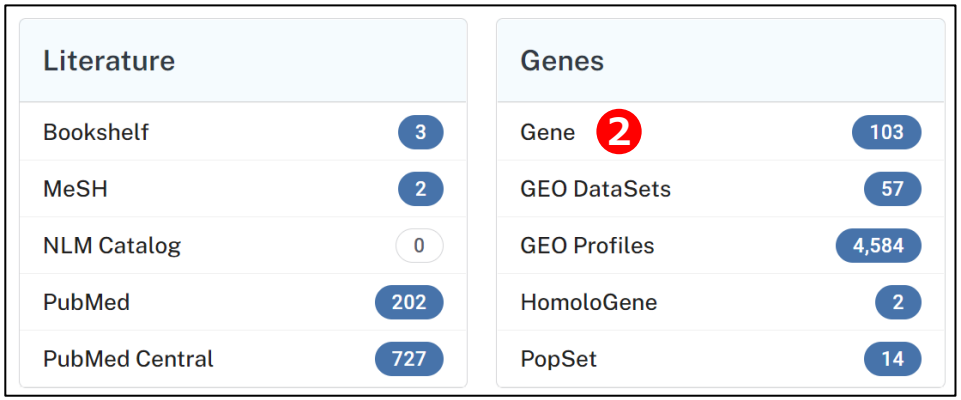
Nucleotide
　塩基配列データベース
Gene
　遺伝子データベース
Genome
　ゲノムデータベース
GEO DataSets
　遺伝子発現データベース

❶「SEP3」と入力して検索する



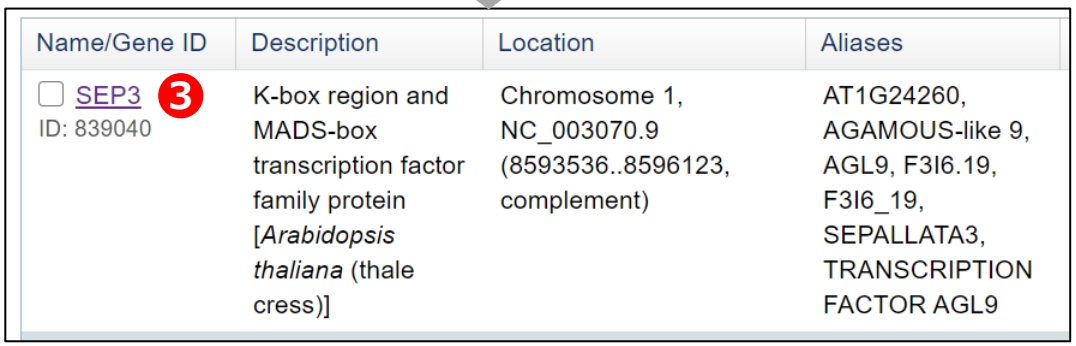
【SEP3】

花の形態形成に関わる転写因子



【Gene】

遺伝子のデータベース



シロイヌナズナ（*Arabidopsis thaliana*）のSEP3を選択する

SEP3 K-box region and MADS-box transcription factor family protein [*Arabidopsis thaliana* (thale cress)]

Download Datasets

Gene ID: 839040, updated on 6-Mar-2021

Summary

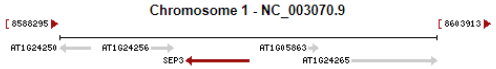
Gene symbol	SEP3
Gene description	K-box region and MADS-box transcription factor family protein
Primary source	Araport:AT1G24260
Locus tag	AT1G24260
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Arabidopsis thaliana (ecotype: Columbia)
Lineage	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis
Also known as	AGAMOUS-like 9; AGL9; F3I6.19; F3I6_19; SEPALLATA3; TRANSCRIPTION FACTOR AGL9
Summary	Member of the MADs box transcription factor family. SEP3 is redundant with SEP1 and 2. Flowers of SEP1/2/3 triple mutants show a conversion of petals and stamens to sepals.SEP3 forms heterotetrameric complexes with other MADS box family members and binds to the CArG box motif.

NEW Try the new Gene table
Try the new Transcript table

周辺の遺伝子マップや、遺伝子産物の機能についての情報が得られる

Genomic context

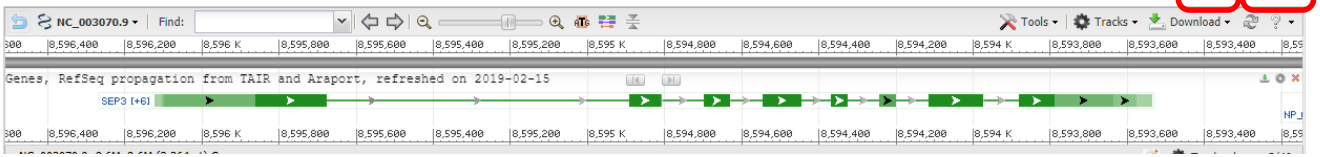
Location: chromosome: 1 — See SEP3 in Genome Data Viewer
Exon count: 8
Sequence: Chromosome: 1; NC_003070.9 (8593536..8596123, complement)



Genomic regions, transcripts, and products

Genomic Sequence: NC_003070.9 — Go to reference sequence details

Go to nucleotide: Graphics FASTA GenBank



FASTAフォーマット

Arabidopsis thaliana chromosome 1 sequence

NCBI Reference Sequence: NC_003070.9

GenBank Graphics

```
>NC_003070.9:c8596123-8593536 Arabidopsis thaliana chromosome 1 sequence
TCTGAGAGTATATTAGAAAGAGAATATTTCAAGTAATGAAGCTGACATGTTTATATGTACTTTGAGAGAA
GTGTTGTGAGATTTGTACAAATGTATATGTACACTTTAAAAAGCAATATAAGATAGATAAAAAAAATATA
AAGAAAAAAAGAAAGAAAGAAAGAAAGAAAGAGAGAGGCTCATATATATATAGAATTGCTTGCAAGGAAA
GAGAGAGAGAGAGATTGAGATATCTTTTGGGAGAGGAGAAAGAAAAAGAAAATGGGAAGAGGGAGAGTAG
AATTGAAGAGGATAGAGAACAAGATCAATAGGCAAGTGACGTTTGCAAAGAGAAGGAATGGTCTTTTGAA
GAAAGCATACGAGCTTTCAGTTCTATGTGATGCAGAAGTTGCTCTCATCATCTTCTCAAATAGAGGAAAG
CTGTACGAGTTTTGCAGTAGTTCGAGGTATATATCTACTTTTGTATATATATTACTTATAACATAAACAT
TTTATATACATATTAAGTAACACAAAAATGTCTTGTATGTATGGGTCTCTCTGTGATGTGTTGTTGTGTC
GTACGTACGTGTTCTATCATATCCTTTTAAAAGAAGCAAAGAGGAAAAAAAATTTGGGATACCCCAAATC
```

