Lancer FASTA et BLAST en ligne de commande

Daniel Gautheret
ESIL, Université de la Méditerranée

V.2006.1 http://www.esil.univ-mrs.fr/~dgaut/Cours

Fasta

- ★ Article original: Lipman and Pearson (1985) Science 227:1435-1441
- ★ Heuristique: recherche d'abord des segments de longueur k exactement semblables (k-mots)
- ★ Raccorde ces segments si présents sur une même diagonale où sur des diagonales proches, puis réaligne la région par programmation dynamique.

- ★ Versions 2.x: 1988 à 1996
- ★ Version 3.x: 1996-
- ★ Actuellement: 3.4

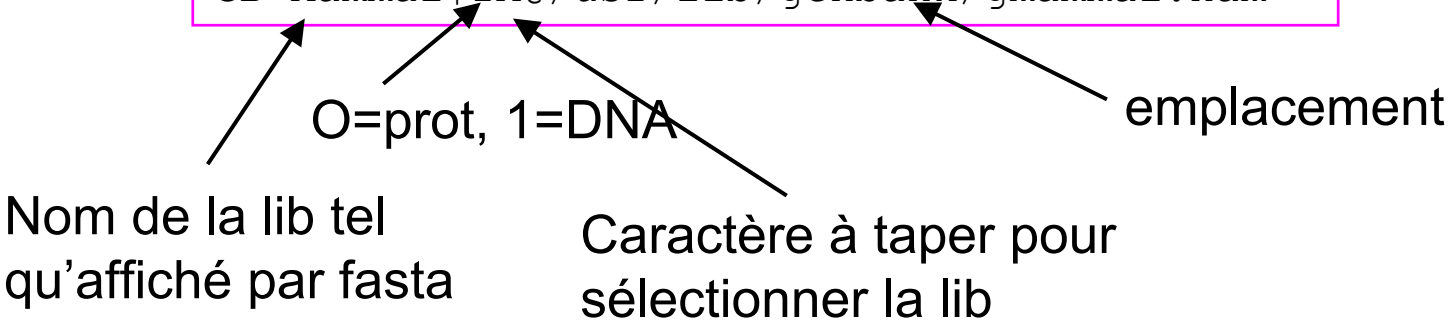
Les différents programmes

- ★ Library search programs: FASTA, FASTX, TFASTA, TFASTX, SSEARCH
- ★ Local homology programs: LFASTA, PLFASTA, LALIGN, PLALIGN, FLALIGN
- ★ Statistical significance: PRDF, RELATE, PRSS, RANDSEQ
- ★ Global alignment: ALIGN

La variable d'environnement FASTLIBS

- All the search programs in the FASTA3 package use the environment variable FASTLIBS to find the protein and DNA sequence libraries.
- The FASTLIBS variable contains the name of a file that has the actual filenames of the libraries.
  - an example of a file that can be referred to by FASTLIBS:

```
NBRF Protein$0P/usr/lib/seq/aabank.lib 0
SWISS PROT 10$0S/usr/lib/vmspir/swiss.seq 5
GB Primate$1P@/usr/lib/genbank/gpri.nam
GB Rodent$1R@/usr/lib/genbank/grod.nam
GB Mammal$1M@/usr/lib/genbank/gmammal.nam
```



Pour affecter la variable: `export FASTLIBS=/usr/lib/fasta/fichier-fastlibs`

Le paramètre ktup

Search speed and selectivity are controlled with the ktup(wordsize) parameter

- ★ For protein comparisons, ktup = 2 by default; ktup =1 is more sensitive but slower.
- ★ For DNA comparisons, ktup=6 by default; ktup=3 or ktup=4 provides higher sensitivity;
- ★ ktup=1 should be used for oligonucleotides (DNA query lengths < 20).

La ligne de commande de Fasta 3.x

- `fasta musplfm.aa prot_test.lib 1`
- `fasta musplfm.aa prot_test.lib > toto.out`
  - (ne pose pas de questions)
- `fasta musplfm.aa prot_test.lib -s pam250.mat`
  - (change la matrice)
- `fasta musplfm.aa prot_test.lib -E 2.0`
  - (remplace E-value par défaut de 10.0)

Banque contenant plusieurs fichiers

- You can specify a group of library files by putting a '@'symbol before a file that contains a list of file names to be searched. For example, if @gmam.nam is in the fastgbs file, the file "gmam.nam" might contain the lines:

```
</seqdb/genbank
gbpri1.seq 1
gbpri2.seq 1
gbpri3.seq 1
gbpri4.seq 1
gbrod.seq 1
gbmam.seq 1
```

