Note du 19 mai 2011

La version de BLAST utilisée dans ce document de mars 2010 est **obsolète**. Les nouveaux utilisateurs de BLAST devraient directement apprendre et utiliser les commandes BLAST+. L'intégralité des commandes est expliquée dans l'aide du site du NCBI *http://www.biomedcentral.com/1471-2105/10/421/additional/*.

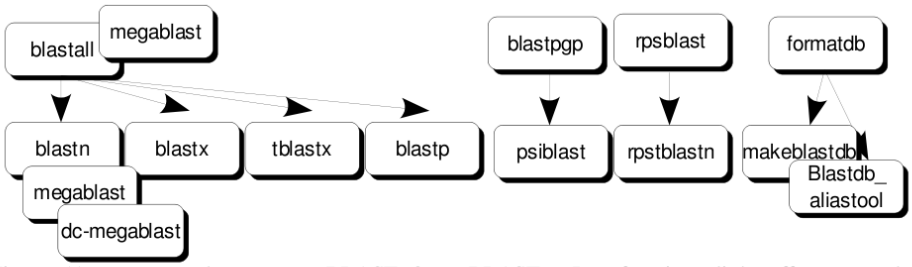


Figure 11 : correspondance entre BLAST 2 et BLAST+. Les fonctionnalités offertes par les applications BLAST+ sont organisées par types de programmes.

Formatage de la banque

Paquet BLAST < *v.*2.2.21 : formatdb -i genome.fasta -t nom_genome -b F -n nom_genome
doit être remplacé par :
Paquet BLAST+

```
makeblastdb -in genome.fasta -dbtype nucl -title nom_genome -out nom_genome_blastdb
```

Recherche par similarité

Paquet BLAST < v.2.2.21

```
blastall -p blastn -i seq.fasta -d mabanque -o resultat -W 11 -y 20 -X 30 -Z 100 -G 5 -E 2 -r 2 -q -3
```

doit être remplacé par :
Paquet BLAST+

```
blastn -query seq.fasta -db mabanque -out resultat -wordsize 11 -xdrop_ungap 20 -xdrop_gap 30
-xdrop_gap_final 100 -gapopen 5 -gapextend 2 -reward 2 -penalty -3
```

A noter également : l'option « f » (le score minimal pour ajouter un mot à la liste de mots voisins pour les protéines) de blastall devient « threshold » dans blastp. L'option « M » (par exemple -M BLOSUM62) devient « matrix » (l'usage devient donc : -matrix BLOSUM62). Enfin, le programme netblast devient obsolète avec l'introduction de l'option « remote » pour effectuer la recherche directement avec le moteur compilé sur les serveurs propres du NCBI, et utiliser leurs bases de données.

Autre option très importante : -soft_masking true, quand utilisé, le masquage dur (recherche + extension) est désactivé au profit du masquage léger (les portions masquées de la requête sont inutilisables pendant la phase de recherche des touches initiales mais sont disponibles pour les phases d'extension). *Soft_masking false* n'existe pas, ce serait une erreur de compréhension de la façon dont est utilisée l'option « soft masking true ».

Depuis la publication de ce document les explications du choix du seuil pour étendre les hits, du X-dropoff pour alignement avec trous, du Z dropoff pour alignements finaux avec gap ont été traités en relation avec l'algorithme de BLAST dans le document **« La galaxie BLAST : 20 ans d'utilisation de la méthode BLAST »**. Une version publique de ce document est disponible ici : *http://gf3.myriapyle.net/gcde/data/BLAST_en_ligne.pdf*.

4.2 Common options

http://www.ncbi.nlm.nih.gov/books/NBK1763/

The following is a listing of options that are common to the majority of BLAST+ applications followed by a brief description of what they do:

4.2.1 best_hit_overhang: Overhang value for Best-Hit algorithm. For more details, see the section Best-Hits filtering algorithm.

4.2.2 best_hit_score_edge: Score edge value for Best-Hit algorithm. For more details, see the section Best-Hits filtering algorithm.

4.2.3 db: File name of BLAST database to search the query against. Unless an absolute path is used, the database will be searched relative to the current working directory first, then relative to the value specified by the BLASTDB environment variable, then relative to the BLASTDB configuration value specified in the configuration file. Multiple databases may be provided as an argument, and they must be separated by a space. Many operating systems now allow spaces in file names and paths, so it is necessary to use quotes. See section 5.15 for details.

4.2.4 dbsize: Effective length of the database.

4.2.5 dbtype: Molecule type stored or to store in a BLAST database.

4.2.6 db_soft_mask: Filtering algorithm ID to apply to the database as soft masking for subject sequences. The algorithm IDs for a given BLAST database can be obtained by invoking blastdbcmd with its -info flag (only shown if such filtering in the BLAST database is available). For more details see the section Masking in BLAST databases.

4.2.7 culling_limit: Ensures that more than the specified number of HSPs are not aligned to the same part of the query. This option was designed for searches with a lot of repetitive matches, but if possible it is probably more efficient to mask the query to remove the repetitive sequences.

4.2.8 entrez_query: Restrict the search of the BLAST database to the results of the Entrez query provided.

4.2.9 evalue: Expectation value threshold for saving hits.

4.2.10 export_search_strategy: Name of the file where to save the search strategy (see section titled BLAST search strategies).

4.2.11 **gapextend**: Cost to extend a gap.

4.2.12 **gapopen**: Cost to open a gap.

