奈良先端科学技術大学院大学　情報科学研究科
平成１８年度「バイオインフォマティクス I」講義資料

ペアワイズアライメントと配列相同性検索

平成１８年７月

川端　猛

１．演習の準備

1.1 演習に必要なファイルのコピー

まず、次のコマンドを入力して、演習に必要なファイルをコピーして、演習用のディレクトリ **BLAST** に移動してください。

```
cd
cp –r /mandara/lecture/takawaba/BLAST .
cd BLAST
```

同様の内容は WEB ページ
http://isw3.naist.jp/IS/Kawabata-lab/LECDOC/BioInfo05/
からも取得できます。

1.1 パスの設定

次に、演習に必要なプログラムが入っているディレクトリにパスを追加してください。もし､パスの追加の方法を知っている方は､以下のディレクトリにパスを追加してください。
/mandara/lecture/takawaba/bin
パスの設定方法が分からない方は BLAST のディレクトリの下にある以下のコマンドを実行してください。

```
./SETUP
```

これにより、もともとの.cshrc が.cshrc.orig_before_bioinfo06 にコピーされ、.cshrc の末尾にパスの設定と、bash を実行する行が書き足されます。

このコマンドは１度だけ実行してください。うまくいかなくても何度も実行しないでください。

この部分は、各人の設定により深刻な問題が生じる可能性があるので、もし、問題が生じるようなら、教官か TA を呼んでください。

２．DNA 配列・アミノ酸配列のファイルフォーマット

演習用のディレクトリの中には、演習用の配列ファイルがいくつか入っています。これらは「FASTA 形式」というファイル形式で入っています。FASTA 形式は、ブラケット(>)で始まるタイトル行に、その遺伝子の名前、コメントなどを記載し、そのあとに１文字表記の配列を記載します。

FASTA 形式のフォーマット

```
>[識別子、タイトルやコメントなど]
[一文字標記の配列。改行で折り返してもかまわない。１行の文字数は任意]
```

>の行（タイトル行）の書き方は、データベースによって様々ですが、[配列名などの識別子]を先頭に書き、スペースを入れて、コメント等を記載するのが通例となっています。例えば、Swissprot の TPIS_ECOLI という配列は以下のようになっています。

FASTA 形式のアミノ酸配列の例

```
>TPIS_ECOLI [P04790] "Triosephosphate isomerase (EC 5.3.1.1) (TIM)”
MRHPLVMGNWKLNGSRHMVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIM
LGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSERRTYHKESDELIAKKFAVLKEQG
LTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP
AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAAELFAQPDIDGALVGGASLKAD
AFAVIVKAAEAAKQA
```

核酸の場合は以下のようになります。Genbank の大腸菌の全ゲノム配列の例です。

FASTA 形式の核酸配列の例

```
>gi|49175990|ref|NC_000913.2| Escherichia coli K12, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
  [途中省略]
AGTGGCAGATGACATAAAACTGGTCGACTGGTTACAACAACGCCTGGGGCTTTTAGAGCAACGAGACACG
GCAATGTTGCACCGTTTGCTGCATGATATTGAAAAAAATATCACCAAATAAAAAACGCCTTAGTAAGTAT
TTTTC
```

多数の配列を収納する配列データベースの場合は、単純にこの形式を積み重ねたものになります。以下に例を示します。

FASTA 形式の複数の配列の例

```
>Y431_METJA [Q57873] Hypothetical UPF0333 protein MJ0431.
MGKMKILKKLLSKKGQLSMEVGVLVAAAVLVAIIAAYFYVKNAKSAVASAGNKSAAFINV
TANKSQEYISNLSNI
>Y420_TREPA [O83435] Hypothetical protein TP0420.
MRRRIYEERGAVRQAGLAHVFEYQGGAAHTGAVQDSDWAVVMRGDIAITLVYAQPVSMPP
VLPLPDFAFQACCSY
>VF07_VARV [P33867] Protein F7.
MTLVMGSCCGRFCDAKNKFKKDDIEEEGEGYCDYKNLNDLDEATRIEFGPLYIINEEKSD
INTLDIKRRYRHAIESVYF
>VCOX_BPHP1 [P51705] Regulatory protein cox.
MSKQNAICINIHMEQPYMTREEFAKKLDVSTRTIDRLRQQGVLKCIKMKNDEGEETERGL
VLVDLVAIAVRNAKNAFQI
```

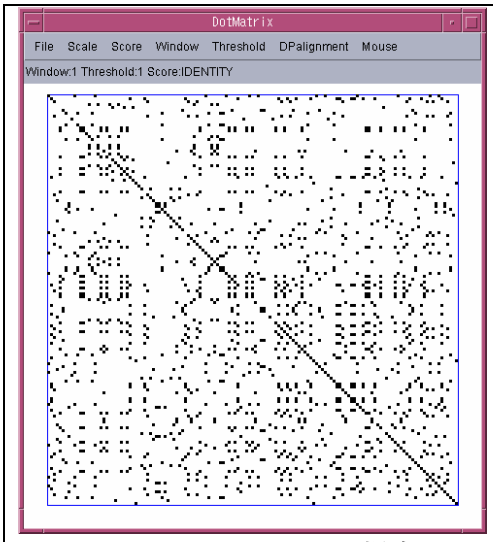
３．ドットマトリックスによるペアワイズアライメント

配列相同性検索行うまえに、ドットマトリックス法による２本の配列のアライメント（ペアワイズアライメント）を行うことで、アライメントの原理について考えてみましょう。演習用のディレクトリに DotMatrix というプログラムが入っています。これは２つの配列を入力して、ドットマトリックスを表示するプログラムで、以下のように使います