GenBankフォーマット

Arabidopsis thaliana chromosome 1 sequence

NCBI Reference Sequence: NC_003070.9

FASTA Graphics

```
LOCUS       NC_003070             2588 bp    DNA     linear   CON 14-FEB-
DEFINITION  Arabidopsis thaliana chromosome 1 sequence.
ACCESSION   NC_003070 REGION: complement(8593536..8596123)
VERSION     NC_003070.9
DBLINK      BioProject: PRJNA116
            BioSample: SAMN03081427
            Assembly: GCF_000001735.4
KEYWORDS    RefSeq.
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
```

データベースカタログ

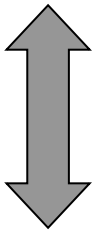
http://integbio.jp/dbcatalog/?lang=ja

相同性検索　（ホモロジー検索）

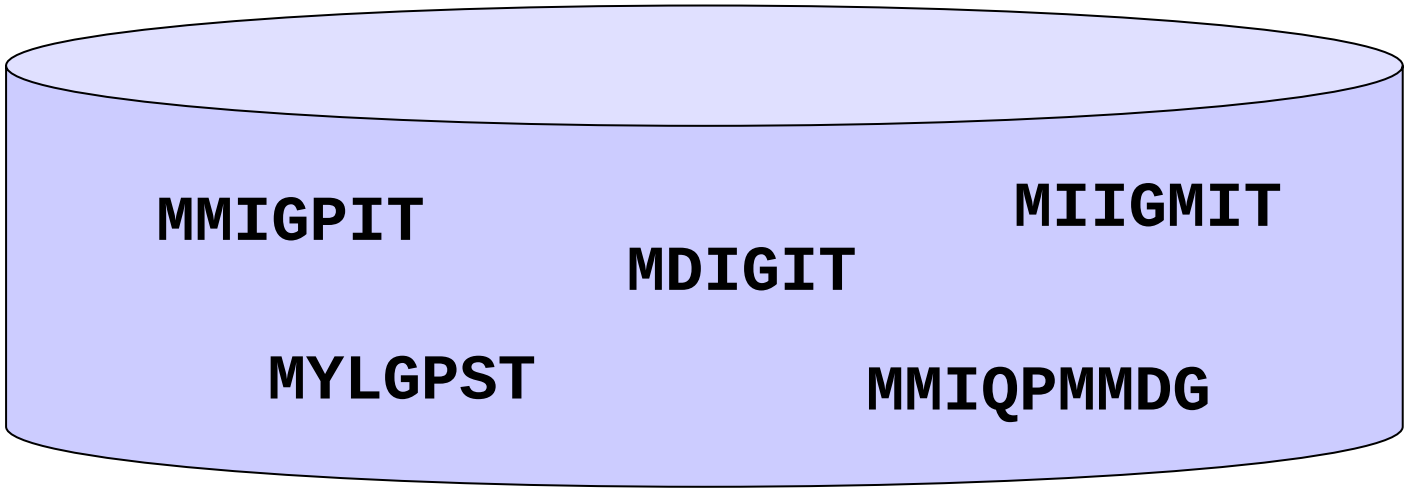
- 相同性検索は，配列の類似性から類縁の遺伝子・タンパク質を検索する方法で，進化・系統分類の解析，機能解析などを目的とした配列解析の最も基本的な手法の一つである.