4.2.13 gilist: File containing a list of GIs to restrict the BLAST database to search. The expect values in the BLAST results are based upon the sequences actually searched and not on the underlying database.

4.2.14 h: Displays the application’s brief documentation.

4.2.15 help: Displays the application’s detailed documentation.

4.2.16 html: Enables the generation of HTML output suitable for viewing in a web browser.

4.2.17 import_search_strategy: Name of the file where to read the search strategy to execute (see section titled BLAST search strategies).

4.2.18 lcase_masking: Interpret lowercase letters in query sequence(s) as masked.

4.2.19 **matrix**: Name of the scoring matrix to use.

4.2.720 max_target_seqs: Maximum number of aligned sequences to keep from the BLAST database.

4.2.21 negative_gilist: File containing a list of GIs to exclude from the BLAST database.

4.2.22 num_alignments: Number of alignments to show in the BLAST output.

4.2.23 num_descriptions: Number of one-line descriptions to show in the BLAST output.

4.2.24 num_threads: Number of threads to use during the search.

4.2.25 out: Name of the file to write the application’s output. Defaults to stdout.

4.2.26 outfmt: Allows for the specification of the search application’s output format. A listing of the possible format types is available via the search application’s -help option. If a custom output format is desired, this can be specified by providing a quoted string composed of the desired output format (tabular, tabular with comments, or comma-separated value), a space, and a space delimited list of output specifiers. The list of supported output specifiers is available via the -help command line option. Unsupported output specifiers will be ignored. This should be specified using double quotes if there are spaces in the output format specification (e.g.: -outfmt "7 sseqid ssac qstart qend sstart send qseq evalue bitscore").

4.2.27 parse_deflines: Parse the query and subject deflines.

4.2.28 query: Name of the file containing the query sequence(s), or ‘-‘ if these are provided on standard input.

4.2.29 query_loc: Location of the first query sequence to search in 1-based offsets (Format: start-stop).

4.2.30 remote: Instructs the application to submit the search to NCBI for remote execution.

4.2.31 searchsp: Effective length of the search space.

4.2.32 seg: Arguments to SEG filtering algorithm (use ‘no’ to disable).

4.2.33 show_gis: Show NCBI GIs in deflines in the BLAST output.

4.2.34 **soft_masking**: Apply filtering locations as soft masks (i.e.: only when finding alignment seeds).

4.2.35 subject: Name of the file containing the subject sequence(s) to search.

4.2.36 subject_loc: Location of the first subject sequence to search in 1-based offsets (Format: start-stop).

4.2.37 strand: Strand(s) of the query sequence to search.

4.2.38 **threshold**: Minimum word score such that the word is added to the BLAST lookup table.

4.2.39 **ungapped**: Perform ungapped alignments only.

4.2.40 version: Displays the application’s version.

4.2.41 window_size: Size of the window for multiple hits algorithm, use 0 to specify 1-hit algorithm.

4.2.42 word_size: Word size for word finder algorithm.

4.2.43 **xdrop_gap (Def. 30)**: X-dropoff value (in bits) for preliminary gapped extensions.

4.2.44 **xdrop_gap_final (Def. 100)**: X-dropoff value (in bits) for final gapped alignment.

4.2.45 **xdrop_ungap (Default is 20)** : X-dropoff value (in bits) for ungapped extensions.

L'extension sans brèche démarre à partir de la seconde touche (voir le document « La galaxie BLAST »). L'extension s'arrête quand le score décroit de xdrop_ungap en dessous du meilleur score observé jusque-là. Sous certaines conditions, le segment va ensuite être étendu par une extension avec brèches.

____

Fin de la note

mer. 10 mars 2010

Utiliser BLAST en ligne de commande

Pouvoir exécuter plusieurs BLAST enchainés est un des avantages de l'utilisation de BLAST en ligne de commande. La modification de paramètres d'une exécution à l'autre de BLAST est très rapide – ce qui permet de se concentrer sur leur ajustement idéal pour un cas d'étude donné. Cela évite de parcourir fastidieusement des pages internet (avec leurs temps de chargements) sur les interfaces proposées en ligne qui sont à réserver à des recherches ponctuelles. Un simple rappel de la commande précédente à la console (flèche du haut à l'invite de commande), suivi de la modification du paramètre incriminé est beaucoup plus rapide. Cela permet aussi de mieux réfléchir à l'utilisation de BLAST que lors de l'utilisation d'une interface web où l'on peut avoir tendance à se contenter des paramètres par défaut.

BLAST peut donc être installé sur votre machine. Sous Windows, il faut exécuter l'invite de commande. ("Exécuter..." puis "cmd"). Il existe des paquets rpm et deb pour linux.

Le principe consiste à ranger dans un fichier (qui sera donné en entrée au programme) votre requête – c'est-à-dire votre séquence nucléotidique ou protéique. La banque de donnée interrogée est une banque présente sur un serveur internet ou bien un fichier spécifique que vous téléchargez sur votre disque dur, par exemple la séquence nucléotidique d'un chromosome au format fasta de l'embl.

