INSTITUT PASTEUR

Recherche de séquences similaires dans les banques de séquences

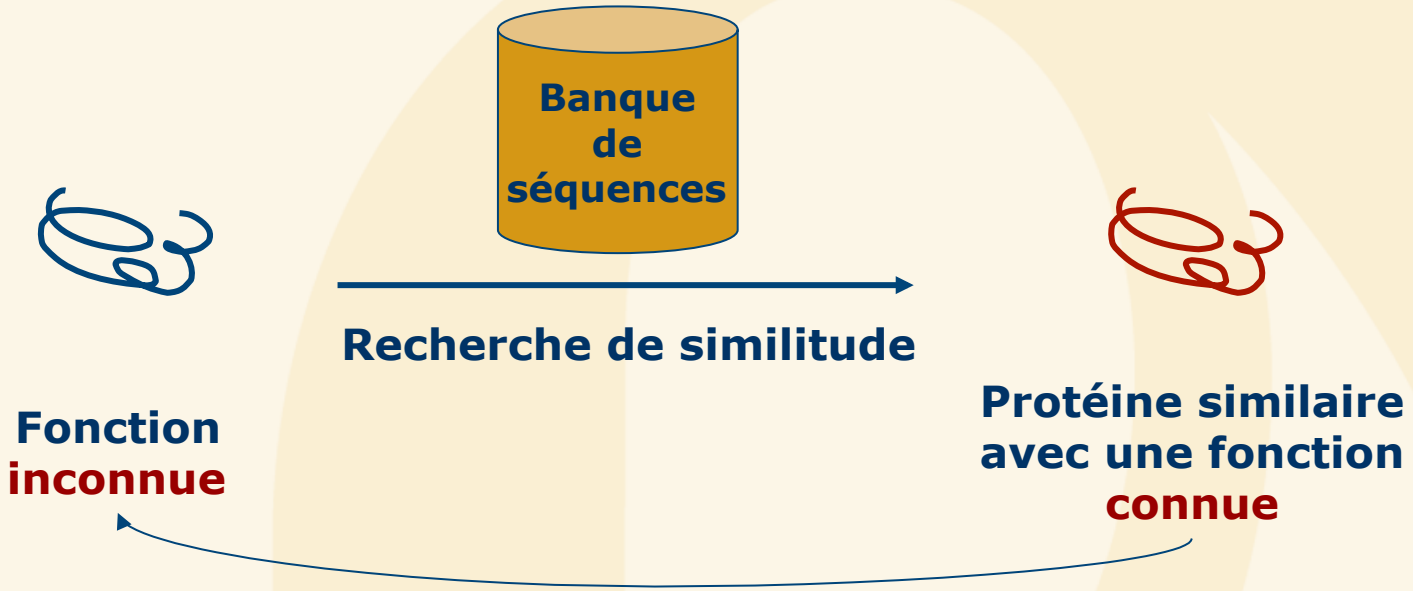
I. Introduction

II. Blast

Pourquoi rechercher des séquences dans les banques?

- Savoir si ma séquence ressemble à d'autres déjà connues.
- Trouver toutes les séquences d'une même famille.
- Identifier des protéines homologues.
- Déterminer si des séquences ont une fonction similaire ou proche.
- Rechercher toutes les séquences qui contiennent un motif donné
- Localiser des régions codantes et non codantes (aligner des séquences génomiques ADN et des séquences exprimées (cDNAs, ESTs)).

Pourquoi?



INSTITUT PASTEUR

Les données

- **Requête:**

  - ✓ADN, ARN, protéine, motifs…..

