Programmes de comparaison de banques de données

FASTA

BLAST

introduction

- Programmes de comparaison de 2 séquences trop longs
- Méthodes heuristiques
- But: filtrer par étapes successives les séquences « intéressantes »
- Etablissement d'un score pour classer les meilleures similitudes locales.
- Les 2 plus utilisés: FASTA et BLAST.

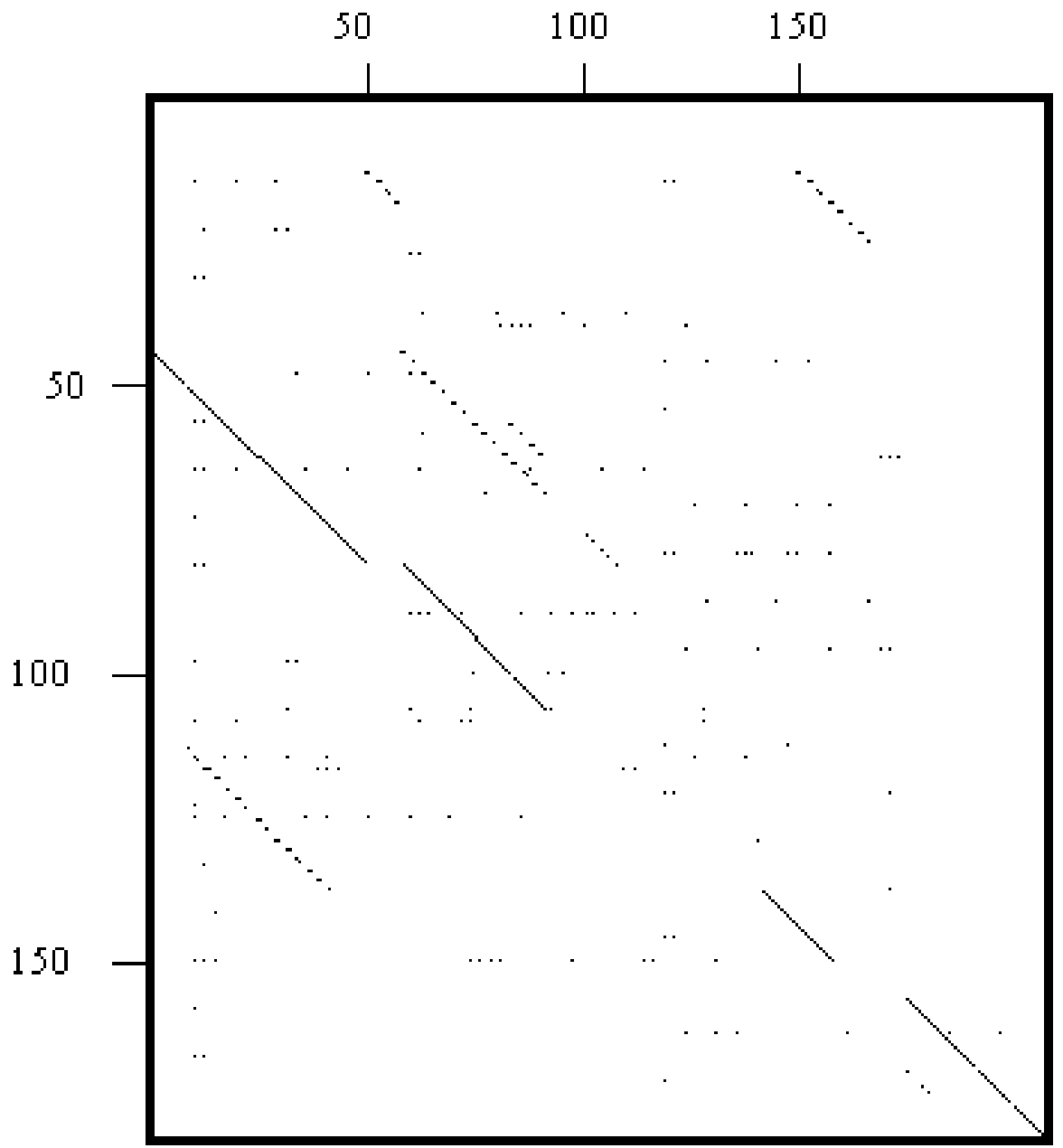
FASTA

*FASTA (pronounced FAST-Aye) stands for **FAST**-**A**ll, reflecting the fact that it can be used for a fast protein comparison or a fast nucleotide comparison*

W. R. Pearson and D. J. Lipman (1988), "Improved Tools for Biological Sequence Analysis", *Proc. Natl. Acad. Sci. USA.* **85**:2444- 2448,
W. R. Pearson (1990) "Rapid and Sensitive Sequence Comparison with FASTP and FASTA" *Methods in Enzymology* **183**:63- 98

