

Recherche par similarité dans les banques/bases de données
La suite Blast (Basic Local Alignment Search Tool)

Algorithme de Blast (version 1)

(Altschul *et al.* (1990) *J. Mol. Biol.*, 215, 403-410)

- ne recherche que les diagonales
- > établir un score pour chaque paire de segments comparée (diagonale)
(séquences ADN : $Se(\text{identité}) = 5$ et $Se(\text{mismatch}) = -4$
séquences protéiques -> matrice de substitution)
- > MSP : Maximal Segment Pair correspond à la paire de segments (diagonale) qui possède le score le plus élevé comparé au score de toutes les autres paires de segments.
- > HSP : High Scoring segment Pair définit une paire de segments comme localement maximale, c'est-à-dire que l'on ne pourra pas augmenter son score soit en la rallongeant, soit en les raccourcissant.
 - > problème : trop de paires de segments, on ne peut pas toutes les retenir.
 - > définition d'un seuil S_{HSP} : Blast ne retient comme HSP que les paires de segments dont le score est supérieur à S_{HSP} .

Algorithme de Blast (version 1)

(Altschul *et al.* (1990) J. Mol. Biol., 215, 403-410)

-> Définition de la valeur du seuil S_{HSP} :

Résultats en statistique de Karling *et al.* permettent d'estimer le score le plus élevé que peut avoir une MSP entre deux séquences dont la similarité ne serait due qu'au hasard

-> ce score est utilisé comme seuil S_{HSP}

-> Sélection des HSP

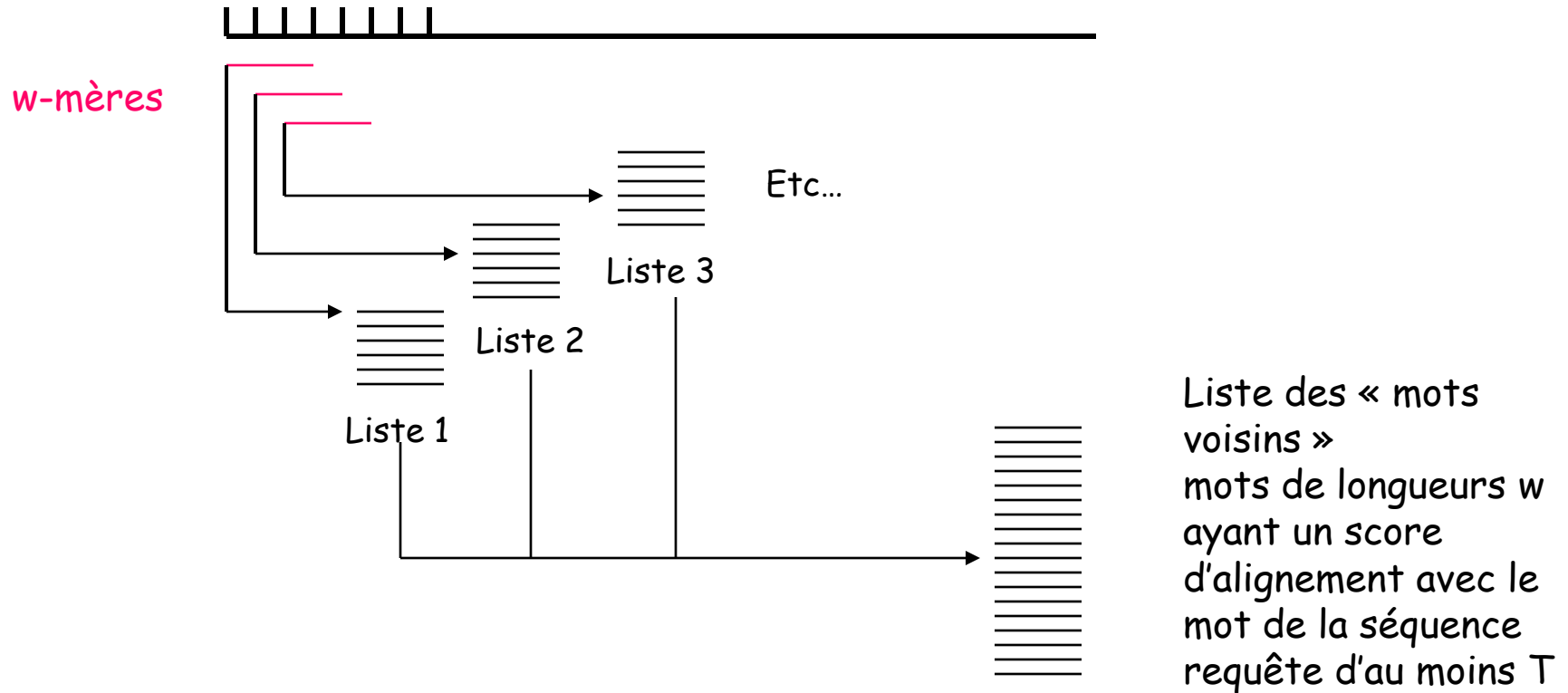
- Paire de segments dont le score est $<$ à S_{HSP} , similarité non significative due au hasard. Cette paire n'est pas retenue.
- Paire de segments dont le score est \geq à S_{HSP} , similarité due à une histoire évolutive commune -> séquence apparentée -> retenue comme HSP