- **Banques de données**

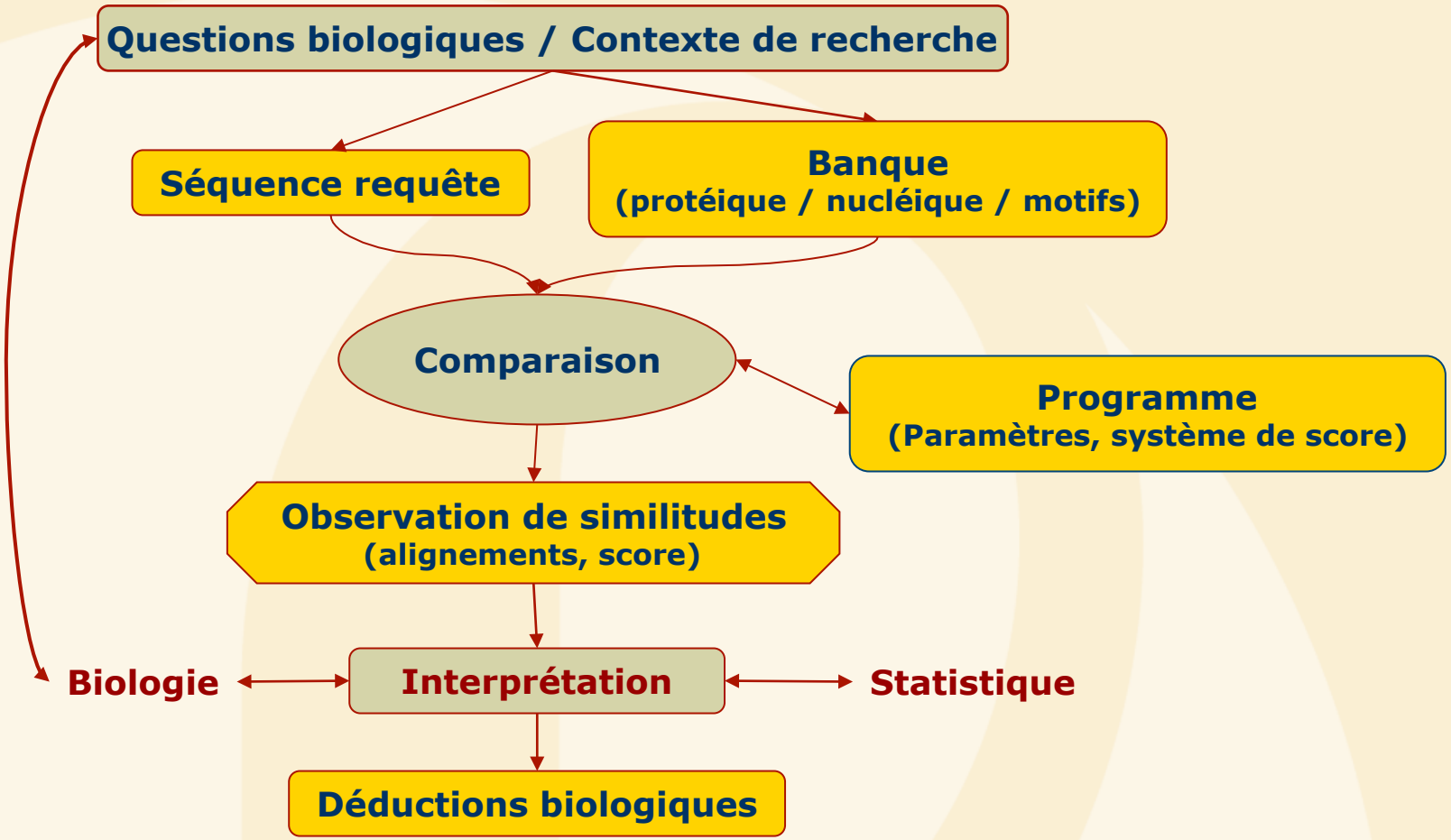
  - ✓ nucléiques : - EMBL, Genbank, DDBJ ….
  - ✓ protéique: - Uniprot/SwissProt, Uniprot/trEMBL, nr …

- **Recherche de similitude**

  - ✓ globale ou locale
  - ✓ plus ou moins significative (statistique)

→ **Nécessité de formaliser et de définir ce que l'on recherche.**

Formaliser



Programme => alignement + statistique → analyse biologique

INSTITUT PASTEUR

Acides aminés ou acides nucléiques

• **Utilise-t-on des séquences protéiques ou des séquences nucléiques pour effectuer une recherche de similitude?**

• **Si vous avez une séquence nucléique:**

- ✓ **faites-vous une recherche uniquement sur une banque nucléique?**
- ✓ **traduisez-vous la séquence nucléique en séquence protéique pour faire une recherche sur une banque protéique?**

• ***Remarque*: en traduisant une séquence nucléique en protéique, il y a une perte de l'information (dégénérescence du code génétique)**

Acides aminés ou acides nucléiques

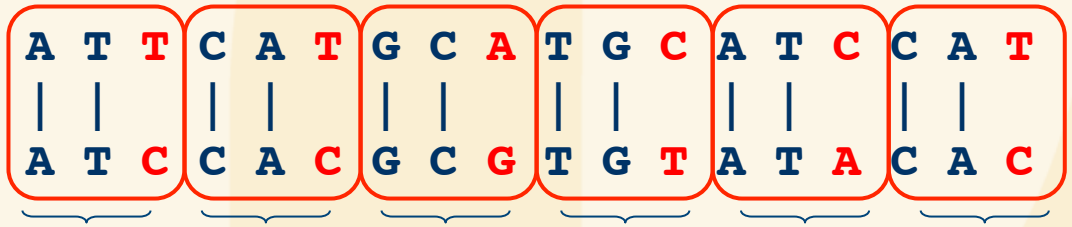
- **ADN** (A, G, C, T)

  ✓ Environ 25% des acides nucléiques de 2 séquences non similaires sont identiques

- **Protéine** (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y)

  ✓ La sensibilité de la comparaison augmente (5% d'identité dû au hasard)

  ✓ 2 protéines avec un fort taux de similitude sont homologues



INSTITUT PASTEUR

Acides aminés ou acides nucléiques

- **Utiliser des séquences protéiques dès que possibles.**

  ✓ **En comparant des séquences d'ADN, on rencontre plus de similitude dues au hasard (25%) par rapport aux séquences protéiques (5%).**

  ✓ **Les banques nucléiques sont beaucoup plus volumineuses et augmentent beaucoup plus rapidement que les banques protéiques**

=> plus la banque est importante plus les résultats dus au hasard sont nombreux

✓ **Pour l'ADN, on utilise souvent la matrice identité.**
✓ **Pour les protéines, on utilise les matrices PAM ou BLOSUM.**

=> plus de sensibilité

Comment effectuer une recherche?