質問配列
（Query）

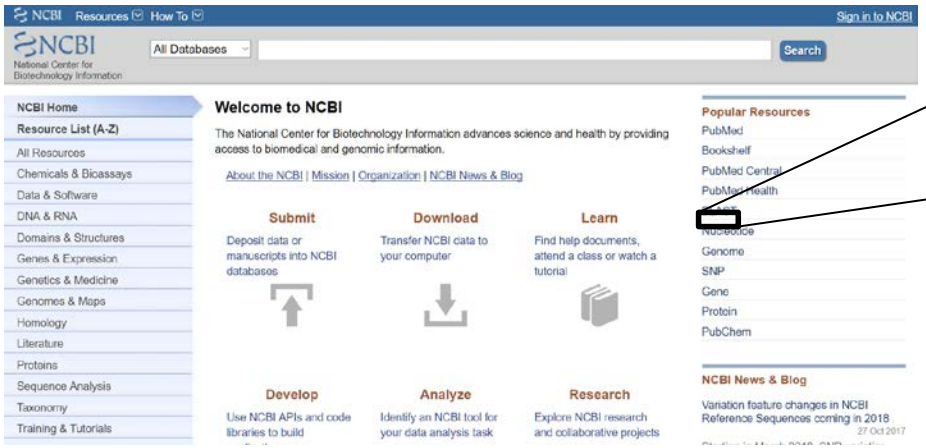
MIGMMIT

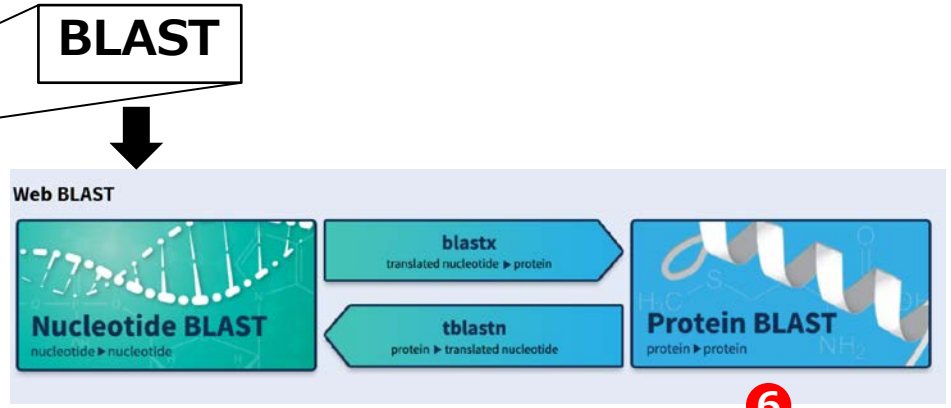


質問配列と**類似した（相同な）配列**を，
データベース上から探索する



BLAST検索のwebページ　https://blast.ncbi.nlm.nih.gov/Blast.cgi



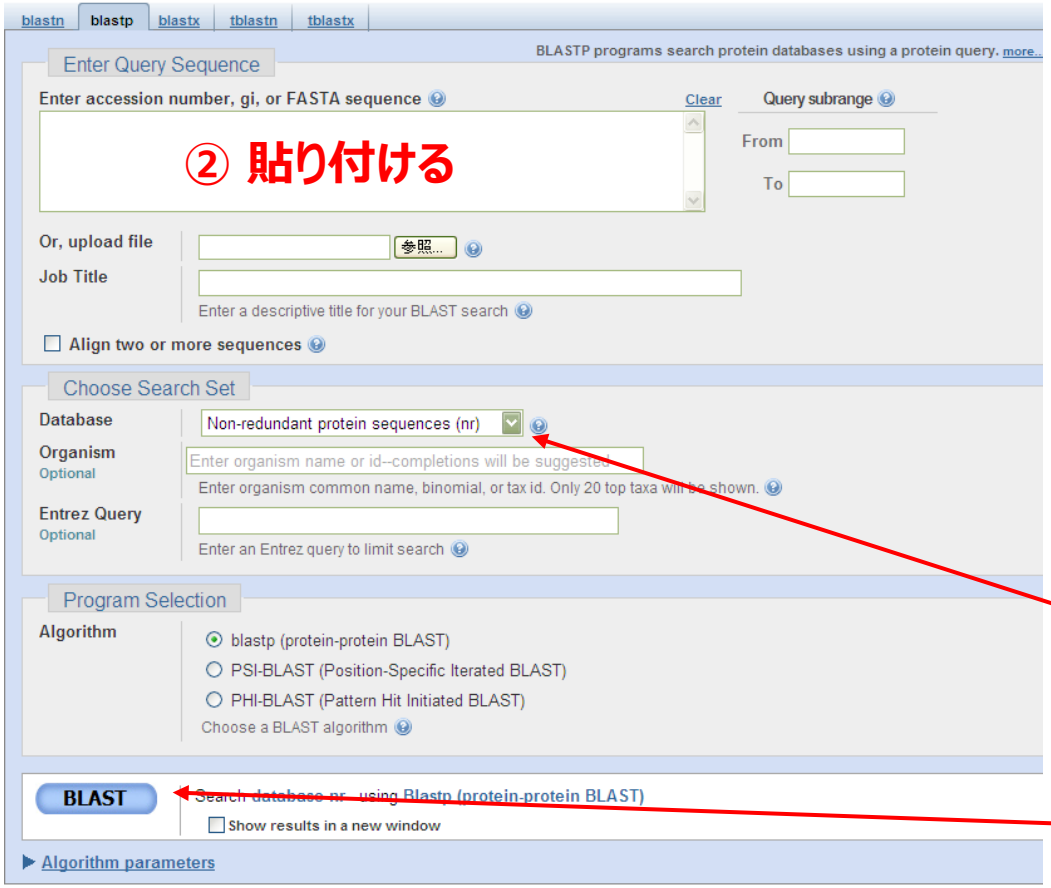


BLAST検索には5種類がある

	プログラム名	質問配列（query）	検索対象
❶	protein blast	アミノ酸配列	アミノ酸配列データベース
❷	blastx	塩基配列	アミノ酸配列データベース
❸	nucleotide blast	塩基配列	塩基配列データベース
❹	tblastn	アミノ酸配列	塩基配列データベース
❺	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




```
>sample1
MNRVFLFGKLSFTPNRLQTKNGTLGATFSMECLDS
SGFNNAKSFIRVTAWGKVASFIVAQNPGVMLFVEG
RLTTYKITNSENKNTYALQVTADKIFHPDEKTTNE
EPIKSTVVDSPFMNPKASVTEAEFEQAFPHQDETD
FNNITPIFENDVQLEEESDD
```

