

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



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

清水謙太郎 (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。

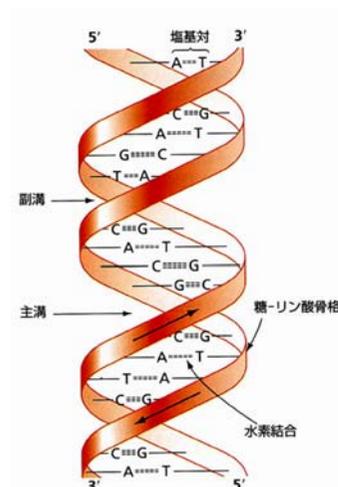
講義日程

講師：大島研郎

- ▶ 2019_生物配列解析基礎_1回目_資料.pdf
- ▶ kiso1 ← 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。
- ▶ Mgenitalium.faa
- ▶ Mpneumoniae.faa
- ▶ parse-blast7.pl
- ▶ test1.seq
- ▶ test2.seq
- ▶ test3.seq
- ▶ Ureaplasma.faa

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 (GATCの並び順) を読むのか?

塩基配列の決定法

= DNAシーケンシング

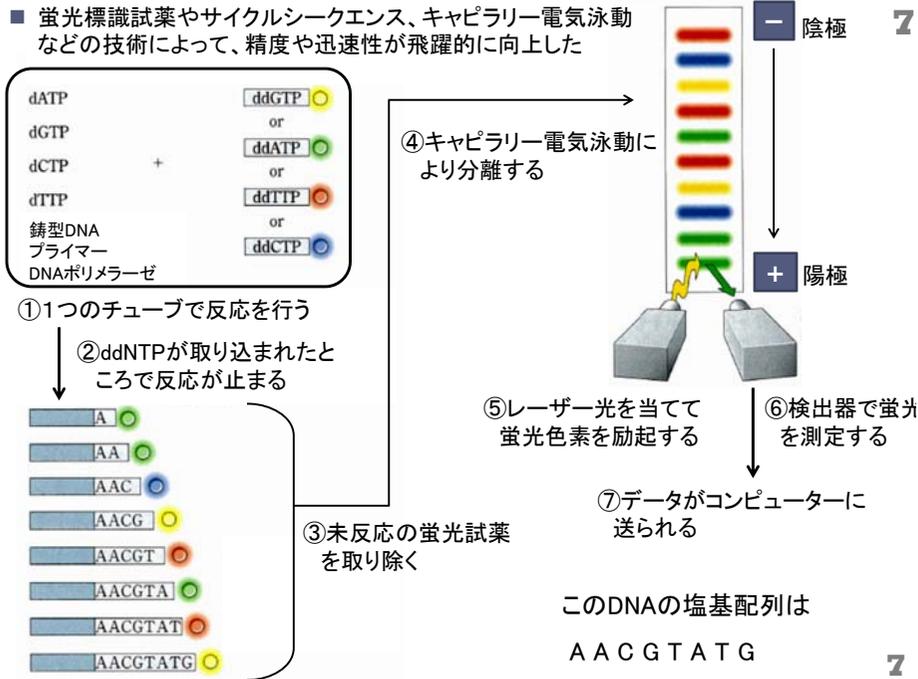
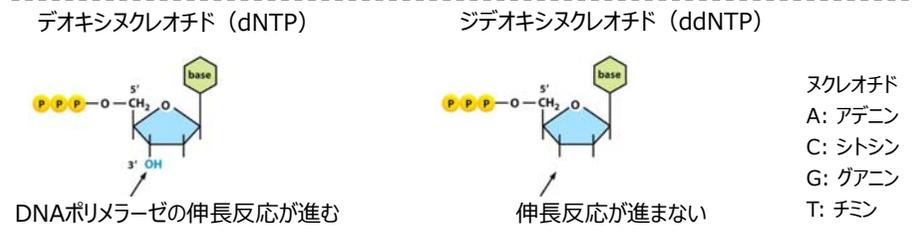
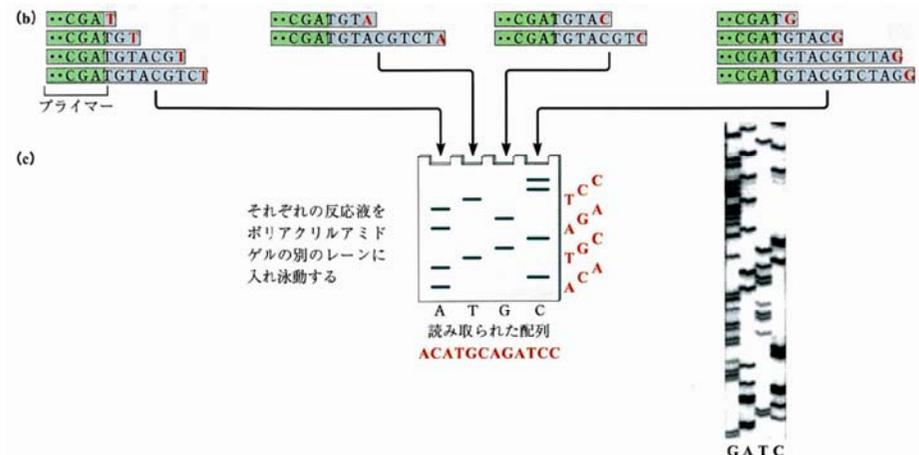
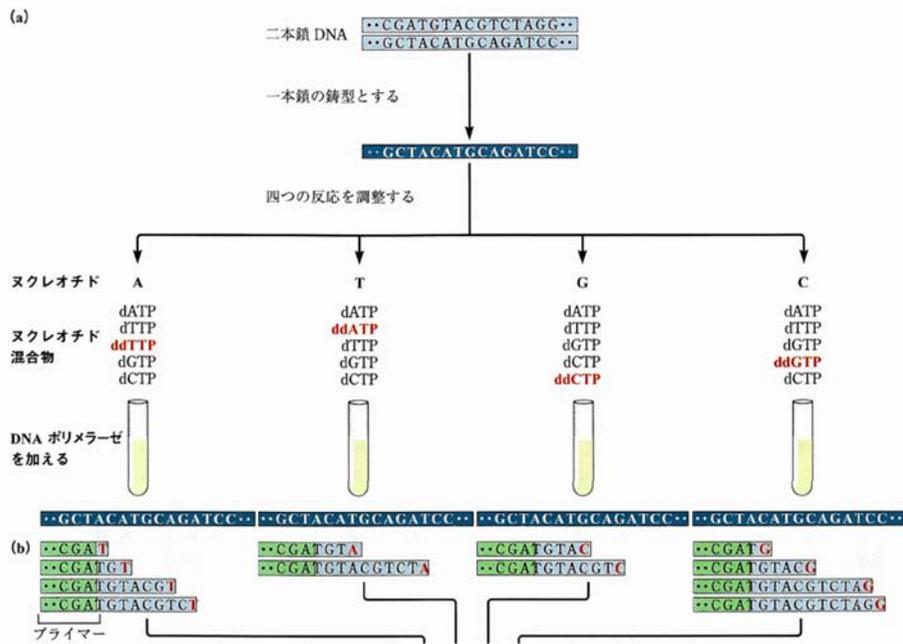
- ジデオキシ法