- **Effectuer un alignement global et/ou local par programmation dynamique sur toutes les séquences de la banque.**

Trop coûteux en temps et en taille mémoire

Utilisation d'algorithmes « heuristiques »

Algorithme heuristique

• **Filtrer les données** de la banque en étapes successives (peu de séquences ont des similitudes avec la séquence requête).

• Les méthodes heuristiques utilisent des **approximations** pour éliminer rapidement les situations sans intérêt et ainsi repérer les séquences de la banque **susceptibles** d'avoir une relation avec la séquence recherchée.

☺ Algorithme très rapide

☹ L'alignement construit n'est pas nécessairement celui de score maximal.

Les principaux logiciels

- **Les deux programmes *heuristiques* les plus utilisés par les biologistes sont:**

  - **FASTA (*Pearson et Lipman, 1988*)**
  - **BLAST (*Altschul et al., 1990, 1997*)**

- **Ces programmes ont une approche différente mais complémentaire pour effectuer des recherches à travers une banque de données.**

- **Ils doivent être utilisés essentiellement comme logiciels permettant de repérer les séquences de la banque susceptibles d'avoir des ressemblances biologiques avec la séquence requête.**

- **Logiciels non optimisés pour comparer deux séquences entre elles.**

- **Les résultats qu'ils procurent devront être confirmés ou renforcés par d'autres logiciels plus spécialisés.**

INSTITUT PASTEUR

Recherche de similitudes dans les banques de séquences

I. Introduction

II. Blast

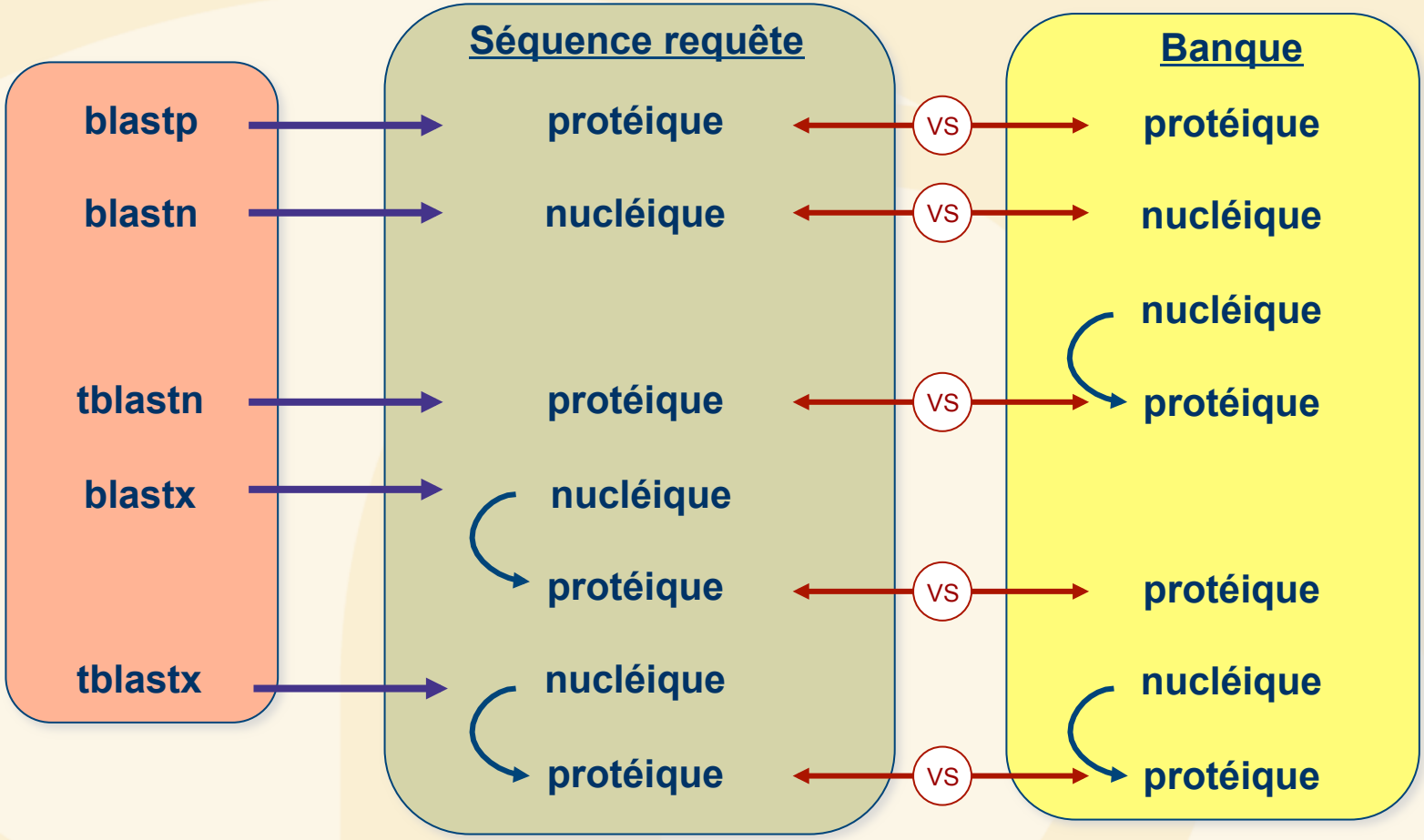
BLAST (Basic Local Alignment Search Tool)

- L'algorithme est basée sur un ***modèle statistique*** (Karlin et Altschul, 1990) qui s'applique aux comparaisons de séquences ***sans*** insertion-délétion.