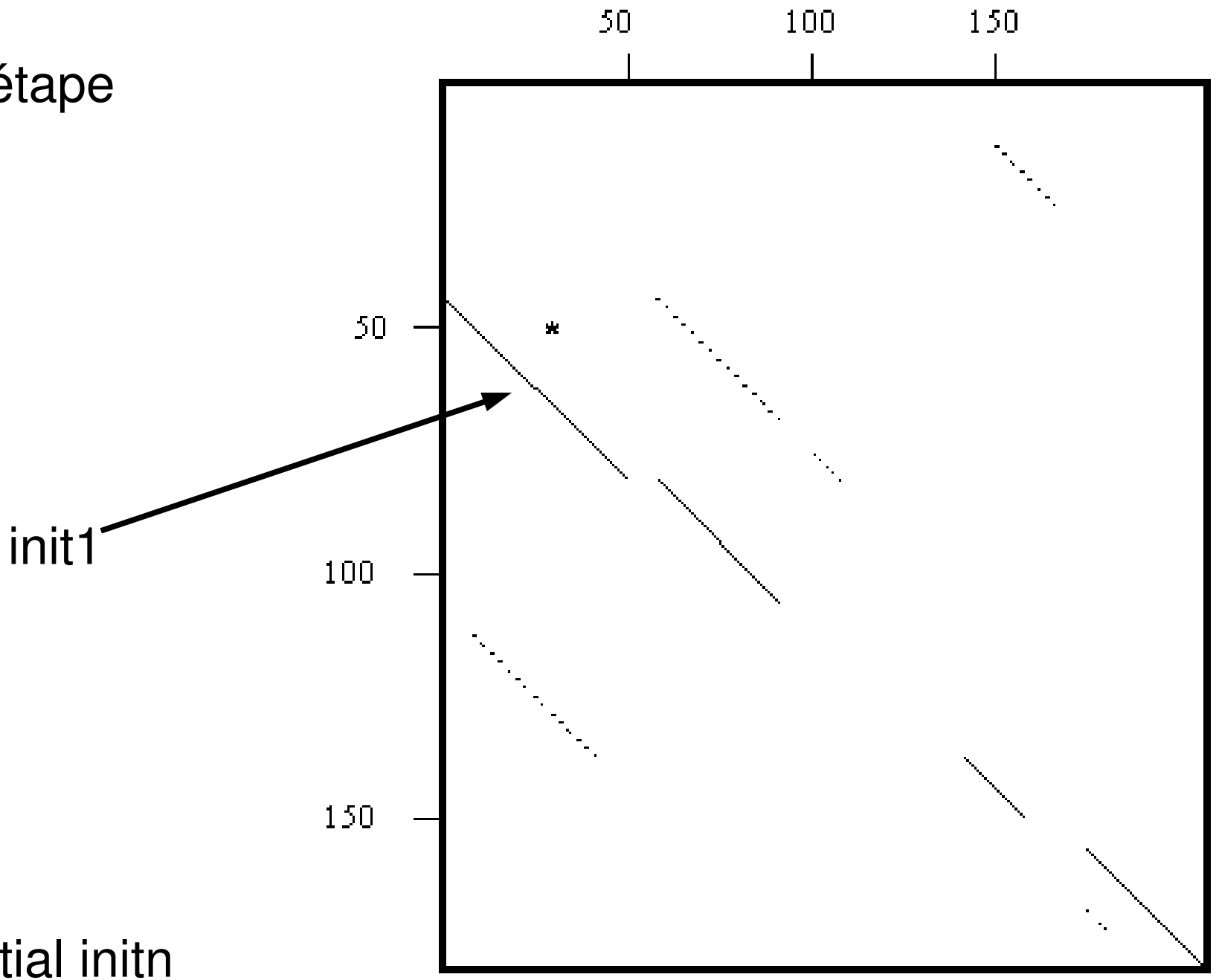
1ère étape

k-tuple
protéine: k = 2
ac.nucléique: k = 4 à 6



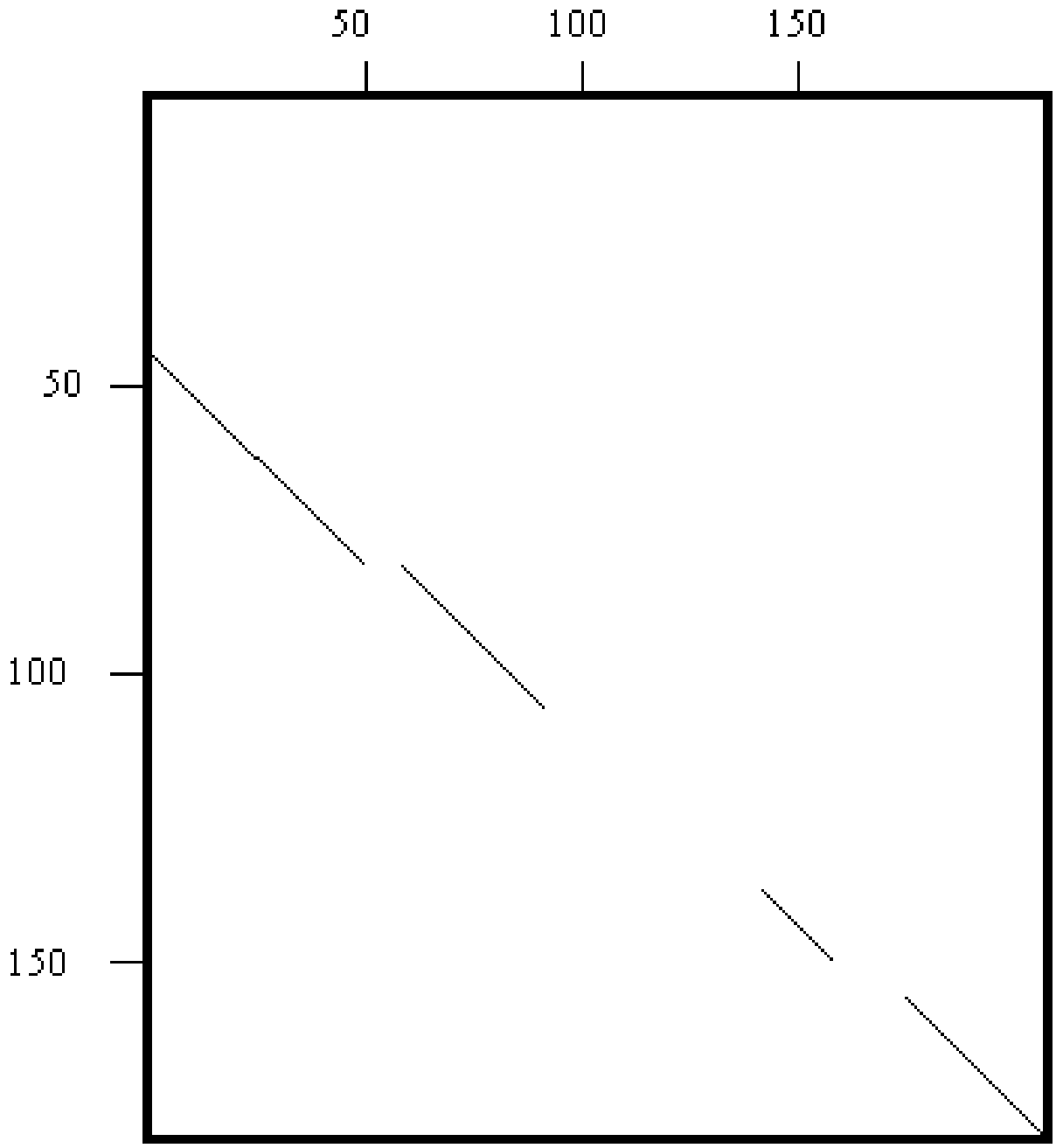
2ème étape

PAM250



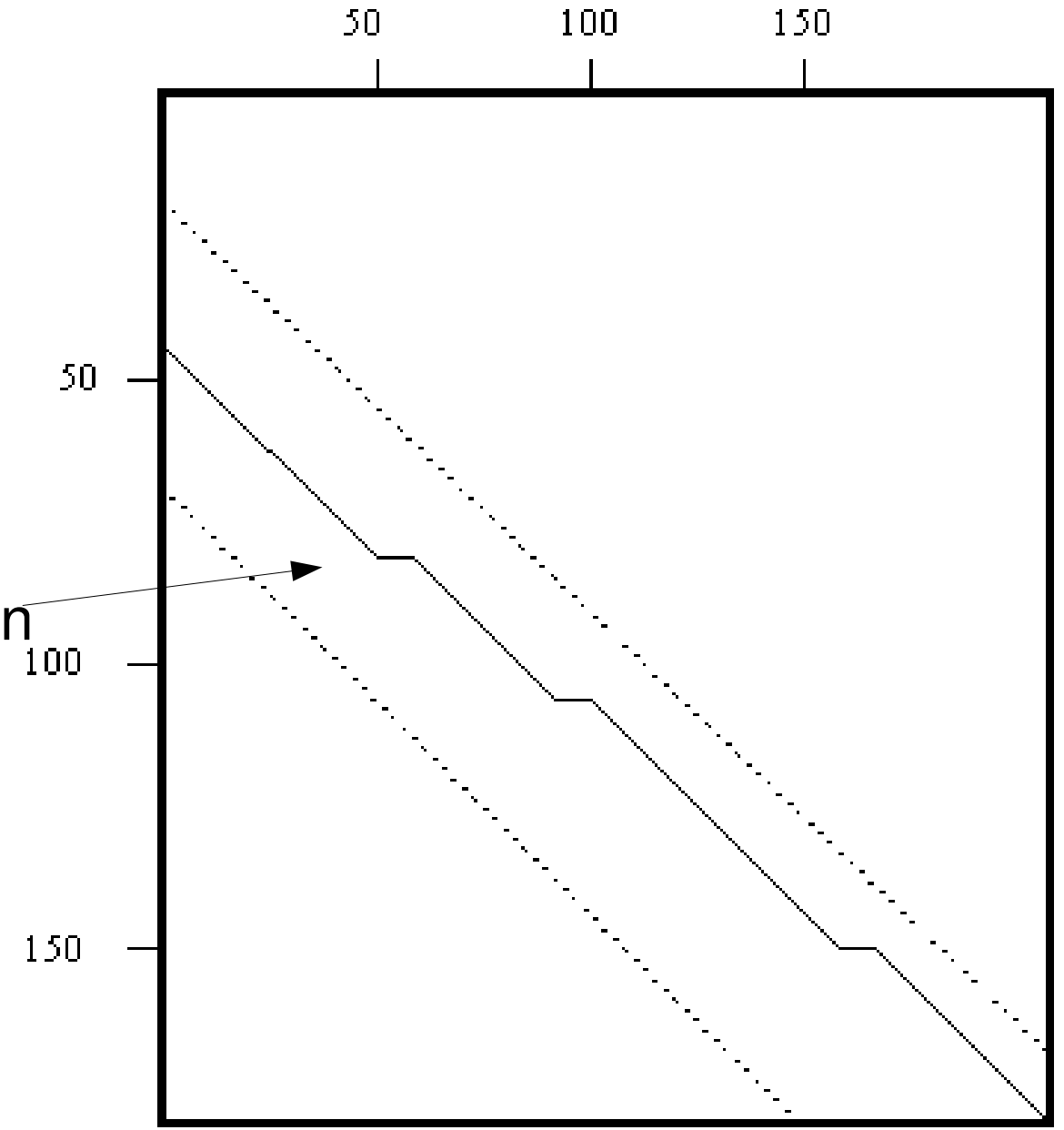
Score initial initn

3ème étape



4ème étape

Score optimal *opt*



bande d'insertion-délétion

Résumé des étapes de calcul

- Recherche des régions à forte identité (par k-tuple).
- Recalcule à l'aide d'une matrice à scores pour les 10 meilleures régions trouvées précédemment (les scores init1 = régions initiales de 1er ordre)
- Joindre les régions. (obtention d'un score initn)
- Alignement optimal des 2 séquences uniquement dans une région délimitée par la meilleure région initiale initn. Est réalisé avec un nombre limité de séquences fixé par l'utilisateur. On obtient un score opt.