別名： サンガー法

- ✓ Sanger et al., 1977
- ✓ DNAポリメラーゼを使って相補鎖を合成する反応を行う
- ✓ 特定のヌクレオチドの位置で反応が停止するようにしておく

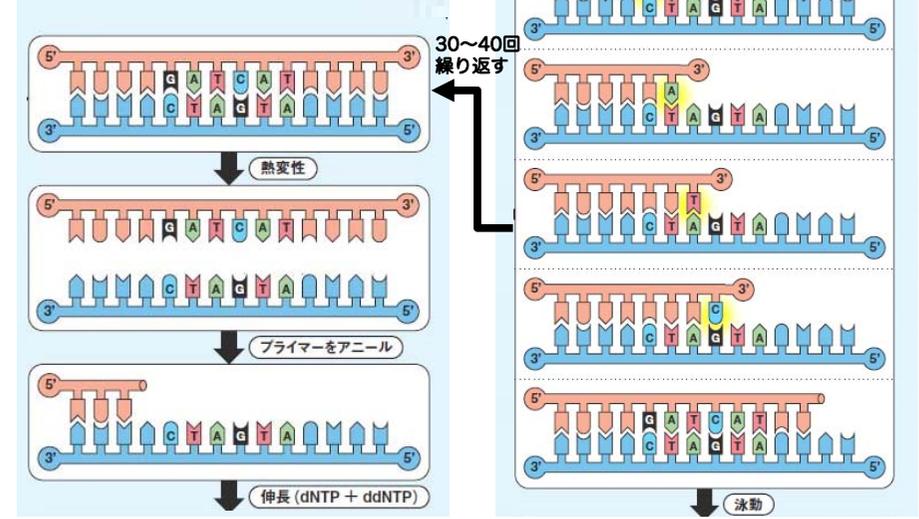
サンガー法の原理

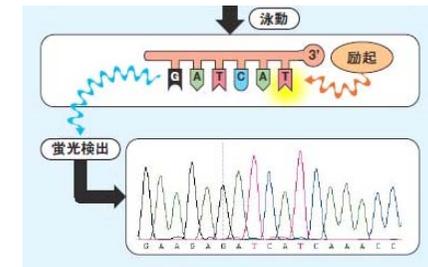
5



サイクルシーケンス法

- ◆ シーケンス反応とPCRとを組み合わせた方法
- ◆ 少量の鋳型DNAでも、塩基配列を決定できる



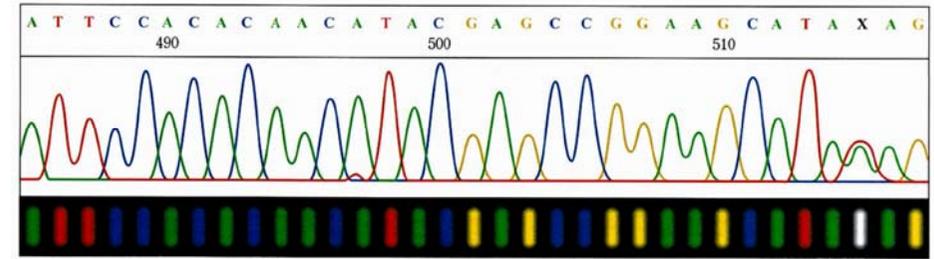


ABI3100シーケンサー



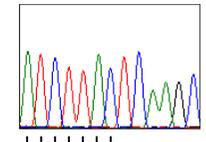
◆当初は平板ゲルで電気泳動していたが、後にキャピラリー電気泳動による機器が普及するようになった

DNAシーケンサーからの出力ファイルの例

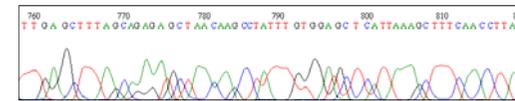


ベースコール

- DNAシーケンサーから出力されたクロマトグラム（波形データ）をもとに、塩基配列を決定する作業。



ATCTTAC...



GATCの文字列にしてしまうとシーケンスの読み具合がわからなくなってしまうので、注意が必要

核酸配列データベース

- 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

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>

アミノ酸配列データベース

UniProt (Universal Protein Resource)
<http://www.uniprot.org/>

- データベースとは、関連性のある一定の情報を集めて、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

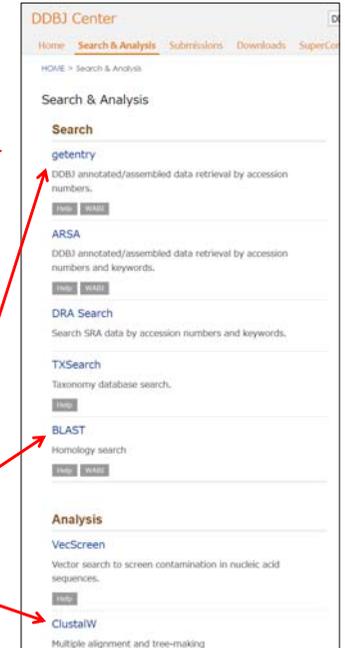
日本DNAデータバンク。GenBankやEMBLと連携して国際塩基配列データベースを構築している。
<http://www.ddbj.nig.ac.jp>



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



AP009356 と入力

```

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

13

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

All Databases

15

- 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

Search NCBI databases

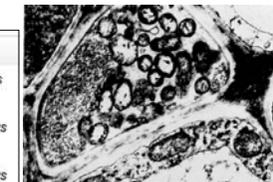
16

replication protein phytoplasma

Results found in 17 databases for "replication protein phytoplasma"

Literature	Genes
Books 1 books and reports	EST 0 expressed sequence tag sequences
MeSH 0 ontology used for PubMed indexing	Gene 62 collected information about gene loci
NLM Catalog 0 books, journals and more in the NLM Collections	GEO DataSets functional genomics studies