- Il existe deux versions de l'algorithme:

  - ***BLAST 1.0*** (1990): **sans** insertion/délétion
  - ***WU-BLAST 2.0*** (1996) et ***NCBI-BLAST2*** (1997): prise en compte des insertions/délétions.

BLAST

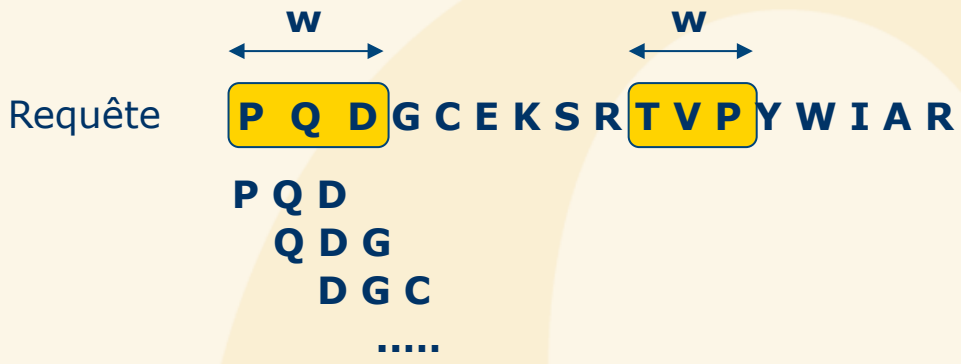


blastall –p blastp –i maSequence –d maBanque

INSTITUT PASTEUR

BLAST1: sans gap

1- Découpage de la séquence requête en mot de longueur *w* (ADN *w* = *11*; protéine *w* = *3*).



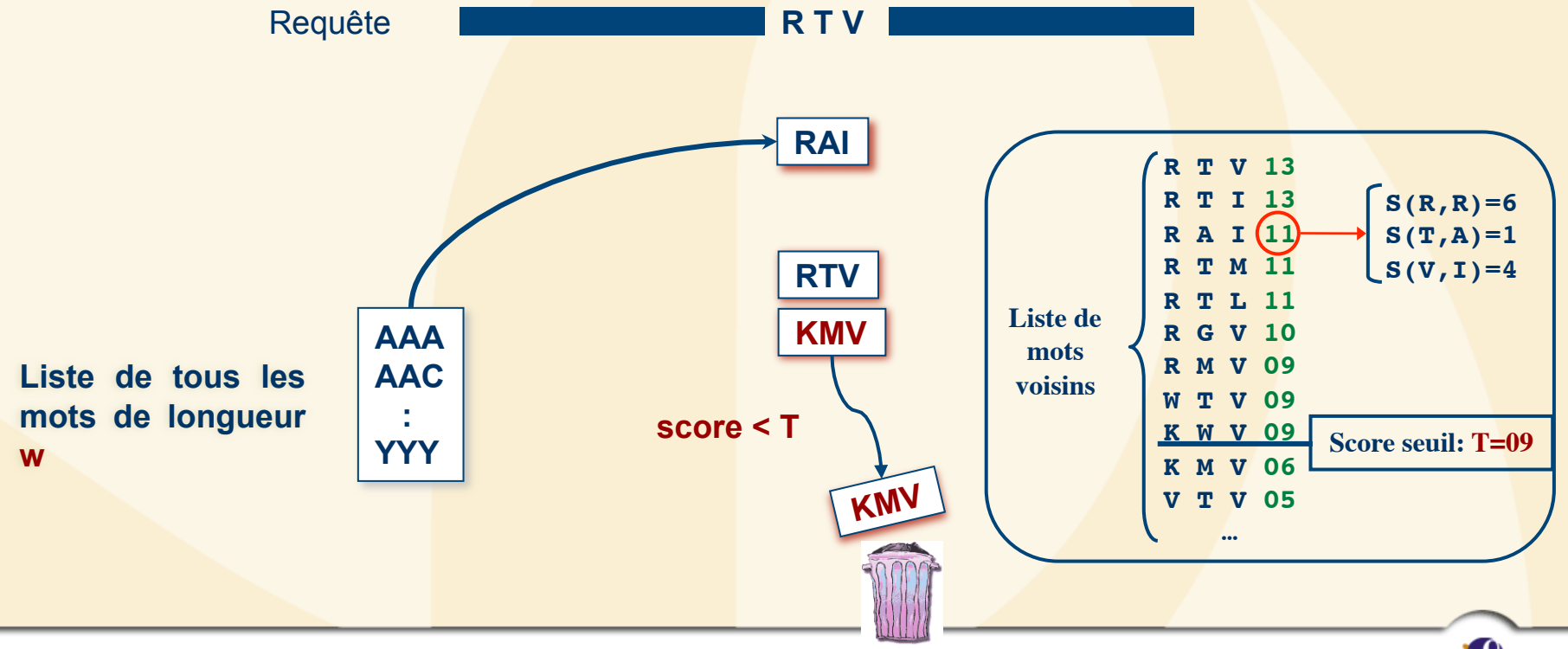
1.b- Elaboration de la liste de tous les mots possibles de longueur *w*.