Algorithme de Blast (version 1)

(Altschul *et al.* (1990) *J. Mol. Biol.*, 215, 403-410)

1ère étape : Établissement d'une liste de « mots voisins »

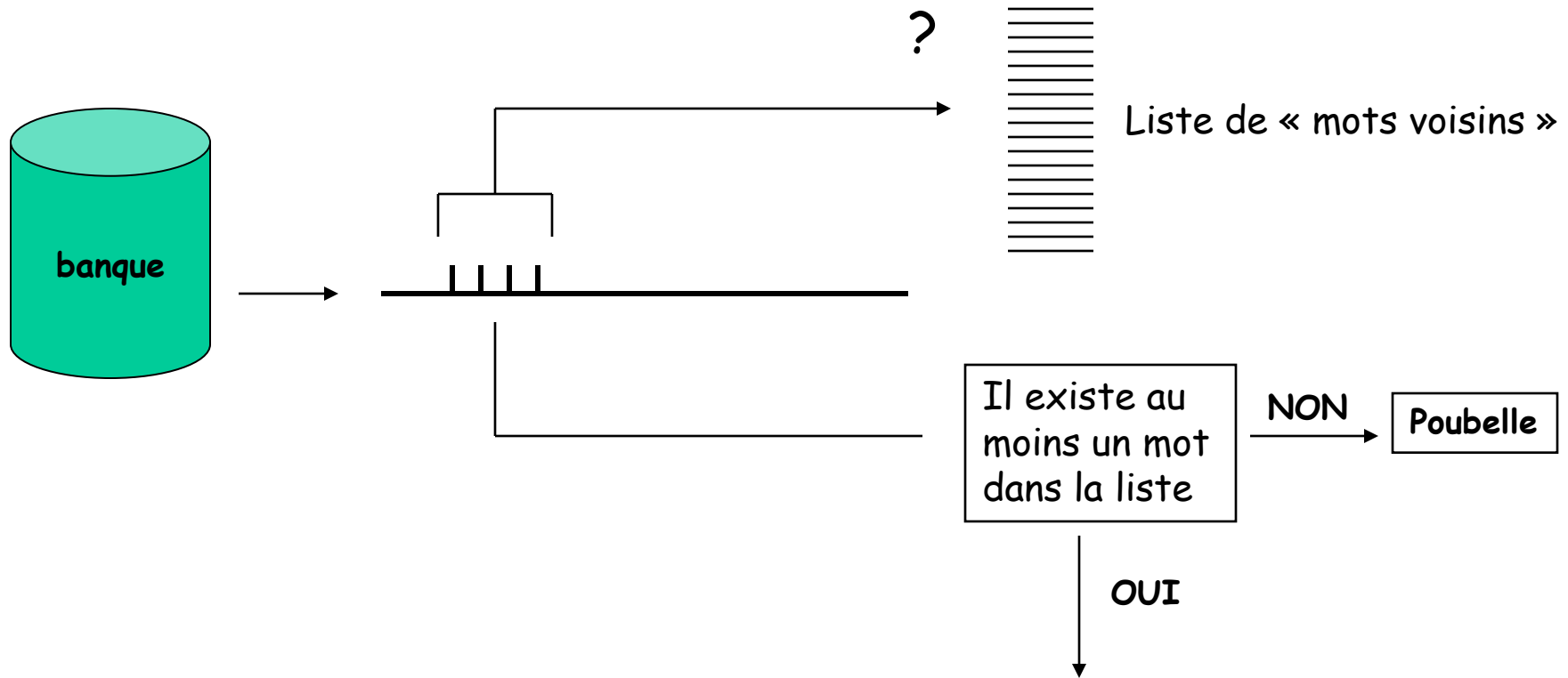
Séquence requête Q



w-mères : mot de taille *w* (*w* = 3 pour les protéines et *w* = 11 pour les séquences d'acides nucléiques)

