W3.4: Stand-alone BLAST を利用した相同性検索

1. Stand-alone BLAST について

Stand-alone BLAST は、ローカルなコンピュータ上で動く BLAST のプログラムである。通常、相同性検索を行う際には Web 上の BLAST を利用するので十分であるが、以下のような解析を行うニーズがあるときには Stand-alone BLAST を利用するのが効果的である。

- 大量のクエリ配列を使って BLAST 検索を行いたいとき
- 自分の持っている未公開のゲノムデータに対して相同性検索を行いたいとき
- 相同性検索を使って比較ゲノム解析を行いたいとき

stand-alone BLAST は以下のサイトから入手することができる。

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

Windows の場合は ncbi-blast-X.XX.X+-win64.exe、Mac OS の場合は ncbi-blast-X.XX.X+-x64-macosx.tar.gz をダウンロードしてインストールする（X はプログラムのバージョンの数字）。

stand-alone BLAST がインストールされているかどうかを確認するには、コマンドプロンプト（Mac OS の場合はターミナル）を立ち上げて、以下のコマンドを入力する。

```
> blastp -help
```

正常にインストールされている場合は、stand-alone BLAST についての説明が表示される。

2. データベースの準備

パソコンの C ドライブの直下に blast という名前のフォルダを作成する（デスクトップなどにフォルダを作成しても良いが、ユーザ名に日本語が使われているときなどにうまく作動しない場合がある）。

以下のコマンドを入力して blast フォルダに移動する。

```
> cd C:¥blast
```

以下のように表示される。

```
C:¥blast>
```

（以降は省略して **>** のみを記載する）

本書では練習用に *Mycoplasma genitalium* のゲノムにコードされるタンパク質のアミノ酸配列データを用いる。

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/027/325/GCF_000027325.1_ASM2732v1/GCF_000027325.1_ASM2732v1_protein.faa.gz

ダウンロードしたファイルを解凍すると、

GCF_000027325.1_ASM2732v1_protein.faa というファイルが生成される。ファイル名が長いので、本書では Mgenitalium.faa に変更して解析作業を行うこととする。blast フォルダの中に Mgenitalium.faa を移動させて、メモ帳やワードパッドなどでファイルを開いて内容を確認する。あるいは、コマンドプロンプト上で以下のコマンドを入力して内容を見ることもできる。

```
> more Mgenitalium.faa
```

more コマンドについて

指定したファイルの内容がコマンドプロンプト上に表示される。[Space]キーを押すことで続きの内容が表示される。1 行ずつ見るには[Enter]キーを押す。終了するには[Q]キー押す。

Mgenitalium.faa には、*Mycoplasma genitalium* のゲノムにコードされるタンパク質のアミノ酸配列データが Multi-FASTA フォーマット形式で記載されている。

```
>WP_009885556.1 DNA polymerase III subunit delta' [Mycoplasma genitalium]
MLTTTHALLIIQRKGSFLKPFLDNYLTSIVCENKNGCKKCINCLEILNNKYNSLYWFDQINPFKRENALQLARIFNRERT
SVNNKNIYLIEEIEKLSSNSINSLLRLVEDSPINSYGIFTTKNESLILSTFLSRVQKVVLKKASKVPFKVSKNDQEIITS
FFTVDEQIEAIENGSFNRFKIILDACLNKKTGTEQIYHAWQIFRDFSNSEIAQLITLIINKTENIDKKSILFNCLKVLPY
NPPKSTLFANLVSW
>WP_009885557.1 dTMP kinase [Mycoplasma genitalium]
MNKGVFVVIEGVDGAGKTALIEGFKKLYPTKFLNYQLTYTREPGGTLLAEKIRQLLLNETMEPLTEAYLFAAARTEHISK
LIKPAIEKEQLVISDRFVFSSFAYQGLSKKIGIDTVKQINHHALRNMMPNFTFILDCNFKEALQRMQKRGNDNLLDEFIK
```

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ないため、BLAST 用のデータベースへ変換する必要がある。以下のコマンドを実行する。