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）

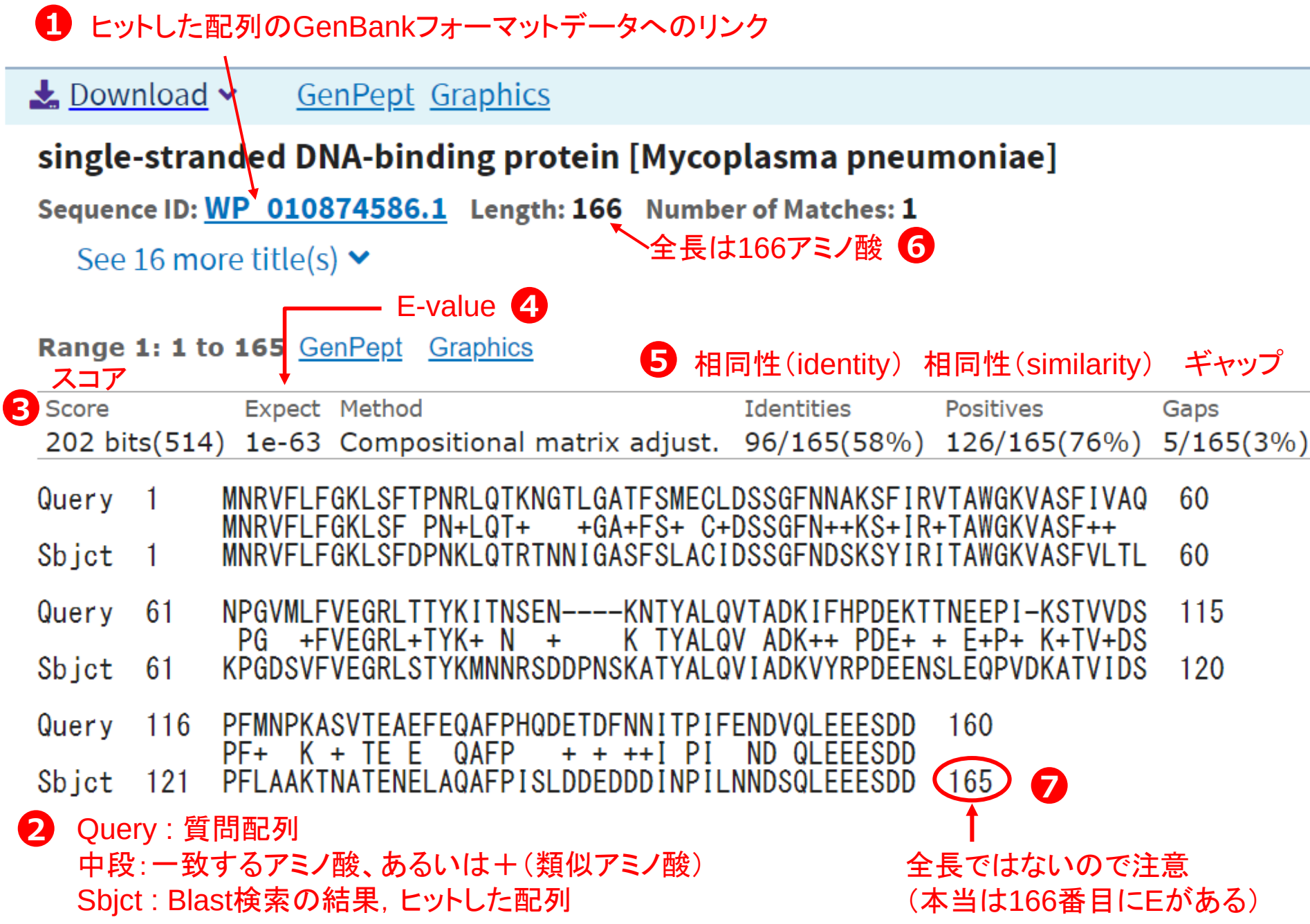
❺ Alignmentsのタブをクリック

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show

select all *100 sequences selected* GenPept Graphic ❷ Distance t ❸ of res ❹ M

❶ Description	Max Score	Total Score	Query Cover	**E value**	Per. Ident
single-stranded DNA-binding protein [Mycoplasma genitalium]	330	330	100%	3e-114	100.00%
single-stranded DNA-binding protein [Mycoplasma pneumoniae]	202	202	100%	1e-63	58.18%
puative 19 kDa protein [Mycoplasma pneumoniae]	70.1	70.1	43%	9e-13	50.00%
single-stranded DNA-binding protein [Brevibacillus borstelensis]	57.4	57.4	75%	4e-07	31.54%
single-stranded DNA-binding protein [Firmicutes bacterium CAG:170]	57.0	57.0	83%	4e-07	25.36%
single-stranded DNA-binding protein [Phorcysia thermohydrogeniphila]	56.2	56.2	68%	6e-07	32.17%
single-stranded DNA-binding protein [Candidatus Colwellbacteria bacterium RIFCSPHIGHO2_02_FULL_43_15]	56.2	56.2	72%	1e-06	33.33%
single-stranded DNA-binding protein [Veillonella seminalis]	54.7	54.7	61%	2e-06	33.98%
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
E-value 4
Range 1: 1 to 165 GenPept Graphics
5 相同性（identity） 相同性（similarity） ギャップ
スコア
3
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
MNRVFLFGKLSF PN+LQT+ +GA+FS+ C+DSSGFN++KS+IR+TAWGKVASF++
Sbjct 1 MNRVFLFGKLSFDPNKLQTRTNNIGASFSLACIDSSGFNDSKSYIRITAWGKVASFVLTL 60
Query 61 NPGVMLFVEGRLTTYKITNSEN----KNTYALQVTADKIFHPDEKTTNEEPI-KSTVVDS 115
PG +FVEGRL+TYK+ N + K TYALQV ADK++ PDE+ + E+P+ K+TV+DS
Sbjct 61 KPGDSVFVEGRLSTYKMNNRSDDPNSKATYALQVIADKVYRPDEENSLEQPVDKATVIDS 120
Query 116 PFMNPKASVTEAEFEQAFPHQDETDFNNITPIFENDVQLEEESDD 160
PF+ K + TE E QAFP + + ++I PI ND QLEEESDD
Sbjct 121 PFLAAKTNATENELAQAFPISLDDEDDDINPILNNDSQLEEESDD 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）

相同性検索のアルゴリズム

（BLASTのサーバの中で行われている手順）

❶ 質問配列とデータベース上の各配列との**アラインメント**を作成する

```
MIGMMIT  ⇒  MI-GMMIT
MIIGMIT     MIIGM-IT

MIGMMIT  ⇒  M-IGMMIT
MDIGIT      MDIG--IT

MIGMMIT  ⇒  M-IGMMIT
MMIGPIT     MMIGP-IT
```

❷ ２つの配列がどのくらい似ているか、アラインメントに点数（**スコア**）をつける

```
MI-GMMIT  ⇒ 21
MIIGM-IT

M-IGMMIT  ⇒ 6
MDIG--IT

M-IGMMIT  ⇒ 11
MMIGP-IT
```

❸ 点数（スコア）の順番に並べて表示する

配列	スコア
MIIGMIT	21
MMIGPIT	11
MDIGIT	6

アラインメント

❶ 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（-）で対応づける

❷ 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
```

❸ 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

❶
```
AFDC    s(A, A) + s(F, E) + s(D, E) + s(C, C) = 8
AEEC       3         -7        3         9
```

❷
```
AFDGC   s(A, A) + s(F, E) + s(D, E) +  gap  + s(C, C) = 0
AEE-C      3         -7        3        -8        9
```

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　BLOSUM: BLOcks amino acid Substitution Matrix

- BLAST検索に使用される**スコア表**
- **同一アミノ酸**や**類似アミノ酸**が縦に並んだ時には高い点数が付くようになっている
- 例えば、アラインメント上でRとKが並んだ場合は、3点が加えられる

【練習】

以下のアラインメントのスコアを計算しなさい。

ただし、ギャップ＝－８とする。