Avantages

- bonne sensibilité car prend en compte les insertions-délétions.
- minimisation des explorations entre les deux séquences
  - étape de programmation dynamique, en ciblant de plus, les régions où l'on doit effectuer la recherche d'alignement.
  - étape d'alignement optimal est réalisée uniquement sur la meilleure région de haute similitude.
- évite en partie le bruit de fond dû à des motifs non significatifs et intrinsèques à la séquence recherchée

Inconvénients

- ne pas pouvoir considérer de grandes insertions durant l'alignement des séquences.
- Fondé sur méthode heuristique.

Améliorations

considère la totalité des diagonales pour effectuer l'algorithme d'alignement local de Smith et Waterman plutôt que d'effectuer l'alignement global de Needleman et Wunsch uniquement sur des portions de séquences protéiques.

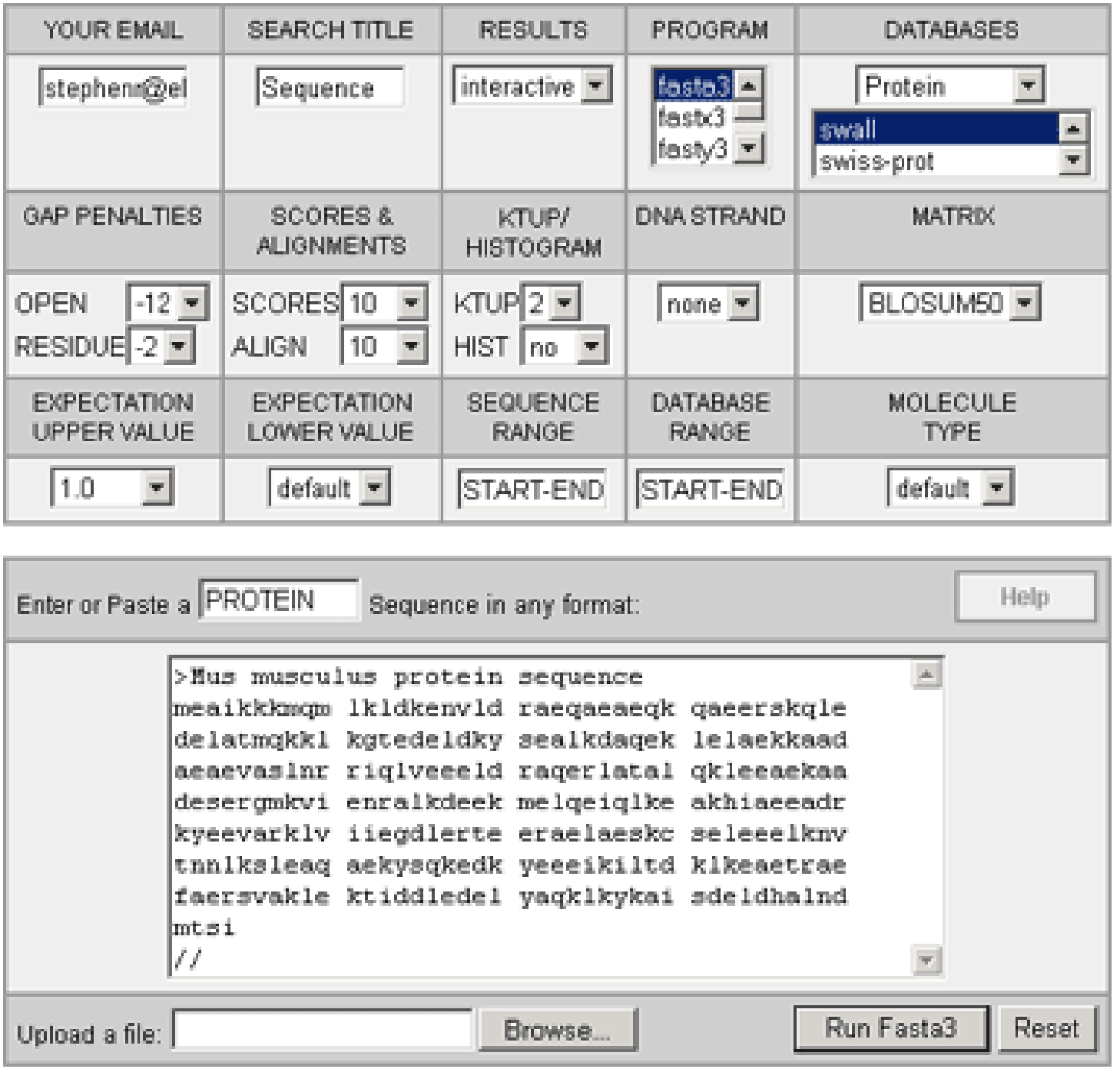
Edition des résultats

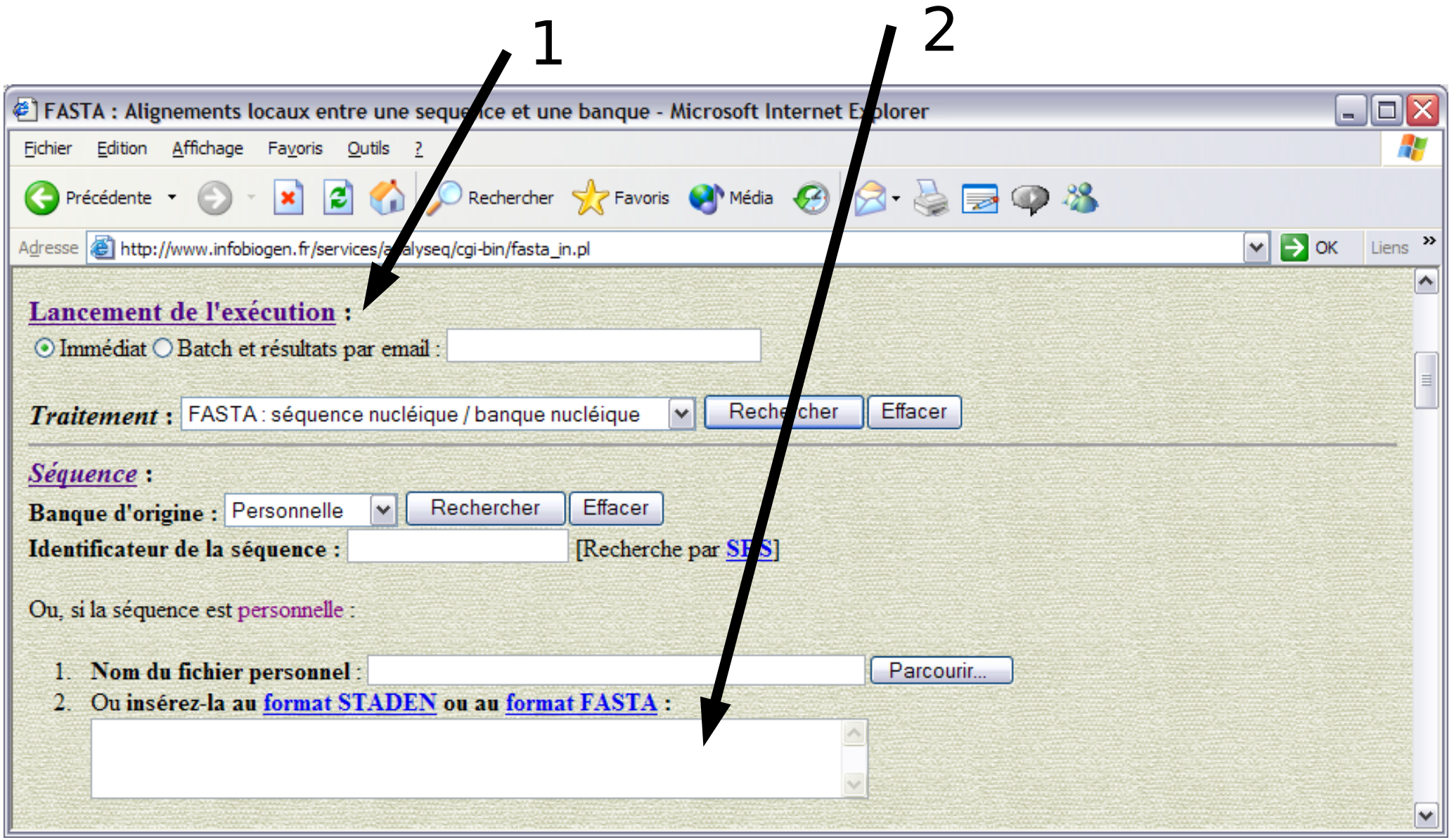
en fonction des scores *opt*.