Algorithme de Blast (version 1)

2ème étape : Recherche de la présence de ces mots dans les séquences de la banque



Extension de l'alignement : 3ème étape

Algorithme de Blast (version 1)

3ème étape : si un « hit » extension de l'alignement

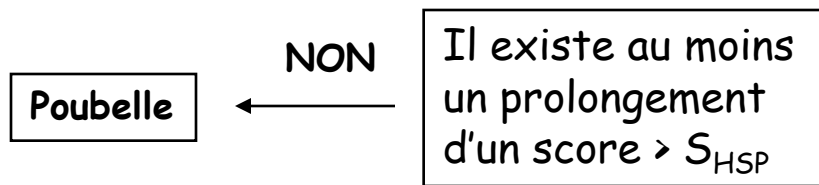
Séquence requête Q



Séquence banque B



Essayer d'étendre l'alignement dans les deux sens (arrêt de l'extension quand le score obtenu décroît au minimum d'une valeur X fixée (drop-off score))



OUI

Meilleur score obtenu par prolongement d'un hit

Score séquence B de la banque

4ème étape : Calcul de la p-value

Statistique Poisson

Classement des séquences de la banque

Normalisation du score

Le score normalisé S' pour un HSP est donné par :

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

Avec :

S score du HSP

λ et K paramètres calculés à partir du système de score et de la composition des séquences

S' est exprimé en une unité appelée *bits*

La p-value

La p-value est la probabilité qu'il existe au moins un HSP, obtenu lors de la comparaison de deux séquences aléatoires (de même longueur et composition que les séquences d'intérêt), dont le score soit supérieur ou égal à celui du HSP issu de la comparaison des séquences d'intérêt.

La p-value

On compare une séquence de longueur n avec une banque de données de longueur m , avec une matrice de comparaison donnée.

On s'intéresse à la distribution du nombre $N(S)$ d'alignements dont le score dépasse la valeur S . Cette distribution suit approximativement une loi de Poisson. On a donc l'expression suivante pour la probabilité qu'il existe un alignement fortuit de score $\geq S$:

$$p(\text{score} \geq S) = 1 - e^{-E(S)}$$

$E(S)$ étant l'espérance mathématique de $N(S)$ et qui s'exprime par :

$$E(S) = Knme^{-\lambda S} = E\text{-value} = \text{nombre attendu d'alignements ayant un score au moins égal à } S$$

Avec λ et K paramètres calculés à partir du système de score et de la composition des séquences

Dans le cas des alignements locaux sans gap, la théorie de Karlin et Altschul fournit donc des formules analytiques dont les paramètres dépendent de ceux de la recherche.

Query= Bsub.RbsA
(493 letters)

Database: Coli
4289 sequences; 1,358,990 total letters

Searching.....done

Sequences producing significant alignments:	Score (bits)	E Value	N
EcolA01.RBSA "ATP-binding component of D-ribose high-affini...	270	e-141	3
EcolA01.MGLA "ATP-binding component of methyl-galactoside t...	164	e-124	4
EcolA01.YJWC "putative ATP-binding component of a transport...	145	e-121	4
EcolA01.XYLG "putative ATP-binding protein of xylose transp...	157	e-120	6

>EcolA01.RBSA "ATP-binding component of D-ribose high-affinity transport system"
Length = 501

Score = 270 bits (587), Expect(3) = e-141
Identities = 115/258 (44%), Positives = 167/258 (64%)

Query: **3** IEMKDIHKTFGKNQVLSGVSFQLMPGEVHALMGENGAGKSTLMNILTGLHKADKQISIN 62
 +++K I K F + LSG + + PG V AL+GENGAGKST+M +LTG++ D G +
 Sbjct: **5** LQLKGIDKAPPGVKALSGAALNVYPGRVMALVGENGAGKSTMMKVLGTGIYTRDAGTLLWL 64