```
MIGMM-IT
M-GMIGIT
```

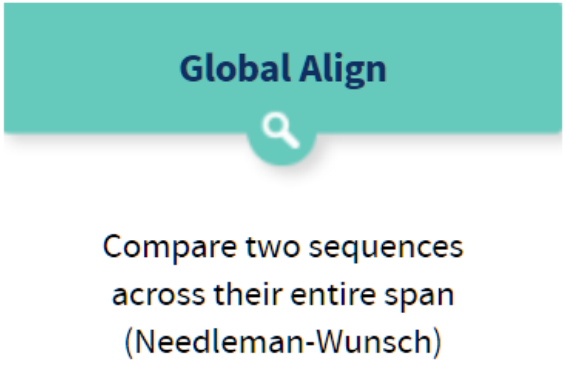
❶ BLOSUM50マトリックス

	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

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



- 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

❶ **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

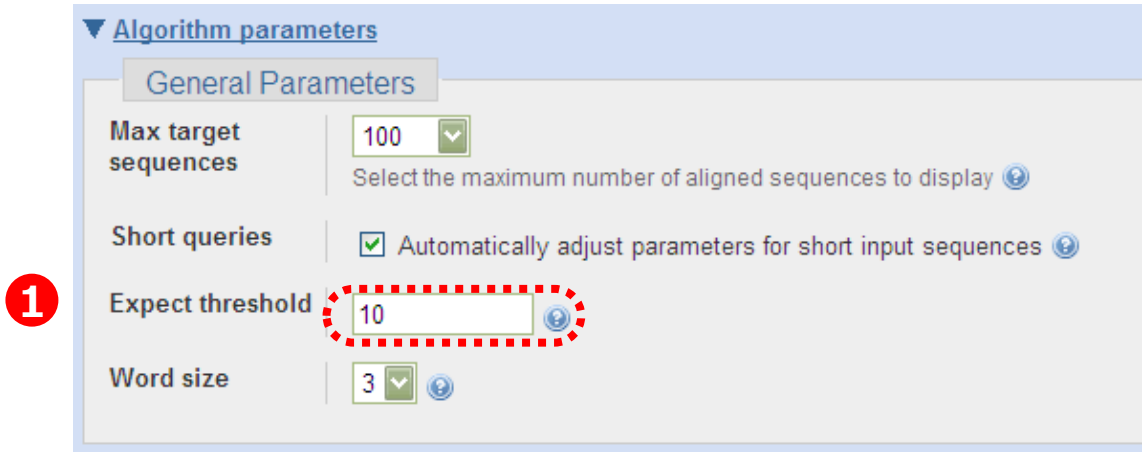
❷ **MSGPVFGIP…**

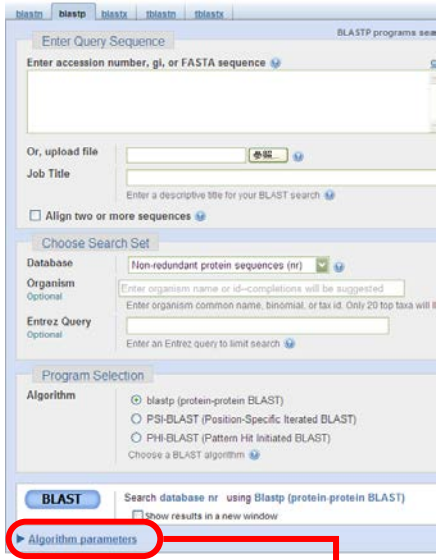
一致したワードを中心にして両方向にアラインメントを伸ばしていく
（類似度が下がってきたらアラインメントを終了する）

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





❶

Algorithm parameters

General Parameters

Max target sequences: 100 — 検索結果の表示件数 ❷
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 ❸

Word size: 3 — BLAST検索時のWordサイズ ❹

Scoring Parameters

Matrix: BLOSUM62 — マトリックスの種類を選ぶ ❺

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

Compositional adjustments: Conditional compositional score matrix adjustment
❼ E-value計算時の設定

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 ❽
☐ 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

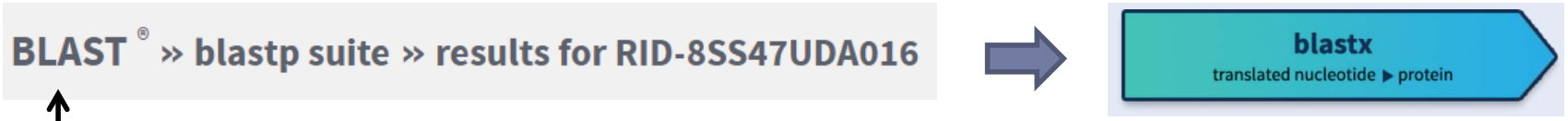
塩基配列を入力

⇩

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

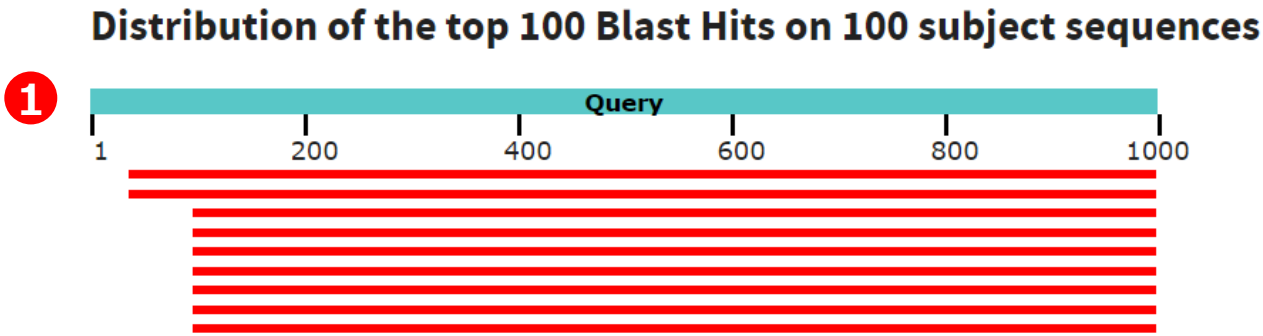
・**non-coding領域**に，タンパク質がコードされていないかどうかを調べたいときなど



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
AATTAGAGAAAACAACAGAGTTGTTATTTCTAGTGATGTTCTTGTTAACAACTAAAC
GAACAATGTTTGTTTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATC
TTACAACCAGAACTCAATTACCCCCTGCATACACTAATTCTTTCACACGTGGTGTTT
ATTACCCTGACAAAGTTTTCAGATCCTCAGTTTTACATTCAACTCAG･･･
```



surface glycoproteinがコードされていることがわかる

blastn (nucleotide blast)