PubMed 19 scientific and medical abstracts/citations	GEO Profiles 0 gene expression and molecular abundance

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



■ ファイトプラズマ
Candidatus
Phytoplasma属細菌

■ 植物の節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

アラインメント

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

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

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

$$\begin{matrix} \text{AFDC} & s(A, A) + s(F, E) + s(D, E) + s(C, C) = 8 \\ \text{AEEC} & \quad 3 \quad \quad -7 \quad \quad 3 \quad \quad 9 \end{matrix}$$

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

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

- 同一ファミリータンパク質のギャップなしでアラインメントされた領域(ブロック)に対し、アミノ酸の置換の頻度を調べて作成
- 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62%一致のパターンを一まとめにしてBLOSUM62を作るのに用いる。

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	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	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	-2	-3	-2	-2	-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	-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	-1	-2	1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-2	-3	-1	-1	-4	-4	-3	15	2	-3	-3	
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	8	-1	-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

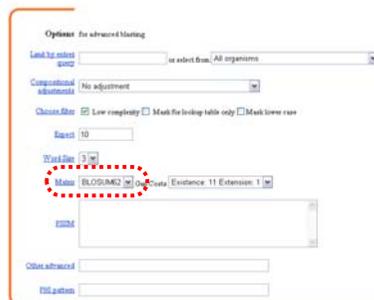
- 2つの配列の最適なグローバルアラインメントを、ダイナミックプログラミング(動的計画法)により求める



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)

25

BLAST検索

26

- 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

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

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

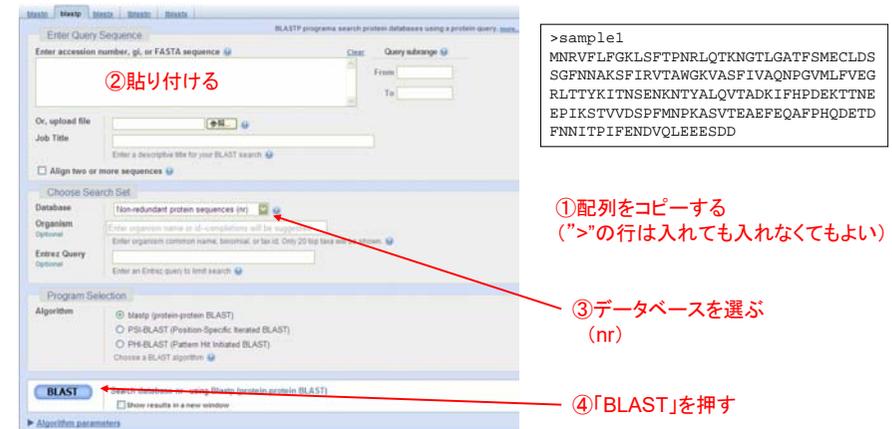
27



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