- In this case, the line beginning with a '<' indicates the directory the files will be found in. The remaining lines name the actual sequence files. So the first sequence file to be searched would be:

```
/usr/lib/genbank/gbpri.seq
```

L'histogramme de Fasta

```
 opt      E()
< 20     0     0:
  22     0     0:           one = represents 3 library sequences
  24     0     0:
  26     0     0:
  28     0     0:
  30     2     2:*
  32     6     9:==*
  34    18    24:====== *
  36    37    50:=============   *
  38    70    82:========================   *
  40   117   114:=====================================*=
  42   144   140:============================================*=
  44   163   154:=================================================*===
  46   154   157:================================================*
  48   176   150:================================================*=========
  50   158   137:============================================*=======
  52   129   120:=======================================*===
  54   115   103:==================================*====
  56    99    86:============================*====
  58    53    71:=================     *
  60    40    57:==============    *
  62    36    46:============   *
  64    38    36:===========*=
  66    21    29:=======  *
  68    23    23:=======*
  70    20    18:=====*=
  72    13    14:====*
  74     6    11:== *
  76     4     8:==*
  78    11     7:==*=
  80     3     5:=*
  82     3     4:=*
  84     3     3:*
  86     0     2:*
  88     3     2:*          inset = represents 1 library sequences
  90     2     1:*
  92     0     1:*         :*
  94     0     1:*         :*
  96     0     1:*         :*
  98     0     1:*         :*
 100     0     0:          *
 102     0     0:          *
 104     0     0:          *
 106     1     0:=         *=
 108     0     0:          *
 110     0     0:          *
 112     1     0:=         *=
 114     0     0:          *
 116     0     0:          *
 118     1     0:=         *=
>120    39     0:==========*=======================================
```

Asterisques=attendu (extreme value distribution)
Tirets=observé

Bien !!

Blast (Lipman, Karlin, Altschul, 1990)

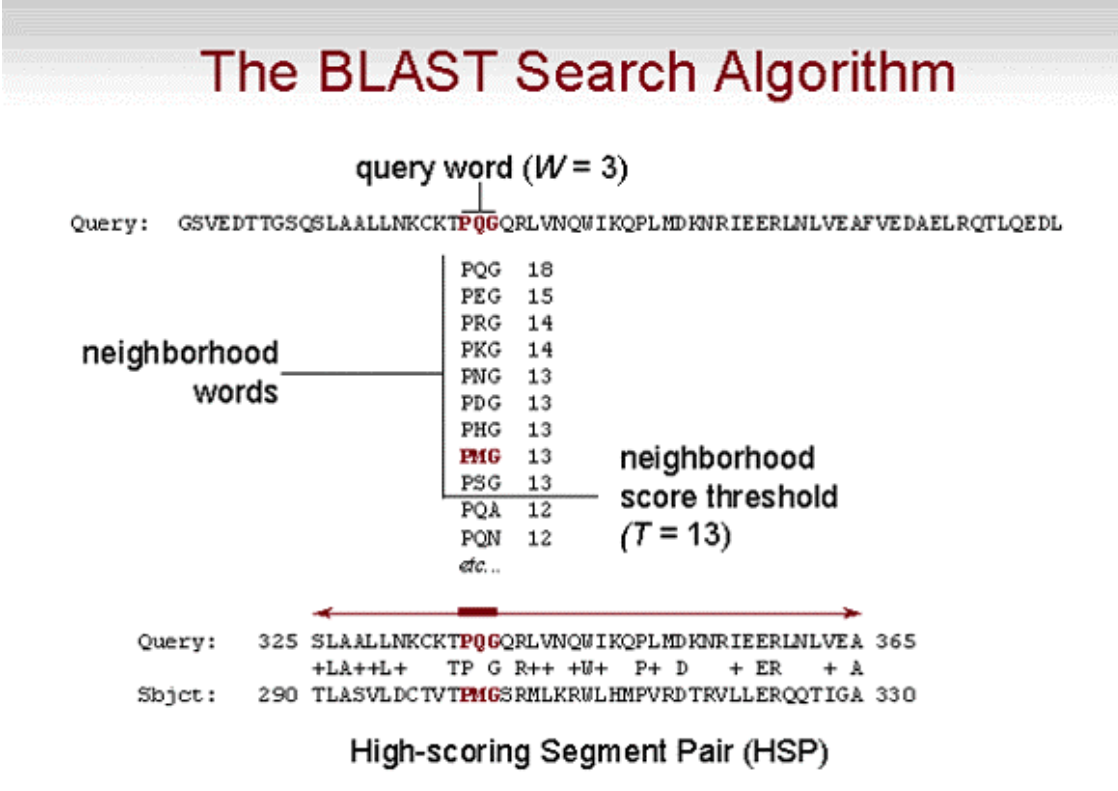
★ Le plus utilisé des programmes d'alignement local

- k-mots également, mots approchés permis au dessus d'un certain score.
- Pré-codage de la base de données et de la requête pour recherche plus rapide des k-mots.
- Version 1: sans Gaps
- Version 2: avec Gaps

★Points forts

•Rapidité

•Calcul de la valeur statistique des scores.