1/ Télécharger et installer le programme

Sous Ubuntu / Debian

1) Activer les sources de logiciels maintenus par la communauté (universe)
2) Installer le logiciel en utilisant un terminal :

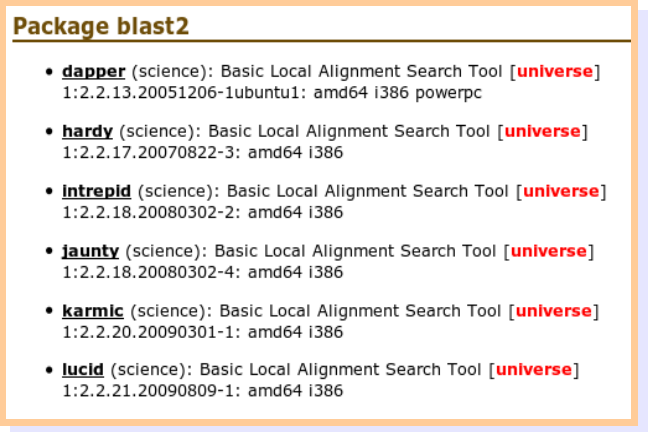
```
apt-cache search blast2
```

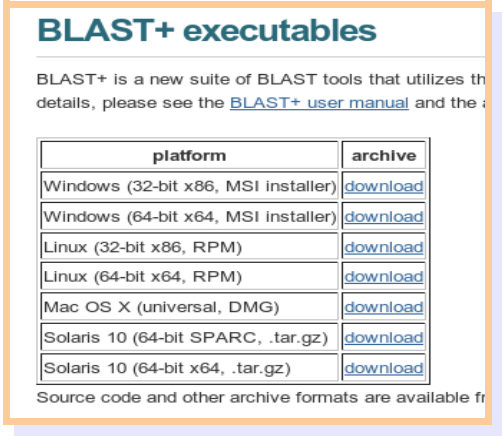
pour vérifier l'existence du paquet deb ;

```
apt-get install blast2
```

pour l'installer.

Vous pouvez aussi rechercher le paquet blast2 sur //packages.ubuntu.com





Sous tous systèmes d'exploitations

télécharger BLAST sur le FTP du NCBI

Lire la page de documentation si votre ordinateur est derrière un pare-feu pour créer le fichier .ncbirc de configuration de connexion.

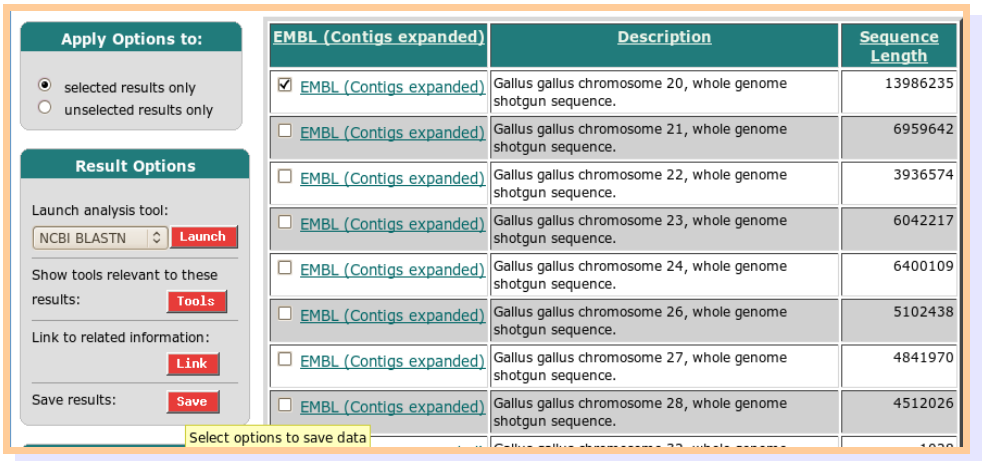
2/ Préparer une banque de données spécifique sur le disque dur

1) Téléchargement

Télécharger les banques de données de Genbank ou des bases de données de génomes ou n'importe quel fichier au format FASTA rapatrié par SRS (comme présenté ci-dessous).

Téléchargement de la séquence nucléotidique du chromosome 20 de gallus gallus au format fasta depuis l'embl.

La banque de données est enregistrée sous le nom *gallus_chrome20.fasta* dans cette exemple.



Le contenu de gallus_chrome20.fasta visualisé avec la commande more

2) Formater la banque

Lorsque la banque est au format fasta il faut la formater pour BLAST avec formatdb, un programme contenu dans l'installation préalablement effectuée.

```
formatdb -i gallus_chrome20.fasta -p F -t Gallus gallus whole genome shotgun sequence -b F -n gallus20
```

-i : indique le fichier d'entrée (entrer plusieurs fichiers est possible)
-p : indique s'il s'agit ou non de séquences protéiques (T pour True, par défaut F pour False)
-t : donne un titre écrit dans les entêtes de la banque
-b : indique si le fichier d'entrée est au format ASN.1 (T, par défaut F)
-n : nomme le fichier de sortie (obligatoire si c'est une concaténation de plusieurs fichiers d'entrée).
Dans un cas simple de formatage d'un seul fichier de nucléotides, la commande suivante suffit :

```
formatdb -i gallus_chrome20.fasta -p F
```

Après le temps d'exécution, trois fichiers sont créés :

```
gallus20.nhr
gallus20.nin
gallus20.nsq
```

Plus d'aide est disponible sur les serveurs du NCBI sur l'utilisation des paramètres de formatdb.

3/ Rechercher une séquence en utilisant BLAST

1) lancer BLAST2 sur une banque locale

```
blastall -p blastn -i genefavori.fasta -d  gallus20 -o resultat_gene_gallus.txt
          blastp
          blastx
          tblastn
          tblastx
```

blastn est utilisé pour les acides nucléiques, blastp est utilisé pour les protéines. Blastx sert à comparer une requête en acide aminée traduite en protéine à une base de données de protéine. tblastn sert à comparer une reqûete de protéine sur un banque nucléique traduite. tblastx sert à comparer une requête nucléique traduite à une banque nucléique traduite.

t---- traduit la banque avant la comparaison
----x traduit la requête avant la comparaison.