Query: 63 GNETYFSNPKEAQHGIAFIHQELNIWPEMTVLENLFIGKEISSKLGVLQTRKMKALAKE 122
 G ET F+ PK +++ GI IHQELN+ P++T+ EN+F+G+E ++ G + + M A A +
 Sbjct: 65 GKETTFTGPKSSQEAGIGIIHQELNLIPQLTIAENIFLGRFVNRFKIDWKTMYAEDK 124

Query: 123 QFDKLSVSLSLDQEAGECSVGGQQMIEIAKALMTNAEVIIMDEPTAALTEREISKLFEVI 182
 KL++ D+ G+ S+G QOM+EIAK L ++VIIMDEPT ALT+ E LF VI
 Sbjct: 125 LLAKLNLRFKSDKLVGDLSIGDQOMVEIAKVLFSFESKVIIMDEPTDALTDTEESLFRVI 184

Query: 183 TALKKNGVSIYIISHRMEEIFAICDRITIMRDGKTVDTTNISETDFDEVVKKMVGRELTE 242
 LK G IVYISHRM+EIF ICD +T+ RDG+ + ++ D +++ MVGR+L +
 Sbjct: 185 RELKSQGRGIVYIISHRMKEIFEICDDVTVFRDQGFIAEREVASLTEDSLIEMMVGRKLED 244

Query: 243 RYPKRTPSLGDKVFEVKN **260**
 +YP + GD +V N
 Sbjct: 245 QYPHLDKAPGDIRLKVDN **262**

Score = 103 bits (220), Expect(3) = e-141
Identities = 43/94 (45%), Positives = 66/94 (69%)

Query: **269** DVSFYVRSGEIVGVSGLMGAGRTEMMRALFGVDRLDTGEIWIAGKKTAIKNPQEA VVKGL 328
 DVSF +R GEI+GVSGLMGAGRTE+M+ L+G +G + + G + ++PQ+ + G+
 Sbjct: **270** DVSFTRKGEILGVSGLMGAGRTEMLKVLYGALPRTSGYVTLDGHEVVTRSPQDGLANGI 329

Query: 329 GFITENRKDEGLLLDTSIRENIALPNLSSFSFKG **362**
 +I+E+RK +GL+L S++EN++L L FS G
 Sbjct: 330 VYISEDRKRDGLVGLMSVKNMSLTALRYFSRAG **363**

Résultat d'une recherche avec la version 1 de BlastP

Score = 165 bits (356), Expect(3) = e-141
Identities = 74/130 (56%), Positives = 92/130 (69%)

Query: **362** GLIDHKREAEFVDLLIKRLTIKTASPETHARHLSGGNQKQVVIKAWIGIGPKVLILDEPT 421
 G + H E + V I+ +K T S E LSGGNQKQV IA+ + PKVLILDEPT
 Sbjct: **364** GSLKHADEQQAVSDFIRLFNVKTPSMEQAIGLLSGGNQKQVAIARGLMTRPKVLILDEPT 423

Query: 422 RGVVDVGAKREIYTLMNELTERGVVAIIMVSSSELPEILGMSDRIIVVHEGRISGEIHAREAT 481
 RGVVDGAK+EIY L+N+ G++II+VSSE+PE+LGMSDRIIV+HEG +SGE +AT
 Sbjct: 424 RGVVDVGAKKEIYQLINQFKADGLSIIIVSSEMPEVLGMSDRIIVMHEGHLSGEFTREQAT 483

Query: 482 QERIMTLATG **491**
 QE +M A G
 Sbjct: 484 QEVLMAAAVG **493**

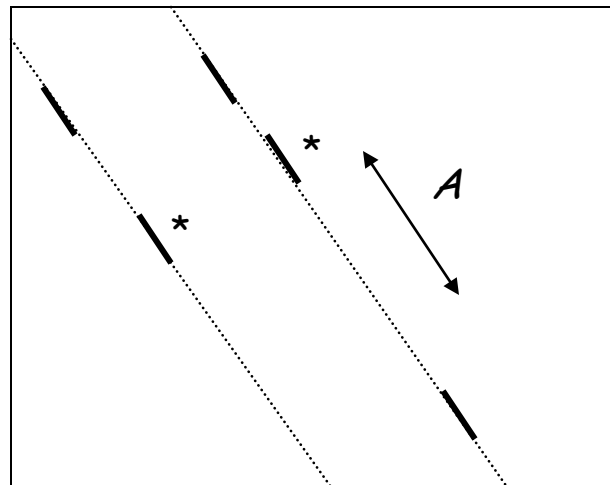
Gapped Blast ou Blast2

(Altschul *et al.* (1997) *Nucleic Acids Res.*, 25, 3389-3402)

Les deux premières étapes identiques à la version 1 de l'algorithme de Blast

Différences :

- étape d'extension des hits (90% du temps d'exécution du Blast)
Rq : une HSP contient souvent plusieurs hits sur la même diagonale et à une distance peu éloignée les uns des autres.
Sélection des hits pour l'étape d'extension : présence de deux hits sur la même diagonale séparés par une distance inférieure ou égale à A .