AAA
AAN
AAR
AAL
....

INSTITUT PASTEUR

BLAST1: sans gap

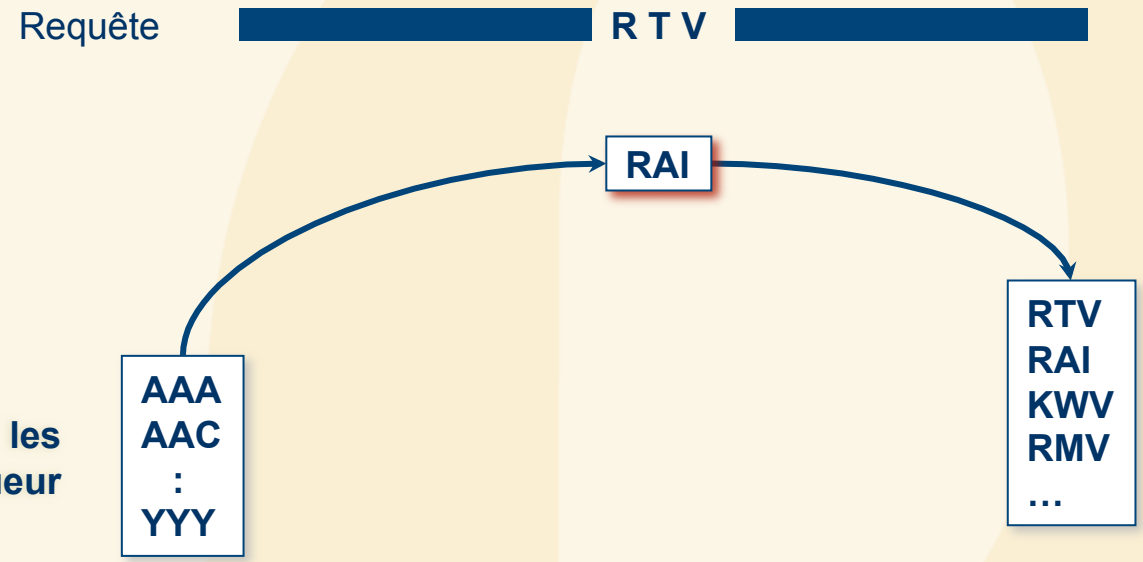
2- Pour chaque mot de la requête, sélection des mots similaires.



INSTITUT PASTEUR

BLAST1: sans gap

2- Pour chaque mot de la requête, sélection des mots similaires.



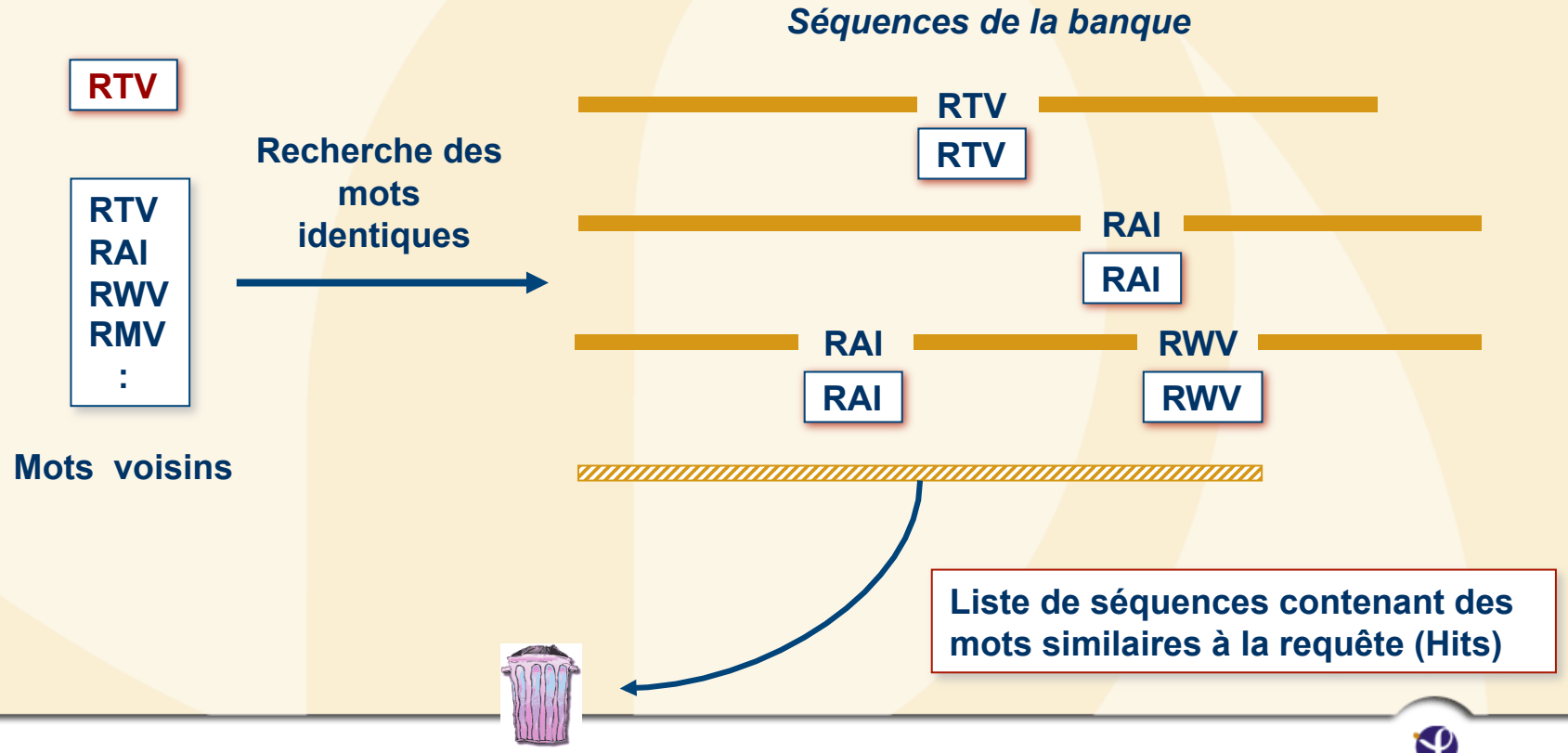
Liste de tous les mots de longueur **w**

