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

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．

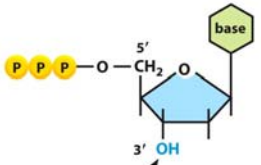
塩基配列の決定　：DNAシークエンス

DNA塩基配列決定のための高速で効率のよい方法が最初に編み出されたのは，1970年代半ばのことである。2つの別々の方法がほぼ同時に公表された。

- **ジデオキシ法**（dideoxy chain termination method; Sanger *et al.*, 1977）。チェーンターミネーター法ともいい，一本鎖DNA分子の配列を決定する方法である。酵素を使って相補的な配列のポリヌクレオチド鎖を合成するが，このとき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。
- **化学分解法**（chemical degradation method; Maxam and Gilbert, 1977）。二本鎖DNA分子の配列を決定する方法である。特定のヌクレオチドの位置で分子を切断する化学的な処理を行う。

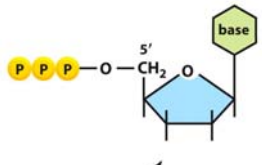
両方法とも最初は同じように広く行われていたが，近年ではジデオキシ法がとくにゲノムの塩基配列決定で多用されている。

デオキシヌクレオチド



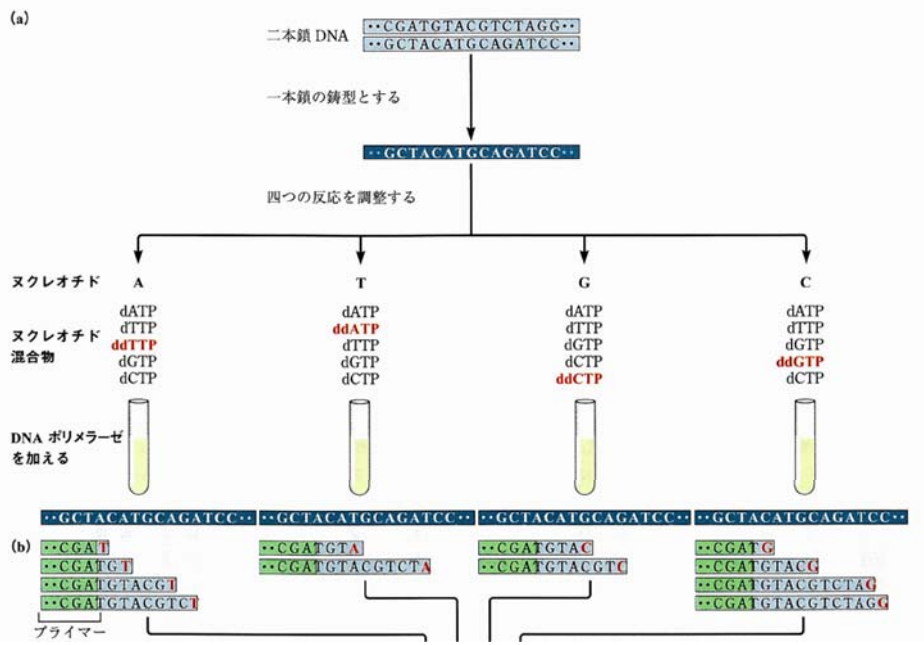
DNAポリメラーゼによって、伸長反応が進む

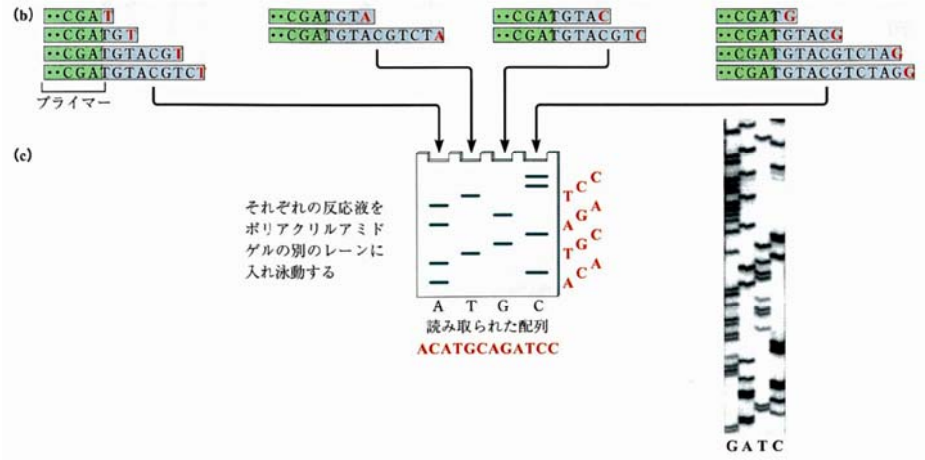