Evaluation des résultats

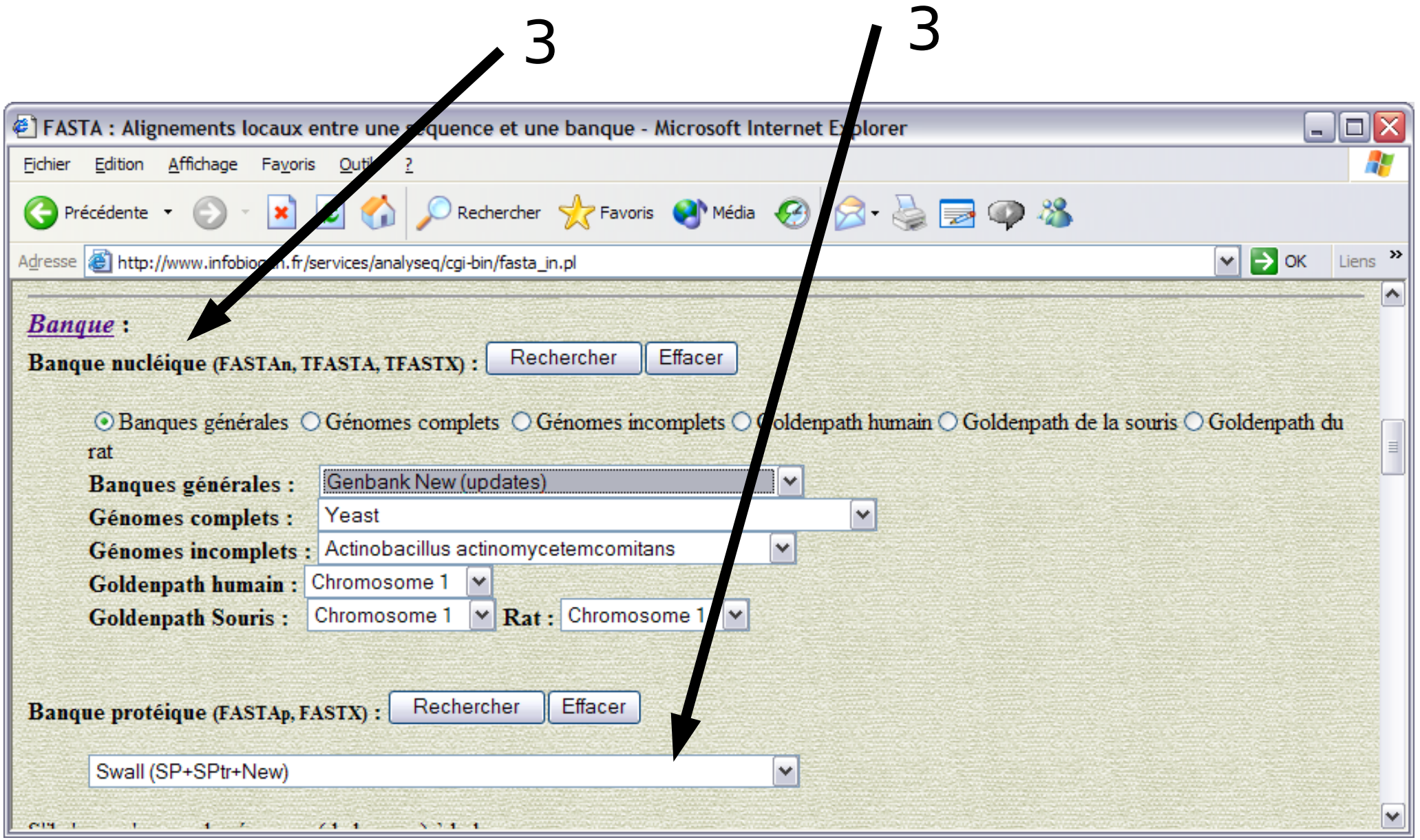
- L'estimation statistique est faite à partir des scores obtenus avec l'ensemble des séquences de la banque.
- programmes PRDF et PRSS (méthode de Monte Carlo) pour estimer la validité d'un score *opt* particulier entre une séquence de la banque et la séquence recherchée.
- PRDF produit des simulations selon l'algorithme de Needleman et Wunsch appliqué localement pour l'étape d'alignement optimal.
- PRSS utilise l'algorithme complet de Smith et Waterman entre deux séquences protéiques.

