6．ホモロジー検索プログラム（FASTA, BLAST）の利用法

目次

1　ホモロジー検索プログラム 6-2
1.1　プログラムの概要 6-2
1.2　問い合わせ配列（FASTA format）の作成 6-3

2　FASTA による検索 6-4
2.1　シェル上からの対話的実行 6-4
2.2　メールサーバー 6-6
2.3　FASTA のアルゴリズム 6-9
2.4　FASTA の結果 6-11
2.5　FASTA の主なオプション 6-13
2.6　まとめ 6-13

3　BLAST による検索 6-14
3.1　シェル上からの起動 6-14
3.2　メールサーバー 6-15
3.3　BLAST のアルゴリズム 6-16
3.4　BLAST の結果 6-17
3.5　BLAST の主なオプション 6-19
3.6　まとめ 6-19

4　検索のヒント 6-20

参考文献 6-20

１．ホモロジー検索プログラム

1.1　プログラムの概要

ホモロジー検索プログラムにはfastaとblastがあり、それぞれ問い合わせ配列とデータベースの種類によって以下のようなプログラムがあります。

FASTA

fasta	蛋白質配列×蛋白質データベース　または 核酸配列×核酸データベース　（自動識別）
tfasta	蛋白質配列×核酸データベース（データベースを翻訳しながら比較）

使用可能なライブラリ

核酸	nr-nt, genbank, genbank-upd, genbank/SUBDIVISIONS, embl, embl/SUBDIVISIONS
蛋白質	nr-aa, swissprot, pir, prf, pdbstr

メールサービスのアドレス　　fasta@genome.ad.jp

BLAST

blastp	蛋白質配列×蛋白質データベース
blastn	核酸配列×核酸データベース
blastx	核酸配列×蛋白質データベース（質問配列を翻訳して比較）
tblastn	蛋白質配列×核酸データベース（データベースを翻訳しながら比較）

使用可能なライブラリ

核酸	nr-nt,genbank,embl
蛋白質	nr-aa,swissprot,pir,prf

メールサービスのアドレス　　blast@genome.ad.jp

各ライブラリの名称は以下のとおりです。

nr-nt	Non-redundant nucleic acid sequence database
genbank	GenBank nucleic acid database
genbank-upd	Cumulative daily updates of GenBank
embl	EMBL nucleic sequence database

※ FASTAではgenbankとemblはsubdivisionごとに検索することもできます。

nr-aa	Non-redundant protein sequence database
swissprot	SWISS-PROT protein sequence database
pir	PIR protein sequence database
prf	PRF protein sequence database
pdbstr	reorganized PDB as a sequence database

オンラインマニュアルが用意されています。man fasta およびman blastで参照できます。メールサービスのヘルプを得るには各メールアドレスに help とだけ書いたメールを送ってください。

1.2　問い合わせ配列 (FASTA format)の作成

問い合わせ配列を以下の形式で記述したファイルを用意してください。配列部分は大文字でも小文字でも構いません。

```
> 名称およびコメント
配列.....
.....
```

問い合わせ配列をデータベースから得るにはbgetコマンドの-fオプションを使います。

PIRからICRT4（RatのInterleukin 4）の配列をとってきてtestseq.aaというファイルに格納します。

```
% bget pir icrt4 -f >testseq.aa
```

```
>ICRT4 (pir)
MGLSPHLAVTLFCFLICTGNGIHGCNDSPLREIINTLNQVTEKGTPCTEMFVPDVLTATR
NTTENELICRASRVLRKFYFPRDVPPCLKNKSGVLGELRKLCRGVSGLNSLRSCTVNEST
LTTLKDFLESLKSILRGKYLQSCTSMS
```

GenBankからHUMSRC01（Humanのc-src gene）の配列をとってきてtestseq.ntというファイルに格納します。

```
% bget genbank humsrc01 -f >testseq.nt
```

```
>HUMSRC01 (genbank)
ctgccaggaccatgggtagcaacaagagcaagcccaaggatgccagccagcggcgccgca
gcctggagcccgccgagaacgtgcacggcgctggcgggggcgctttccccgcctcgcaga
cccccagcaagccagcctcggccgacggccaccgcggccccagcgcggccttcgccccg
cggccgccgagcccaagctgttcggaggcttcaactcctcggacaccgtcacctccccgc
agagggcgggcccgctggccggtcagtgcgc
```