```
>sample1
MNRVFLFGKLSFTPNRLQTKNGTLGATFSMECLDS
SGFNNAKSFIRVTANGKVASFIVAQNPVMLFVEG
RLTTYKIINSENKNYALQVTADKIFHPDEKTTNE
EPISKSTVVDSPFMNPKASVTEAEFEQAFPHQDET
FNNITPIFENDVQLEESDD
```

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

28

Download ▾ GenPept Graphics

single-stranded DNA-binding protein [Mycoplasma pneumoniae]
 Sequence ID: WP_010874586.1 Length: 166 Number of Matches: 1
 ▶ See 57 more title(s)

Range 1: 1 to 165 E-value
 GenPept Graphics 相同性 (identity) 相同性 (similarity) ギャップ

Score スコア	Expect	Method	Identities	Positives	Gaps
202 bits(514)	6e-64	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 MNRVFLFGKLSFDPNKLOTRTNIGASFLACIDSSGFNDSKSYIRITAWGKVASFVLT 60

Query 61 NPGVMLFVEGRLTTYKITNSN---KNTYALQVTADKIFHPDEKTTNEEPI-KSTVVD 115
 PG +FVEGRL+TYK+ N + K TYALQV ADK++ PDE+ + E+P+ K+TV+DS
 Sbjct 61 KPGDSVFVEGRLSTYKMNRRSDDPNKATYALQVIADKVRPDEENSLEQPVDKATVIDS 120

Query 116 PFMNPKASVTEAEFEQAFPHQDETFNNTPIFENDVQLEEEESDD 160
 PF+ K+ TE E QAFP +++I PI ND QLEEEESDD
 Sbjct 121 PFLAAKTATENELAQAFPISLDDEDDINPILNDSQLEEEESDD 165

アラインメント
 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

全長は166アミノ酸

全長ではないので注意 (本当は166番目にEがある)

E-value

- E-valueは、現在のデータベースにおいて、全く偶然に同じスコアになる配列の数の期待値であり、E valueが小さいほど偶然には起こり得ないことを示している。
- 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の検索結果しか出力されなくなる。

Algorithm parameters

General Parameters

Max target sequences: 100
 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

Word size: 3
 BLAST検索時のWordサイズ

Algorithm parameters

General Parameters

Max target sequences: 100 検索結果の表示件数

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10 E-valueのしきい値

Word size: 3 BLAST検索時のWordサイズ

Scoring Parameters

Matrix: BLOSUM62 マトリックスの種類を選ぶ

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

Compositional adjustments: Conditional compositional score matrix adjustment 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
 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

```

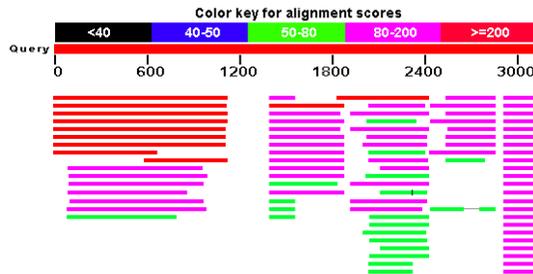
S S * E V K E E K D I H H K L V K K Q
F K V R S E R G * * Y S A K S G K K S
H V K S * K R K R I L I I S * F R K K I
TACTTGAATGAGATGAAAGAGAAGGAATAGTTATACTACGAAATCTTGGAAAAAACTA
      10      20      30      40      50      60
ATGAACTTTACTCTACTTTCTCTTCCCTTATCAATATGATGCTTTAGAACCCTTTTTTGTGAT
M N F T L L S L P Y Q Y D A L E P F F D
* T L L Y F L F L I N M M L * N L F L I
E L Y S T F S S L S I * C F R T F F * Y
    