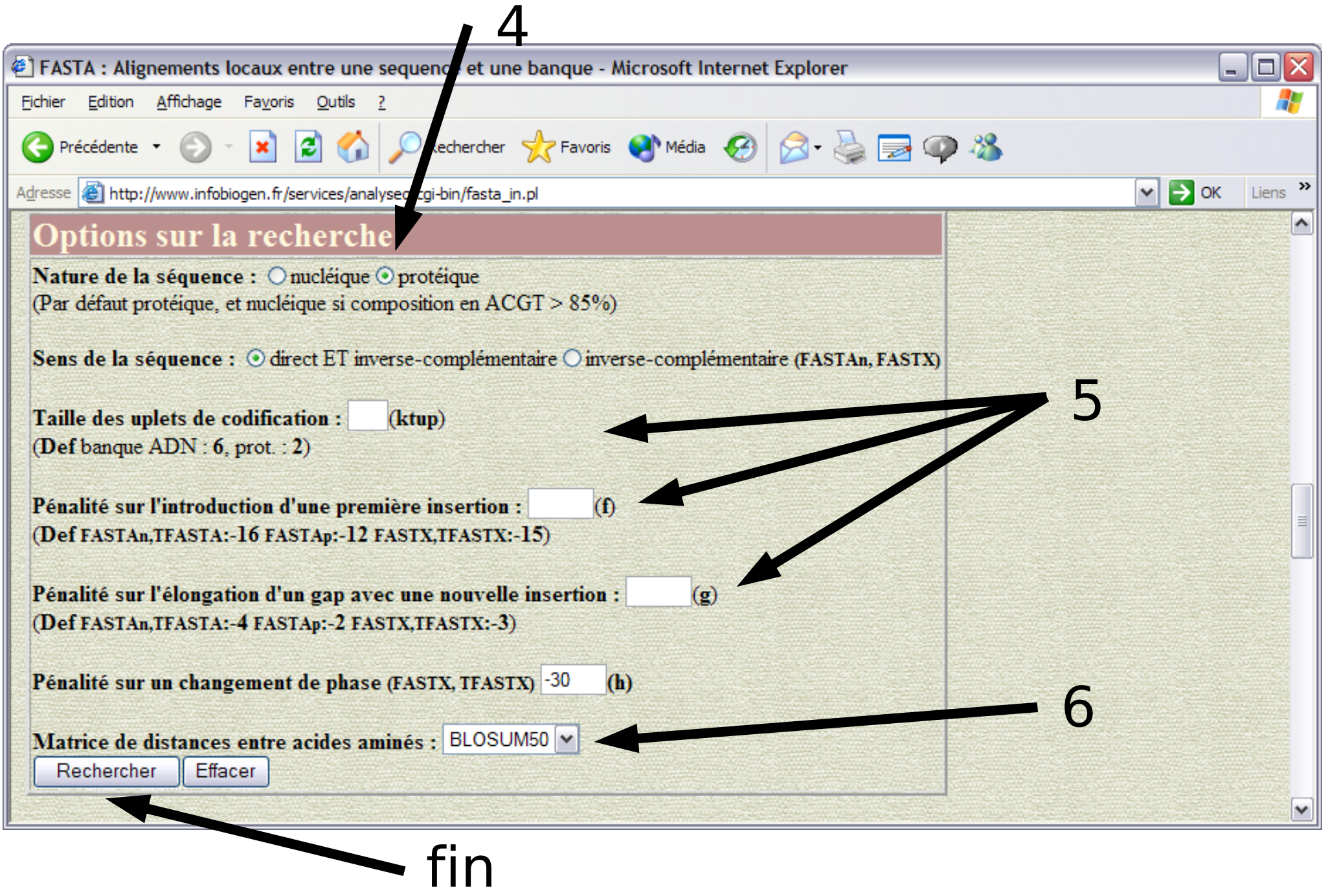
FASTA version 3





FASTA version 2


3
3
FASTA : Alignements locaux entre une séquence et une banque - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse
http://www.infobiogen.fr/services/analyseq/cgi-bin/fasta_in.pl
OK
Liens
Banque :
Banque nucléique (FASTAn, TFASTA, TFASTX) :
Rechercher
Effacer
Banques générales
Génomes complets
Génomes incomplets
Goldenpath humain
Goldenpath de la souris
Goldenpath du rat
Banques générales :
Genbank New (updates)
Génomes complets :
Yeast
Génomes incomplets :
Actinobacillus actinomycetemcomitans
Goldenpath humain :
Chromosome 1
Goldenpath Souris :
Chromosome 1
Rat :
Chromosome 1
Banque protéique (FASTAp, FASTX) :
Rechercher
Effacer
Swall (SP+SPtr+New)


4
FASTA : Alignements locaux entre une sequenc et une banque - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse
http://www.infobiogen.fr/services/analyseq/cgi-bin/fasta_in.pl
OK
Liens
Options sur la recherche
Nature de la séquence : nucléique protéique
(Par défaut protéique, et nucléique si composition en ACGT > 85%)
Sens de la séquence : direct ET inverse-complémentaire inverse-complémentaire (FASTAn, FASTX)
Taille des uplets de codification : (ktup)
(Def banque ADN : 6, prot. : 2)
Pénalité sur l'introduction d'une première insertion : (f)
(Def FASTAn,TFASTA:-16 FASTAp:-12 FASTX,TFASTX:-15)
Pénalité sur l'élongation d'un gap avec une nouvelle insertion : (g)
(Def FASTAn,TFASTA:-4 FASTAp:-2 FASTX,TFASTX:-3)
Pénalité sur un changement de phase (FASTX, TFASTX) -30 (h)
Matrice de distances entre acides aminés : BLOSUM50
Rechercher
Effacer
5
6
fin

Program	Function
fasta	scan a protein or DNA sequence library for similar sequences
fastx/y	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfastx/y	compares a protein to a translated DNA data bank.

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool

Karlin S. and Altschul S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA*, **87**, 2264-2268.

Karlin S. and Altschul S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA,* **90**, 5873-5877

- Conception fondée sur modèle statistique

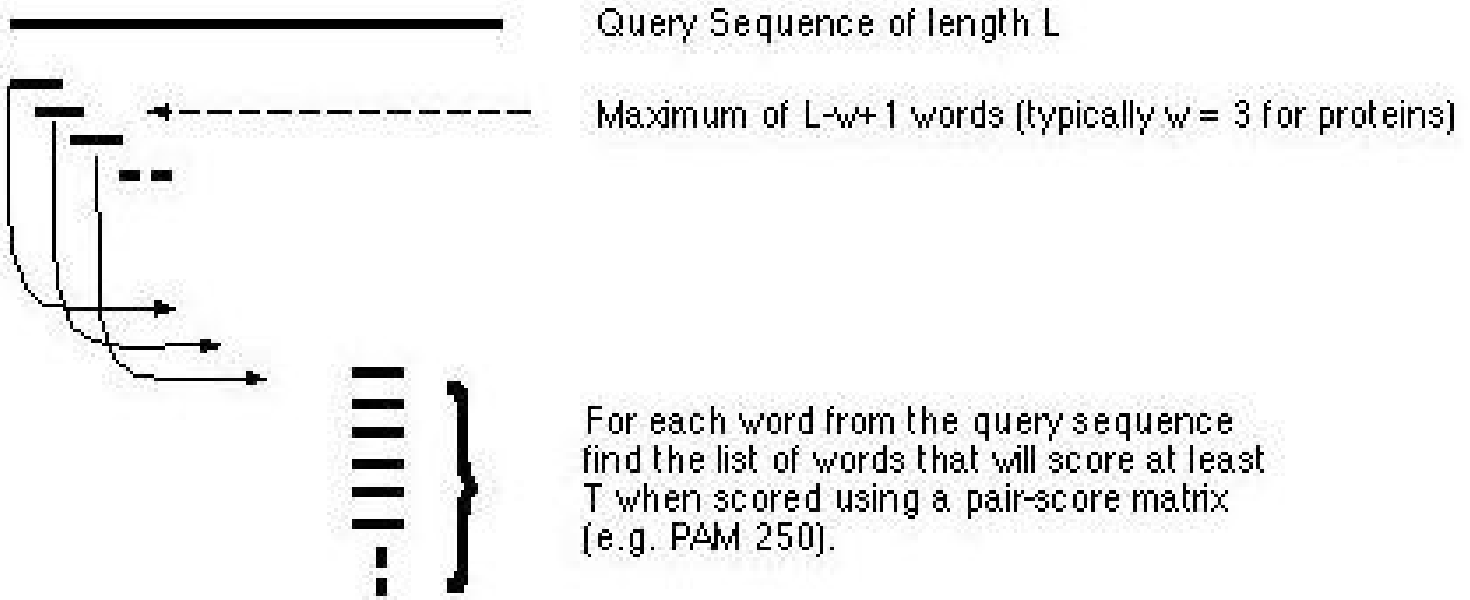
- HSP : High-scoring Segment Pair: un segment commun, de score significatif et le + long possible entre 2 séquences correspondant à une similitude sans insertion-délétion.

- MSP (Maximal-scoring Segment Pair). Le meilleur score obtenu parmi tous les couples de fragments possibles que peuvent produire 2 séquences.

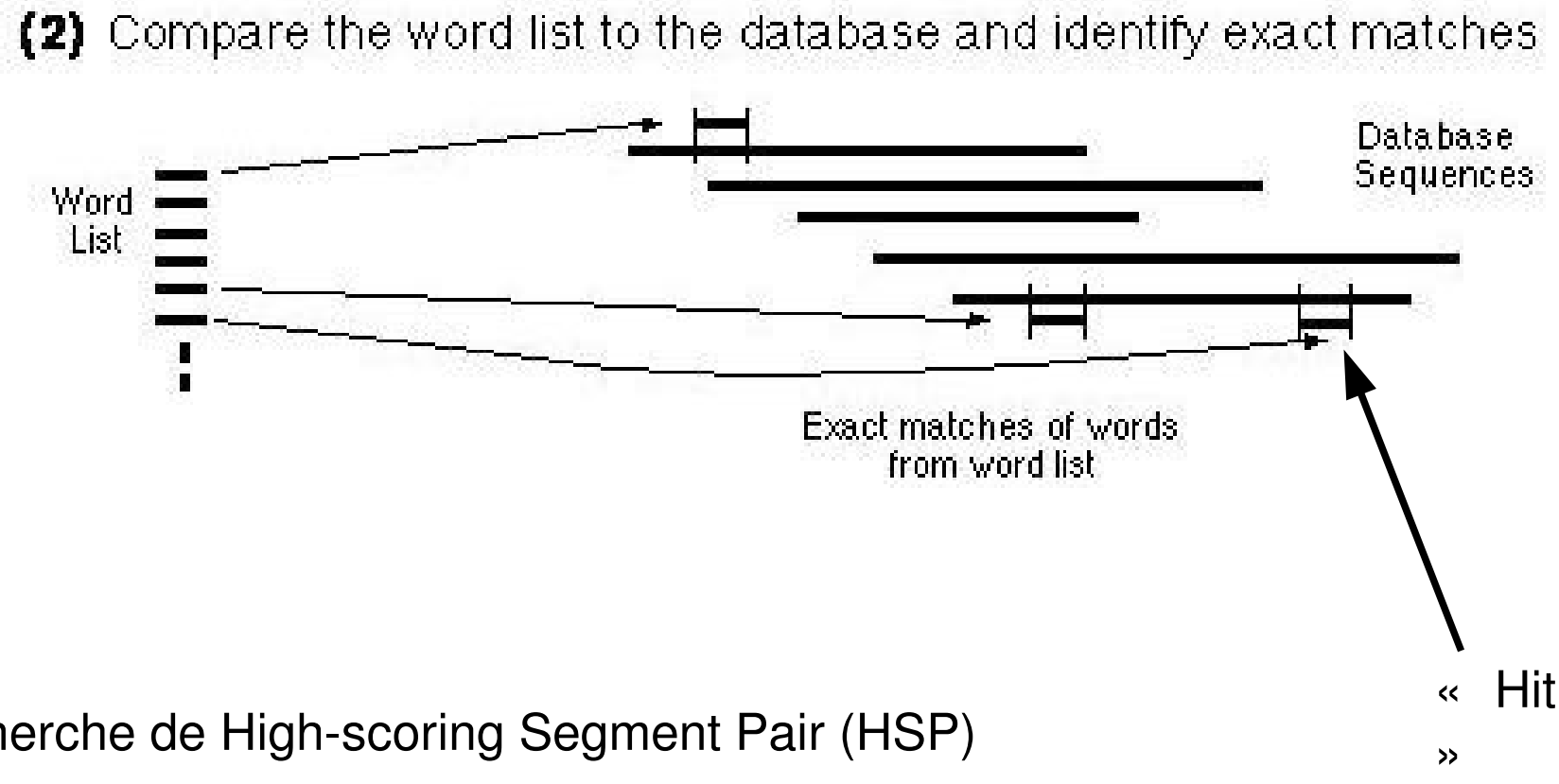
- Méthodes statistiques de BLAST permettent la détermination de la signification biologique des MSPs.