ジデオキシヌクレオチド



伸長反応が進まない

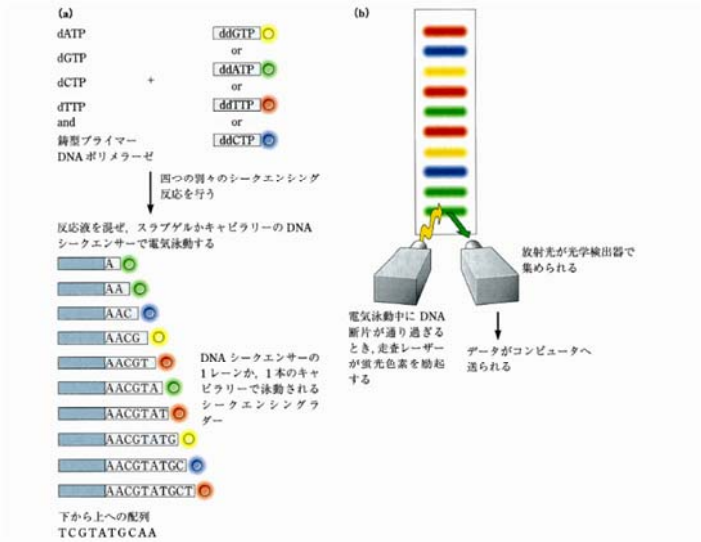
Sanger法





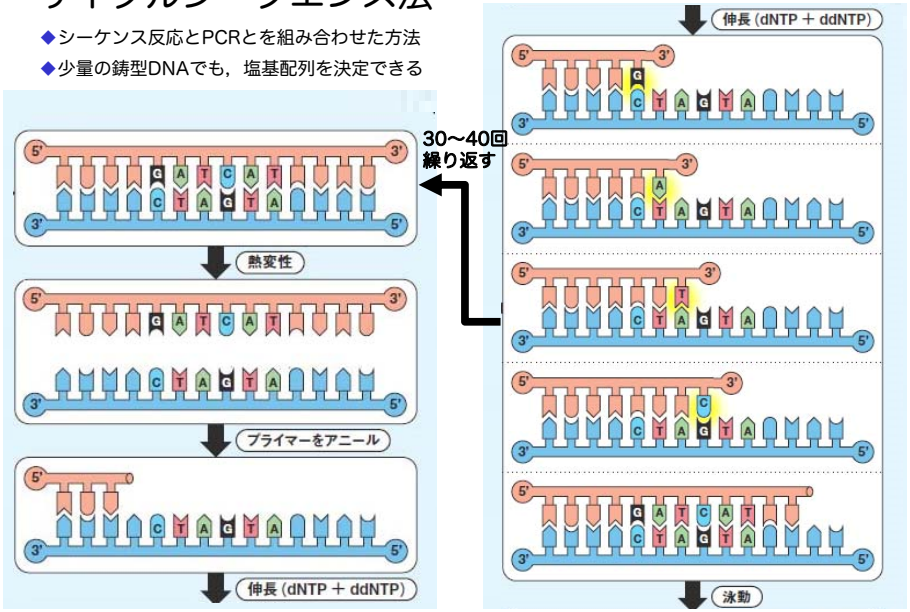
Sanger ジデオキシ DNA シークエンシング法. 4 種の塩基それぞれに 2′,3′-ジデオキシヌクレオチドを準備する．これらの分子は正常な 5′-三リン酸をもつため DNA ポリメラーゼによって伸長する DNA 鎖に取り込まれる．しかし，伸長する DNA 鎖に取り込まれると，ジデオキシヌクレオチド(ddNTP)は次に入ってくる dNTP とリン酸ジエステル結合を形成することができない．その DNA 鎖の伸長は停止する．(a) Sanger 法シークエンシング反応液は，配列を決定するための DNA，その鎖の末端と相補的な短い DNA 断片(プライマー)，正常な dNTP と濃度が注意深く調整された一つの特異的 ddNTP，およびそれ以外の dNTP を含む．後から DNA 分子をオートラジオグラフィーで検出するために，少量の 1 種またはそれ以上の放射線標識された dNTP も同時に含まれている．(b) DNA ポリメラーゼが加えられると，プライマーから正常な重合が始まる．ddNTP が偶然取り込まれると，その鎖の伸長は停止する．もし ddNTP：dNTP の濃度が適切であれば，標識された一連の鎖が合成され，その長さは DNA の末端から特定の塩基の位置までなので，異なっている．(c) その結果生じた，標識された断片はアクリルアミドゲルで，大きさに従って分離され，オートラジオグラフィーが行われる．断片のパターンから DNA 配列が読み取られる．一般に最近のシークエンシング法では，反応産物は鎖停止 ddNTP と結