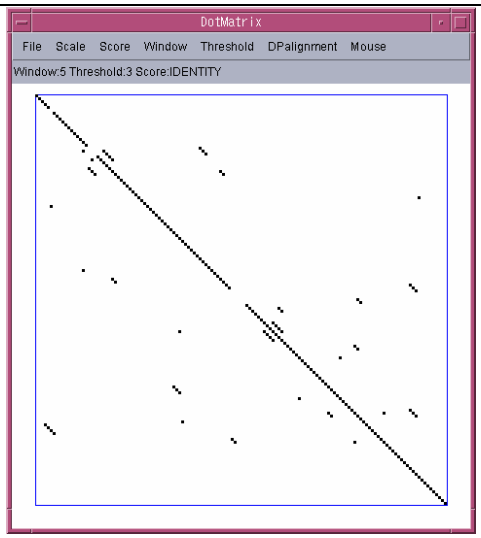
```
./DotMatrix ［配列ファイル１(横)］ ［配列ファイル２(縦)］
```

例えば、ヒトのヘモグロビンα鎖とマウスのヘモグロビンα鎖の配列を以下のように入力するとウィンドウが表示されます。

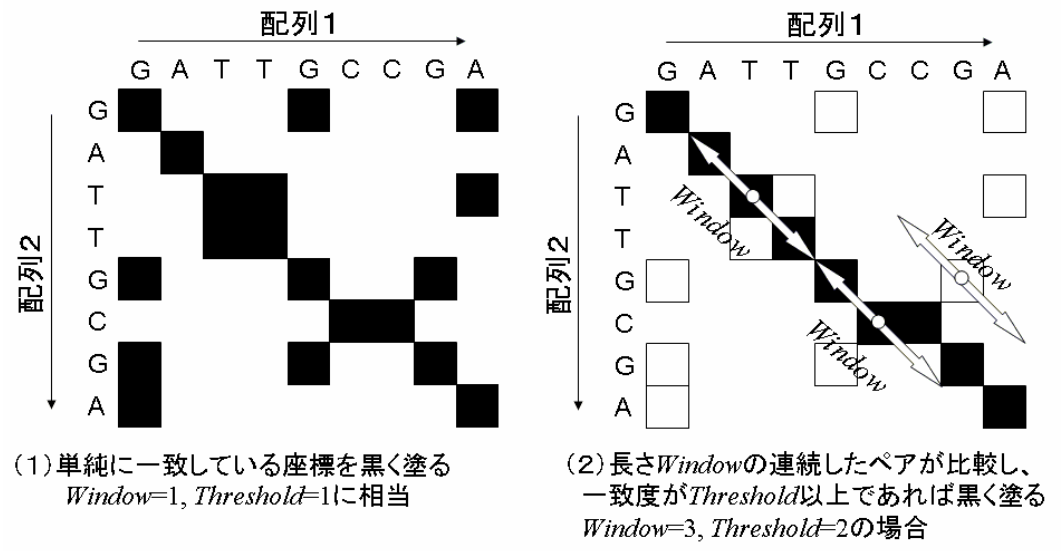
```
./DotMatrix HBA_HUMAN HBA_MOUSE
```



Window =1, Threshold=1 の場合



Window=5, Threshold=1 の場合



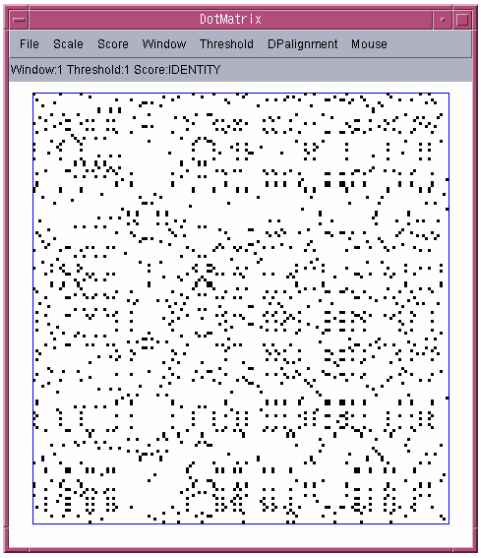
（1）単純に一致している座標を黒く塗る
Window=1, *Threshold*=1に相当

（2）長さ*Window*の連続したペアが比較し、
一致度が*Threshold*以上であれば黒く塗る
Window=3, *Threshold*=2の場合

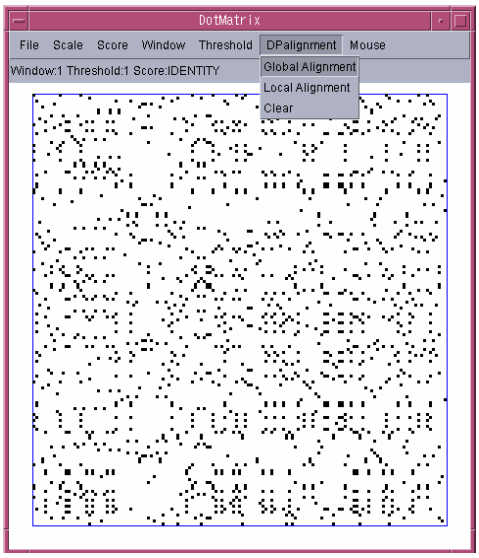
ドットマトリックスの描画の原理

次により類似度が低いヒトのヘモグロビンα鎖とβ鎖を試してみます。これらに対して、ドットマトリックス法だけでなく、動的計画法によるアライメントも実行してみましょう。