Les hits marqués d'une * subiront une extension sans gaps analogue à la 3ème étape du Blast version 1

Gapped Blast ou Blast2

Différences :

- Les HSP sélectionnés qui auront un score supérieur ou égal à S_{HSP} servent ensuite de points d'ancrage à une recherche d'alignement local optimal par programmation dynamique. La matrice de programmation dynamique est explorée dans les deux directions à partir d'un résidu aligné (graine).

Choix de la graine : on choisit le long d'une HSP le segment de 11 résidus ayant le meilleur score et le résidu central est utilisé comme graine.

La recherche du chemin optimal est limité aux cellules de la matrice telles que le score de l'alignement ne devienne pas inférieur de plus de Xg au score maximal atteint jusque là (modification de l'algorithme de Smith et Waterman)

Significativité des alignements : La E-value

Dans le cas des alignements locaux avec gaps, il n'existe pas de théorie décrivant la distribution attendue des scores.

La E-value (Expect value), le nombre attendu d'alignement qui par chance aurait un score $\geq S$, est évaluée en regardant les scores des alignements générés par comparaison de séquences aléatoires ayant même longueur et même composition que la séquence requête.

La suite Blast

Un ensemble de programmes :

programme	séquence requête	Banque
BlastN	nucléique	nucléique
BlastP	protéique	protéique
BlastX	nucléique (séquence traduite dans les 6 cadres de lecture)	protéique
tblastN	protéique	nucléique (séquences de la banque traduites dans les 6 cadres de lecture)
tblastX	nucléique (séquence traduite dans les 6 cadres de lecture)	nucléique (séquences de la banque traduites dans les 6 cadres de lecture)

La suite Blast

Possibilité d'appliquer des filtres et masques (paramètres de l'algorithme) :

- masquer les séquences de faible complexité (proposé pour l'ensemble des programmes de la suite Blast)
- dans le cas d'une recherche avec une séquence d'acides nucléiques contre une banque de séquences nucléiques (BlastN), masquer les séquences répétées (ex: les séquences Alu chez les primates).

Une première analyse compare la séquence sonde à une banque de séquences d'éléments répétés. Les zones de la séquence sonde s'alignant avec les séquences d'éléments répétés sont masquées pour la recherche de similarité dans la banque.

Exemple d'interface du programme BlastP (site NCBI)

The image shows a screenshot of the NCBI BLASTP web interface in a Mozilla Firefox browser. The browser's address bar shows the URL: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blastp`. The page title is "Protein BLAST: search protein databases using a protein query - Mozilla Firefox".

The interface includes a navigation bar with "Home", "Recent Results", "Saved Strategies", and "Help". Below this, there are tabs for "blastn", "blastp", and "blastx", with "blastp" selected. The main heading is "BLASTP programs search protein databases using a protein query. [more...](#)".

The "Enter Query Sequence" section contains a large text input field for the query sequence, a "Clear" button, and a "Query subrange" section with "From" and "To" input fields. A blue arrow points from the text "Votre séquence sonde" to the main query input field.

Below the query input is a section for "Or, upload file" with a file selection button labeled "Parcourir...". There is also a "Job Title" input field with a placeholder "Enter a descriptive title for your BLAST search".

The "Choose Search Set" section includes a "Database" dropdown menu set to "Non-redundant protein sequences (nr)", an "Organism" input field with a placeholder "Enter organism name or id--completions will be suggested", and an "Entrez Query" input field with a placeholder "Enter an Entrez query to limit search". A blue arrow points from the text "Choix de la banque" to the "Database" dropdown menu.

The "Program Selection" section shows the "Algorithm" dropdown menu set to "blastp (protein-protein BLAST)". Other options include "PSI-BLAST (Position-Specific Iterated BLAST)" and "PHI-BLAST (Pattern Hit Initiated BLAST)".