新しい配列を検索にかけるときは、emacsなどのエディタで配列ファイルを作成してください。

2. FASTAによる検索

2.1 シェル上からの対話的実行

2.1.1 fastaの起動

FASTAを対話的に実行するには、UNIXのプロンプトの状態から単に'fasta'とだけ入力します。

```
hgc{lect}99: fasta
 fasta 1.6b2 [Jan, 1992] searches a sequence data bank

 test sequence file name: testseq.aa

 Choose sequence library:

    P: NBRF PIR Database (Release 35.0)
    S: Swiss-Prot (Release 24)
    F: PRF Sequence Database(Release 93.1)

 Enter library filename (e.g. prot.lib), letter (e.g. P)
 or a % followed by a list of letters (e.g. %PN): s
 ktup? (1 to 2) [2] (return)
>ICRT4 (pir) : 147 aa
 vs Swiss-Prot (Release 24) library
 searching /bio/db/fasta/db/swissprot/swissprot library
```

test sequence file nameは用意した問い合わせ配列のファイル名を入力します。
sequence libraryは使用可能なライブラリの一覧を表示しますので記号で選択します。
%のあとに記号を続けて入力することにより、複数のライブラリを選択することもできます。また、オリジナルのライブラリファイルを検索したい場合は、直接ライブラリのファイル名を入力してください。
ktupは検索の速さと正確さを決めるパラメータで、核酸の場合１－６、蛋白質の場合１、２の中から選べます。[ ]内にデフォルトの値（核酸６、蛋白質２）が表示されますので、変更しないときはリターンキーを入力するだけで結構です。

（参考）
ここで入力したファイル名などはUNIXのプロンプト状態から一度に入力することもできます。

```
hgc{lect}99: fasta testseq.aa s 2
```

また、オプションを指定したい場合は、プログラム名に続いて入力してください。

```
hgc{lect}100: fasta -o
```

2.1.2　結果の保存

検索が終了すると結果を保存するファイル名や保存する内容を尋ねてきます。
以下ではスコア上位20個とアライメント 5 個をファイル「 result.fasta.aa 」に保存します。

```
  initn   init1
<  2    25    25:=============
   4     3     3:==
   … (ヒストグラムの表示) …
  80     0     0:
> 80     4     3:==
9545403 residues in 28154 sequences
 statistics exclude scores greater than 72
 mean initn score:  24.5 (7.35)
 mean init1 score:  24.0 (6.15)
 4504 scores better than 30 saved, ktup: 2, variable pamfact
 joining threshold: 27  scan time:  0:00:47
 Enter filename for results : result.fasta.aa
 How many scores would you like to see? [20] (return)
The best scores are:                                      initn init1   opt
IL4_RAT INTERLEUKIN-4 PRECURSOR (IL-4) (B-CELL STIMULA  747   747   747
    … (スコア上位20個表示) …
CIC1_CYPCA L-TYPE CALCIUM CHANNEL ALPHA-1 SUBUNIT        68    36    42
 More scores? [0] (return)
 Display alignments also? y
 number of alignments [20]? 5
Library scan:  0:00:47  total CPU time:  0:00:48
hgc{lect}100:
```

2.2　メールサーバー

2.2.1　リクエストの形式

fasta@genome.ad.jpに以下のような内容のメールを送ってください。

```
program fasta
datalib swissprot
ktup 1
-b 20 -d 5 -o
begin
>ICRT4 (pir)
MGLSPHLAVTLFCFLICTGNGIHGCNDSPLREIINTLNQVTEKGTPCTEMFVPDVLTATR
NTTENELICRASRVLRKFYFPRDVPPCLKNKSGVLGELRKLCRGVSGLNSLRSCTVNEST
LTTLKDFLESLKSILRGKYLQSCTSMS
```

以下のキーワードを指定して下さい。省略した場合は下線部がデフォルトの値になります。デフォルト値のないキーワードは省略できません。

program　　fasta or tfasta

datalib　　nr-aa,pir,swissprot,prf,pdbstr(蛋白質)
nr-nt,genbank,genbank-upd,genbank/SUBDIVISION,
embl,embl/SUBDIVISION(核酸)

ktup　　1 - 6 (核酸), 1 or 2 (蛋白質)

(options)

begin

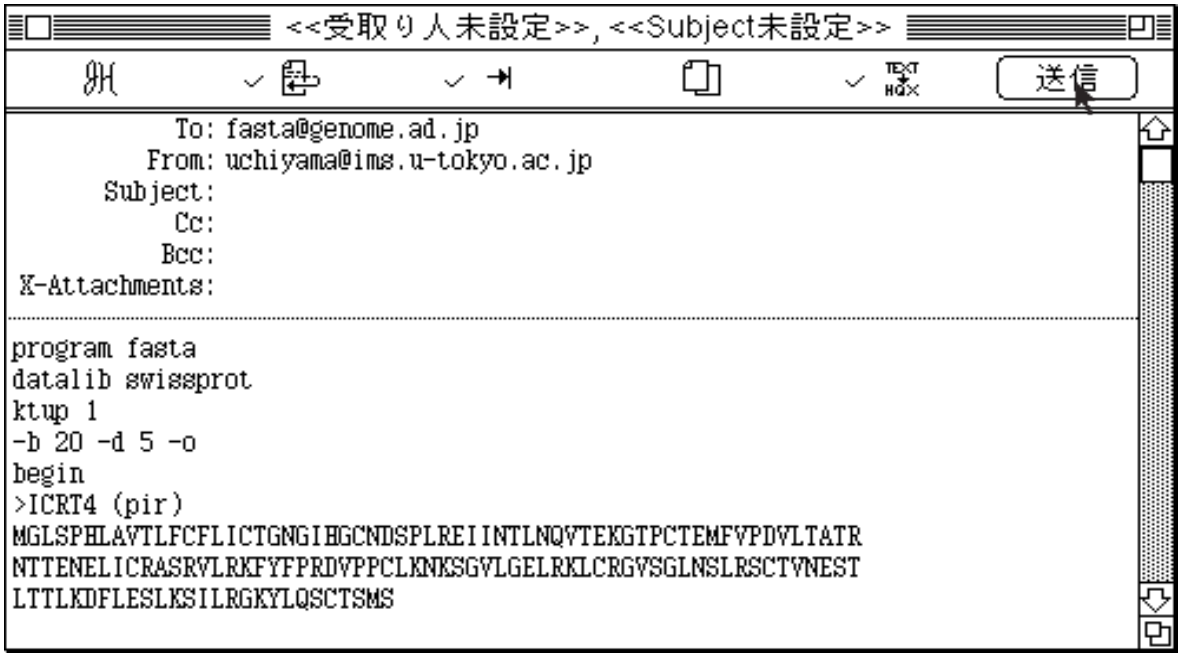
(sequence)