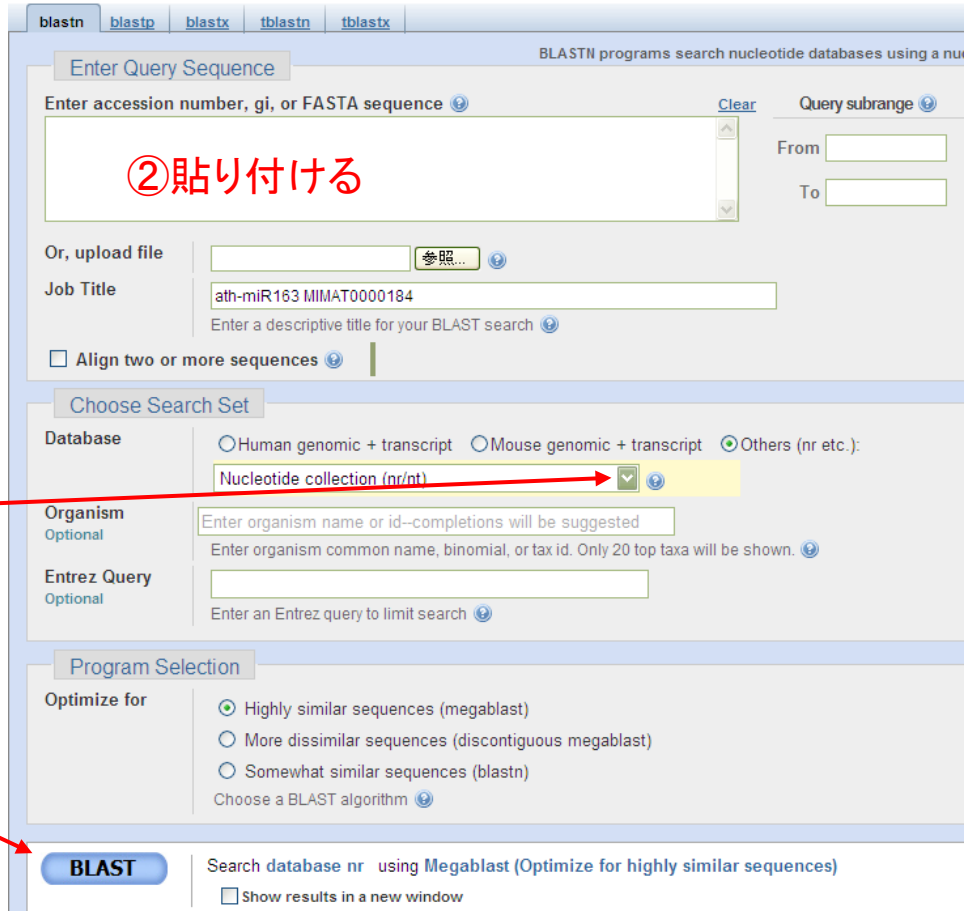
上部のBLASTをクリックし、blastn のページへ

>sample3
TTGAAGAGGACTTGGAACTTCGAT

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



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

⇩

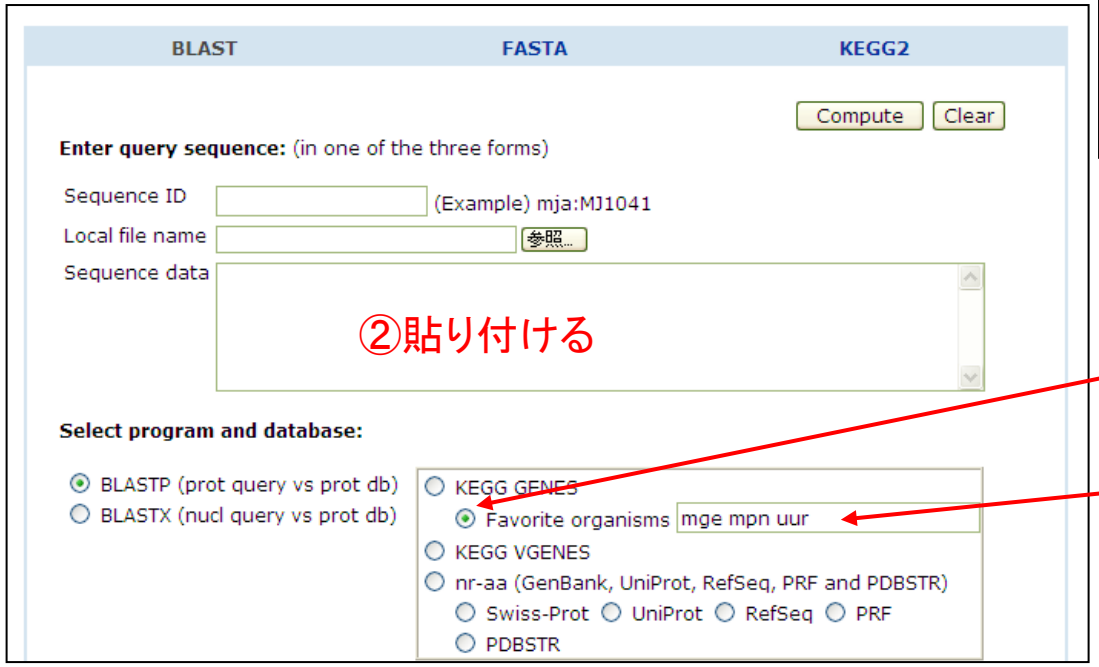
データベース上の塩基配列も，６通りのreading frameのすべてについて翻訳し，このアミノ酸配列データに対して検索

・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
MDENETQFNKLNQVKNKLKIGVFGIGGAGNNIVDASLYHYPN
LASENIHFYAINSDLQHLAFKTNVKNKLLIQDHTNKGFGAGG
DPAKGASLAISFQEQFNTLTDGYDFCILVAGFGKGTGTGATP
VFSKILKTKKILNVAIVTYPSLNEGLTVRNKATKGLEILNKA
TDSYMLFCNEKCTNGIYQLANTEIVSAIKNLIELITIPLQQN
IDFEDVRAFFQTKKTNQDQQLFTVTHPFSFSFDSKDSIEQFA
KQFKNFEKVSYFDHSIVGAKKVLLKANINQKIVKLNFKQIQD
IIWTKIDNYQLEIRLGVDFVTTIPNIQIFILSEHKNPVSLPI
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）

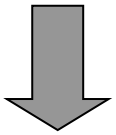


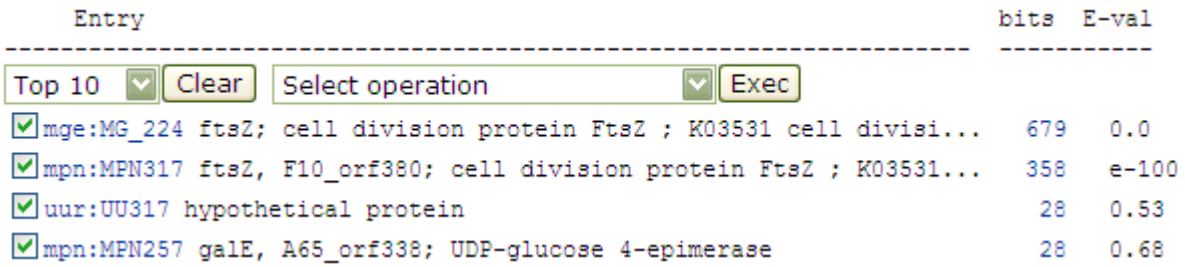
③Favorite organisms を選択

④「mge mpn uur」と入力

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

⑤「Compute」を押す





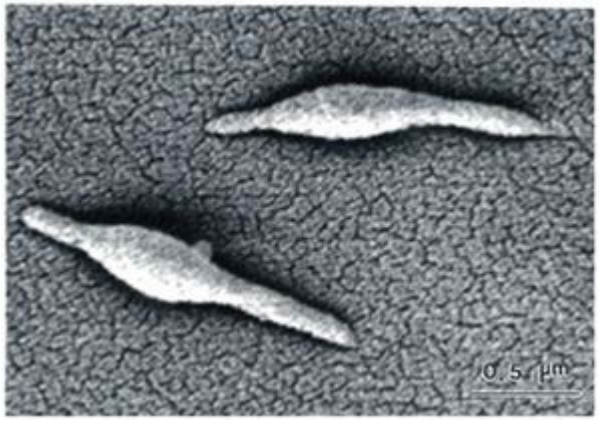
*Ureaplasma*は，*ftsZ*を持っていないことがわかる

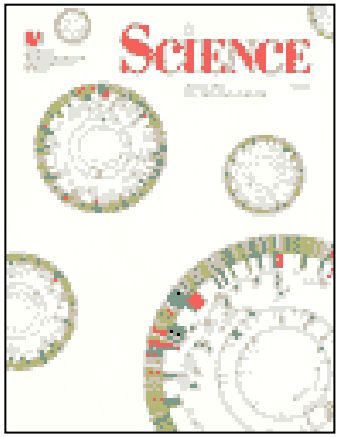
- **大量の質問配列**についてBLAST検索を行いたい
- 自分の持っている**未公開のデータ**に対して検索したい
- ホモロジー検索を用いて**比較ゲノム解析**を行いたい