Protéine W = 3
Ac.nucléiques W = 11

(1) For the query, find the list of high scoring words of length w

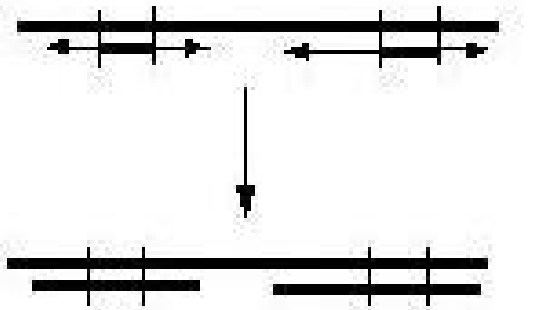


Pour les protéines
construction d'une liste de **mots similaires**.
Mot similaire : mot obtenant un score > seuil par matrice de substitution



Recherche de High-scoring Segment Pair (HSP)

(3) For each word match, extend the alignment in both directions to find alignments that score greater than a threshold of value S



Maximal Segment Pairs (MSPs)

Extension s'arrête si:
la fin d'une des 2 séquences est atteinte
le score cumulé ≤ 0
le score cumulé < max - x

Figure from Barton, G.J. Protein Sequence Alignment and Database Scanning (University of Oxford, Laboratory of Molecular Biophysics)

Evaluation des résultats

Score (bits) $S' = (\lambda S - \ln K) / \ln 2$

λ et **K :** constantes statistiques du système de scores

S : score brut

m,n : longueurs des 2 séquences

E-value: $E = mn2^{-S'}$

Avantages

- Algorithme fondé sur critères statistiques
- Recherche des fragments identiques mais aussi similaires (pour protéines).
  - Via la matrice de substitution, intègre des critères biologiques.

- Résultats triés selon plusieurs critères (ex: signification statistique et non pas seulement valeur de score)
- Très rapide (optimisation du programme, précodification de la banque)

Inconvénients

- bruit de fond important lors de l'identification des segments. Si séquence possède des régions répétées ou des segments de basse complexité (segments non spécifiques d'une caractéristique biologique mais communs à plusieurs familles).
  - Filtres: SEG ou XNU

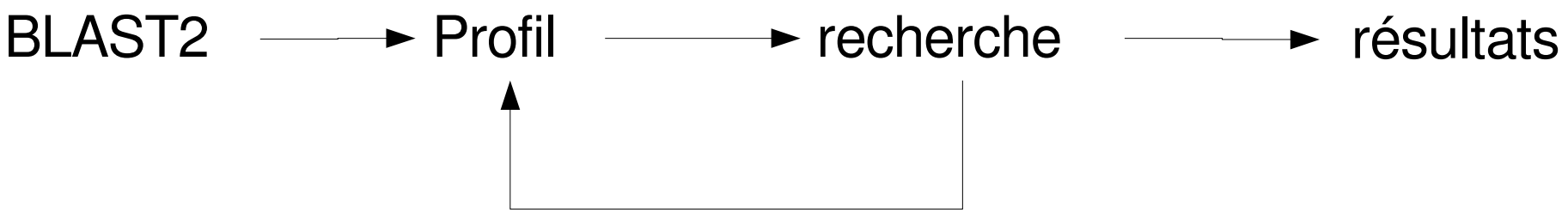
Program	Function
BLASTn	DNA sequence vs DNA sequence library
BLASTp	protein sequence vs protein sequence library
BLASTX	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
TBLASTN	compare a protein sequence to a nucleic sequence database translated in forward and reverse frames.