オプションについては後述しますが、ここではスコアを２０個、アライメントを５個表示して、最適化されたスコアでランクづけするようなオプションを指定しています。

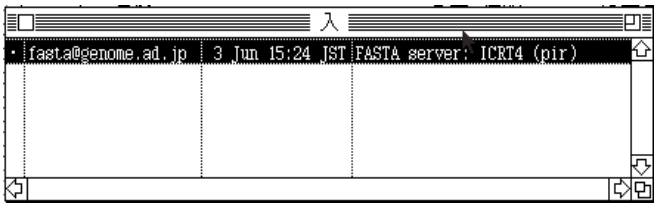
2.2.2　リクエストメールの発送と結果の確認

・Eudoraを使う場合

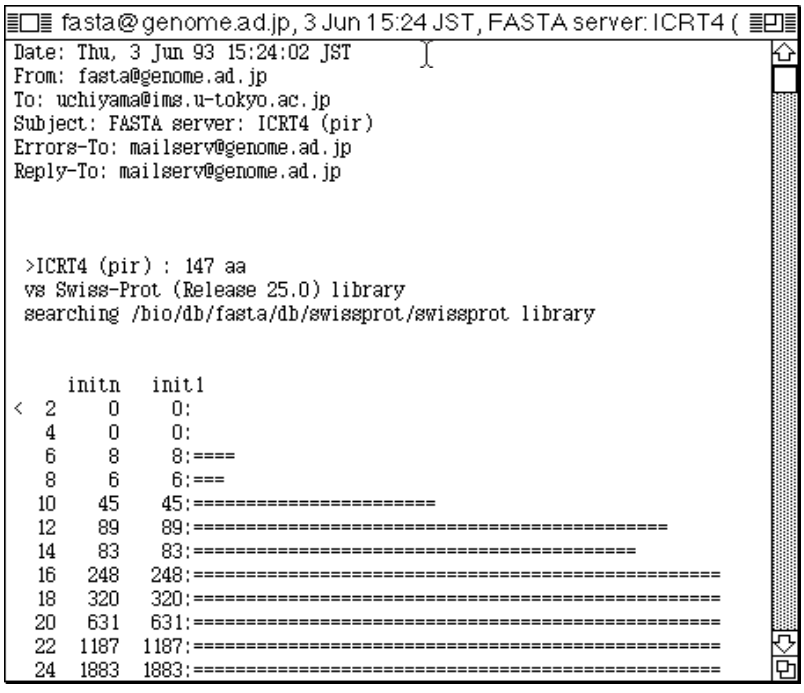
「メッセージ」のプルダウンメニューから「新規メッセージ」を選択し、以下のようなウインドウが表示されたら、To: フィールドにfasta@genome.ad.jp、点線以下にリクエストを入力してください。



送信をクリックしてリクエストを送信します。



しばらくすると返事のメールの到着を知らせるメッセージが表示され、続いて左図のようなウインドウが現われます。今届いた検索結果のメールが反転していますのでこれをダブルクリックすると内容を見ることができます。



・MHを使う場合

```
hgc{lect}% emacs                    ( emacs エディタの起動 )
(ESC) x mh-smail                    ( MHの起動 )
To: fasta@genome.ad.jp              ( メールサーバーのアドレス )
Cc: (return)
Subject: (return)
ここでTo:, cc:, Subject: が一覧表示されますので----以下にリクエストを入力して下さい。

program fasta                       |
datalib swissprot                   |
ktup 1                              |( リクエストの入力 )
-b 20 -d 5 -o                       |
begin                               |
(Control-x) i testseq.aa            (配列をファイルから取り込む)
(Control-c) (Control-c)             (メールの発送 )
```

※配列をファイルから取り込むのではなく、新しく入力する場合は、 beginに続けて入力して下さい。
※ Subjectは特に入力する必要はありませんが、何か入力しておくと返送される検索結果のメールのSubjectが「Re:（リクエストメールのSubject）」となりますので、検索結果を整理する際に便利です。