```

- 塩基配列を決定したが、何が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、調べたいときなど

>sample2

```
ATGAAATTAAGAATCTGCGAACTTGTATTATAAAAACCTTAATTACTAAAACATAAA
ATAGAAACTATTTTAGAACTAAAAAAGCCATTCAAATATGCCTATATTTG
CATGATAAAGATATTTATCAAAATGATAAAGAGGCTCAATTGAATGGTAAAAAGTA
GGAGATATAAAGCTCCTCATTGGCATATATATTTAAGATTTAATTATTCACATGAT
ACAAAAATATCGCTCAATGGTTAATACTGAGGATAATTTGTTCCAAAAATAAAA
GGTAGATTTAGTGATGCCTTAATGTATATGATTCATGCTAATAGGTC
```

blastx検索



33

blastn (nucleotide blast)

>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

34

tblastn

35

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有効

BLAST検索 (GenomeNet)

```
>sample5
MDENETQFNKLNQVKNLKIIVGFGI GGAGNNIVDASLYHYPN
LASENIHPYA INSDLQHLAFKINVKNKLLIQDHTNKGFGAGG
DPKAGSLAISPFQEQNTLTDGYDFCLVAGFGKGTGTGATP
VFSKILKTKKILNVAIVTYP SLNEGLTVRNKATKGLEILNKA
TDSYMLFCNEKCTNGIYQLANTEIVSAIKNLELITIPLQON
IDFPEDVRAFPQTKKTNDQQLFTVTHPFSFSDSKDSIEQFA
KQPKNPKVSYFDHSIVGAKKVLKANKINQKIVKLNPKQIQD
IITWKIDNYQLEIRLGVDFVTTIPNIQIPILSEHKNPVS LPI
DNKSTENNQNKLLDELKELGKMYVKHQNIY
```



Entry	bits	E-val
Top 10		
Clear		
Select operation		
Exec		
✓ mge:MG_224 ftsZ; cell division protein FtsZ; K03531 cell divisi...	679	0.0
✓ mpn:MPN317 ftsZ, F10_orf380; cell division protein FtsZ; K03531...	358	e-100
✓ uur:UU317 hypothetical protein	28	0.53
✓ mpn:MPN257 galE, A65_orf336; 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Query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)

37

- コマンドプロンプトを立ち上げてください

スタート → すべてのプログラム → アクセサリ → コマンドプロンプト

```
C:¥Users¥iu>
```

- 以下、省略して

```
>
```

と記述します

- 「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

38

stand-alone BLASTのダウンロード

- 以下のFTPサイトにアクセスします。
<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>

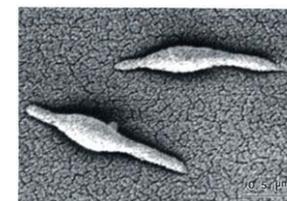
名前	サイズ
ファイル: ChangeLog	1 KB
ファイル: ncbi-blast-2.7.1+-1.src.rpm	18531 KB
ファイル: ncbi-blast-2.7.1+-1.src.rpm.md5	1 KB
ファイル: ncbi-blast-2.7.1+-1.x86_64.rpm	185089 KB
ファイル: ncbi-blast-2.7.1+-1.x86_64.rpm.md5	1 KB
ファイル: ncbi-blast-2.7.1+-src.tar.gz	22710 KB
ファイル: ncbi-blast-2.7.1+-src.tar.gz.md5	1 KB
ファイル: ncbi-blast-2.7.1+-src.zip	26322 KB
ファイル: ncbi-blast-2.7.1+-src.zip.md5	
ファイル: ncbi-blast-2.7.1+-win64.exe	
ファイル: ncbi-blast-2.7.1+-win64.exe.md5	
ファイル: ncbi-blast-2.7.1+-x64-linux.tar.gz	228821 KB
ファイル: ncbi-blast-2.7.1+-x64-linux.tar.gz.md5	1 KB

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

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

40

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

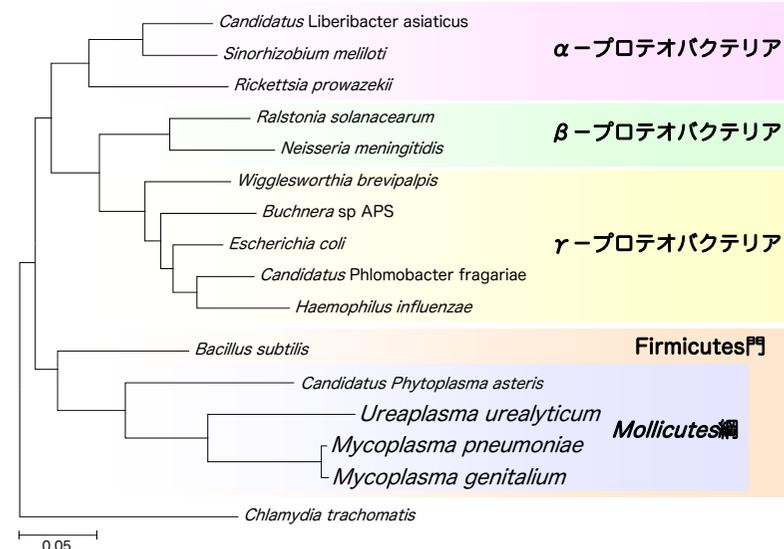
講師：大島研郎

- ▶ 2017_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

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

41

マイコプラズマの系統学的位置



42

- blastフォルダに移動します

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

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
.
.
.
```