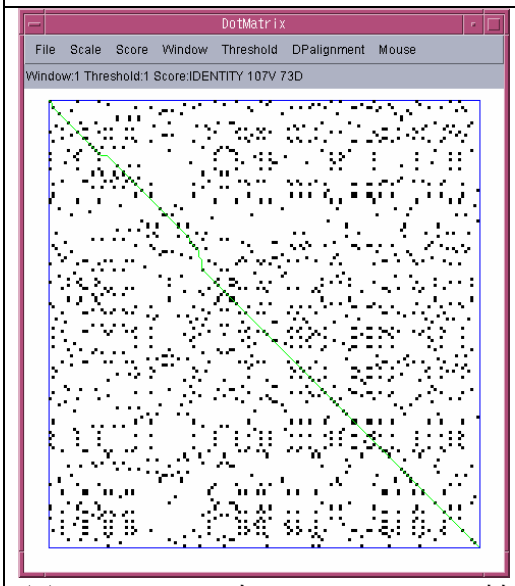
```
./DotMatrix HBA_HUMAN HBB_HUMAN
```



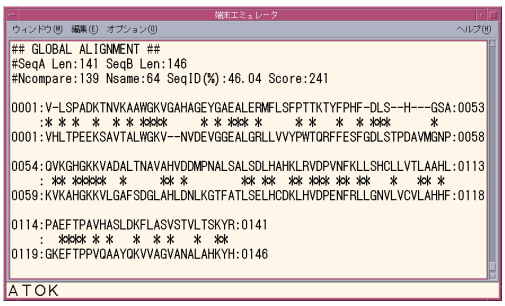
(1)Window=1, Threshold=1 の場合



(2)[DPalignment]メニューから、[Global Alignment]を選択



(3)マトリックス上にアライメントに対応するパスが引かれます。



(4)立ち上げたターミナルに、アライメントが表示されます。

４．アミノ酸配列の配列相同性解析による機能推定（blastp）

blast は非常に多機能なプログラムですが、本演習では最も使い方が単純なアミノ酸配列どうしの比較(blastp)だけを行います。より詳細な使い方は付録を参照してください。

4.1 クエリとする配列

本演習では、あるバクテリアのゲノムの DNA 配列が決まったことを想定し、遺伝子発見の手続きも終了して、既に我々の手元に何本かのアミノ酸配列があるとして、それらのタンパク質の機能のアノテーションを行いたいと考えます。演習用のディレクトリの中に、**X01、X02,…,X08** の８本のアミノ酸配列が FASTA 形式で収納されています。これらの配列をクエリ配列として、blast を実行することで、このタンパク質の機能推定をすることを目的とします。

4.2 設定ファイルの準備

BLAST を実行するには**.ncbirc** というファイルにスコア行列などの data ディレクトリと配列データベースを収納するディレクトリを記載する必要があります。本演習では、演習用のディレクトリ **BLAST** に既に設定された**.ncbirc** があります。作業ディレクトリを **BLAST** で実行すれば正常に動作するはずです。

4.3 データベース等の準備

検索対象とするデータベースは４つ使います。これらは既にダウンロード済みでディレクトリ**/home/is/takawaba/BLASTDB** に置いてあります。フォーマットなどの設定は終了しているので、皆さんが特別にすることはありません。４つのデータベースは以下のもので、全てアミノ酸配列のデータベースです。

データベース名	配列数	説明
swissprot	170000	Swissprot データベースに登録されているタンパク質のアミノ酸配列。様々な生物種のタンパク質が含まれている。
ecoli_aa	4237	大腸菌(*Escherichia coli*)のゲノムにコードされている全タンパク質のアミノ酸配列
bsubt_aa	4105	枯草菌(*Bacillus subtilis*)のゲノムにコードされている全タンパク質のアミノ酸配列
mgeni_aa	484	マイコプラズマ(*Mycoplasma genitalium*)のゲノムにコードされている全タンパク質のアミノ酸配列

4.4 blastp の実行

今回、演習で用いる解析はクエリ配列もデータベースもアミノ酸配列なので、**blastall** というプログラムを **blastp** というオプションで実行します。基本的な入力コマンドは以下のようになります。

```
blastall -p blastp -d [データベース名] -i [クエリ配列名] -o [出力ファイル名]
```

例えば、X01 という配列ファイルをクエリ配列にして、大腸菌のアミノ酸配列に対する検索を行うには以下のようなコマンドを打ちます。

```
blastall -p blastp -d ecoli_aa -i X01 -o X01.ecoli
```

出力結果を less で確認してください。ファイルフォーマットについては、次ページに説明があります。

```
less X01.ecoli
```

同様に、ライブラリを変えて、以下のような計算を行ってください。

```
blastall -p blastp -d ecoli_aa -i X01 -o X01.ecoli
```

```
blastall -p blastp -d bsubt_aa -i X01 -o X01.bsubt
```