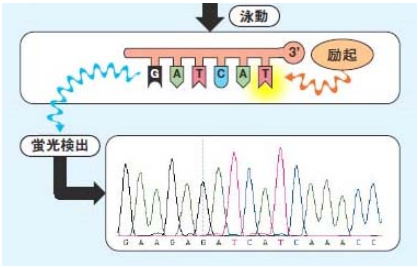
蛍光標識と自動化機械による電気泳動が DNA シークエンシングに革命をもたらす



サイクルシークエンス法

- ◆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

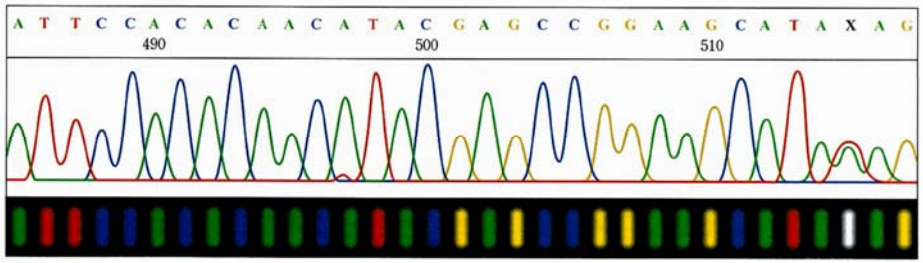


図 10-13
自動 DNA シークエンサーからの読取りデータを示す配列出力ファイル． Phred にコールされた配列が，図の上段に左から右に示してある．電気泳動を行った部分に対する 4 色のシグナルの分布が，ヒストグラムで示されている．出力は X 軸に異なる色でそれぞれの塩基を示し，シグナルの強度を Y 軸に示している．DNA 断片には，非常にはっきりとしたシャープなピークを示すものや，横に広がって隣のシグナルと重なり合ってしまうものもある．この例では，出力ファイルの右端の部分では，ピークの高さが低く，他と重なり合っていることからもわかるように，配列の品質が非常に悪い．515 の位置では(X で印した)，二つのピークが重なっているために，Phred はその塩基をコールすることができなかった．

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

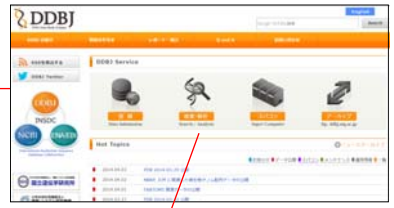
UniProt （Universal Protein Resource）
http://www.uniprot.org/

- 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．大量の情報を高速に処理することができる．

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

塩基配列の登録



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

DDBJ DNA Data Bank of Japan

getentry --アクセッション番号等によるエントリ検索-- [Engli

ID : 検索

DNA データベース : DDBJ / GenBank / EMBL WGS　出力形式 : フラットファイル(DDBJ)

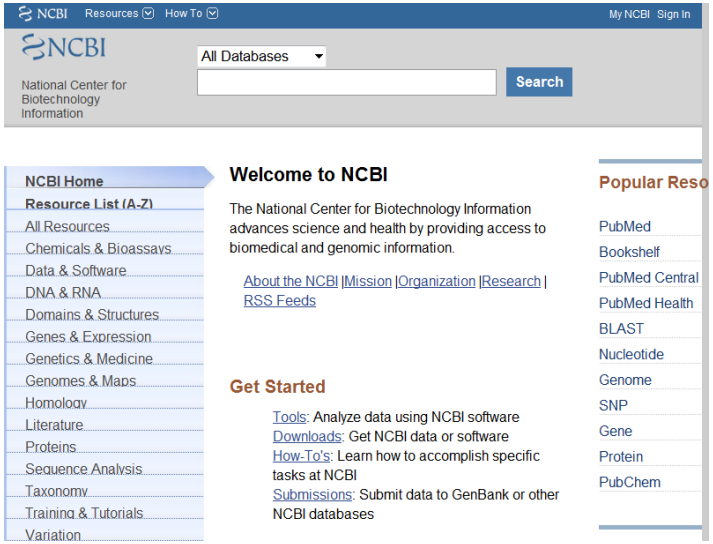
Protein データベース : UniProt PDB DAD Patent　出力形式 : default