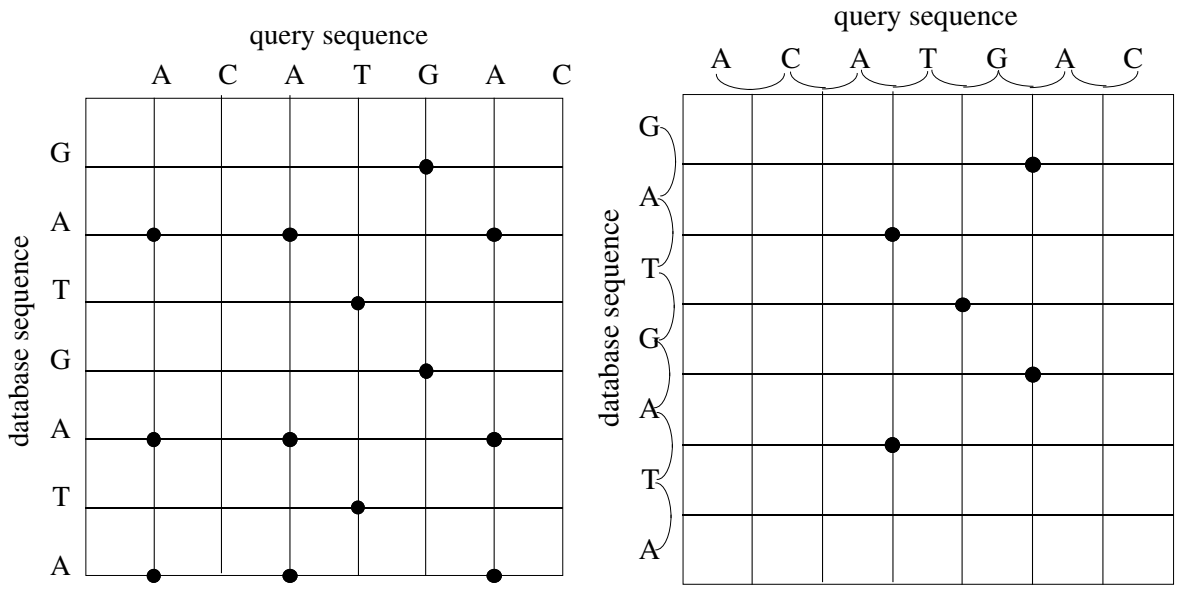
メールが返ってきたら、inc, showで確認します。

```
hgc{lect}107: inc
Incorporating new mail into inbox...

   1+ 04/21 fasta@genome.ad.jp   ICRT4 (pir)
hgc{lect}108: show
Date: Wed, 21 Apr 93 17:45:47 JST
From fasta@genome.ad.jp Tue Apr 20 08:50:57 1993
Message-Id: <9304210917.AA27366@ss05>
To: lect@ims.u-tokyo.ac.jp
Subject: FASTA server: ICRT4 (pir)
Errors-To: mailserv@genome.ad.jp
Reply-To: mailserv@genome.ad.jp

 >ICRT4 (pir) : 147 aa                                    |
vs Swiss-Prot (Release 24) library                        |
searching swissprot library                               |
                                                          |
     initn   init1                                        |     ( FASTAの結果 )
<  2     0      0:                                        |
   4     0      0:                                        |
   6     7      7:====                                    |
   8     4      4:==                                      |
  10    43     43:======================                  |
```

2.3　FASTAのアルゴリズム

2.3.1 ktup

FASTAでは検索の際パラメータとしてktup値を指定します。
ktup値は初期検索を行なう際に照合の単位とする文字列の長さで、ktup値が大きいほど粗い検索を行なうことになります。



Lookup Table (ktup = 1)

A	1-3-6
T	4
G	5
C	2-7

Lookup Table (ktup = 2)

AA	
AT	3
AG	
AC	1-6
TA	
TT	
:	:

ktup2のときはホモロジーの低い領域が無視されるので、検索は高速になります。

2.3.2　スコアの計算

FASTAでは 3 つのスコアが定義されています。

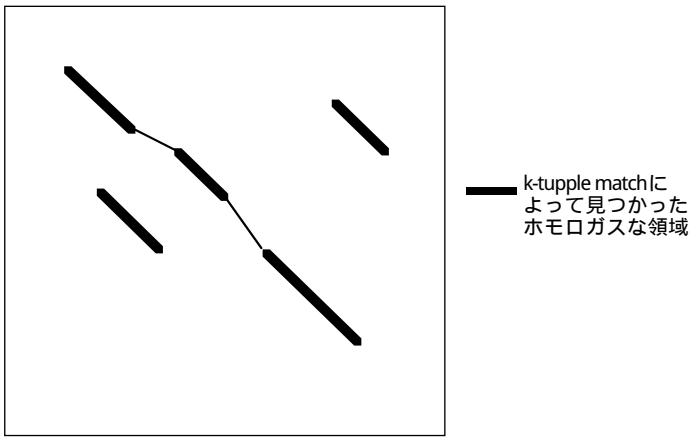
init1　　初期検索で見つかった領域のうち最も高いスコアを持つ領域のスコア

initn　　初期検索で見つかった領域を複数つなぎ合わせて得られる領域のうち最も高いスコアを持つ領域のスコア

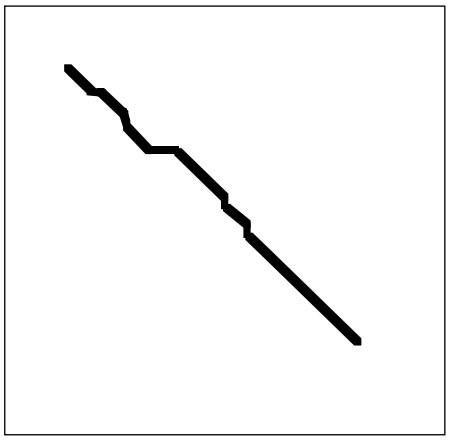
opt　　ダイナミックプログラミング法によって得られたギャップを入れた最適なアライメントのスコア

optスコアが最も正確なホモロジースコアですが、通常はinitnのスコアに基づいてランキングし、上位のエントリーのみについてアライメントを行ないます。
-o オプションを指定するとoptスコアに基づいたランクづけを行ないます。（精度は上がりますが時間が 2 倍程度かかります）

ホモロガスな領域の結合 (initn score)



最適アライメント (opt score)



2.4　FASTA の結果

FASTAの結果はスコアのヒストグラム、ホモロジーのランキングとアライメントからなります。

2.4.1　ヒストグラムとホモロジーのランキング

```
     initn   init1
<  2    25    25:=============
   4     3     3:==
   6    15    15:========
   8    31    31:================
  10   123   123:==================================================
  12   416   416:==================================================
  14   333   333:==================================================
     :
     :
  70     3     0:++
  72     1     0:+
  74     6     0:+++
  76     0     0:
  78     2     0:+
  80     0     0:
> 80     4     3:==
9545403 residues in 28154 sequences
 statistics exclude scores greater than 72
 mean initn score:  24.5 (7.35)
 mean init1 score:  24.0 (6.15)
 4504 scores better than 30 saved, ktup: 2, variable pamfact
 joining threshold: 27  scan time:  0:00:27
The best scores are:                                      initn init1   opt
IL4_RAT INTERLEUKIN-4 PRECURSOR (IL-4) (B-CELL STIMULA  747   747   747
IL4_MOUSE INTERLEUKIN-4 PRECURSOR (IL-4) (B-CELL STIMU  439   439   451
IL4_HUMAN INTERLEUKIN-4 PRECURSOR (IL-4) (B-CELL STIMU  300   146   314
RPB3_HUMAN DNA-DIRECTED RNA POLYMERASE II 33 KD POLYPE   91    65    68
ACHG_TORCA ACETYLCHOLINE RECEPTOR PROTEIN, GAMMA CHAIN   77    39    49
CO6_HUMAN COMPLEMENT C6 PRECURSOR                        77    47    47
ACHE_MOUSE ACETYLCHOLINE RECEPTOR PROTEIN, EPSILON CHA   74    39    52
ACHE_BOVIN ACETYLCHOLINE RECEPTOR PROTEIN, EPSILON CHA   74    39    50
```

ヒストグラムはデータベース中でのinit1, initnのスコアの値の分布を示します。
ヒストグラム中「+」はinitn、「-」はinit1の分布を表わします。initnとinit1の数が等しいときは「=」で分布が表わされます。

ランキングはinitnの値によって行なわれていますが、optの方が一般にはより正確なスコアです。特にinitnに比べてoptの方が大分大きくなっている場合は、ランクの割に類似性が高いと考えられますので注意してください。

2.4.2　アライメント

```
IL4_HUMAN INTERLEUKIN-4 PRECURSOR (IL-4) (B-CELL STIMU  300  146  314
  43.5% identity in 154 aa overlap

               10         20          30         40         50
ICRT4  MGLSPHLAVTLFCFLICTGNGIHGCN-DSPLREIINTLNQVTEKGTPCTEMFVPDVLTAT
       :::...:  .:: .: :.:: .:: . : .:.X:::::: .:::. : :::. :.:...:.
IL4_HU MGLTSQLLPPLFFLLACAGNFVHGHKCDITLQEIIKTLNSLTEQKTLCTELTVTDIFAAS
               10         20         30         40         50         60

      60         70         80                 90        100        110
ICRT4  RNTTENELICRASRVLRKFYFPRD-VPPCL-------KNKSGVLGELRKLCRGVSGLNSL
       .:::::.: .::::. ::::.:X ...  ..::       .... ..  :..: :.. ::..:
IL4_HU KNTTEKETFCRAATVLRQFYSHHEKDTRCLGATAQQFHRHKQLIRFLKRLDRNLWGLAGL
               70         80         90        100        110        120

              120        130        140
ICRT4  RSCTVNESTLTTLKDFLESLKSILRGKYLQSCTSMS
       .::.:.:.. .::..:::.::.:.:.::  .:.:
IL4_HU NSCPVKEANQSTLENFLERLKTIMREKY-SKCSS
               130        140        150
```