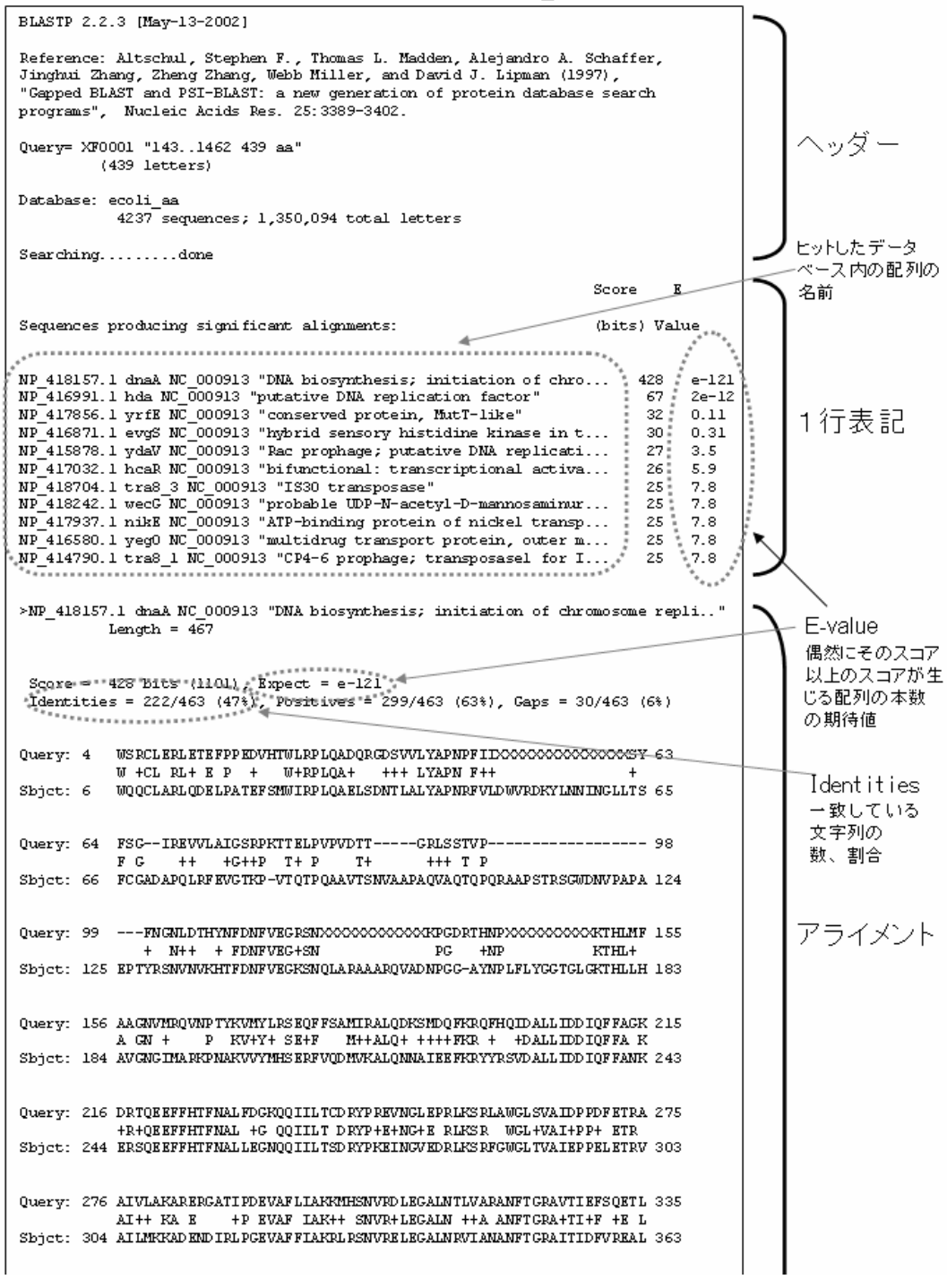
```
blastall -p blastp -d mgeni_aa -i X01 -o X01.mgeni
```

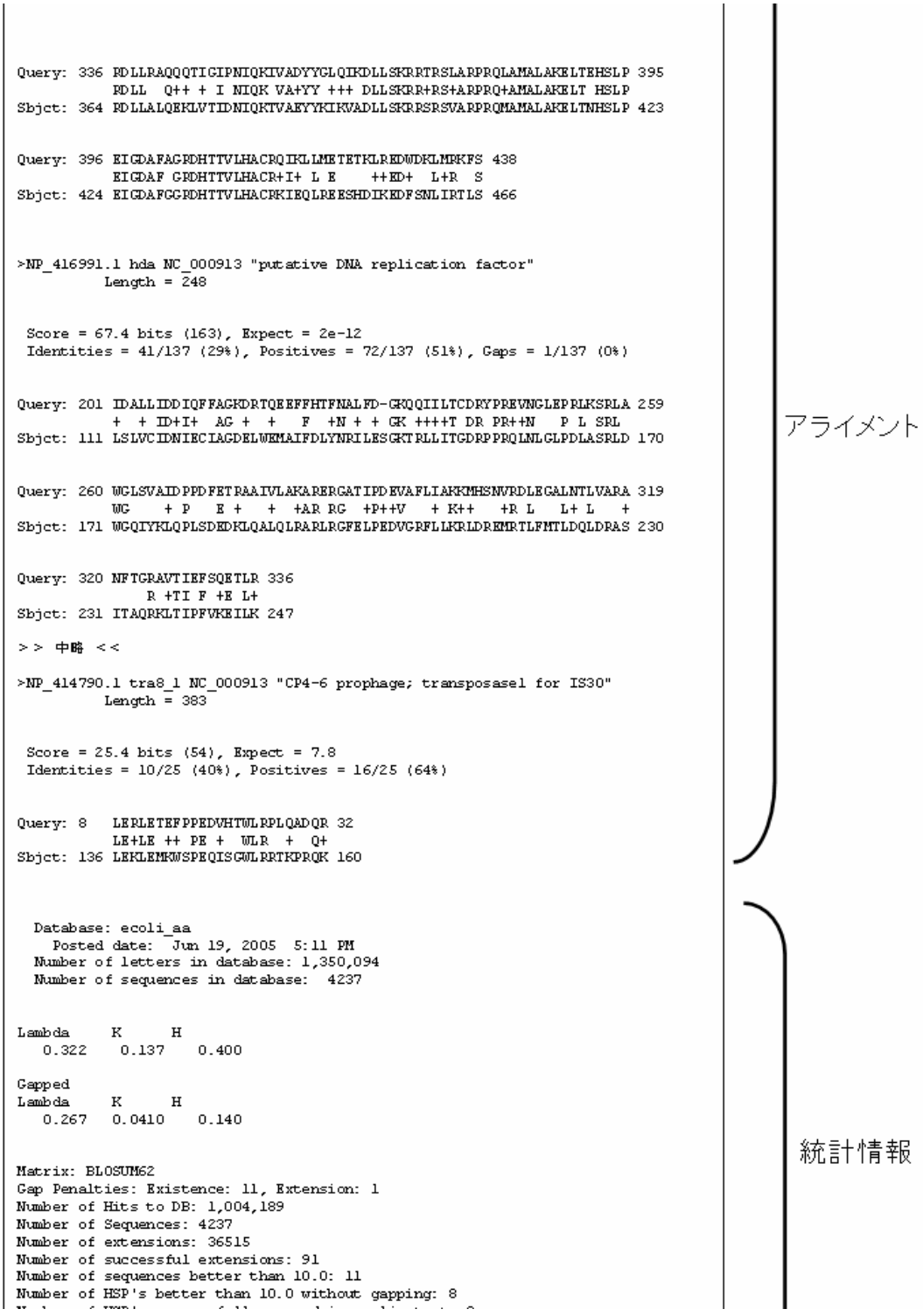
```
blastall -p blastp -d swissprot -i X01 -o X01.sws
```