取得方法 : html　上限 : 10 件

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
```

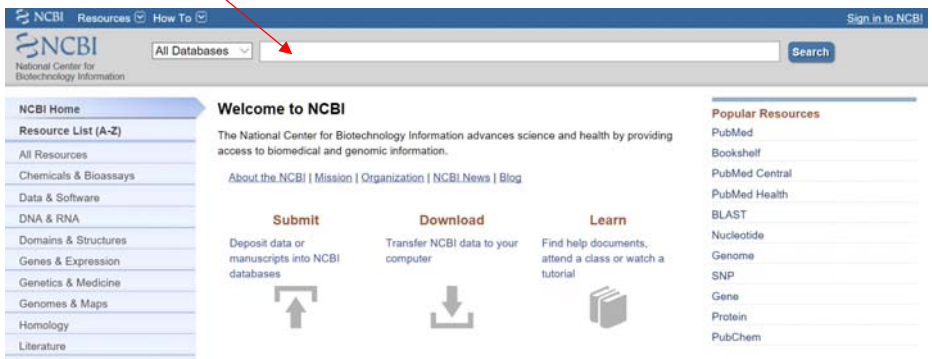

- National Center for Biotechnology Information
  http://www.ncbi.nlm.nih.gov/

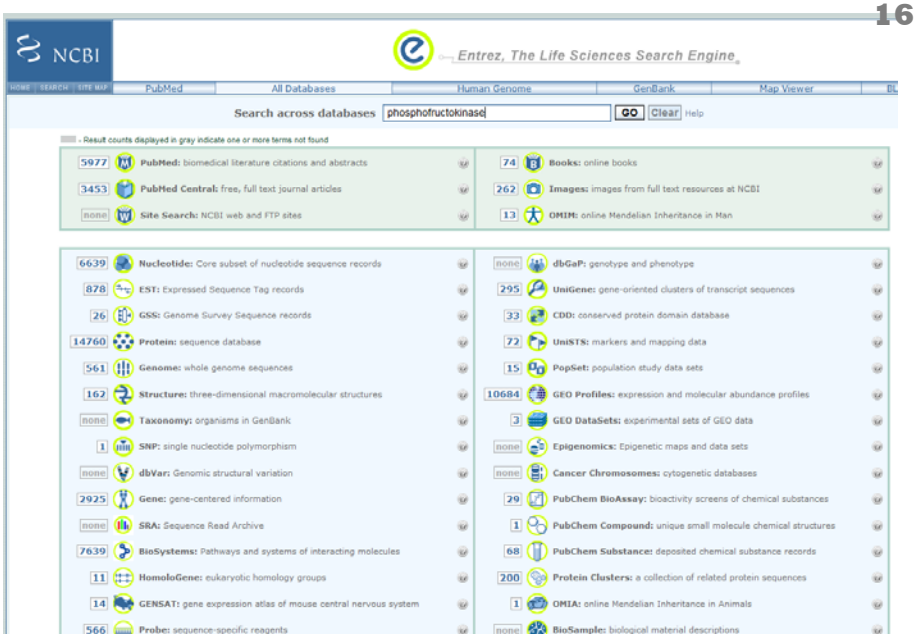


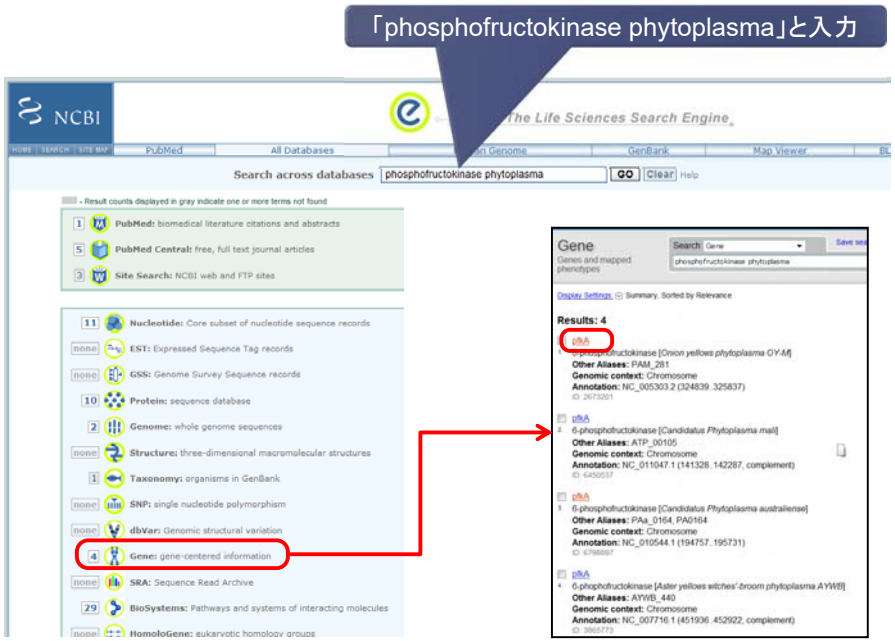
All Databases

- 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・ヌクレオチドシークエンスデータベース・タンパク質シークエンスデータベース・ゲノムシークエンスデータベース・3D高分子構造データベース等．それぞれのデータベースは，関連付けがされており一度に多くのことが調べられる．

例えば「phosphofructokinase」と入力してみる







データベースカタログ

http://lifesciencedb.jp/lsdb.cgi?pg=0

データベース検索（ホモロジー検索）

- ホモロジー検索は，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

http://ssearch.ddbj.nig.ac.jp/top-j.html

FASTA

http://fasta.genome.jp/

BLAST

http://blast.genome.jp/

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

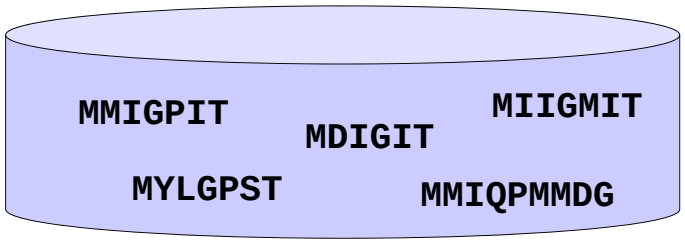
http://blast.ddbj.nig.ac.jp/top-j.html

ホモロジー検索（相同性検索）とは？

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



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
  - スペースの挿入にはペナルティを適用

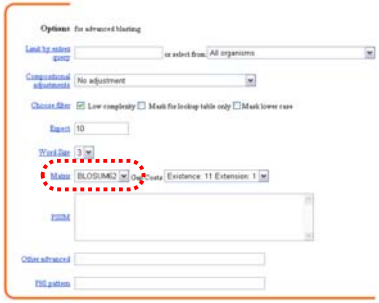
```
AFDC   s(A, A) + s(F, E) + s(D, E) + s(C, C) = 8
AEEC      3         -7        3         9

AFDGC  s(A, A) + s(F, E) + s(D, E) + space + 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（ 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	-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	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

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

- ダイナミックプログラミングによる方法は、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
- 効率的な手法の利用