Options concernant la séquence-requête

-i : pour donner le fichier d'entrée. Le fichier d'entrée peut contenir une seule séquence ou plusieurs séquences au format fasta.

- L, pour chercher avec la portion de 100 à 400 d'une reqûete, utiliser : -L "100,400"

Options concernant la banque interrogée

-d : pour donner le nom de la banque interrogée (locale ou à distance). La banque de séquence doit avoir été préalablement formatée pour BLAST avec "formatdb".

Options de comptabilisation du score

- G : coût d'ouverture de gap (par défaut 5 pour blastn, 10 pour blastp, blastx et tblastn)
- E : coût d'extension de gap (par défaut 2 pour blastn, 1 pour blastp, blastx et tblastn)
- M : matrice de correspondance à utiliser pour les protéines
    choix : BLOSUM62,BLOSUM45,BLOSUM80,PAM30,PAM70

-q : pénalité pour un mismatch de nucléotide lors d'un blastn (-3 par défaut)
-r : récompense pour une concordance nucléotidique exacte lors d'un blastn (-1 par défaut)

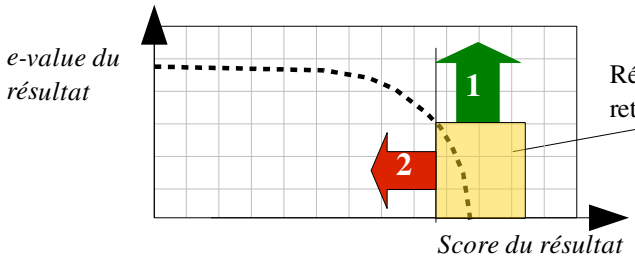
Options de sélection

- W : taille du mot (0 invoque le comportement par défaut : 11 pour blastn, 3 pour les autres)
-f : seuil pour étendre les hits
    Par défaut :
    0 pour blastn
    11 pour blastp
    12 pour blastx
    13 pour tblastn et tblastx
    Lorsque le score du mot est en dessous de cette valeur l'extension du hit n'est pas lancée.
-X : X dropoff pour les alignements avec gap
    Par défaut 30 pour blastn, 0 pour tblastx, 15 pour les autres
-Z : X dropoff value pour les alignements finaux avec gap
    Par défaut : 30 pour blastn, 0 pour tblastx, 25 pour les autres
-g : permettre un alignement contenant des gaps

Options de filtrage

-e : filtrer selon l'expected value (le seuil le plus utilisé pour limiter le nombre d'HSP. Pour des protéines, on le fixe généralement entre 10^{-2} et 10^{-3} (avec blosum62) pour n'obtenir que des HSP significatifs.)

Le paramètre E (Expected value) est une mesure statistique du nombre de hits auxquels on peut s'attendre du fait du simple hasard pour une recherche dans une banque de taille donnée. Il rend ainsi compte du "bruit de fond" qui affecte les recherches d'alignements. Il constitue donc un moyen commode de fixer les résultats et ne retenir que les alignements significatifs. Sa valeur décroît exponentiellement en fonction du score requis pour les alignements (l'obtention d'un alignement de score élevé a très peu de chances d'être le fait du simple hasard). **Lorsque sa valeur de filtrage augmente, la liste des résultats jugés (numériquement) significatifs s'allonge en incorporant des résultats de score plus faible**.



1) augmenter la valeur de filtrage c'est
2) augmenter la liste des résultats retenus avec des résultats de plus faibles scores.

-F : un filtre de basse complexité est utilisé par défaut, il faut donc penser à le retirer si on ne veut pas l'utiliser (-F F)

Il existe une option "Filtre" qui permet de ne pas prendre en compte les domaines de faible complexité (séquences répétées, segments comme CGCGCGC, ATATAATATAAA, ...). Il faut essayer avec ET sans filtre pour comparer les résultats . Par défaut, les filtres sont activés.

Options de traduction

-Q : code génétique utilisé pour la traduction de la requête

1 : standard
2 : mitochondrial de vertébrés
3 : mitochondrial de levure
4 : Mold, protozoan, coelanterate mitochondrial and mycoplasma / spiroplasma
5 : mitochondrial d'invertébrés
6 : macronoyau de ciliés et dasycladacés
9 : mitochondrial d'échinodermes
10 : noyau d'euplotes
11 : bactérien
12 : noyau de levure alternatif
13 : mitochondrial d'ascidies
14 : mitochondrial de vers plats
15 : macronoyau de blespharismes

-D : code génétique utilisé pour la traduction de la banque de donnée

Mêmes options

-S : recherche sur les brins

1 : sens ; 2 : antisens ; 3 : les deux.

Le -S 1 est utile dans le cas d'un CDS comparé à une banque de protéines.

Options de formatage du rapport de sortie

-o : la sortie est par défaut redirigée sur l'écran mais l'utilisateur peut spécifier un nom de fichier de sortie

-v : nombre de descriptions d'une ligne à montrer (mettre 500 pour commencer)

-b : nombre de séquences dont les alignements sont montrés (250)

-m : Options de formatage de la présentation des alignements

Le mode de sortie le plus utilisé est le classique (-m 1), par défaut, mais on peut également utiliser le mode tableau (-m 8) ou le mode xml (-m 7) quand les résultats sont interprétés par un autre programme et non par une personne.