同様な計算を、**X02,X03,....X08** についても行ってください。結果を less で確認し、回答シートにまとめてください。

BLAST(blastp)の出力例

クエリ配列: XF0001,ライブラリがecoli_aaの場合




Query: 336 RDLLRAQQQTIGIPNIQKIVADYYGLQIKDLLSKRRTRSLARPRQLAMALAKELTEHSLP 395
RDLL Q++ + I NIQK VA+YY +++ DLLSKRR+RS+ARPRQ+AMALAKELT HSLP
Sbjct: 364 RDLLALQEKLVTIDNIQKTVAEYYKIKVADLLSKRRSRSVARPRQMAMALAKELTNHSLP 423
Query: 396 EIGDAFAGRDHTTVLHACRQIKLLMETETKLREDWDKLMRKFS 438
EIGDAF GRDHTTVLHACR+I+ L E ++ED+ L+R S
Sbjct: 424 EIGDAFGGRDHTTVLHACRKIEQLREESHDIKEDFSNLIRTLS 466
>NP_416991.1 hda NC_000913 "putative DNA replication factor"
Length = 248
Score = 67.4 bits (163), Expect = 2e-12
Identities = 41/137 (29%), Positives = 72/137 (51%), Gaps = 1/137 (0%)
Query: 201 IDALLIDDIQFFAGKDRTQEEFFHTFNALFD-GKQQIILTCDRYPREVNGLEPRLKSRLA 259
+ + ID+I+ AG + + F +N + + GK ++++T DR PR++N P L SRL
Sbjct: 111 LSLVCIDNIECIAGDELWEMAIFDLYNRILESGKTRLLITGDRPPRQLNLGLPDLASRLD 170
Query: 260 WGLSVAIDPPDFETRAAIVLAKARERGATIPDEVAFLIAKKMHSNVRDLEGALNTLVARA 319
WG + P E + + +AR RG +P++V + K++ +R L L+ L +
Sbjct: 171 WGQIYKLQPLSDEDKLQALQLRARLRGFELPEDVGRFLLKRLDREMRTLFMTLDQLDRAS 230
Query: 320 NFTGRAVTIEFSQETLR 336
R +TI F +E L+
Sbjct: 231 ITAQRKLTIPFVKEILK 247
> > 中略 < <
>NP_414790.1 tra8_1 NC_000913 "CP4-6 prophage; transposasel for IS30"
Length = 383
Score = 25.4 bits (54), Expect = 7.8
Identities = 10/25 (40%), Positives = 16/25 (64%)
Query: 8 LERLETEFPPEDVHTWLRPLQADQR 32
LE+LE ++ PE + WLR + Q+
Sbjct: 136 LEKLEMKWSPEQISGWLRRTKPRQK 160
Database: ecoli_aa
Posted date: Jun 19, 2005 5:11 PM
Number of letters in database: 1,350,094
Number of sequences in database: 4237
Lambda K H
0.322 0.137 0.400
Gapped
Lambda K H
0.267 0.0410 0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 1,004,189
Number of Sequences: 4237
Number of extensions: 36515
Number of successful extensions: 91
Number of sequences better than 10.0: 11
Number of HSP's better than 10.0 without gapping: 8
アライメント
統計情報

相同性検索から分子機能をアノテーションするための一般的な注意

「相同性を十分期待させるほど類似性が高い場合、お互いの分子機能も近いはずだ」が基本論理です。ですから、まず相同であるかのチェック、相同であるなら分子機能も近いかどうかをチェックする必要があります。以下に主要なチェック項目を挙げます。

(1)E-value が十分低いこと

経験的には、少なくとも E-value が 0.01〜0.0001 より小さい値であることが、相同であるための条件になります。

(2)同一残基率（Sequence Identities）が十分高いこと

経験的には Sequence Identities が 30%より高いことが相同性であることの条件になります。30%を下回る場合は、必ず E-value が十分低いことを確認してください。

(3)アライメントされている領域が十分に長いこと

クエリの配列とほぼ同じ長さがアラインされていることが理想です。部分的な一致の場合には、そのドメインの機能は共通していても、蛋白質全体としてその分子機能が対応しないことがあるので注意が必要です。

(4)ヒットした相手の分子機能がよくわかっていること

ライブラリ配列の中には、分子機能がよくわかっていない蛋白質、蛋白質として発現しているかどうかすら確証のない蛋白質が含まれている場合があります。”Hypothetical protein”,“Function-unknown protein”,“Probable xxx protein”などという記述の場合がそうです。当然のことながら、これらの配列と相同であったとしても機能予測は不可能です。

(5)遠縁のホモログの場合、詳細な機能の一致は期待しない

いくら相同であっても、特に遠い弱い類似性の場合、機能の詳細は食い違う場合があります。例えば、キナーゼという酵素であっても、リン酸基の転移するという反応は同じだが、転移させる相手（基質）は異なるということが生じても不思議ではありません。