Liste des mots voisins qui, alignés avec la requête, ont un score > T

INSTITUT PASTEUR

BLAST1: sans gap

3- Pour chaque liste de mots voisins (score > T), identification de tous les mots identiques dans la banque.



INSTITUT PASTEUR

BLAST1: sans gap

4- Pour chaque «hit», extension sans gap de l'alignement dans les deux sens.

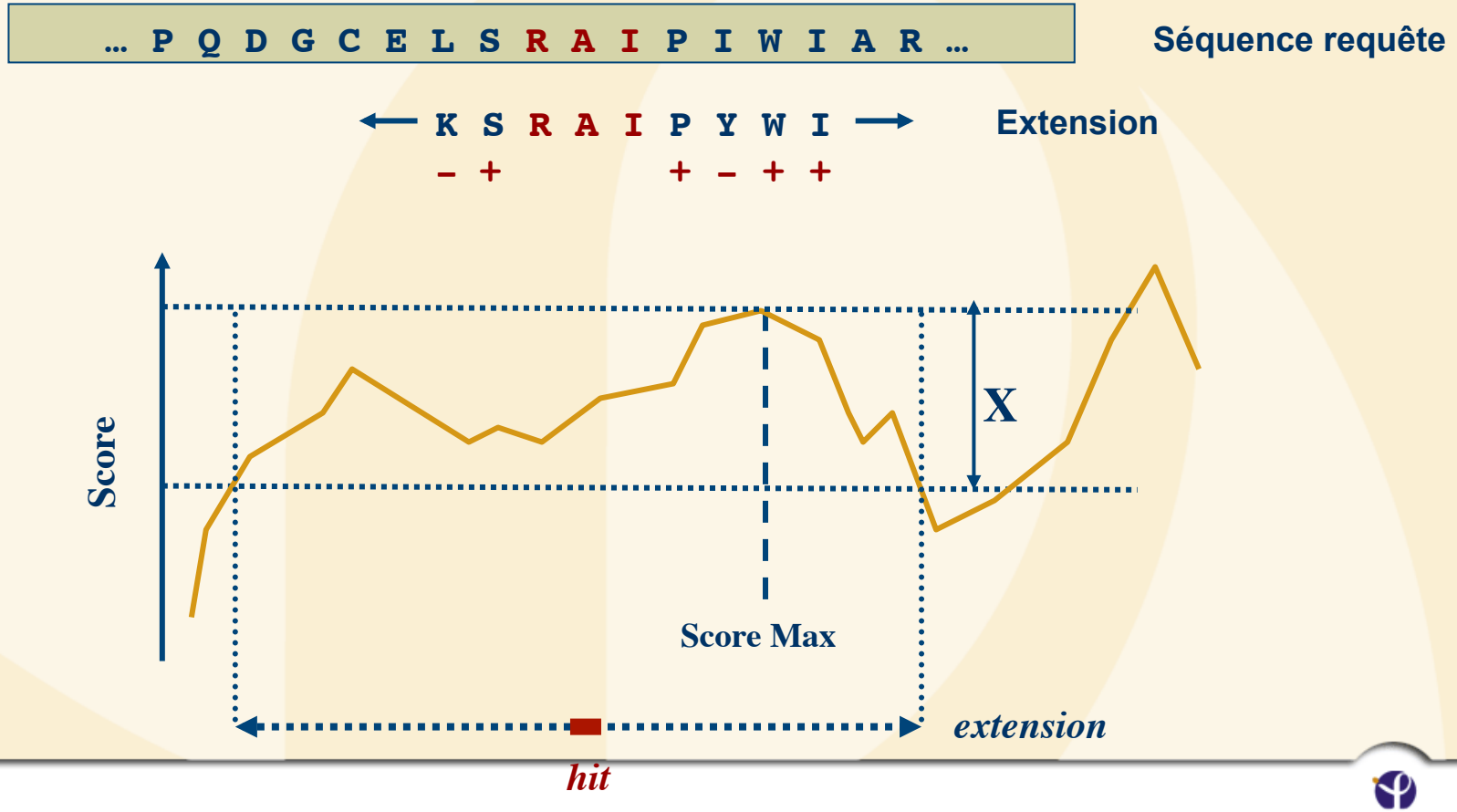
… P Q D G C E L S R A I P I W I A R … *Séquence requête*

← K S R A I P Y → *HIT*

… P Q D G C E K S R A I P Y W I A R … *Séquence de la banque*

BLAST1: sans gap

5- On stoppe l'extension sans gap quand le score S a diminué de plus de X par rapport au score maximum atteint jusque là.