0	Pairwise
1	**query-anchored showing identities**
2	query-anchored no identities
3	flat query-anchored, show identities
4	flat query-anchored, no identities
5	query-anchored no identities and blunt ends
6	flat query-anchored, no identities and blunt ends
7	**XML BLAST output**
8	**tabular (not post processing)**
9	tabular with comment lines (post-processed, sorted)
10	ASN, text
11	ASN, binary

Variation de la ligne de commande de BLAST :

Imaginons que les résultats d'une première recherche n'aient pas été significatifs. On peut affiner la recherche en n'utilisant pas les filtres et en limitant les résultats à des valeurs d'e-value inférieures à un certain seuil.

```
blastall -p  blastn  -i genefavori.fasta -d gallus20 -e1E-3 -F F -m 8 -o resultat.tab
```

effectue un BLAST de la séquence nucléotidique genefavori.fasta sur la banque formatée gallus20, limite les résultats aux valeurs d'e-value inférieures à 10^{-3}, n'utilise pas les filtre de complexité , réalise un fichier de sortie sous la forme d'un tableau et range le résultat du BLAST dans un fichier resultat.tab.

```
blastall -p  blastn  -i genefavori.fasta -d gallus20 -e1E-3 -F F -m 7 -o resultat.xml
```

Dans la ligne de commande ci-dessus le résultat sera formaté en xml.

```
blastall -p  blastn  -i 1.fasta -d bank -F F -o resultat.txt
```

peut également être obtenu avec la commande :

```
blastall -p  blastn  -i 1.fasta -d bank -F F > resultat.txt
```

(redirection de la sortie dans un nouveau fichier avec ">").

Si je désire mener un nouveau BLAST et *ajouter* le résultat à un fichier resultat.txt pré-existant :

```
blastall -p  blastn  -i 1.fasta -d bank -F F >> résultat.txt
```

Pour une interprétation manuelle, "html4blast" peut être utilisé :

```
html4blast -g -o resultat_blast1.html resultat_blast1
```

ouvrir le résultat avec un interpréteur html (navigateur internet).

2) lancer BLAST2 sur une banque d'un serveur de l'Internet

La commande suivante doit être utilisée :

```
blastcl3 -p blastn -i genefavori.fasta -d embl     -o resultat_gene_gallus.txt
            blastp                     gbvrt
            blastx
            tblastn
            tblastx
```

Les banques disponibles sont :

Banques nucléotidiques du NCBI (pré-formatées)

embl (dernière sortie + les mises à jour)
embl_new (les mises à jour)
genbank (dernière sortie + les mises à jour)
genbank_new (les mises à jour)
gbbct : genbank bacteria
gbpri : primates
gbmam : other mammals
gbrod : rodents
gbvrt : autres vertébrés
gbinv : invertébrés
gbpln : plantes + levure
gbvrl : virus
gbphg : phages
gbest : EST
gbsts : STS (Sequence tagged site)
gbsyn : synthetic
gbpat : brevetés (patented)
gbuna : non annotés
gbgss : Genome Survey Sequences
gbhtg : (High throughput genomic sequencing)
imgt : IMGT/LIGM-DB, ImMunoGeneTics sequence database
borrelia : borrelia burgdorferi (génome complet)
ecoli : Escherichia Coli (génome complet)
genitalium : Mycoplasma Genitalium (génome complet)
pneumoniae : Mycoplasma Pneumoniae (génome complet)
hpylori : Helicobacter pylori (génome complet)
bsubtilis : Bacillus Subtilis (génome complet)
tuberculosis : Mucobacterium tuberculosis (génome complet)

Banques protéiques du NCBI (pré-formatées)

uniprot : Universal Protein Resource
uniprot_sprot : swissprot
uniprot_trembl : TrEmbl
nrprot : NCBI non-redondante, traduction des CDS de Genbank + PDB + Swissprot + PIR
nrprot_month : NCBI non-redondante mensuelle
genpept : traduction de Genbank (dernière sortie + maj)
genpept_new : les mises à jour de genpept
gpbct : bactéries genpept
gbpri : primates
gbmam : autres mammifères
gprod : rongeurs
gpvrt : autres vertébrés
gpinv : invertébrés
gppln : plantes (incluant les levures)
gpvrl : virus
gpphag : phages
gpsts : STS
gpsyn : synthetic
gprpat : brevetés (patented)
gpguna : unatotated
gphtg : HTGSequencing
sbase : domaines annotés

ypestis : Yersinia pestis (génome non fini)
yeast : chromosomes de la levure
pfalciparum : Plasmodium falciparum 3D7

Table des matières

Mots : 2036 Caractères : 12517 Nom : tuto-blast Pages : 9

Table S1: Options common to all BLAST+ search applications. An option of type "flag" takes no argument, but if present is true. Some options are valid only for a local search ("remote" option not used), others are valid only for a remote search ("remote" option used).