(5)遠縁の生物種のホモログには注意が必要

E-value が十分低くても、極端に遠い生物種の組み合わせの場合(例えば、バクテリアとヒトなど単細胞生物と多細胞生物の場合など)、いくらホモログ、オーソログであっても、対応する分子機能がシフトする場合があります。こういった場合、できるだけ下位レベルの分子機能の一致を期待し、上位の細胞レベルの機能については慎重に判断する必要があります。

付録：BLAST の使い方

本稿では、WEB ページを用いずにローカルで BLAST を動かすときに必要な情報をまとめました。

1.インストールの方法

演習用のマシンには既に BLAST はインストールされています。この章は、皆さんがお使いのマシンで BLAST をインストールする必要がある場合に参考にしてください。

BLAST はフリーウエアであり、プログラムソース、様々なプラットフォームの実行形式のファイルが FTP サーバから自由にダウンロードすることができます。

(1) FTP サーバ `ftp.ncbi.nlm.nih.gov` に anonymous でアクセスする。あるいは、 ブラウザで、`ftp://ftp.ncbi.nlm.nih.gov` にアクセスする。

(2) blast/executable/LATEST に移動する。

次のような、様々な OS、CPU 用の実行形式のファイルがあるはずです。ファイル名のバージョン・日付等は、逐次更新されるので、これらとは多少異なる可能性があります。

`blast-20041205-ia32-linux.tar.gz` LINUX (CPU:intel 32bit)
`blast-20041205-ia32-solaris.tar.gz` Solaris(CPU:intel 32bit)
`blast-20041205-ia32-win32.exe` Windows (CPU:intel 32bit)
`blast-20041205-ppc32-maxosx.tar.gz` MaxOSX(CPU:PowerPC 32bit)

(3)必要なファイルをダウンロードし、しかるべきディレクトリにおく。

(4)解凍・展開する。

展開の方法は、OS によって異なります。UNIX 系の OS の場合、適当なディレクトリを作ってダウンロードしたファイルを置いたあと、

```
gunzip blast-20041205-ia32-linux.tar.gz
tar xvf blast-20041205-ia32-linux.tar
```

で展開でき、blast-2.2.10 などのディレクトリが生成されて、その下に実行形式のバイナリファイル、データ、ドキュメントが展開されます。

Windows 系の場合は適当なフォルダをつくってダウンロードしたファイルを'blastz.exe'を置き、ダブルクリックすると、自己解凍して、そのディレクトリに実行ファイルを展開します。

2.BLAST のプログラムの種類

BLAST をインストールすると、いくつかのバイナリが展開されます。主なコマンドの概要を以下にまとめます。

blastall	標準的な BLAST プログラムです。オプションによって、塩基配列、アミノ酸配列の両方に対応できます。本演習では主としてこのコマンドを用います。
formatdb	配列データベースの前処理を行なうためのコマンド
bl2seq	2つの配列をペアワイズアライメントするためのコマンド
fastacmd	配列データベースから、一つの配列だけを取り出すときに必要なコマンド
blastpgp	高感度のアミノ酸配列検索プログラム PSI-BLAST,PHI-BLAST を使うためのコマンド。本演習では詳細は説明しません。
rpsblast	PSI-BLAST で作成されたプロフィールをクエリとして検索する Reverse PSI-BLAST を使うためのコマンド。本演習では詳細は説明しません。
blastclust	BLAST のスコアをもとに配列のクラスタリングを、Single Linkage Clustering のアルゴリズムを用いて行うコマンド。本演習では詳細は説明しません。

３.“.ncbirc”ファイルの編集

自分のホームディレクトリか、カレントディレクトリに'.ncbirc'という設定ファイルを必ず置く必要があります。演習ではコピーしていただいた BLAST というディレクトリに既に'.ncbirc'というファイルがあるはずなので、これをそのまま使ってもらいます。他のディレクトリから BLAST を実行したい場合、自分自身で配列ライブラリを作りたい場合は、この'.ncbirc'ファイルを編集する必要があります。\footnote{Windows の場合は特別なディレクトリに'.ncbirc'のファイルを置く必要があります。詳細はプログラムに付属の README.blt を読んでください。

.ncbirc の書式

```
[NCBI]
DATA=[BLAST をインストールしたディレクトリ]/data

[BLAST]
BLASTDB=[配列データベースを置くディレクトリ]
```

.ncbirc の例。演習で配布するもの

```
[NCBI]
DATA=/auto/home/is/takawaba/tool/blast02Jan18/data

[BLAST]
BLASTDB=/auto/home/is/takawaba/BLASTDB
```

4. 配列データベースの準備

対象配列データは FASTA 形式の塩基配列、アミノ酸配列であればどこからダウンロードしても、自分で用意してもかまいません。NCBI で標準的な配列については FTP サーバで毎日最新のものが提供されています。

4.1 NCBI の FTP サーバから、フォーマット済みの配列データベースを入手する場合

(1) FTP サーバ ftp.ncbi.nlm.nih.gov に anonymous でアクセスする。あるいは、 ブラウザで、ftp://ftp.ncbi.nlm.nih.gov にアクセスする。
(2) blast/db に移動する。
この下にいろいろな種類の配列データベースがあります。主なものを挙げます。ここで、「非冗長」とはデータベース内に完全に同一の配列が存在しないことです。

核酸配列データベース

データベース名	ファイル名	配列数	文字数	内容
nr	**nt.00.tar.gz, nt.01.tar.gz, nt.02.tar.gz**	310 万	141 億	核酸配列のデータベース。GenBank, EMBL,DDBJ のデータベースを融合し、bulk divisions gss, sts, pat, est, and htg divisions を除いたもの。非冗長ではない。