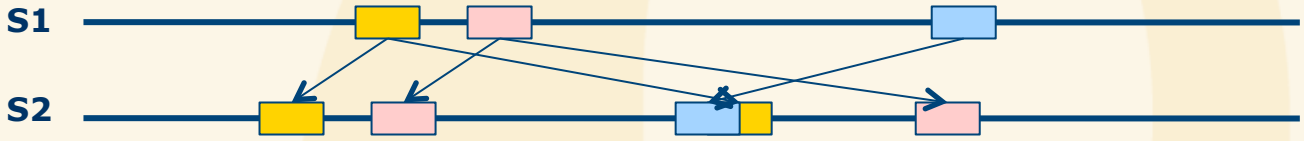


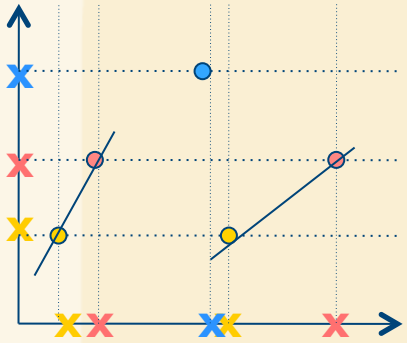
INSTITUT PASTEUR

BLAST2: AVEC gap

- La procédure d'extension des hits est responsable de **90%** du temps d'exécution de BLAST.

- **Hypothèse**: Un segment de score optimal est probablement beaucoup plus long qu'un simple mot de trois lettres. Il doit exister plusieurs mots de trois lettres très proche. Et en particulier, des mots qui peuvent être joint sans insertion de gap (ie: sur la même diagonale).



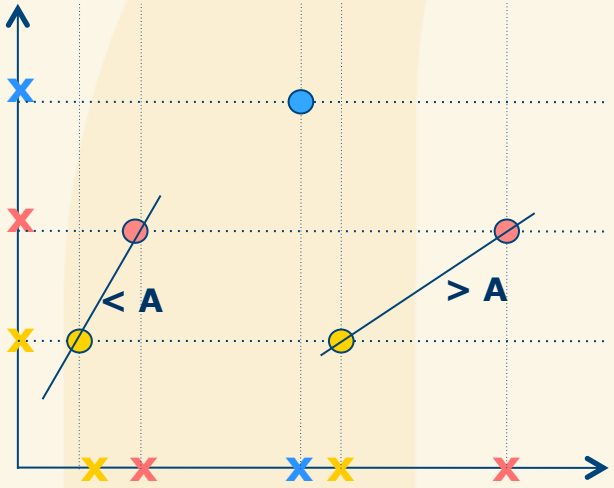


INSTITUT PASTEUR

BLAST2: AVEC gap

- **L'élongation des hits s'effectue seulement lorsque deux hits sont joignables sans insertion de gap et distant d'au plus de A résidus.**

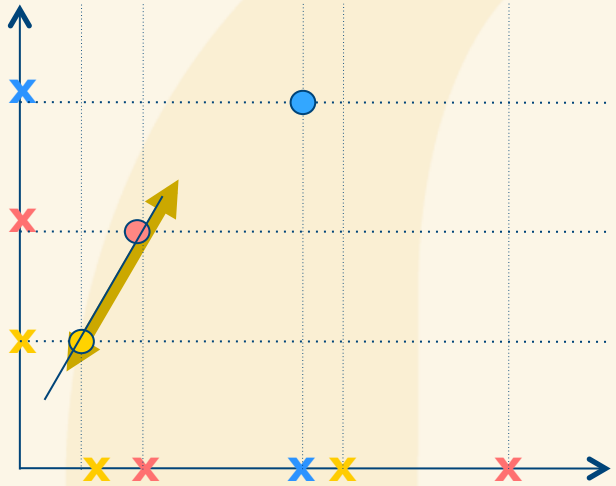
- **Le nombre de séquences ayant deux hits sur la même diagonale diminue considérablement.**



INSTITUT PASTEUR

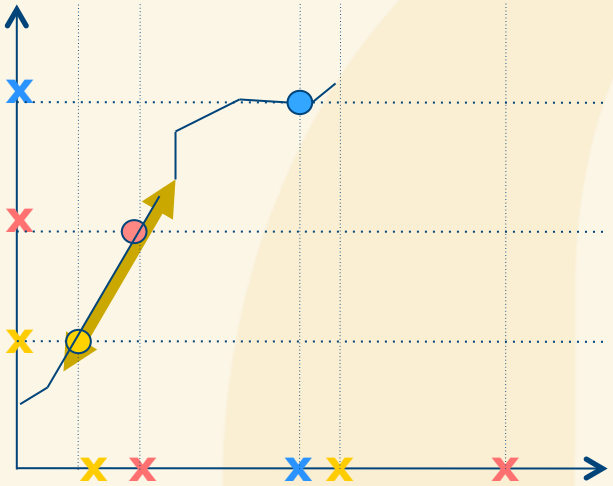
BLAST2: avec gap

- **Les « hits » sélectionnés subissent une extension *sans gap*.**



BLAST2: avec gap

- Les **extensions** générées servent de point d'encrage à une recherche d'alignement local optimal par programmation dynamique SW avec gap limitée à une région.