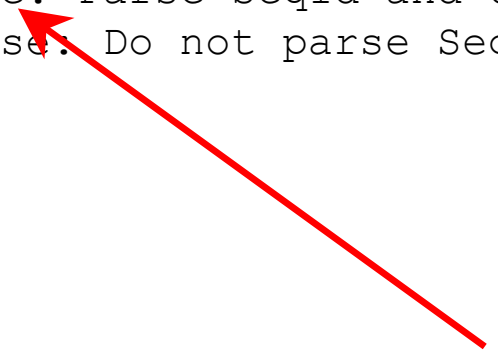
Le fichier .ncbirc

```
[NCBI]
Data=/export/homes/personnels/gbma/dgaut/Bin
[BLAST]
BLASTDB=.
```

La commande formatdb

```
formatdb 2.2.6   arguments:

  -i  Input file(s) for formatting [File In]  Optional
  -p  Type of file
         T - protein
         F - nucleotide [T/F]  Optional
    default = T
  -o  Parse options
         T - True: Parse SeqId and create indexes.
         F - False: Do not parse SeqId. Do not create indexes.
 [T/F]  Optional
    default = F
```



Cet index optionnel permet ensuite de récupérer des séquences à partir de leur « id » (par fastacmd)

La commande blastall (sous-ensemble des options)

```
blastall 2.2.6   arguments:

  -p  Program Name [String]
  -d  Database [String]
    default = nr
  -i  Query File [File In]
    default = stdin
  -e  Expectation value (E) [Real]
    default = 10.0
  -F  Filter query sequence (DUST with blastn, SEG with others)
    default = T
  -v  Number of database sequences to show one-line descriptions for
    default = 500
  -b  Number of database sequence to show alignments for (B) [Integer]
    default = 250
  -M  Matrix [String]
    default = BLOSUM62
  -W  Word size, default if zero (blastn 11, megablast 28, all others 3)
    default = 0
  -n  MegaBlast search [T/F]
    default = F
```

MEGABLAST

- ★ Megablast est une version de BLASTN jusqu'à 10x plus rapide, adaptée à l'alignement de séquences très semblables (par ex. ne différent que par des erreurs de séquençage)
- ★ Différences algorithmiques:
  - Taille des mots par défaut=28 (au lieu de 11)
  - Greedy algorithm au lieu de programmation dynamique pour extension
  - Calcul des gaps plus rapide

BLASTZ ★ Schwartz et al. Genome Res. 2003

★ Utilisé pour aligner de longues régions génomiques, avec des parties peu semblables

★ Même principe que BLAST, avec modification suivantes:

Recherche des paires de 12-mer agencées dans le même ordre

1. Remove lineage-specific interspersed repeats from both sequences.
2. For all pairs of spaced 12-mers (one from each sequence) that are identical except perhaps for one transition, do the following.
   - 2.1 Extend the induced alignment in each direction, not allowing gaps. Stop extending when the score decreases more than some threshold.
   - 2.2 If the gap-free alignment scores more than 3000 (say) then
     - 2.2.1. Repeat the extension step, but allow for gaps.
     - 2.2.2. Retain the alignment if it scores above 5000 (say).
3. Between each pair of adjacent alignments from step 2, repeat step 2, but using a more sensitive seeding procedure (e.g., 7-mer exact matches) and lower score thresholds both for gap-free alignments (say, 2000 instead of 3000) and for gapped alignments (say, 2000 instead of 5000).
4. Adjust sequence positions in the resulting alignments to make them refer to the original sequences (i.e., account for step 1).
5. Filter the alignments as appropriate for particular purposes. For many uses we apply axtBest, which finds a best way to align each aligned human position. For other studies, such as mapping segmental duplications, other strategies are appropriate.

Figure 1 BLASTZ in a nutshell.

★ BLASTZ utilisé dans serveurs Web d'alignement de 2 génomes (pipmaker) ou N génomes (multipipmaker)