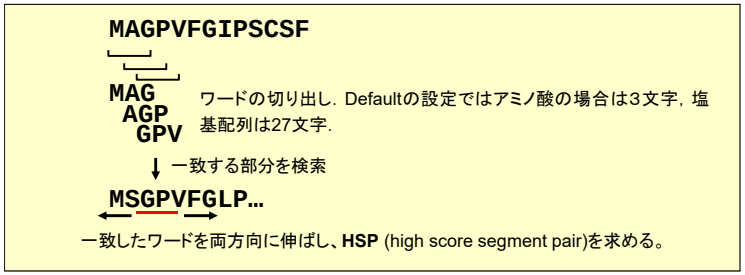
- FASTA
  - 一致する配列の断片を高速に検索、限られた候補に対して精確な手法を適用
  - Lipman and Pearson (1985)

- BLAST
  - 局所的に類似の部分配列を高速に検索
  - Altschul (1990)

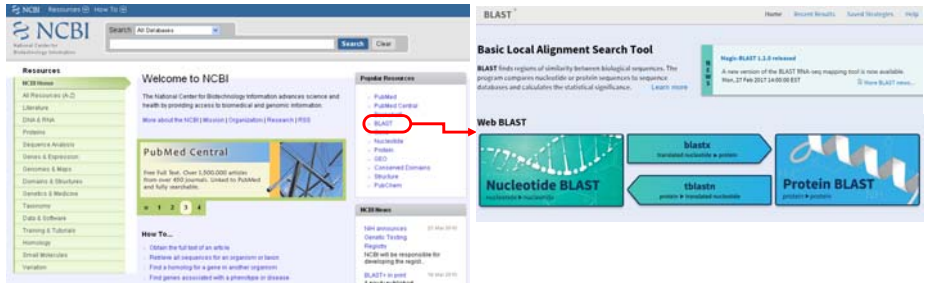
BLAST検索

- 配列を固定長の断片（ワード）に区切り，ワード単位で類似する断片を検索する．
- これらを類似度が最大になるまで両方向に伸ばして局所的なアラインメントを行い，最後にこれらを結合して，最終的なアラインメントを行う手法．

- 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．



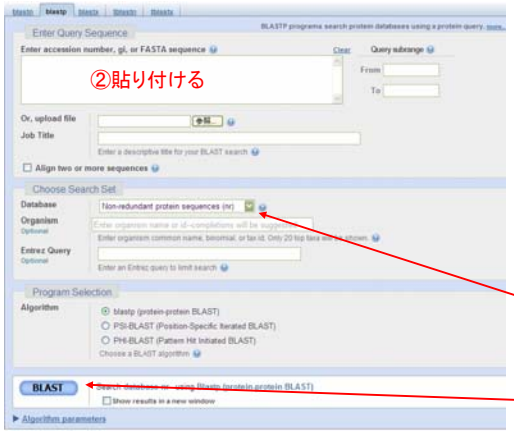
http://www.ncbi.nlm.nih.gov/



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
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

NCBIのref_seq番号

Geneデータベース

```
>ref|NP_109917.1|  single-stranded DNA binding protein [Mycoplasma pneumoniae M129]
 sp|P75542.1|SSB_MYCPN  RecName: Full=Single-stranded DNA-binding protein; Short=SSB;
AltName: Full=Helix-destabilizing protein
 gb|AAB96250.1|  single-stranded DNA binding protein [Mycoplasma pneumoniae M129]
Length=166

 GENE_ID: 877060 ssb | single-stranded DNA binding protein
[Mycoplasma pneumoniae M129] (10 or fewer PubMed links)
 Score =  202 bits (514),  Expect = 7e-51, Method: Compositional matrix adjust.
 Identities = 96/165 (58%), Positives = 126/165 (76%), Gaps = 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
```

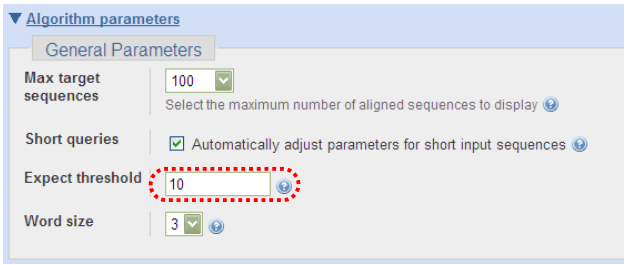
アラインメント
Query : 質問配列
Sbjct : Blast検索の結果，ヒットした配列