At the bottom, there is a "BLAST" button and a search bar containing the text "Search database nr using Blastp (protein-protein BLAST)". The browser's status bar at the very bottom shows "Rechercher : complexity", navigation buttons for "Précédent" and "Suivant", and a "Terminé" status.

Les paramètres « cachés »

Algorithm parameters

General Parameters

Max target sequences

100

Nombre de séquences cibles

Select the maximum number of aligned sequences to display

Short queries

Automatically adjust parameters for short input sequences

Expect threshold

10

Seuil E-value

Word size

3

Taille des mots pour construire la liste des mots voisins

Max matches in a query range

0

Scoring Parameters

Matrix

BLOSUM62

Choix de la matrice de substitution

Gap Costs

Existence: 11 Extension: 1

Pondération des gaps : ouverture et extension

Compositional adjustments

Conditional compositional score matrix adjustment

Filters and Masking

Filter

Low complexity regions

Mask

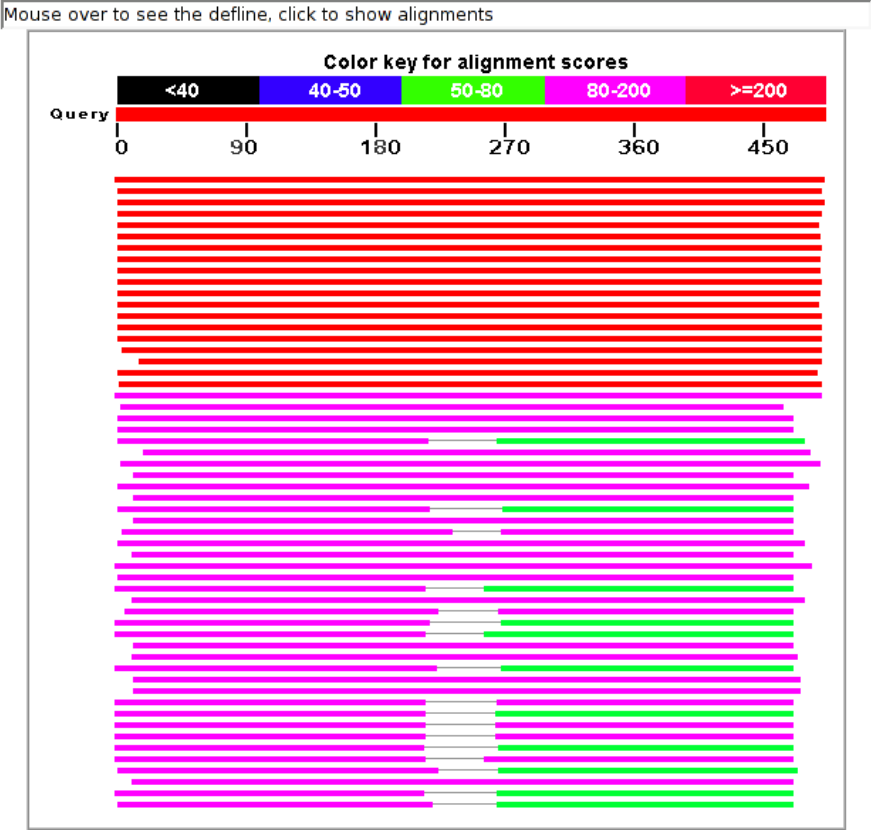
Mask for lookup table only

Mask lower case letters

Query= Bsub.RbsA
Length=493

Résultat d'une recherche avec BlastP2 sur le site du NCBI

Distribution of 205 Blast Hits on the Query Sequence



[Distance tree of results](#) NEW

Sequences producing significant alignments:	Score (Bits)	E Value
gi 7404442 sp P36947 RBSA_BACSU Ribose transport ATP-binding pro	932	0.0
gi 132119 sp P04983 RBSA_ECOLI Ribose transport ATP-binding prot	432	2e-120
gi 1172865 sp P44735 RBSA_HAEIN Ribose transport ATP-binding pro	416	1e-115
gi 6016558 sp Q56342 MGLA_TREPA Galactoside transport ATP-bindin	400	8e-111
gi 731981 sp P32721 ALSA_ECOLI D-allose transport ATP-binding pr	395	2e-109
gi 2506104 sp P23924 MGLA_SALTY Galactoside transport ATP-bindin	395	2e-109
gi 1170944 sp P44884 MGLA_HAEIN Galactoside transport ATP-bindin	391	3e-108
gi 20137563 sp Q9S472 ARAG_BACST L-arabinose transport ATP-bindi	390	5e-108
gi 77416578 sp P0AAG8 MGLA_ECOLI Galactoside transport ATP-bi...	390	6e-108