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

-in オプション：後ろにデータベースとなるファイル名を指定する。

-dbtype オプション：データベースがアミノ酸配列の場合は prot、データベースが塩基配列の場合は nucl を指定する。

blast フォルダに新たに 7 つのファイルが生成されているはずである。

3. stand-alone BLAST の実行

Query（クエリ配列）には test1.seq を用いる。以下のコマンドを入力して内容を確認する。

```
> more test1.seq
```

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

MANELTWHDVLAEEKQQPYFLNTLQTVASERQSGVTIYPPQKDVFNAFRFTELGDVKVVILGQDPYHGPGQA
HGLAFSVRPGIAIPPSLLNMYKELENTIPGFTRPNHGYLESWARQGVLLLNTVLTVRAGQAHSHASLGWETF
TDKVISLINQHREGVVFLLWGSHAQKKGAIIDKQRHHVLKAPHPSPLSAHRGFFGCNHFVLANQWLEQRGET
```

※　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る。

test1.seq をクエリ配列として用いて、Mgenitalium.faa のデータベースに対して blastp 検索を行うために、以下のコマンドを入力する。

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

-db：後ろにデータベースを指定する

-query：後ろにクエリ配列（query）を指定する

下図のように、検索結果がコマンドプロンプト上に表示される。

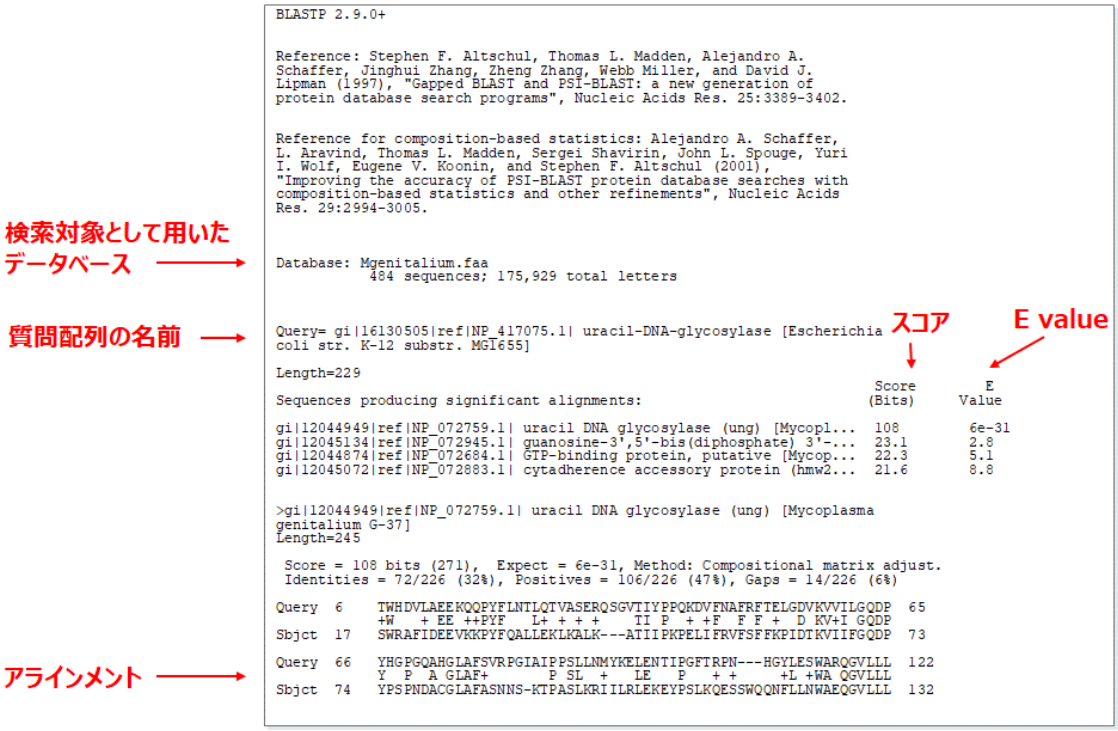
```
Database: Mgenitalium.faa
         511 sequences; 181,545 total letters