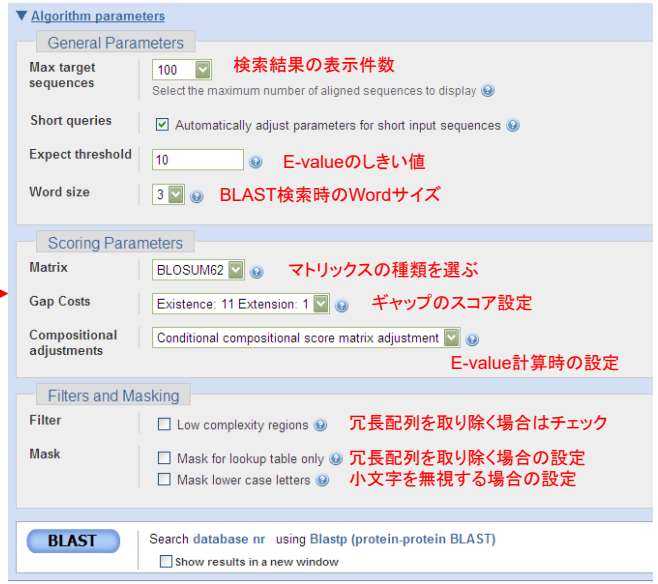
全長ではないので注意
（本当は，…SDDEまで続く）

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．





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
TACTTGAAATGAGATGAAAGAGAAGGAATAGTTATACTACGAAATCTTGGAAAAAAACTA
          10        20        30        40        50        60
ATGAACTTTACTCTACTTTCTCTTCCTTATCAATATGATGCTTTAGAACCTTTTTTTGAT
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
```

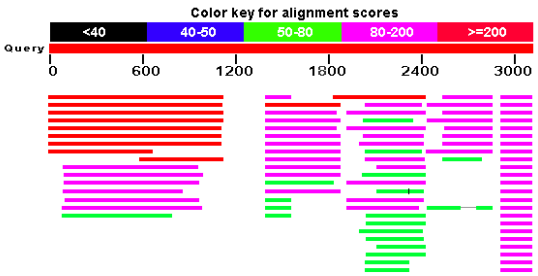
・塩基配列を決定したが，何がコードされているかわからないとき
・non-coding領域に，タンパク質がコードされていないかどうか，調べたいときなど


```
>sample2
ATGAAATTAAGAATCTGCGAACTTGTTATTAATAAAACTTTAATTACTAAAACTAAAATAGAAACTATTTTAGAAACTAAAAA
AAAAGCCATTCAAAATTATGCCTATATTTTGCATGATAAAGATATTTATCAAAATGATAAAGAGGCTCAATTGAATGGTAAAA
AAGTAGGAGATATAAAAGCTCCTCATTGGCATATATATTTAAGATTTAATTATTCACATGATACAAAAAATATCGCTCAATGG
TTTAATACTGAGGATAATTTTGTTTCCAAAATAAAAGGTAGATTTAGTGATGCCTTAATGTATATGATTCATGCTAATAGGTC
                                    .
                                    .
                                    .
```

↓

blastx検索

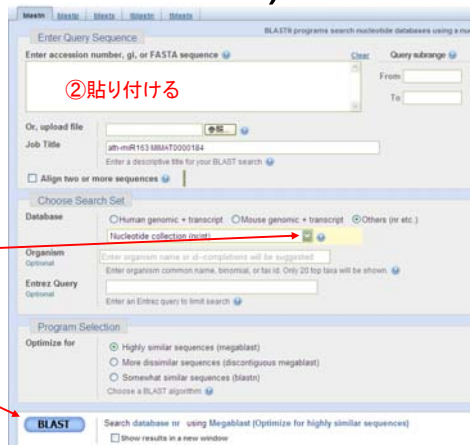
↓



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）



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

↓

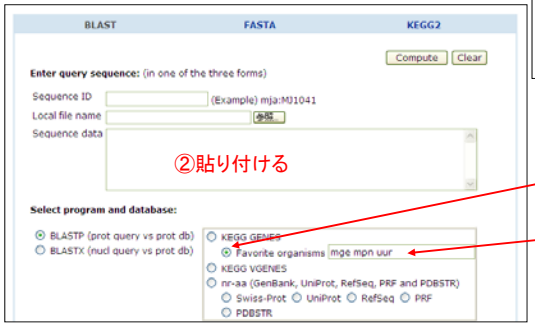
6通りのreading frameのすべてについて翻訳

↓

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有効

BLAST検索（GenomeNet）




```
>sample5
MDENETQFNKLNQVKNKLKIGVFGIGGAGNNIVDASLYHYPN
LASENIHFYAINSDLQHLAFKTNVKNKLLIQDHTNKGFGAGG
DPAKGASLAISFQEQFNTLTDGYDFCILVAGFGKGTGTGATP
VFSKILKTKKILNVAIVTYPSLNEGLTVRNKATKGLEILNKA
TDSYMLFCNEKCTNGIYQLANTEIVSAIKNLIELITIPLQQN
IDFEDVRAFFQTKKTNQDQQLFTVTHPFSFSFDSKDSIEQFA
KQFKNFEKVSYFDHSIVGAKKVLLKANINQKIVKLNFKQIQD
IIWTKIDNYQLEIRLGVDFVTTIPNIQIFILSEHKNPVSLPI
DNKSTENNQNKLKLLDELKELGMKYVKHQNQIY
```

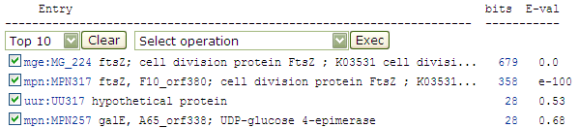
①配列をコピーする
（">"の行は入れても入れなくてもよい）

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）

- 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

stand-alone BLASTのダウンロード

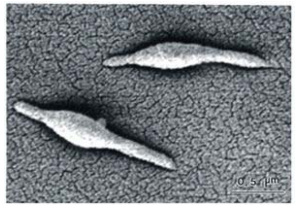
- 以下のFTPサイトにアクセスします．
ftp://ftp.ncbi.nih.gov/blast/executables/LATEST

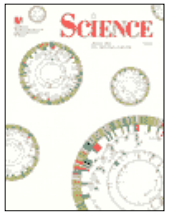
ダウンロードしたファイルをダブルクリックして、
インストールします

通常は，C:¥Program Files¥NCBI¥blast-2.2.26+　にインストールされます

細菌の全ゲノム解読

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

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，
C:¥Users¥iu¥Desktop¥blast
に入れてください

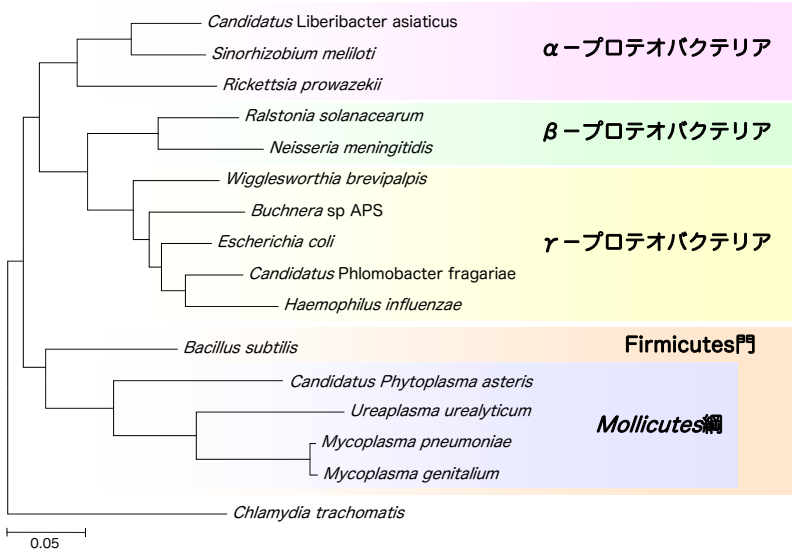
講義日程（平成25年度）

1．平成25年04月09日（PC使用）

講師：大島研郎

- 130409
- Mgenitalium.faa
- Mpneumoniae.faa
- parse-blast7.pl
- test1.seq
- test2.seq
- test3.seq
- Ureaplasma.faa

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
```