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

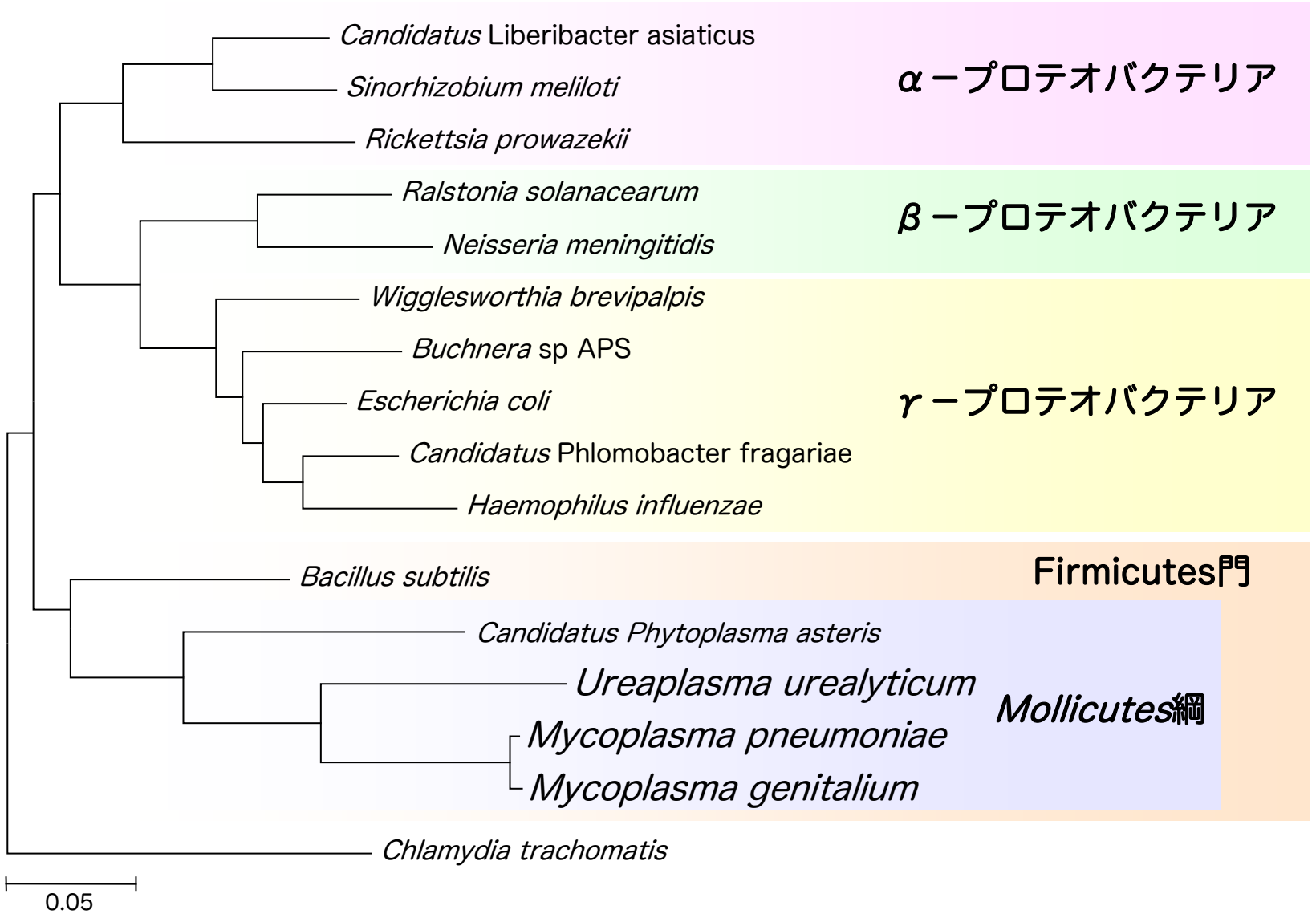
生物種	ゲノムサイズ（Mbp）	全ゲノム解読された年
Haemophilus influenzae	1.83	1995
★ *Mycoplasma genitalium*	0.58	1995
★ *Mycoplasma pneumoniae*	0.82	1996
⋮		
Bacillus subtilis	4.21	1997
Escherichia coli	4.67	1997
⋮		
★ *Ureaplasma parvum*	0.75	2000
⋮		





◆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**のファイルがおいてあります
- 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

データが塩基配列の場合は nucl

上記の方法でエラーが出る場合は、以下の方法を試してみてください

```
makeblastdb -in Mgenitalium.faa -dbtype prot –max_file_sz 1000000
```

※ ユーザ名に日本語が使われているときなどに、デスクトップではうまく作動しない場合があります。その場合は、blastフォルダを C:ドライブの直下に移動させてみてください。

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]
MANELTWHDVLAEEKQQPYFLNTLQTVASERQSGVTIYPPQKDVFNAFRFTELG
DVKVVILGQDPYHGPGQAHGLAFSVRPGIAIPPSLLNMYKELENTIPGFTRPNH
GYLESWARQGVLLLNTVLTVRAGQAHSHASLGWETFTDKVISLINQHREGVVFL
LWGSHAQKKGAIIDKQRHHVLKAPHPSPLSAHRGFFGCNHFVLANQWLEQRGET
PIDWMPVLPAESE
```

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を質問配列として用いて、Mgenitalium.faaデータベースに対してblastp検索を行うために、以下のコマンドを実行します

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

❶ ❷

-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

※ 講義資料で改行するところにはスペースが入ります

```
> 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 →

スコア

E value

```
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

                                                                  Score       E
Sequences producing significant alignments:                      (Bits)     Value

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    TWHDVLAEEKQQPYFLNTLQTVASERQSGVTIYPPQKDVFNAFRFTELGDVKVVILGQDP  65
            +W   + EE ++PYF   L+ + + +    TI P  + +F  F F +  D KV+I GQDP
Sbjct  17   SWRAFIDEEVKKPYFQALLEKLKALK---ATIIPKPELIFRVFSFFKPIDTKVIIFGQDP  73

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

E value設定

- E-valueが小さいほど、配列同士の相同性が高いことを示します
- 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 **l（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
  –evalue 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
PPGVKLEKLTQLISAAQKQGIRCIVDSSGEALSAALAIGNIELVKPNQKELSALVNRELTQPDDVRKAAQ
EIVNSGKAKRVVVSLGPQGALGVDSENCIQVVPPPVKSQSTVGAGDSMVGAMTLKLAENASLEEMVRFGV
AAGSAATLNQGTRLCSHDDTQKIYAYLSR

>gi|16132212|ref|NP_418812.1| phosphoglyceromutase 2 [Escherichia coli K12]
MLQVYLVRHGETQWNAERRIQGQSDSPLTAKGEQQAMQVATRAKELGITHIISSDLGRTRRTAEIIAQAC
GCDIIFDSRLRELNMGVLEKRHIDSLTEEEENWRRQLVNGTVDGRIPEGESMQELSDRVNAALESCRDLP
QGSRPLLVSHGIALGCLVSTILGLPAWAERRLRLRNCSISRVDYQESLWLASGWVVETAGDISHLDAPAL
DELQR

>gi|16131851|ref|NP_418449.1| glucosephosphate isomerase [Escherichia coli K12]
MKNINPTQTAAWQALQKHFDEMKDVTIADLFAKDGDRFSKFSATFDDQMLVDYSKNRITEETLAKLQDLA
KECDLAGAIKSMFSGEKINRTENRAVLHVALRNRSNTPILVDGKDVMPEVNAVLEKMKTFSEAIISGEWK
GYTGKAITDVVNIGIGGSDLGPYMVTEALRPYKNHLNMHFVSNVDGTHIAEVLKKVNPETTLFLVASKTF
TTQETMTNAHSARDWFLKAAGDEKHVAKHFAALSTNAKAVGEFGIDTANMFEFWDWVGGRYSLWSAIGLS
IVLSIGFDNFVELLSGAHAMDKHFSTTPAEKNLPVLLALIGIWYNNFFGAETEAILPYDQYMHRFAAYFQ
QGNMESNGKYVDRNGNVVDYQTGPIIWGEPGTNGQHAFYQLIHQGTKMVPCDFIAPAITHNPLSDHHQKL
LSNFFAQTEALAFGKSREVVEQEYRDQGKDPATLDYVVPFKVFEGNRPTNSILLREITPFSLGALIALYE
HKIFTQGVILNIFTFDQWGVELGKQLANRILPELKDDKEISSHDSSTNGLINRYKAWRG
```