option	**type**	**default value**	**description and notes**
db	string	none	BLAST database name.
query	string	stdin	Query file name.
query_loc	string	none	Location on the query sequence (Format: start-stop)
out	string	stdout	Output file name
evalue	real	10.0	Expect value (E) for saving hits
subject	string	none	File with subject sequence(s) to search.
subject_loc	string	none	Location on the subject sequence (Format: start-stop).
show_gis	flag	N/A	Show NCBI GIs in report.
num_descriptions	integer	500	Show one-line descriptions for this number of database sequences.
num_alignments	integer	250	Show alignments for this number of database sequences.
html	flag	N/A	Produce HTML output
gilist	string	none	Restrict search of database to GI's listed in this file. Local searches only
negative_gilist	string	none	Restrict search of database to everything except the GI's listed in this file. Local searches only.
entrez_query	string	none	Restrict search with the given Entrez query. Remote searches only.
culling_limit	integer	none	Delete a hit that is enveloped by at least this many higher-scoring hits.
best_hit_overhang	real	none	Best Hit algorithm overhang value (recommended value: 0.1)
best_hit_score_edge	real	none	Best Hit algorithm score edge value (recommended value: 0.1)
dbsize	integer	none	Effective size of the database
searchsp	integer	none	Effective length of the search space
import_search_strategy	string	none	Search strategy file to read.
export_search_strategy	string	none	Record search strategy to this file.
parse_deflines	flag	N/A	Parse query and subject bar delimited sequence identifiers (e.g., gi|129295).
num_threads	integer	1	Number of threads (CPUs) to use in blast search.
remote	flag	N/A	Execute search on NCBI servers?
oufmt	string	0	alignment view options: 0 = pairwise, 1 = query-anchored showing identities, 2 = query-anchored no identities, 3 = flat query-anchored, show identities,

			4 = flat query-anchored, no identities, 5 = XML Blast output, 6 = tabular, 7 = tabular with comment lines, 8 = Text ASN.1, 9 = Binary ASN.1 10 = Comma-separated values Options 6, 7, and 10 can be additionally configured to produce a custom format specified by space delimited format specifiers. The supported format specifiers are: qseqid means Query Seq-id qgi means Query GI qacc means Query accesion sseqid means Subject Seq-id sallseqid means All subject Seq-id(s), separated by a ';' sgi means Subject GI sallgi means All subject GIs sacc means Subject accession sallacc means All subject accessions qstart means Start of alignment in query qend means End of alignment in query sstart means Start of alignment in subject send means End of alignment in subject qseq means Aligned part of query sequence sseq means Aligned part of subject sequence evalue means Expect value bitscore means Bit score score means Raw score length means Alignment length pident means Percentage of identical matches nident means Number of identical matches mismatch means Number of mismatches positive means Number of positive-scoring matches gapopen means Number of gap openings gaps means Total number of gap ppos means Percentage of positive-scoring matches frames means Query and subject frames separated by a '/' qframe means Query frame sframe means Subject frame When not provided, the default value is: 'qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore', which is equivalent to the keyword 'std'

Table S2: Options for the blastn application. The blastn application searches a nucleotide query against nucleotide subject sequences or a nucleotide database. An option of type "flag" takes no arguments, but if present the argument is true. Options marked "experimental" may be removed or changed with little or no notice. Four different tasks are supported: 1.) "megablast", for very similar sequences (e.g, sequencing errors), 2.) "dc-megablast", typically used for inter-species comparisons, 3.) "blastn", the traditional program used for inter-species comparisons, 4.) "blastn-short", optimized for sequences less than 30 nucleotides.

option	**task(s)**	**type**	**default value**	**description and notes**
word_size	megablast	integer	28	Length of initial exact match.
word_size	dc-megablast	integer	11	Number of matching nucleotides in initial match. dc-megablast allows non-consecutive letters to match.
word_size	blastn	integer	11	Length of initial exact match.
word_size	blastn-short	integer	7	Length of initial exact match.
gapopen	megablast	integer	0	Cost to open a gap.
gapextend	megablast	integer	none	Cost to extend a gap. This default is a function of reward/penalty value.
gapopen	blastn, blastn-short, dc-megablast	integer	5	Cost to open a gap.
gapextend	blastn, blastn-short, dc-megablast	integer	2	Cost to extend a gap.
reward	megablast	integer	1	Reward for a nucleotide match.
penalty	megablast	integer	-2	Penalty for a nucleotide mismatch.
reward	blastn, dc-megablast	integer	2	Reward for a nucleotide match.
penalty	blastn, dc-megablast	integer	-3	Penalty for a nucleotide mismatch.
reward	blastn-short	integer	1	Reward for a nucleotide match.

penalty	blastn-short	integer	-3	Penalty for a nucleotide mismatch.
strand	all	string	both	Query strand(s) to search against database/subject. Choice of both, minus, or plus.
dust	all	string	20 64 1	Filter query sequence with dust.
filtering_db	all	string	none	Mask query using the sequences in this database.
window_masker_taxid	all	integer	none	Enable WindowMasker filtering using a Taxonomic ID. NOTE: experimental.
window_masker_db	all	string	none	Enable WindowMasker filtering using this file. NOTE: experimental.
soft_masking	all	boolean	true	Apply filtering locations as soft masks.
lcase_masking	all	flag	N/A	Use lower case filtering in query and subject sequence(s)?
db_soft_masking	all	integer	none	Filtering algorithm ID to apply to the BLAST database as soft masking.
perc_identity	all	integer	0	Percent identity cutoff.
template_type	dc-megablast	string	coding	Discontiguous MegaBLAST template type. Allowed values are coding, optimal and coding_and_optimal.
template_length	dc-megablast	integer	18	Discontiguous MegaBLAST template length.
use_index	megablast	boolean	false	Use MegaBLAST database index.
index_name	megablast	string	none	MegaBLAST database index name.
xdrop_ungap	all	real	20	Heuristic value (in bits) for ungapped extensions.
xdrop_gap	all	real	30	Heuristic value (in bits) for preliminary gapped extensions.
xdrop_gap_final	all	real	100	Heuristic value (in bits) for final gapped alignment.
no_greedy	megablast	flag	N/A	Use non-greedy dynamic programming extension.
min_raw_gapped_score	all	integer	none	Minimum raw gapped score to keep an alignment in the preliminary gapped and trace-back stages. Normally set based

				upon expect value.
ungapped	all	flag	N/A	Perform ungapped alignment.
window_size	dc-megablast	integer	40	Multiple hits window size, use 0 to specify 1-hit algorithm