>gi|132119|sp|P04983|RBSA_ECOLI Ribose transport ATP-binding protein rbsA
Length=501

Score = 432 bits (1110), Expect = 2e-120, Method: Composition-based stats.
Identities = 232/490 (47%), Positives = 325/490 (66%), Gaps = 2/490 (0%)

```
Query 3 IEMKDIHKTFGKNQVLSGVVSFQLMPGEVHALMGENGAGKSTLMNILTGLHKADKQISIN 62
+++K I K F + LSG + + PG V AL+GENGAGKST+M +LTG++ D G +
Sbjct 5 LQLKGIDKAFPGVKALSQAALNVYPGRVMALVGENGAGKSTMMKVLTGIYTRDAGTLLWL 64

Query 63 GNETYFSPNPKAEAEQHGIAFIHQELNIWPEMTVLENLFIGKEISSKLGVLQTRKMKALAKE 122
G ET F+ PK +++ GI IHQELN+ P++T+ EN+F+G+E ++ G + + M A A +
Sbjct 65 GKETTFTGPKSSQEAGIGIIHQELNLIPQLTIAENIFLGRFVNRFGKIDWKTMAYAEADK 124

Query 123 QFDKLSVSLSLDQEAGECSVGGQQMIEIAKALMTNAEVIIMDEPTAALTEREISKLFEEVI 182
KL++ D+ G+ S+G QQM+EIAK L ++VIIMDEPT ALT+ E LF VI
Sbjct 125 LLAKLNLRFKSDKLVGDLSIGDQQMVEIAKVLSFESKVIIMDEPTDALTDTETESLFRVI 184

Query 183 TALKKNGVSIYIYISHRMEEIFAICDRITIMRDGKTVDTTNISETFDEVVKKMVGRELTE 242
LK G IVYISHRM+EIF ICD +T+ RDG+ + ++ D +++ MVGR+L +
Sbjct 185 RELKSQGRGIVYIYISHRMKEIFEICDDVTVFRDQGQFIAEREVASLTEDSLIEMMVGRKLED 244

Query 243 RYPKRTPSLGDKVFEVKNASVKGSFEDVSFYVRSGEIVGVSGLMGAGRTEMMRALFGVDR 302
+YP + GD +V N G DVSF +R GEI+GVSGLMGAGRTE+M+ L+G
Sbjct 245 QYPHLDKAPGDIRLKVNDLCPG-VNDVSFTLRKGEILGVSGLMGAGRTELMKVLYGALP 303

Query 303 LDTGEIWIAGKKTAIKNPQEAVKKGLGFITENRKDEGLLLDTSIRENIALPNLSSFS-PK 361
+G + + G + ++PQ+ + G+ +I+E+RK +GL+L S++EN++L L FS
Sbjct 304 RTSGYVTLDGHEVVTRSPQDGLANGIVYISEDKRKRDGLVLGMSVKENMSLTALRYFSRAG 363

Query 362 GLIDHKREAEFVDLLIKRLTIKTASPETHARHLSGGNQQKVVIKWIIGPKVLILDEPT 421
G + H E + V I+ +KT S E LSGGNQQKV IA+ + PKVLILDEPT
Sbjct 364 GSLKHADEQQAVSDFIRLFNVKTPSMEQAIGLLSGGNQQKVAIARGLMTRPKVLILDEPT 423

Query 422 RGVVDVGAKREIYTLMNELTERGVAIIMVSSELPEILGMSDRIIVVHEGRISGEIHAREAT 481
RGVVDVGAK+EIY L+N+ G++II+VSSE+PE+LGMSDRIIV+HEG +SGE +AT
Sbjct 424 RGVVDVGAKKEIYQLINQFKADGLSIIILVSSEMPEVLGMSDRIIVMHEGHLSGEFTREQAT 483

Query 482 QERIMTLATG 491
QE +M A G
Sbjct 484 QEVLMAAAVG 493
```

Résultat d'une recherche
avec BlastP2 sur le site
du NCBI (suite)