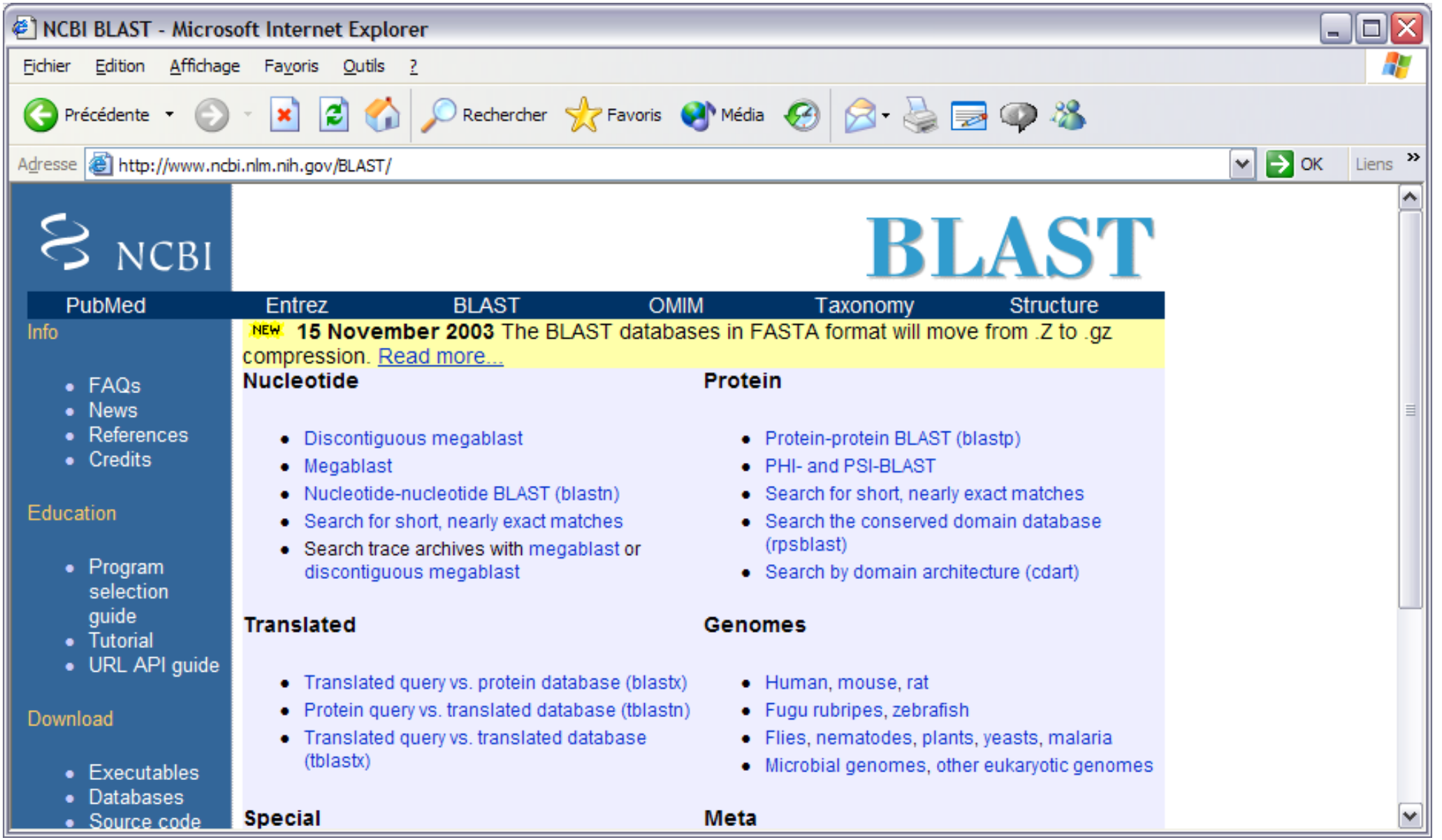
BLAST2

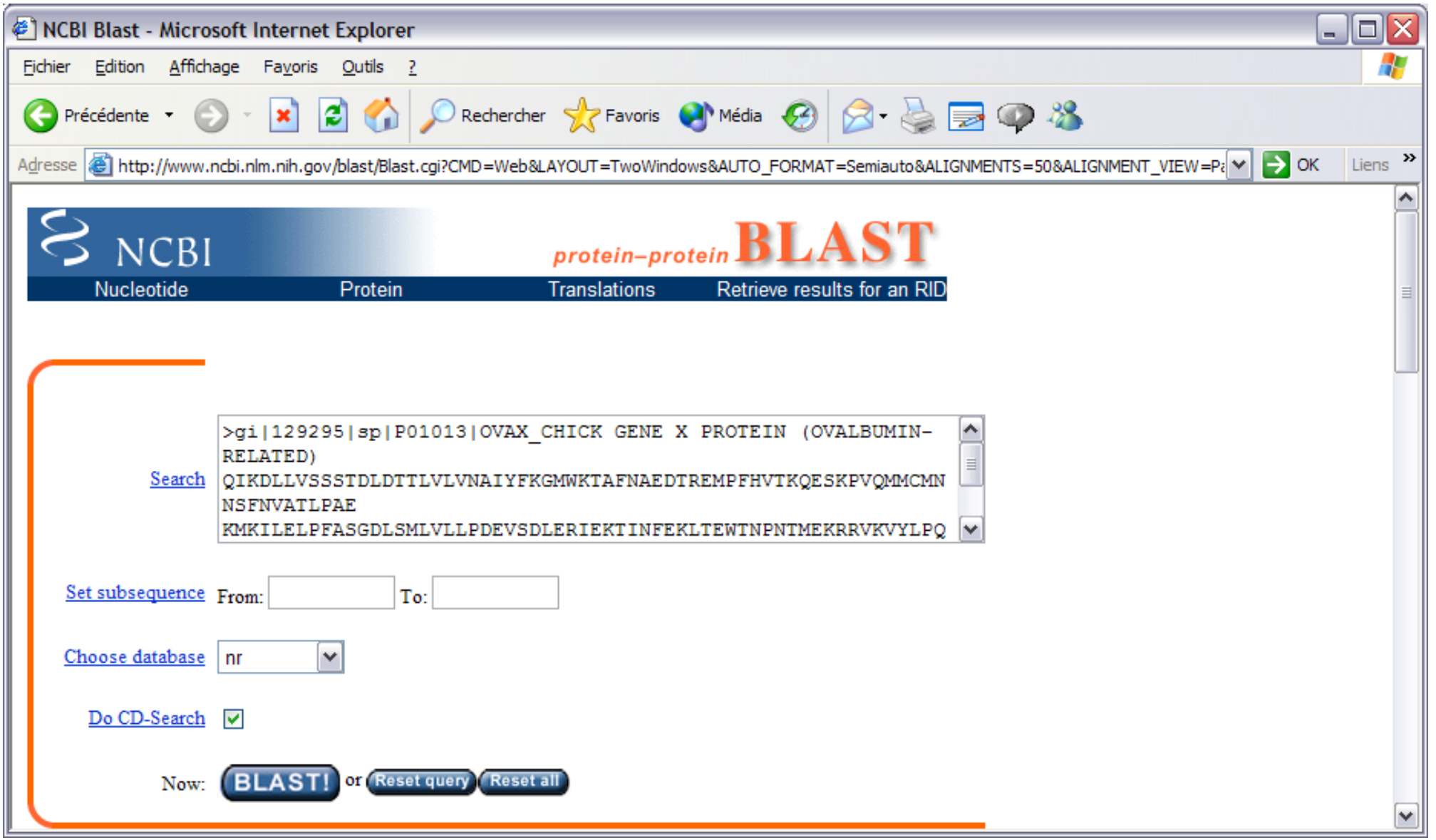
- 3x + rapide que BLAST1
- phase d'extension n'a lieu que si 2 hits sont sur la même diagonale.

PSI-BLAST (Position Specific Iterative)

- sensibilité accrue
- construction d'un profil à partir d'un 1er BLAST classique
- utile pour rechercher membres d'une même famille.
- Déduire fonctions de protéines hypothétiques

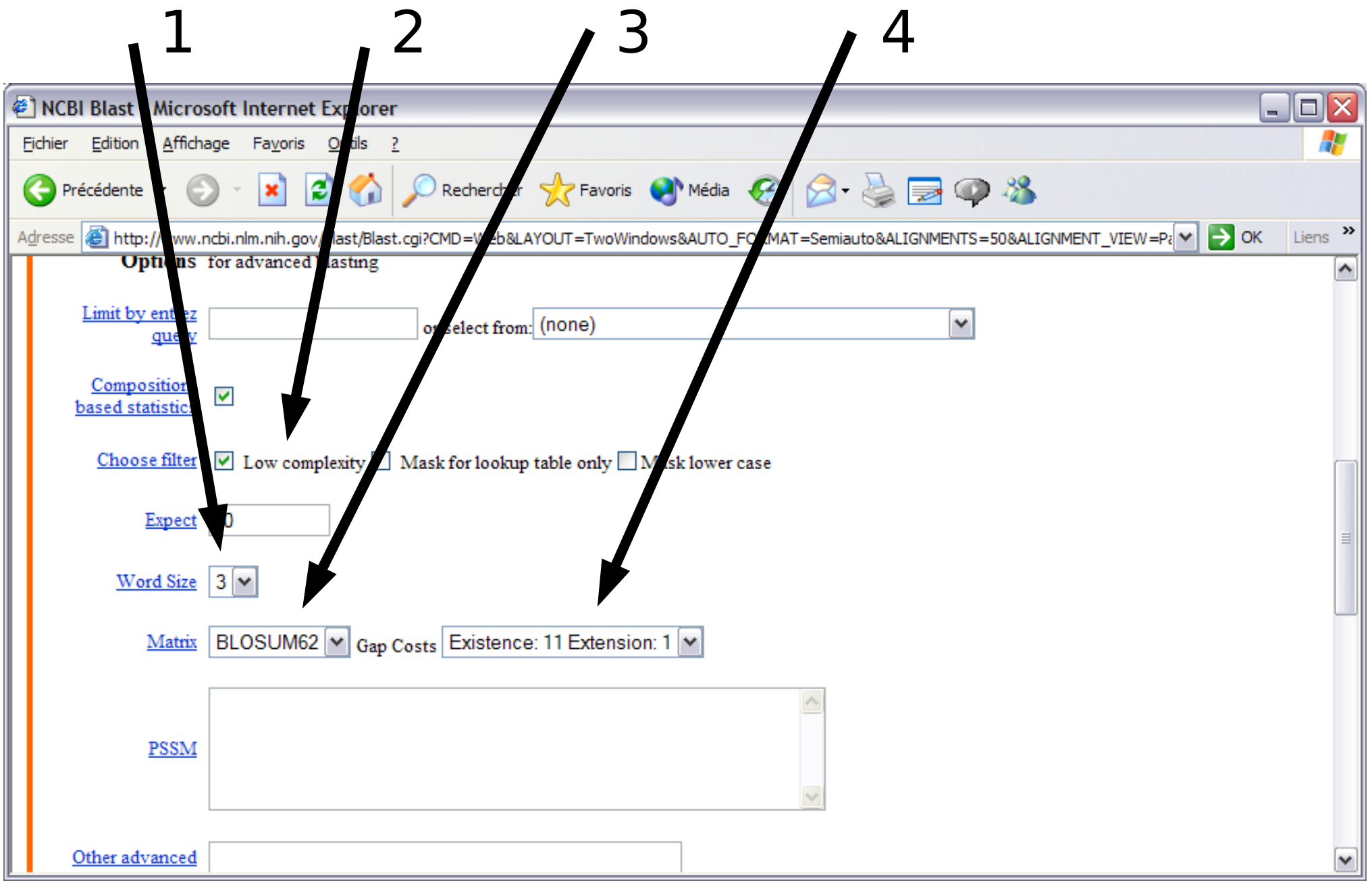



NCBI BLAST - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse
http://www.ncbi.nlm.nih.gov/BLAST/
OK
Liens
NCBI
BLAST
PubMed Entrez BLAST OMIM Taxonomy Structure
Info
FAQs
News
References
Credits
Education
Program selection guide
Tutorial
URL API guide
Download
Executables
Databases
Source code
NEW 15 November 2003 The BLAST databases in FASTA format will move from .Z to .gz compression. Read more...
Nucleotide
Discontiguous megablast
Megablast
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
PHI- and PSI-BLAST
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Search by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Human, mouse, rat
Fugu rubripes, zebrafish
Flies, nematodes, plants, yeasts, malaria
Microbial genomes, other eukaryotic genomes
Special
Meta