43

データベースの準備

- 練習用にMycoplasma genitaliumゲノムデータを用います。dbフォルダの中にMgenitalium.faaというMulti-FASTAフォーマットと呼ばれる形式のファイルがおります。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] キーを押します。

dbフォルダ内のファイルを、メモ帳等で開いてもOKです

44

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
```

-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 (prot)
or 塩基配列 (nucl)

45

stand-alone BLASTの実行

46

- 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]  
MANELTWHDVLAEEKQQPYFLNTLQTVASERQSGVTIYPPQKDFNAFRFTELG  
DVKVVILGQDPYHGPGQAHGLAFSVRPGIAIPSSLNMYKELENTIPGFTRPNH  
GYLESWARQGVLLLLNTVLTVRAGQAHSHASLGWETFDDKVISLINQHREGVVFL  
LWGSHAQKKGAIIDKQRHHVCLKAPHPSPLSAHRGFFGFCNHFVLANQWLEQRGET  
PIDWMPVLPAAESE
```

ファイル名 (例えばtest1.seqなど) を入力するときに、「t」や「test」などを入力した後、Tabを押すことで、その文字から始まるファイル名を表示させることができます

stand-alone BLASTの実行

- test1.seqをquery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
```

-db : データベース指定

-query : 質問配列 (query) 指定

47

stand-alone BLASTの実行

- 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

↑(上矢印)を押すと、過去に入力したコマンドが出てきます

- リダイレクトを使って出力することもできます

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

48

```

BLASTP 2.2.10 [Oct-19-2004]

Reference: Altschul, Stephen F., 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.

Query: hsa:7100.TLR5: toll-like receptor 5 (A)
      (858 letters)

Database: nr-aa: Non-redundant protein sequence database Release
05-04-11      1,952,394 sequences; 634,153,439 total letters
Searching.....done

Sequences producing significant alignments:

                                Score   E
                                (bits) Value
Top 100Top 50Top 20Top 10Top 5 Select operationCLUSTALWMAFFTPTRRNDraw alignmentSearch common
motifs(pfam)Search common motifs(prosite)
sp:AB060695_1 [AB060695] Toll-like receptor 5 [Homo sapiens] 1666 0.0
sp:TLR5_HUMAN [O60602] Toll-like receptor 5 precursor (Toll/inte... 1662 0.0
sp:AX590493_1 [AX590493] Sequence 5 from Patent WO02085933; [Hom... 1659 0.0
spu:AB208697_1 [AB208697] Toll-like receptor 5 [Sus scrofa] 1303 0.0
sp:TLR5_MOUSE [Q9ULF7] Toll-like receptor 5 precursor.prf:26102... 1203 0.0
tr:Q8CB40_MOUSE [Q8CB40] Mus musculus adult female vagina cDNA, ... 1193 0.0
tr:Q5DAS9_CHICK [Q5DAS9] Toll-like receptor 5. 866 0.0
tr:Q5GR02_CHICK [Q5GR02] Toll-like receptor 5 precursor.>gpu:A76... 848 0.0
tr:Q5U5B1_XENLA [Q5U5B1] LOC495313 protein.>gpu:BC084773_1 [BC08... 738 0.0
sp:Q870716_1 [Q870716] Sequence 9 from Patent EP1433792; [unid... 734 0.0
pfi:302356A membrane-toll-like receptor - Oncorhynchus mykiss (... 585 e-165
tr:Q5H720_FUGRU [Q5H720] TLR5.>gpu:AC156437_1 [AC156437] TLR5 [T... 548 e-154
sp:AX590495_1 [AX590495] Sequence 7 from Patent WO02085933; [syn... 437 e-121
tr:Q7ZT61_ONCMY [Q7ZT61] Toll-like receptor5.>gpu:AB062504_1 [AB0... 366 1e-99
.
.
.
>gp:AB060695_1 [AB060695] Toll-like receptor 5 [Homo sapiens] Top
      length = 858
Score = 1666 bits (4315), Expect = 0.0
Identities = 827/844 (97%), Positives = 827/844 (97%)

Query: 15 AGPVFIPGSCSFDGRIAFYRPNLTQVQVNLNTERLLLSFNIRIVTASSFFPKXXXXX 74
      AGPVFIPGSCSFDGRIAFYRPNLTQVQVNLNTERLLLSFNIRIVTASSFF
Sbjct: 15 AGPVFIPGSCSFDGRIAFYRPNLTQVQVNLNTERLLLSFNIRIVTASSFFLEQLQL 74