アミノ酸配列（タンパク質配列）データベース

データベース名	ファイル名	配列数	文字数	内容
nr	**nr.tar.gz**	240 万	8.4 億	非冗長なタンパク質の配列データベース。GenPept, Swissprot, PIR, PDF, PDB,NCBI RefSeq をまとめて作成
swissprot	**swissprot.tar.gz**	17 万	6300 万	Swissprot データベースのタンパク質配列（最新のメジャーアップデート版）
pdbaa	**pdbaa.tar.gz**	2 万	480 万	PDB（立体構造データベース）のタンパク質配列

(3)必要なファイルをダウンロードし、gunzip で展開する。

(4) formatdb を実行する
BLAST では、配列データを高速検索するために、formatdb というコマンドを実行する必要があります。

塩基配列の場合は、以下のようにコマンドを入力してください。

formatdb -i [塩基配列ファイル名] **–p F -o T**

アミノ酸配列の場合は同様ですが、以下のように'-p T'オプションをつけて実行してください。

formatdb -i [アミノ酸配列ファイル名] **-p T -o T**

5. blastall の使い方

5.1 必要最小限のオプション

blastall にはたくさんのオプションがありますが、必要最小限のオプションは、-p, -i, -d, -o の4つだけです。

(1) **-p** [Program Name]

比較するクエリ配列とデータベース配列のタイプを指定します。以下の5つから選択します。

Program Name	クエリ配列	データベース配列	比較回数	典型的な使用法
blastn	核酸	核酸	2回 相補鎖にした DB 核酸配列とも比較	ゲノム DNA のアノテーション。cDNA のゲノムへのマッピング。非コーディング領域の比較。
blastp	アミノ酸	アミノ酸	1回	タンパク質配列からの比較的遠縁のホモログの発見。
blastx	核酸 (を翻訳したアミノ酸)	アミノ酸	6回 クエリから6通りのアミノ酸配列を生成して比較（3通りのフレームx相補鎖も）	ゲノム DNA から遺伝子（タンパク質をしている領域）の発見
tblastn	アミノ酸	核酸 (を翻訳したアミノ酸)	6回 DB から6通りのアミノ酸配列を生成して比較（3通りのフレームx相補鎖も）	タンパク質のゲノムへのマッピング
tblastx	核酸 (を翻訳したアミノ酸)	核酸 (を翻訳したアミノ酸)	36回 クエリ、DB の双方とも可能な6通りのアミノ酸配列について比較	やや遠縁の生物種間のゲノム転写物の比較。タンパク質 DB に登録されていない遺伝子のヒットを期待。

(2) **-i** [Query File]
クエリ配列の FASTA 形式のファイル名を指定します。

(3) **-d** [Database]
配列データベースファイルのファイル名を指定します。このファイルは'.ncbirc'の BLASTDB で指定したディレクトリになければなりません。

(4) **-o** [Output File] (Optional)
結果の出力ファイルを指定します。デフォルトは標準出力です。

5.2 実行例

qdna という FASTA 形式の核酸配列をクエリにして、nt データベースを検索する場合

```
blastall -p blastn -i qdna -d nt -o resultfile
```

qprotein という FASTA 形式のアミノ酸配列をクエリにして、nr データベースを検索する場

```
blastall -p blastp -i qprotein -d nr -o resultfile
```

5.3 その他のオプション

-e [E-value]

Evalue のカットオフの値。デフォルトは 10 です。

-F [T or F]

クエリ配列をフィルタリングするかどうかのオプションで、T(true)か F(false)を指定します。デフォルトは T(true)になっています。アミノ酸で使われるフィルタは SEG と呼ばれる低複雑性領域をフィルタです。低複雑性領域とは、'RRRSRRRRKRRR'や'GSGGSGSSSGSG'のように重複や強い機能的な要求により同じようなアミノ酸が密集している領域のことです。これらは、配列相同性解析の基本となる「置換、挿入、削除による進化モデル」から逸脱するので、深刻なエラーを生じることがあります。これらは、フィルタ後には'X'という文字に置換されます。以下に例を示します。

LEKHLRTYWEDFSTSSTSSSTSSTSSSTSGHRYSEKG：フィルタ前
LEKHLRTYWEDFxxxxxxxxxxxxxxxxxxGHRYSEKG：フィルタ後

-m [0 or 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11]

アライメントの表示オプションです。デフォルトは 0 です。0 から 9 まで 10 通りの値を指定できます。

```
-m 0 : pairwise
-m 1 : query-anchored showing identities
-m 2 : query-anchored  no identities
-m 3 : flat query-anchored, show identities
-m 4 : flat query-anchored, no identities
-m 5 : query-anchored no identities and blunt ends
-m 6 : flat query-anchored, no identies and blunt ends
-m 7 : XML Blast output
-m 8 : tabular
-m 9 : tabular ith comment lines
-m 10: ASN,text
-m 11: ASN,binary
```