Table S3: Options for the blastp application. The blastp application searches a protein sequence against protein subject sequences or a protein database. An option of type "flag" takes no arguments, but if present the argument is true. Two different tasks are supported: 1.) "blastp", for standard protein-protein comparisons, 2.) "blastp-short", optimized for query sequences shorter than 30 residues. This table reflects the 2.2.23 BLAST+ release. On earlier releases the blastp-short task was not implemented.

option	**task**	**type**	**default value**	**description and notes**
word_size	blastp	integer	3	Word size of initial match.
word_size	blastp-short	integer	2	Word size of initial match.
gapopen	blastp	integer	11	Cost to open a gap.
gapextend	blastp	integer	1	Cost to extend a gap.
gapopen	blastp-short	integer	9	Cost to open a gap.
gapextend	blastp-short	integer	1	Cost to extend a gap.
matrix	blastp	string	BLOSUM62	Scoring matrix name.
matrix	blastp-short	string	PAM30	Scoring matrix name.
threshold	blastp	integer	11	Minimum score to add a word to the BLAST lookup table.
threshold	blastp-short	integer	16	Minimum score to add a word to the BLAST lookup table.
comp_based_stats	blastp	string	2	Use composition-based statistics: D or d: default (equivalent to 2) 0 or F or f: no composition-based statistics 1: Composition-based statistics as in NAR 29:2994-3005, 2001 2 or T or t : Composition-based score adjustment as in Bioinformatics

				21:902-911, 2005, conditioned on sequence properties 3: Composition-based score adjustment as in Bioinformatics 21:902-911, 2005, unconditionally
comp_based_stats	blastp-short	string	0	Use composition-based statistics : D or d: default (equivalent to 2) 0 or F or f: no composition-based statistics 1: Composition-based statistics as in NAR 29:2994-3005, 2001 2 or T or t : Composition-based score adjustment as in Bioinformatics 21:902-911, 2005, conditioned on sequence properties 3: Composition-based score adjustment as in Bioinformatics 21:902-911, 2005, unconditionally
seg	all	string	no	Filter query sequence with SEG (Format: 'yes', 'window locut hicut', or 'no' to disable).
soft_masking	blastp	boolean	false	Apply filtering locations as soft masks
xdrop_ungap	all	real	7	Heuristic value (in bits) for ungapped extensions
xdrop_gap	all	real	15	Heuristic value (in bits) for preliminary gapped extensions.
xdrop_gap_final	all	real	25	Heuristic value (in bits) for final gapped alignment/
window_size	blastp	integer	40	Multiple hits window size, use 0 to specify 1-hit algorithm.
window_size	blastp-short	integer	15	Multiple hits window size, use 0 to specify 1-hit algorithm.
use_sw_tback	all	flag	N/A	Compute locally optimal Smith-Waterman alignments?

Table S4: Options for the blastx application. The blastx application translates a nucleotide query and searches it against protein subject sequences or a protein database.

option	**type**	**default value**	**description and notes**
word_size	integer	3	Word size for initial match.
gapopen	integer	11	Cost to open a gap.
gapextend	integer	1	Cost to extend a gap.
matrix	string	BLOSUM62	Scoring matrix name.
threshold	integer	12	Minimum score to add a word to the BLAST lookup table.
seg	string	12 2.2 2.5	Filter query sequence with SEG (Format: 'yes', 'window locut hicut', or 'no' to disable).
soft_masking	boolean	false	Apply filtering locations as soft masks.
xdrop_ungap	real	7	Heuristic value (in bits) for ungapped extensions.
xdrop_gap	real	15	Heuristic value (in bits) for preliminary gapped extensions.
xdrop_gap_final	real	25	Heuristic value (in bits) for final gapped alignment.
window_size	integer	40	Multiple hits window size, use 0 to specify 1-hit algorithm.
strand	string	both	Query strand(s) to search against database/subject. Choice of both, minus, or plus.
query_genetic_code	integer	1	Genetic code to translate query, see ftp://ftp.ncbi.nih.gov/entrez/misc/data/gc.prt
frame_shift_penalty	integer	0	Frame shift penalty (for use with out-of-frame gapped alignment). NOTE: statistics may not be correct with the option
max_intron_length	integer	0	Length of the largest intron allowed in a translated nucleotide sequence when linking multiple distinct alignments (a negative value disables linking).