Query: 75 XXXGSVPTPLTIDKEAFNRLNLRILDLGSSKIYFLHDPDFQGLHFLFELRLYFCGLSDA 134
      CSQYPTPLTIDKEAFNRLNLRILDLGSSKIYFLHDPDFQGLHFLFELRLYFCGLSDA
Sbjct: 75 LELGSVPTPLTIDKEAFNRLNLRILDLGSSKIYFLHDPDFQGLHFLFELRLYFCGLSDA 134

```

質問配列の名前

検索対象として用いたデータベース

アラインメント

E value設定

- E valueは、現在のデータベースにおいて、全く偶然に同じスコアになる配列の数の期待値であり、E valueが小さいほど偶然には起こり得ないことを示しています
- BLAST検索の際にE valueの閾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
- 閾値を設定するには、`-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」と「1」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

```

大量Queryのホモロジー検索法

- stand-alone BLASTは、Multi-FASTA形式のquery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query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ひと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]
MVRITYTLTAPSLDSATITPQIYPEGKLRCTAPVFEFGGGGINVARALAHGGSATAIFPAGGATGHEHLV
SLLDADENVPVATVEAKDWRQNLNHHVHVEASGEQYRFPMPGALNEDEFRLQLEESQVLEIESGAILVIGSL
PPGVKLEKLTQLISAAQKQGIKRCIVDSSGEALSAAALIGNIELVKPNQKLSALVRELTQPDVVRKAAQ
EIVNSGKAKRVVSLGQGALGVDSENCIQVPPVVKSQSTVGGAGDSMVGAMTLKLAENASLEEMRVFQV
AAGSAATLNQGTSLCSHDDTKIYAYLSR

>gi|16132212|ref|NP_418812.1| phosphoglyceromutase 2 [Escherichia coli K12]
MLQVLYVRHGETQWNAERRIQQSDSLPTAKGEQQAMQVATRAKELGITHIISDLGRTRRTAEIIAQAC
GCDIIFDSRLRELNGVLEKRRHIDSLTEEEENWRRQLVNGTVDGRIPPEGESMQELSDRVNALESRCRDLF
QGSRLPLVSHGIALGLCVSTLILGLPAWAERRRLRNCISIRVDYQESLWLAGSWGVTAGDISHLDPAL
DELQR

>gi|16131851|ref|NP_418449.1| glucosephosphate isomerase [Escherichia coli K12]
MKKINFPQTAAWQALQKHFDEMKDVTIADLFARIGDRFSEFSAFDDQMLVDYSKNRITESTLAKLQDLA
KEDCLAGAIKSMFSGEKLNRTEKRAVLKLVKLNSTFILLVDKDVMPVAVLSEKMKFSEALISGSEWK
GYTGKALTDVNIIGIGSDLGPYVVEALRFPYKHNHNFVSNVDGTHIAEVLKKNPETHLPLVASKTF
TTQETMTNAHSARDWFLKAGDEKHKVAKFAALSTNAKAVGEFGIDTANMFEFVDWGGVRSYLSWASIGLS
IVLSIGFDFVELLSGAHAMKHFSTTPEAKNLVLLALIGIWNFFGAETEAALPYDQYMRFAAYFQ
QGMMESNGKYVDRNGNVVDYQTGPFIWGEPEGTNGHAFYQLIHQGTKMPVCDPIAPAITHNPLSDHHQKL
LSNFFAQTEALLAFGKSRVVEQYRDQKDPATLDLVVFPVKFEGNRPTNSILLREITFPFSLGALIALIYE
HKIFTQGVILNIFTFDQWVLELQKLANRILPELKDDEKISSDSSTNGLNIRYKAWRG

```

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
```

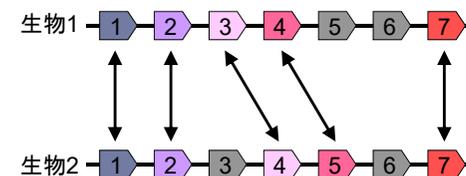
- これらと相同なアミノ酸配列がMgenitalium.faa内にあるかどうかを調べるために、以下のコマンドを実行してください

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

53

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
- このような、非常に類似性が高く、おそらく共通の祖先遺伝子から派生したと考えられるタンパク質をコードする遺伝子のことを、「オーソログ遺伝子」と呼びます
- 片方の生物種の遺伝子（あるいはアミノ酸）配列をqueryとして用いて、相手のゲノム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同定できます