アライメントは局所的に類似性の高い領域のみで行なわれます。アライメントされた領域の実際の配列中での位置は配列の上または下に示されています。
':'はアミノ酸が一致していることを、'.'は類似したアミノ酸であること (スコアが 0 以上) であることを表わしています。
'X'は初期検索で見つかった領域 (init1)の始まりと終わりを表わしています。

(参考)
スコアはアミノ酸置換行列 (PAM250) の加算により計算されます。

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X
A  2
R -2  6
N  0  0  2
D  0 -1  2  4
C -2 -4 -4 -5 12
Q  0  1  1  2 -5  4
E  0 -1  1  3 -5  2  4
G  1 -3  0  1 -3 -1  0  5
H -1  2  2  1 -3  3  1 -2  6
I -1 -2 -2 -2 -2 -2 -2 -3 -2  5
L -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6
K -1  3  1  0 -5  1  0 -2  0 -2 -3  5
M -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6
F -3 -4 -3 -6 -4 -5 -5 -5 -2  1  2 -5  0  9
P  1  0  0 -1 -3  0 -1  0  0 -2 -3 -1 -2 -5  6
S  1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  2
T  1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -3  0  1  3
W -6  2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4  0 -6 -2 -5 17
Y -3 -4 -2 -4  0 -4 -4 -5  0 -1 -1 -4 -2  7 -5 -3 -3  0 10
V  0 -2 -2 -2 -2 -2 -2 -1 -2  4  2 -2  2 -1 -1 -1  0 -6 -2  4
B  0 -1  2  3 -4  1  3  0  1 -2 -3  1 -2 -4 -1  0  0 -5 -3 -2  3
Z  0  0  1  3 -5  3  3  0  2 -2 -3  0 -2 -5  0  0 -1 -6 -4 -2  2  3
X  0 -1  0 -1 -3 -1 -1 -1 -1 -1 -1 -1 -1 -2 -1  0  0 -4 -2 -1 -1 -1 -1
```

2.5　FASTA の主なオプション

-b＃　　表示するスコアの数
-d＃　　表示するアライメントの数
-o　　最適化されたスコア(OPT)に基づいてランク付けする
-c＃　　-oで最適化する際のinitn値の閾値の設定
-n　　問い合わせ配列をDNA配列と見なす
-s file　　置換のスコア行列の設定

詳しくは man fastaでマニュアルを参照してください。

2.6　まとめ

使用法　　（[]内は省略可能）

```
fasta [options]
```

または

```
fasta [options] query_file data_library [ktup]
```

検索の特徴

初期検索はk-tuple matchを用いており、KTUP値で精度、速度の調節が可能。
複数の領域を結合して求めたinitn scoreをもとにしてホモロジーのランクづけを行なう。
最終的にはダイナミックプログラミング法でギャップを入れた最適なアライメントを行なう。

３．BLASTによる検索

3.1　シェル上からの起動

blastには４種類のプログラムがありますがいずれも以下の形式で起動します。

プログラム名　データベース名　問い合わせ配列ファイル名

blastの結果は標準出力に出力されますので、リダイレクション(>)でファイルに書き出すようにして下さい。

blastp　　蛋白質配列×蛋白質データベース

```
blastp pir testseq.aa > result.blastp
```

blastn　　核酸配列×核酸データベース

```
blastn genbank testseq.nt >result.blastn
```

blastx　　核酸配列×蛋白質データベース

```
blastx swissprot testseq.nt >result.blastx
```

tblastn　　蛋白質配列×核酸データベース

```
tblastn embl testseq.aa >result.tblastn
```

（参考）
オプションを指定するときは問い合わせ配列ファイル名のあとで指定します。

```
blastp pir testseq.aa V=20 B=10
```

（スコアを２０個、アライメントを１０個表示します）

3.2　メールサーバー

blast@genome.ad.jpに以下のような内容のメールを送ってください。