Table S5: Options for the tblastn application. The tblastn application searches a protein query against nucleotide subject sequences or a nucleotide database translated at search time.

option	**type**	**default value**	**description and notes**
word_size	integer	3	Word size for initial match.
gapopen	integer	11	Cost to open a gap.
gapextend	integer	1	Cost to extend a gap.
matrix	string	BLOSUM62	Scoring matrix name.
threshold	integer	13	Minimum score to add a word to the BLAST lookup table.
seg	string	12 2.2 2.5	Filter query sequence with SEG (Format: 'yes', 'window locut hicut', or 'no' to disable).
soft_masking	boolean	false	Apply filtering locations as soft masks.
xdrop_ungap	real	7	Heuristic value (in bits) for ungapped extensions.
xdrop_gap	real	15	Heuristic value (in bits) for preliminary gapped extensions.
xdrop_gap_final	real	25	Heuristic value (in bits) for final gapped alignment.
window_size	integer	40	Multiple hits window size, use 0 to specify 1-hit algorithm.
db_gen_code	integer	1	Genetic code to translate subject sequences, see ftp://ftp.ncbi.nih.gov/entrez/misc/data/gc.prt
frame_shift_penalty	integer	0	Frame shift penalty (for use with out-of-frame gapped alignment. NOTE: statistics may not be correct with the option
max_intron_length	integer	0	Length of the largest intron allowed in a translated nucleotide sequence when linking multiple distinct alignments (a negative value disables linking).
comp_based_stats	string	D	Use composition-based statistics for tblastn: D or d: default (equivalent to 2) 0 or F or f: no composition-based statistics 1: Composition-based statistics as in NAR 29:2994-3005, 2001 2 or T or t : Composition-based score adjustment as in Bioinformatics 21:902-911, 2005, conditioned on sequence properties 3: Composition-based score adjustment as in Bioinformatics 21:902-911, 2005, unconditionally Default = `2'

Table S6: Options for the tblastx application. The tblastx application searches a translated nucleotide query against translated nucleotide subject sequences or a translated nucleotide database An option of type “flag” takes no arguments, but if present the argument is true. This table reflects the 2.2.23 BLAST+ release.

option	**type**	**default value**	**description and notes**
word_size	integer	3	Word size for initial match.
matrix	string	BLOSUM62	Scoring matrix name.
threshold	integer	13	Minimum word score to add the word to the BLAST lookup table.
seg	string	12 2.2 2.5	Filter query sequence with SEG (Format: 'yes', 'window locut hicut', or 'no' to disable).
soft_masking	boolean	false	Apply filtering locations as soft masks.
xdrop_ungap	real	7	Heuristic value (in bits) for ungapped extensions.
window_size	integer	40	Multiple hits window size, use 0 to specify 1-hit algorithm.
strand	string	both	Query strand(s) to search against database subject sequences. Choice of both, minus, or plus.
query_genetic_code	integer	1	Genetic code to translate query, see ftp://ftp.ncbi.nih.gov/entrez/misc/data/gc.prt
db_gen_code	integer	1	Genetic code to translate subject sequences, see ftp://ftp.ncbi.nih.gov/entrez/misc/data/gc.prt
max_intron_length	integer	0	Length of the largest intron allowed in a translated nucleotide sequence when linking multiple distinct alignments (a negative value disables linking)

Table S7: Options for the makeblastdb application. This application builds a BLAST database. An option of type "flag" takes no arguments, but if present the argument is true.

option	**type**	**default value**	**description**
in	string	stdin	Input file/database name; the data type is automatically detected, it may be any of the following: FASTA file(s) and/or BLAST database(s)
dbtype	string	prot	Molecule type of input, values can be nucl or prot.
title	string	none	Title for BLAST database. If not set the input file name will be used.
parse_seqids	flag	N/A	Parse bar delimited sequence identifiers (e.g., gi|129295) in FASTA input.
hash_index	flag	N/A	Create index of sequence hash values.
mask_data	string	none	Comma-separated list of input files containing masking data as produced by NCBI masking applications (e.g. dustmasker, segmasker, windowmasker).
out	string	input file name	Name of BLAST database to be created. Input file name is used if none provided. This field is required if input consists of multiple files.
max_file_size	string	1GB	Maximum file size to use for BLAST database.
taxid	integer	none	Taxonomy ID to assign to all sequences.
taxid_map	string	none	File mapping sequence IDs to taxonomy IDs.
logfile	string	none	Program log file (default is stderr).

Table S8: Options for blastdbcmd application. This application reads a BLAST database and produces reports.

option	type	default value	description and notes
db	string	nr	BLAST database name.
dbtype	string	guess	Molecule type stored in BLAST database, one of nucl, prot, or guess.
entry	string	none	Comma-delimited search string(s) of sequence identifiers: e.g.: 555, AC147927, 'gnl|dbname|tag', or 'all' to select all sequences in the database
entry_batch	string	none	Input file for batch processing (Format: one entry per line)
pig	integer	none	PIG (protein identity group) to retrieve.
info	flag	N/A	Print BLAST database information.
range	string	none	Range of sequence to extract (Format: start-stop).
strand	string	plus	Strand of nucleotide sequence to extract. Choice of plus or minus.
mask_sequence_with	string	none	Produce lower-case masked FASTA using the algorithm IDs specified.
out	string	stdout	Output file name.
outfmt	string	%f	Output format, where the available format specifiers are: %f means sequence in FASTA format %s means sequence data (without defline) %a means accession %g means gi %o means ordinal id (OID) %t means sequence title %l means sequence length %T means taxid %L means common taxonomic name %S means scientific name %P means PIG %mX means sequence masking data, where X is an optional comma-separated list of integers to specify the algorithm ID(s) to display (or all masks if absent or invalid specification). Masking data will be

			displayed as a series of 'N-M' values separated by ';' or the word 'none' if none are available. For every format except '%f', each line of output will correspond to a sequence.
target_only	flag	N/A	Definition line should contain target GI only.
get_dups	flag	N/A	Retrieve duplicate accessions.
line_length	integer	80	Line length for output.
ctrl_a	flag	N/A	Use Ctrl-A as the non-redundant definition line separator.