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
```

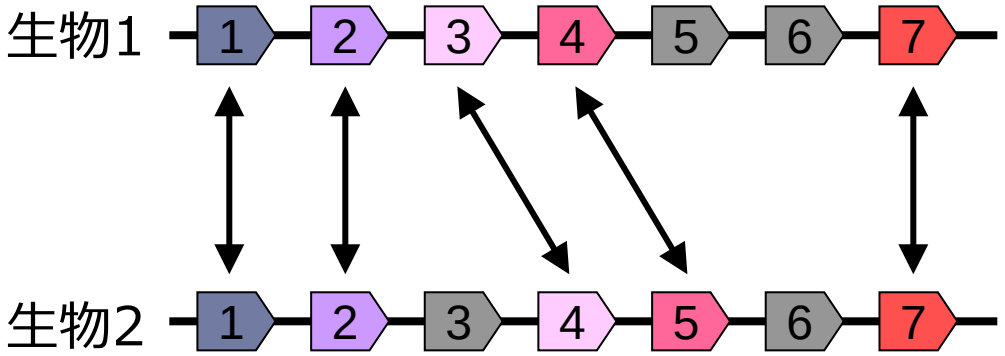
- これらと相同なアミノ酸配列がMgenitalium.faa内にあるかどうかを調べるために，以下のコマンドを実行してください

```
> 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
  –evalue 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

Python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
- **Python**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; Pgi1p	554	641	0	73%
16131757	NP_010335.1	triosephosphate isomerase; Tpi1p	248	192	4.00E-50	60%
16131754	NP_011756.1	phosphofructokinase alpha subunit; Pfk1p	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; Pgk1p	416	255	7.00E-69	57%
16130826	NP_012863.1	aldolase; Fba1p	359	352	4.00E-98	68%
16130686	NP_011770.1	enolase I; Eno1p	437	359	1.00E-100	62%
16130106	NP_009965.1	ribokinase; Rbk1p	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%

- ”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-blast.py** を用意しておきました．

※ Stand-alone BLASTのバージョンが古いと parse-blast.pyがうまく作動しないことがあります。その場合は、parse-blast2.pyで試してみてください

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

Length=380

                                                                  Score        E
Sequences producing significant alignments:                      (Bits)     Value

gi|12044851|ref|NP_072661.1| DNA polymerase III, subunit beta (dn...  525        0.0


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

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

Query  17   LNNVIVSNNKMKPYHSYLLIEATEKEINFYANNEYFSAKCTLAENIDVLEEGEVIVKGKI  76
            +NNVI+SNNK+KP+HSY LIEA EKEINFYANNEYFS KC L +NID+LE+G +IVKGKI
Sbjct  1    MNNVIISNNKIKPHHSYFLIEAKEKEINFYANNEYFSVKCNLNKNIDILEQGSLIVKGKI  60
```

- Python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て，result4.txtを構文解析します
  → list1.txt というファイルが新たに出来上がります

```
> python parse-blast.py -in result4.txt -out 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

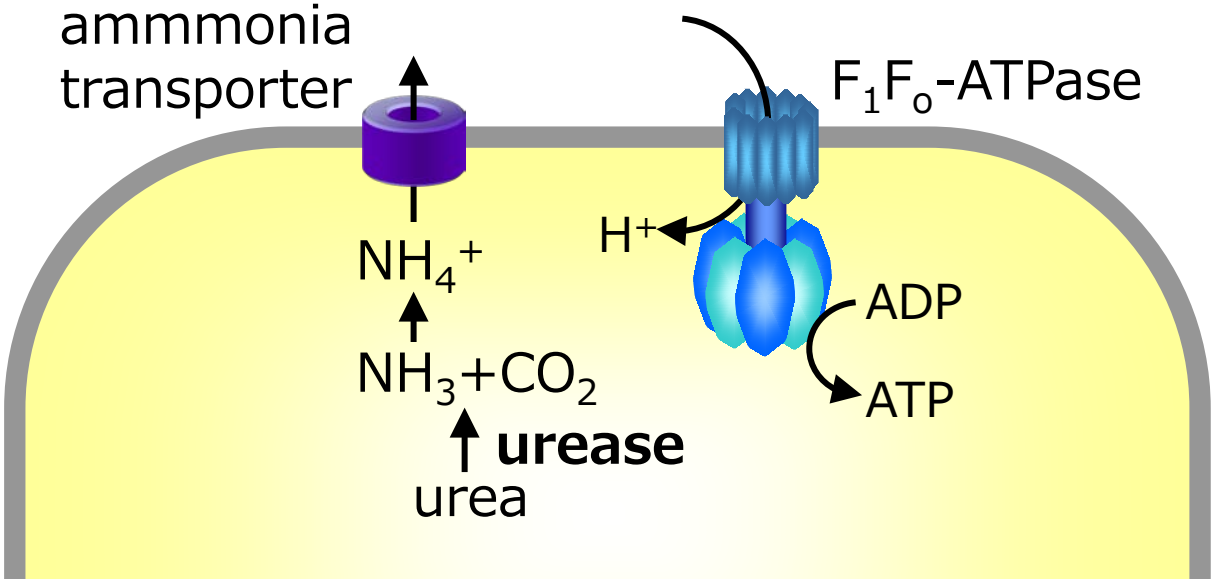
2 BLAST検索でヒットした配列の情報
（ヒットしなかった場合は空欄）

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 b	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 oxidation	NP_072668.1	thiophene and furan oxidation	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-blast.py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[at]hosei.ac.jp です（[at]を@に変換）
- メールの件名は「**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**」「**本日の講義の感想**」を記載してください

氏名：〇 〇　〇 〇

所属：× × × ×専攻　△ △ △ △研究室

学生証番号：□ □ □ □ □

講義の感想：