```
program blastp
datalib nr-nt
H=1 V=20 B=5
begin
>ICRT4 (pir)
MGLSPHLAVTLFCFLICTGNGIHGCNDSPLREIINTLNQVTEKGTPCTEMFVPDVLTATR
NTTENELICRASRVLRKFYFPRDVPPCLKNKSGVLGELRKLCRGVSGLNSLRSCTVNEST
LTTLKDFLESLKSILRGKYLQSCTSMS
```

以下のキーワードを指定してください。

program　　blastp,blastn,blastx,tblastn
datalib　　nr-aa,pir,swissprot,prf(蛋白質)
　　nr-nt,genbank,embl(核酸)
(options)
begin
(sequence)

オプションについては後述しますが、ここではヒストグラムを表示し、スコア20個、アライメント5個を表示するようなオプションを指定しています。

メールの発送と結果の確認についてはFASTAの場合と同じです。

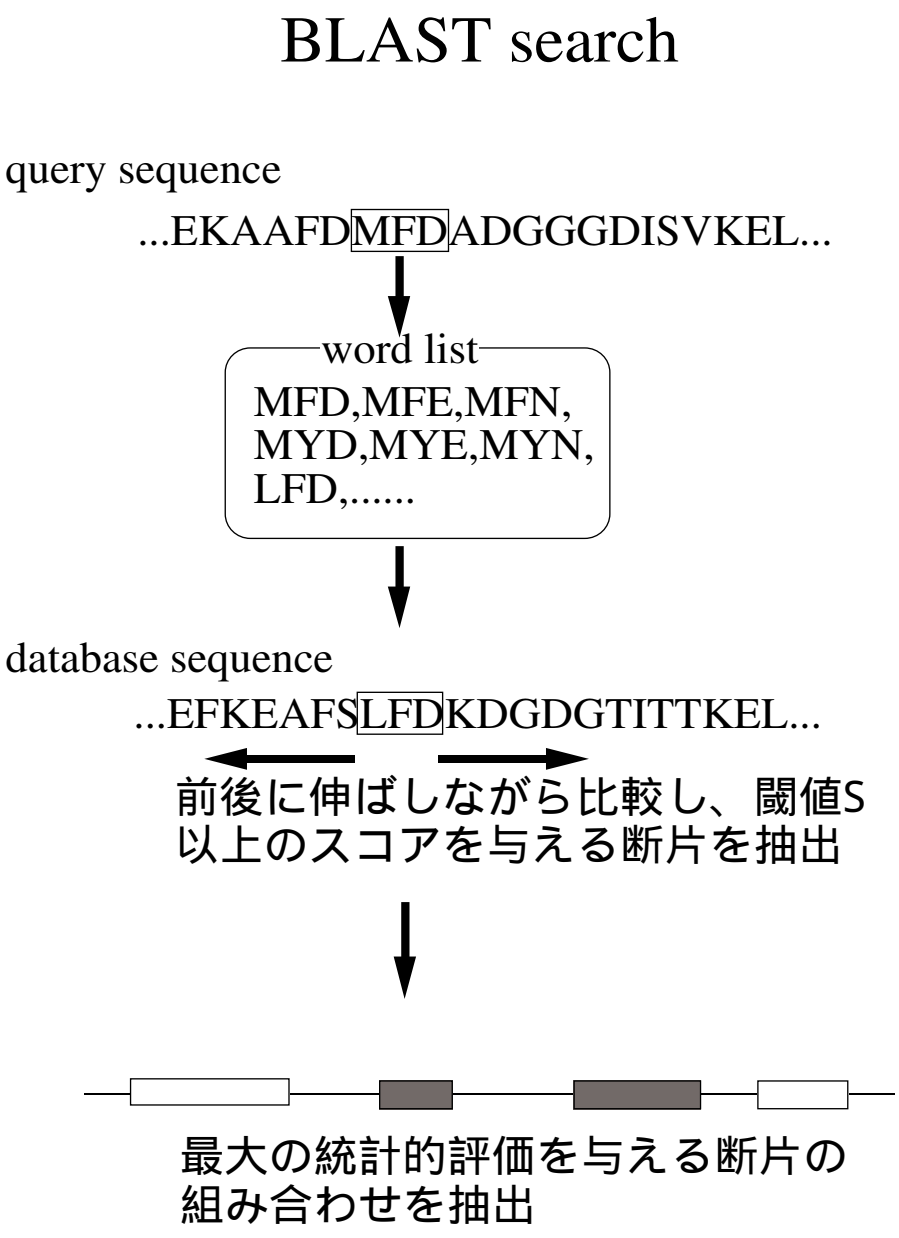
3.3 BLASTのアルゴリズム

BLASTではあらかじめ問い合わせ配列から固定長の配列断片（ワード）を抜き出し、各ワードごとに類似のワードを列挙してワードのリストを生成します。初期検索はこれらのワードの検索になります。

ワードが見つかったら前後に領域を伸ばしながら類似性スコアを計算して、スコアが閾値を超える領域を抽出します。このようにして抽出された、ギャップなしで類似性の高い配列断片の対をHigh-scoring Segment Pair (HSP) と呼んでいます。

全体の類似性の評価は、見つかったHSPの組に対する統計的評価で行なわれます。これは、それ以上のスコアを持つ場合が偶然に起こる確率をもとにして行なわれます（確率が小さいほど有意に類似しているということになります）。複数のHSPが見つかった場合は、最大の統計的評価を与えるような領域の組み合わせを求めます。

なお、BLASTでは最後までギャップを入れたアライメントは行ないません。



3.4　BLAST の結果

3.4.1　スコア表示

BLASTでは類似性の評価に"High score"、"Smallest Poisson Probability"の 2 つの値を用いています。

High scoreは最も類似性の高い領域のスコアです。

Smallest Poisson Probabilityは、最も高い統計的評価を与える断片対について、断片対の数Nと、そのスコアが偶然にN個以上出現する確率P(N)で表わされています。

ランキングはP(N)の小さい順に行なわれます。

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Poisson Probability P(N)	N
ICRT4 Interleukin-4 precursor - Rat | 960.0 1.0 2.0 1.0 1.0	772	4.9e-103	1
IIMSG1 Interleukin-4 precursor - Mouse | 960.0 1.0 1.0 1....	414	1.1e-52	1
S03872 *Interleukin-4 precursor - Mouse	258	5.8e-30	1
A30546 Interleukin-4 - Human	156	6.6e-18	2
A25946 Lymphocyte stimulatory factor 1 precursor - Human	156	6.6e-18	2
JH0684 *Interleukin 4 precursor - Bovine	116	2.1e-11	3
GNVSTV Genome polyprotein - Tobacco vein mottling virus |...	58	0.78	1
A36264 *DNA-directed RNA polymerase II 33K chain - Human ...	38	0.79	3
VCBED6 Major capsid protein - Equid herpesvirus 1 (strain...	41	0.86	3
B39987 *Hemagglutinin precursor - Influenza A virus (stra...	37	0.90	4
F33157 *Hemagglutinin precursor - Influenza A virus (stra...	37	0.90	4
TVHUFF Protein-tyrosine kinase fes/fps - Human #EC-number...	52	0.92	2
A40937 Bullous pemphigoid antigen precursor - Human	56	0.95	1
A40882 *Bullous pemphigoid antigen - Human (fragment)	56	0.95	1

（参考）

BLASTでは統計的評価はデータベースの大きさや残基の出現頻度などから理論的に計算されます。表示されるスコアの閾値も、検索の結果まったく偶然に見つかる数の期待値が１０個であるようなスコアの値が理論的に計算されて設定されます。

3.4.2　アライメント

各High-scoring Segment Pairに対応するアライメントが統計的評価値とともに複数表示されます。