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．

db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
```

-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）

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
```

ファイル名（例えばtest1.seqなど）を入力するときに，「t」や「test」などと入力した後，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として用い， 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

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．

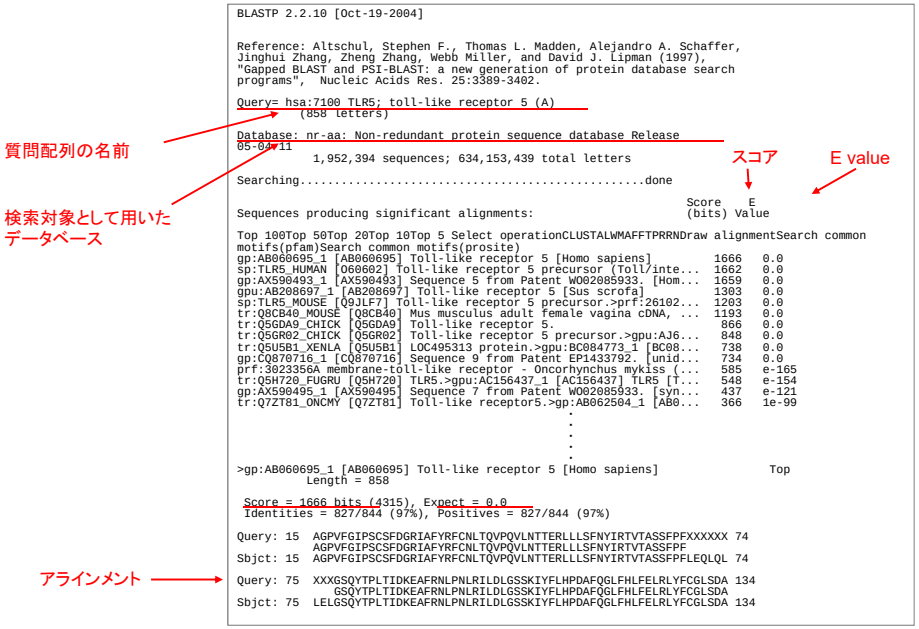
```
> blastp -db Mgenitalium.faa -query test1.seq
  -out result1.txt
> more result1.txt
```

-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
```


```
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
05-07-11
           1,952,394 sequences; 634,153,439 total letters