nr

All non-redundant GenBank CDS translations+RefSeq Proteins + PDB + SwissProt +PIR+PRF

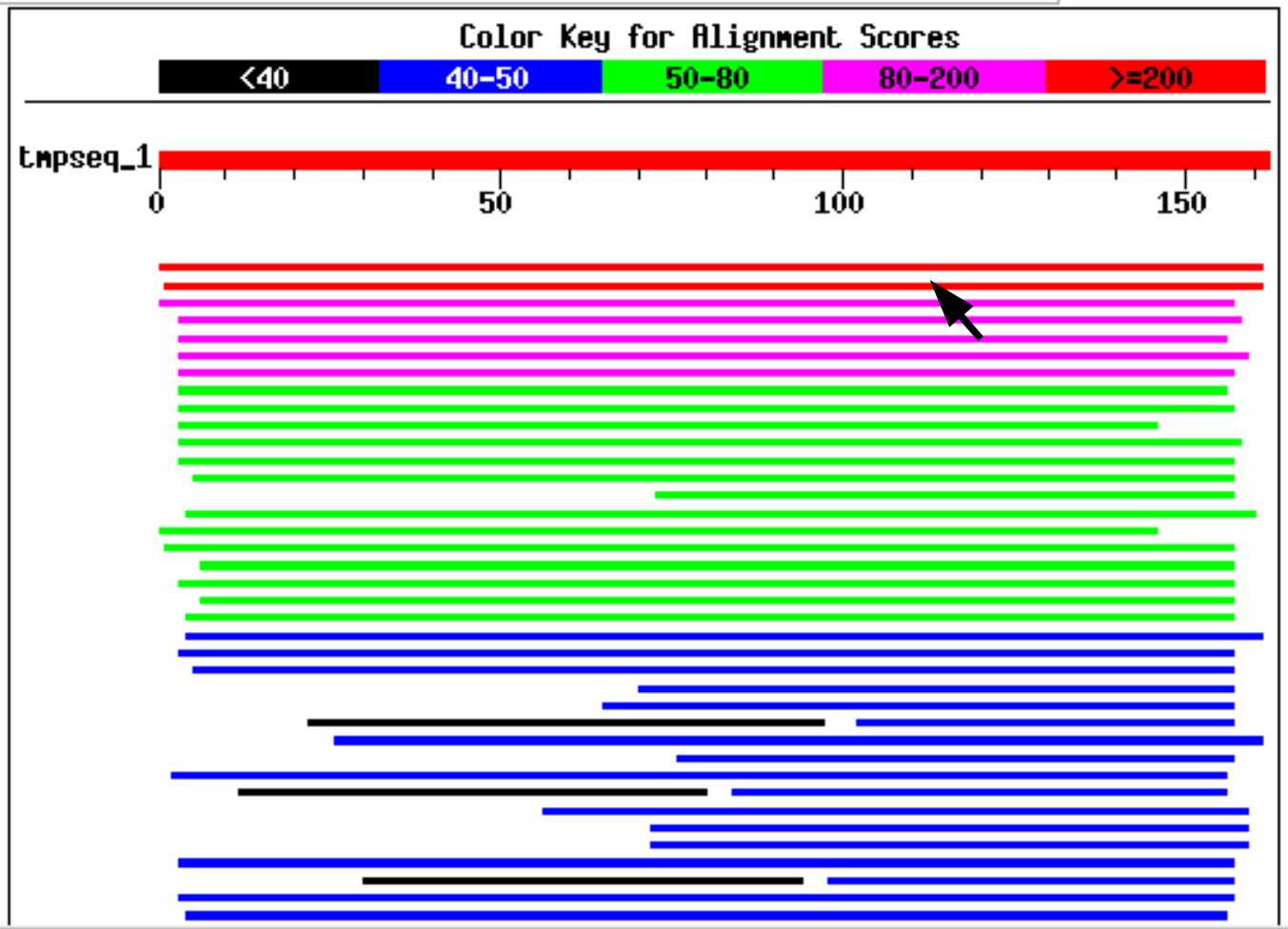

1
2
3
4
NCBI Blast - Microsoft Internet Explorer
Fichier Edition Affichage Favoris Outils ?
Précédente
Rechercher
Favoris
Média
Adresse
OK
Liens
Options for advanced blasting
Limit by entrez query
or select from: (none)
Composition based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
PSSM
Other advanced

BLAST tutorial (part 2) - Microsoft Internet Explorer

see the new alignments, use the "format results" button once again.

Distribution of 95 Blast Hits on the Query Sequence

1MJH Structure-Based Assignment Of The Biochemical Function O..S= 272 E=2e-72



Sequences producing significant alignments: Score E Value

(bits)

1.	sp|Q57997|Y577_METJA PROTEIN MJ0577 >gi|2128018|pir||A64372...	314	2e-85
2.	pdb|1MJH| Structure-Based Assignment Of The Biochemical F...	272	1e-72
3.	dbj|BAA29916| (AP000003) 170aa long hypothetical protein [P...	107	6e-23
4.	sp|Q57951|Y531_METJA HYPOTHETICAL PROTEIN MJ0531 >gi|212801...	91	4e-18
5.	gi|2622094 (AE000872) conserved protein [Methanobacterium t...	85	4e-16
6.	gi|2621993 (AE000865) conserved protein [Methanobacterium t...	81	4e-15
7.	gi|2621194 (AE000803) conserved protein [Methanobacterium t...	80	7e-15

```
sp|Q57951|Y531_METJA HYPOTHETICAL PROTEIN MJ0531 >gi|2128015|pir||C64366
hypothetical
           protein homolog MJ0531 - Methanococcus jannaschii
           >gi|1591234 (U67502) conserved hypothetical protein
           [Methanococcus jannaschii]
           Length = 170

 Score = 91.3 bits (223), Expect = 4e-18
 Identities = 59/156 (37%), Positives = 88/156 (55%), Gaps = 14/156 (8%)

Query: 4   MYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVAGLN 63
           +YKKI+ PTD S+ +  A KH          EV  ++V+D          S  +G+
Sbjct: 25  LYKKIVIPTDGSDVSLEAAKHAINIAKEFDAEVYAIYVVD---------VSPFVGLPA-- 73

Query: 64  KSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEGVDI 123
           +   E  +EL   L EE +  ++ +KK  E+ G K+   ++ G+P  EIV+ AE +  D+
Sbjct: 74  EGSWELISEL---LKEEGQEALKKVKKMAEEWGVKIHTEMLEGVPANEIVEFAEKKKADL 130

Query: 124 IIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKR 159
           I+MG+ GKT L+ ILLGSV E VIK ++ PVLVVK+
Sbjct: 131 IVMGTTGKTGLERILLGSVAERVIKNAHCPVLVVKK 166
```

Limites de la recherche de similarités pour déterminer une fonction

- Gènes inconnus : cas des gènes « orphelins »
- Erreurs
- Gènes orthologues et paralogues
- Évolution
  - Épissage alternatif: élimination différente des introns → ARNm différents
  - Association de fragments de gènes différents → fonctions nouvelles
- Maturation post-traductionnelle
  - Protéine codée par le génome ↔ protéine mature

- **Vérification expérimentale**