```
>A30546 Interleukin-4 - Human
        Length = 153

 Score = 156 (72.9 bits), Expect = 1.3e-14, P = 1.3e-14
 Identities = 30/53 (56%), Positives = 39/53 (73%)

Query:    27 DSPLREIINTLNQVTEKGTPCTEMFVPDVLTATRNTTENELICRASRVLRKFY 79
             D  L+EII TLN +TE+ T CTE+ V D+  A++NTTE E  CRA+ VLR+FY
Sbjct:    28 DITLQEIIKTLNSLTEQKTLCTELTVTDIFAASKNTTEKETFCRAATVLRQFY 80

 Score = 101 (47.2 bits), Expect = 1.1e-06, Poisson P(2) = 6.6e-18
 Identities = 21/42 (50%), Positives = 30/42 (71%)

Query:    98 LRKLCRGVSGLNSLRSCTVNESTLTTLKDFLESLKSILRGKY 139
             L++L R + GL  L SC V E+  +TL++FLE LK+I+R KY
Sbjct:   107 LKRLDRNLWGLAGLNSCPVKEANQSTLENFLERLKTIMREKY 148

 Score = 61 (28.5 bits), Expect = 0.57, Poisson P(3) = 3.8e-12
 Identities = 12/24 (50%), Positives = 14/24 (58%)
```

BLASTでは核酸配列の場合、相補鎖も同時に検索します。プラス・マイナスどちらのストランドで見つかったかはアライメントの前に表示されます。

```
>gb:GIRRNATR Giardia intestinalis DNA for rRNA tandem repeat unit
            Length = 5566

  Minus Strand HSPs:

 Score = 123 (34.0 bits), Expect = 0.60, P = 0.45
 Identities = 35/48 (72%), Positives = 35/48 (72%), Strand = Minus

Query:   194 GGGCTCGGCGGCCGCGGGGGCGAAGGCCGCGCTGGGGCCGCGGTGGCC 147
             || | | | |||| |||||||||||| ||||   |||| ||  || |
Sbjct:  2107 GGCCGCAGAGGCCCCGGGGGCGAAGGCGGCGCGCAGGCCCCGCCGGAC 2154
```

3.5　BLASTの 主なオプション

H=#	ヒストグラム表示 (H≠ 0のとき)
V=#	スコア表示数
B=#	アライメント表示数
S=#	カットオフスコア
E=#	スコアの見つかる個数の期待値(default 10) （カットオフスコアSが自動的に設定される）
W=#	初期検索に用いるword のサイズ
T=#	word list に加えるスコアの閾値
M=file	matrix file の指定
C=#	genetic codeの指定

詳しくは man blast でマニュアルを参照してください。

3.6　まとめ

使用法

```
blast database query_file [options]
```

検索の特徴

初期検索は置換を許したword search
ギャップなしで閾値を越えるスコアを与える領域を抽出。
複数の領域が見つかったときは組み合わせをとって統計的評価を行ない、それをもとにランク付けする。
一般にFASTAより高速。特に核酸データベースの検索は速い。

４．検索のヒント

・一般的にはより高速なBLASTをまず用いる。
・特に弱いホモロジーの検出をしたい場合はFASTAのKTUP=1で-oオプションをつけた検索を試してみる。（蛋白質配列の場合）
・FASTAで核酸配列データベースを検索する場合、あまり低いKTUP値は使用しない。（通常４～６の範囲を用いる）
・MOTIF検索と組み合わせて使用するのも効果的。

参考文献

FASTA
"Improved tools for biological sequence comparison"
W.R. Pearson and D.J. Lipman (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444-2448
BLAST
"Basic Local alignment search tool"
S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman (1990) *J. Mol. Biol.* **215**, 403-410