Searching.....................................................done

                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

Top 100Top 50Top 20Top 10Top 5 Select operationCLUSTALWMAFFTPRRNDraw alignmentSearch common
motifs(pfam)Search common motifs(prosite)
gp:AB060695_1 [AB060695] Toll-like receptor 5 [Homo sapiens]          1666   0.0
sp:TLR5_HUMAN [Q60602] Toll-like receptor 5 precursor (Toll/inte...   1662   0.0
gp:AX590493_1 [AX590493] Sequence 5 from Patent WO02085933. [Hom...   1659   0.0
gpu:AB208697_1 [AB208697] Toll-like receptor 5 [Sus scrofa]           1303   0.0
sp:TLR5_MOUSE [Q9JLF7] Toll-like receptor 5 precursor.>prf:26102...   1203   0.0
tr:Q8CB40_MOUSE [Q8CB40] Mus musculus adult female vagina cDNA, ...   1193   0.0
tr:Q5GDA9_CHICK [Q5GDA9] Toll-like receptor 5.                         866   0.0
tr:Q5GR02_CHICK [Q5GR02] Toll-like receptor 5 precursor.>gpu:AJ6...    848   0.0
tr:Q5U5B1_XENLA [Q5U5B1] LOC495313 protein.>gpu:BC084773_1 [BC08...    738   0.0
gp:CQ870716_1 [CQ870716] Sequence 9 from Patent EP1433792. [unid...    734   0.0
prf:3023356A membrane-toll-like receptor - Oncorhynchus mykiss (...    585   e-165
tr:Q5H720_FUGRU [Q5H720] TLR5.>gpu:AC156437_1 [AC156437] TLR5 [T...    548   e-154
gp:AX590495_1 [AX590495] Sequence 7 from Patent WO02085933. [syn...    437   e-121
tr:Q7ZT81_ONCMY [Q7ZT81] Toll-like receptor5.>gp: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  AGPVFGIPSCSFDGRIAFYRFCNLTQVPQVLNTTERLLLSFNYIRTVTASSFPFXXXXXX 74
           AGPVFGIPSCSFDGRIAFYRFCNLTQVPQVLNTTERLLLSFNYIRTVTASSFPF
Sbjct: 15  AGPVFGIPSCSFDGRIAFYRFCNLTQVPQVLNTTERLLLSFNYIRTVTASSFPFLEQLQL 74

Query: 75  XXXGSQYTPLTIDKEAFRNLPNLRILDLGSSKIYFLHPDAFQGLFHLFELRLYFCGLSDA 134
               GSQYTPLTIDKEAFRNLPNLRILDLGSSKIYFLHPDAFQGLFHLFELRLYFCGLSDA
Sbjct: 75  LELGSQYTPLTIDKEAFRNLPNLRILDLGSSKIYFLHPDAFQGLFHLFELRLYFCGLSDA 134
```

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「l」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]
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

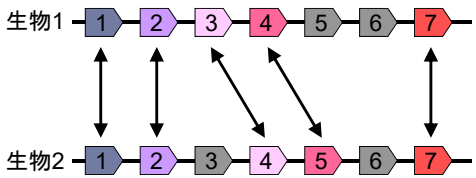
```
> more test3.seq
```

- 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
> more result3.txt
```


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



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
```

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; Cdc19 p	500	40.8	2.00E-04	50%
16130827	NP_009938.1	3-phosphoglycerate kinase; Pgk1 p	416	255	7.00E-69	57%
16130826	NP_012663.1	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; Cdc19 p	500	247	3.00E-66	49%
16129733	NP_012483.1	Glyceraldehyde-3-phosphate dehydrogenase	332	427	1.00E-120	77%


```
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



                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

ref|NP_009755.1| Glucose-6-phosphate isomerase; Pgi1p                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; Pgi1p
          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  TELPAWSKLQKIYESQGKTLSVKQEFQKDAKRFEKLNKTFTNYDGSKILFDYSKNLVNDE 72

Query: 62  TLAKLQDLAKECDLAGAIKSMFSGEKINRTENRAVLHVALRNRSNTPILVDGKDVMPEVN 121
            +A L +LAKE ++ G   +MF GE IN TE+RAV HVALRNR+N P+ VDG +V PEV+
Sbjct: 73  IIAALIELAKEANVTGLRDAMFKGEHINSTEDRAVYHVALRNRANKPMYVDGVNVAPEVD 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

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

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 [Myc	NP_072665.1	seryl-tRNA synthetase (serS	417	669	0	76%
gi|13507745	thymidylate kinase [Mycoplasi	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 oxidatior	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%

*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*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」を質問配列として用いて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）」

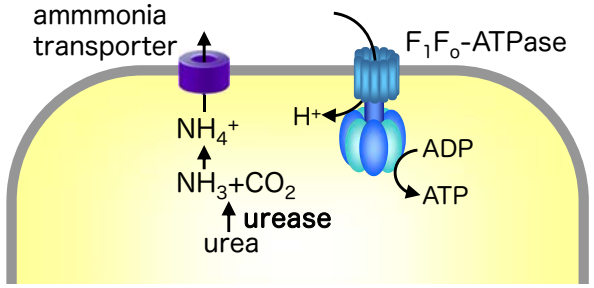
送信先: kenro@hosei.ac.jp ← kenro@hosei.ac.jpを選ぶ
件名: BLAST課題 ← 「BLAST課題」と入力
氏名:
所属:
学生証番号:
E-mail: Ccを送る
「氏名」「所属」「学生証番号」「メールアドレス」を入力

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

確認 リセット

◆*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|13357899	urease complex component[Ureaplasma parvum		
gi|13357900	urease complex component[Ureaplasma parvum		
gi|13357898	urease complex component[Ureaplasma parvum		
gi|13357901	urease complex component[Ureaplasma parvum		
gi|13357902	urease subunit alpha [Ureaplasmaparvum serovar		
gi|13357903	urease complex component[Ureaplasma parvum		
gi|13357897	urease complex component[Ureaplasma parvum		
gi|13357896	ferrichrome transport ATP-bindingprotein [Ureap	NP_109882.1	cobalt transport ATP-binding protein [Myc
gi|13357895	hemolysin [Ureaplasma parvumserovar 3 str. ATC		
gi|13358000	hypothetical protein UU437[Ureaplasma parvum	NP_110226.1	UV protection protein MucB [Mycoplasma
gi|13358001	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