54

ホモロジー検索による比較ゲノム

- Mpneumoniae.faaには、Mycoplasma pneumoniaeがゲノムにコードす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 Mpneumoniae.faa
-evalue 1e-10 -out result4.txt
> more result4.txt
```

55

perlを用いたデータ処理

- 大量のQueryに対して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
- Queryのアクセス番号や、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_009338.1	3-phosphoglycerate kinase; Pgg1 p	416	255	7.00E-69	57%
16130826	NP_012863.1	aldolase; Fba1 p	359	352	4.00E-98	68%
16130686	NP_011770.1	eno1ase I; Eno1 p	437	359	1.00E-100	62%
16130106	NP_009965.1	ribo kinase; 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%

56

BLASTP 2.2.5 [Nov-16-2002]

Reference: Altschul, Stephen F., 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.

Query= gi|16131851|ref|NP_418449.1| glucosephosphate isomerase [Escherichia coli K12] (549 letters)

Database: yeast.aa 6298 sequences; 2,974,038 total letters

Sequences producing significant alignments:
Score E (bits) Value
ref|NP_009755.1| Glucose-6-phosphate isomerase; Pgilp 641 0.0
ref|NP_011646.1| Ygr130cp 30 0.98
ref|NP_013146.1| spindle pole body component; Stu2p 29 1.7
ref|NP_013847.1| (putative) involved in cell wall biogenesis; Ec... 28 3.7
ref|NP_013523.1| Ylr419wp 28 3.7

>ref|NP_009755.1| Glucose-6-phosphate isomerase; Pgilp
Length = 554
Score = 641 bits (1654), Expect = 0.0
Identities = 326/549 (59%), Positives = 401/549 (73%), Gaps = 16/549 (2%)

Query: 7 TQTAAWQALQKHFDEM-KDVTIADLFAKDGDRFSKFSATFDD---QMLVDYSKNRITEE 61
T+ AW LQK ++ K +++ F KD RF K + TF + ++L DYSKN + +E
Sbjct: 13 TELPAWSKLQKIYESQGKTL SVKQEFQKDAKRFELKNTFTNYDGSKILFDYSKNLVNDE 72
Query: 62 TLAKLQDLAKECDLAGAIKSMFSGEKINRTENRAVLHVALRNRNTPILVDGKDVMPVEV 121
+A L +LAKE ++ G +MF GE IN TE+RAV HVALRNR+N P+ VDG +V PEV+
Sbjct: 73 IIAALIELAKEANVTGLRDAMPFGHEINSTE DRAVYHVALRNRANKPMYVDGNNVAPEVD 132

- "Query=" で始まる行に質問配列の情報が, ">" 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.
これらの行だけ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を処理して, list1.txtを生成します.

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

「スタート」

すべてのプログラム

Microsoft Office → Microsoft Office Excel

「list1.txt」をExcel上にドラッグ & ドロップ

Table with columns: 質問配列の情報, 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), スコア, E-value, Identity. Includes a note: 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

<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」を質問配列に用いてBLAST検索を行い, Ureaplasmaがコードする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の閾値は, 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

「受講生の方へ」のページ

↓
「課題提出用Web mailページへ（講義室のみからアクセス可）」

送信先: ← kenro@hosei.ac.jpを選ぶ

件名: ← 「BLAST課題」と入力

氏名:

所属:

学生証番号:

E-mail: Ccを送る

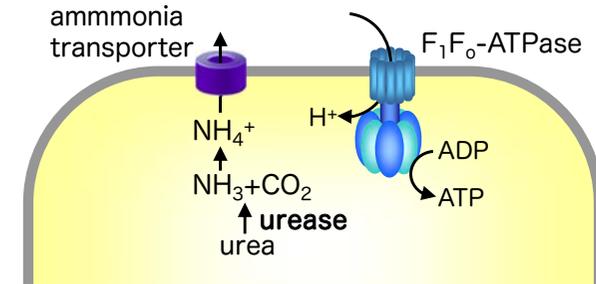
「氏名」「所属」「学生証番号」「メールアドレス」を入力

本文:

← 本日の講義の感想を、ご記入ください

添付ファイル: 次の確認画面で指定して下さい

◆ *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、約95%のATPを合成する



Query GI	Query	Hit Ref No	Hit Function
gi 335789	urease complex component [Ureaplasma parvum]		
gi 335789	urease complex component [Ureaplasma parvum]		
gi 335789	urease complex component [Ureaplasma parvum]		
gi 335789	urease complex component [Ureaplasma parvum]		
gi 335789	urease subunit alpha [Ureaplasma parvum serovar]		
gi 335789	urease complex component [Ureaplasma parvum]		
gi 335789	urease complex component [Ureaplasma parvum]		
gi 335789	urease complex component [Ureaplasma parvum]		
gi 335789	ferriochrome transport ATP-binding protein [Ureapl.	NP_109882.1	cobalt transport ATP-binding protein [Myc
gi 335789	hemolysin [Ureaplasma parvum serovar 3 str. ATC		
gi 335800	hypothetical protein UU437 [Ureaplasma parvum]	NP_110226.1	UV protection protein MucB [Mycoplasma
gi 335800	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