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


WP_014894390.1 uracil-DNA glycosylase [Mycoplasma genitalium]        108     6e-31
WP_009885909.1 RelA/SpoT family protein [Mycoplasma genitalium]      23.1    2.9
WP_009885916.1 redox-regulated ATPase YchF [Mycoplasma genitalium]   22.3    5.3
WP_010869377.1 terminal organelle tip protein HMW2 [Mycoplasma ge...  21.6    9.1
```

検索結果をテキストファイルとして出力するには，-out オプションを用いる。-out の後ろに任意の出力ファイル名を指定する。ここでは result1.txt というファイル名で出力する。

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

※　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くる。

メモ帳やワードパッドを使って result1.txt を開き、出力結果を確認する。




```
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 が小さいほど、配列同士の相同性が高いことを示す。

stand-alone BLAST で検索の際に 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 の検索結果のみを出力させることができる。しきい値を設定するには，以下のように-evalue オプションを用いる。

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

blastp と同様に blastx 検索や blastn 検索を行うことができる。test2.seq には塩基配列データが入っている。これをクエリ配列に使って blastx を実行するには、以下のコマンドを入力する。

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

stand-alone BLAST は，Multi-FASTA 形式のクエリ配列にも対応している。例えば，下のような複数の配列を含むファイルをクエリ配列として用いると，それぞれをクエリ配列として 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る。

```
>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
```

例として、100 個分のアミノ酸配列をクエリ配列に用いた blastp 検索を行う。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る。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は、以下のコマンドを実行する。

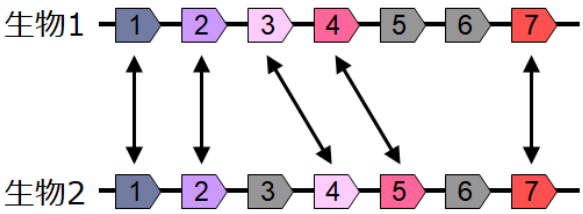
```
> blastp -db Mgenitalium.faa -query test3.seq -evalue 1e-10 -out result3.txt
```

メモ帳やワードパッドを使って result3.txt を開いて、結果を確認する。

4. 相同性検索を用いた比較ゲノム解析

アミノ酸配列が相同なタンパク質は、機能も似ていることが推測される。相同性が高く、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」と呼ぶ。２つの生物種について、それぞれのゲノムにコードされる遺伝子の塩基配列（あるいはタンパク質のアミノ酸配列）を比較してオーソログがあるかどうかを調べることによって、遺伝子の並び順を調べたり、片方の生物が特異的に有している遺伝子を特定することができる。

例えば、片方の生物種のすべてのタンパク質をクエリ配列、もう一方の生物種のすべてのタンパク質をデータベースとして用いて相同性検索を行い、最も相同性の高いタンパク質を特定することで、オーソログを網羅的に調べることができる。なお、ここでは２つの生物のうちの片方をクエリ配列にする例を示すが、実際にはクエリ配列とデータベースを入れ替えて相同性検索を行い、互いに最も相同性の高いものをオーソログとすることが望ましい。



例として、*Mycoplasma genitalium* に加えて、*Mycoplasma pneumoniae* のゲノムにコードされるタンパク質のアミノ酸配列データを用いる。

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/900/660/465/GCF_900660465.1_50648_A01-3/GCF_900660465.1_50648_A01-3_protein.faa.gz

ダウンロードしたファイルを解凍すると、GCF_900660465.1_50648_A01-3_protein.faa というファイルが生成される。ファイル名が長いので、本書では Mpneumoniae.faa に変更して解析作業を行うこととする。blast フォルダの中に Mpneumoniae.faa を移動させる。

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る。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には、以下のコマンドを実行する。

```
> 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、結果を確認する。

大量のクエリ配列を使って BLAST 検索を行うと，結果が羅列した形で出力される。Python などのプログラミング言語を用いることで，クエリ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、必要な情報だけを取り出してリストを作成することができる。