この中で、比較的便利なのは、マルチプルアライメント風の-m 4 です。-m 4 にすると例えば以下のようにアライメントが出力されます。

```
QUERY 1   AAIVKLGGDDGSLAFVPNNITVGAGESIEFINNAGFPHNIVFDEDAVPA-GVD--ADAI- 56
7pcy- 1   AAIVKLGGDDGSLAFVPNNITVGAGESIEFINNAGFPHNIVFDEDAVPA-GVD--ADAI- 56
1iuz- 1   AqIVKLGGDDGSLAFVPSKISVAAGEAIEFVNNAGFPHNIVFDEDAVPA-GVD--ADAI- 56
2plt- 2    AtVKLGADSGALeFVPKTLTIkSGETVNFVNNAGFPHNIVFDEDAIPS-GVN--ADAI- 56
1pla- 1    AeVKLGSDDGGLVFsPSSFTVAAGEKItFkNNAGFPHNIVFDEDeVPA-GVN--AEkI- 55
1ag6- 3      VlLGGDDGSLAFLPGDFSVASGEEIvFkNNAGFPHNVVFDEDeIPS-GVD--AakIs 56
1plc- 3      VlLGADDGSLAFVPSEFSISpGEKIvFkNNAGFPHNIVFDEDSIPS-GVD--ASkIs 56
9pcy- 3      VlLGSgDGSLVFVPSEFSVpSGEKIvFkNNAGFPHNVVFDEDeIPA-GVD--AvkI- 55
1bypA 1    AeVlLGSSDGGLAFVPSDLSIASGEKItFkNNAGFPHNdLFDKkeVPA-GVD--VtkI- 55
1pcs- 3    AtVKMGSDSGALVFePSTVTIkAGEEVKWVNNklsPHNIVFDaDGVPA-dT---AakL- 56
1b3iA 5      IKMGtDkyAplYePKaLSISAGDTVEFVmNkvgPHNVIFDK--VPA-GeS--ApAL- 55
1bxvA 5      IKMGADNGmLAFePSTIeIqAGDTVQWVNNklaPHNVVvE---------g--qpeL- 49
1bawA 5      VKMGADSGlLqFePaNVTVhpGDTVKWVNNklpPHNILFDDkqVPG-AskelADkL- 59
1kdj- 1    AkVEVGdEvGNFkFyPDSITVSAGEAVEFtlvGetgHNIVFDipAgap-GTv--ASeL- 55
1fa4A 5      VKLGSDkGlLVFePaKLTIkpGDTVEFLNNkvpPHNVVFDaalnPAkSADl-AkSL- 59
```

-m 4 の出力の例

もう一つ便利なのは、-m 8 の出力です。これは、ホモログの情報が、1行ずつタブ区切りで出力されます。タブ区切りのファイルは、計算機で読み込む易く、Excelなどのの表計算ソフトで読み込むこともできます。フィールドは順に、[Query id], [Subject id], [% identity], [alignment length], [mismatch], [gap_openings], [q.start], [q.end], [s.start], [s.end], [e-value], [bit score]の順番に表示されます。

```
7pcy-   7pcy-    100.00   98   0   0    1   98    1   98  2e-53    200
7pcy-   1iuz-     85.71   98  14   0    1   98    1   98  1e-46    178
7pcy-   2plt-     62.89   97  36   0    2   98    2   98  2e-33    134
7pcy-   1pla-     63.54   96  35   0    2   97    1   96  8e-31    125
7pcy-   1ag6-     60.42   96  36   1    4   97    3   98  5e-28    116
7pcy-   1plc-     59.38   96  37   1    4   97    3   98  8e-28    115
7pcy-   9pcy-     59.38   96  37   1    4   97    3   98  7e-27    112
7pcy-   1bypA     54.08   98  43   1    2   97    1   98  5e-24    103
7pcy-   1pcs-     45.36   97  52   1    2   98    3   98  1e-20   92.0
7pcy-   1b3iA     42.11   95  53   1    4   98    5   97  6e-17   79.7
7pcy-   1bxvA     38.95   95  50   1    4   98    5   91  1e-15   75.5
7pcy-   1bawA     41.00  100  54   2    4   98    5  104  2e-15   75.1
7pcy-   1kdj-     41.18  102  55   1    2   98    1  102  3e-15   74.3
7pcy-   1fa4A     44.44   99  50   3    4   97    5  103  8e-15   72.8
```

-m 8 の出力の例

5．その他のコマンドの使い方

(1)bl2seq ： 2つの配列のアライメントを行うコマンド

blastall による配列相同性検索は、クエリ配列とライブラリの中のそれぞれの配列とのペアワイズ・アライメントを繰り返し行って似ている配列を出力するプログラムですが、bl2seq は2つの配列だけを入力して、アライメントを行うコマンドです。既に相同性が確認されている2つの配列を比較するときに便利です。オプションは、blastall と似ています。**-p** によるプログラムの指定以外に、**-i** と**-j** のオプションで、2つの配列を指定します。例えば、比較したいアミノ酸配列のファイル名の一つが qprotein1、もう一つが qprotein2 であったとすると、以下のようなコマンドで、ペアワイズアライメントを得ることができます。

```
bl2seq  -p blastp -i qprotein1 –j qprotein2 -o resultfile
```

(2)fastacmd ： ライブラリからある配列だけを取り出すコマンド

ライブラリファイルの中から、ある配列だけを取り出すときに使うコマンドです。Blastall でヒットした配列を取り出して、clustalw など他のプログラムで、丁寧にアライメントしなおしたいときに便利なコマンドです。Fastacmd で必要なオプションは2つで、一つは-d によるデータベース名の指定、もう一つは、-s による検索文字列を指定します。検索は、原則的に FASTA ファイルの先頭が>の行の、最初のワードを検索します。例えば、swissprot というデータベースから、RECA_ECOLI という検索文字列を含む配列を、抽出するには、以下のようなコマンドを打ちます。

```
fastacmd –d swissprot -s RECA_ECOLI
```

2つ以上の検索文字列を使用する場合は、カンマ区切りで入力します。例えば、RECA_ECOLI と RECA_BACSU の両方を取り出したい場合は、以下のように入力します。

```
fastacmd –d swissprot -s RECA_ECOLI,RECA_BACSU
```

また、このコマンドを使うには、ライブラリを formatdb でフォーマットするとき、Parse オプション **–o T** を付けて、実行しておく必要があります。

参考文献

·Ian Korf, Mark Yandell, Joseph Bedell ”BLAST”. Oreilly & Associates, 2003.

·BLAST WEBページ http://www.ncbi.nlm.nih.gov/BLAST/

·BLAST Information Page
http://www.ncbi.nlm.nih.gov/Education/BLASTinfro/information3.html