- La matrice est explorée dans les 2 directions à partir du milieu du segment de 11 résidus ayant le meilleur score dans l'extension.

- La recherche du chemin optimal est limitée aux cellules de la matrice tel que le score de l'alignement ne doit pas diminuer de plus de **Xg** par rapport au score maximum atteint jusque là.

BLAST2: avec gap

- **Toutes les extensions conservées sont des HSP: « Hight scoring Segment Pair ».**

- **Le HSP obtenant le meilleur score S est appelé MSP: « Maximal Segment Pair ».**

- **Les HSP conservés sont présentés par ordre de E-value croissantes.**

E-value = nombre de HSP obtenus par hasard ayant un score supérieur ou égal à S

- **WU-Blast: S'il existe plusieurs HSP entre la séquence requête et une séquence de la banque, les scores de ces HSP sont additionnés pour déterminer la E-value.**

Blast out: liste des hits

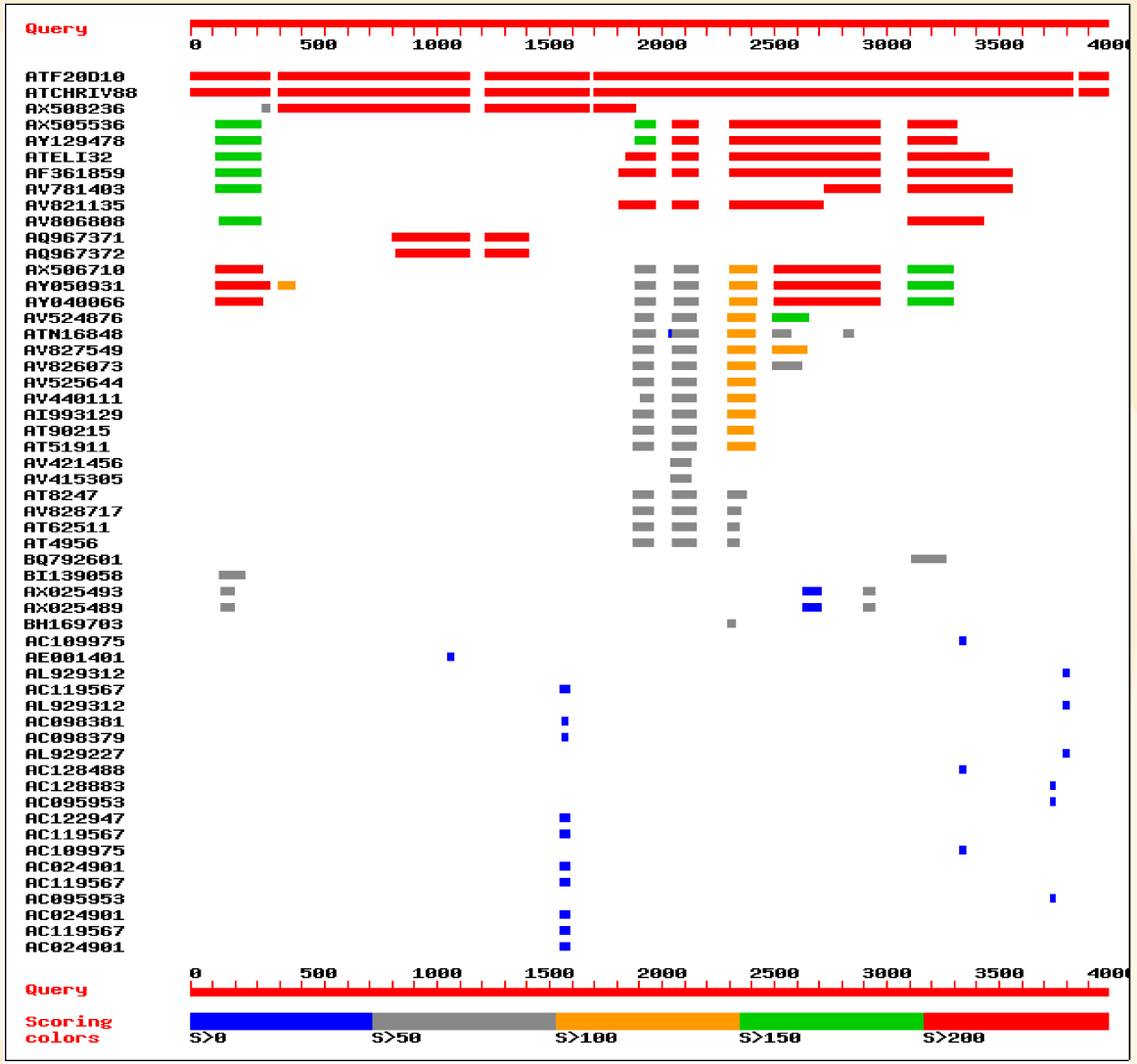
BLASTN 2.2.10 [Oct-19-2004]

Query= , 4080 bp.
(4080 letters)
Database: Embl version 85
73,606,078 sequences; 129,213,367,758 total letters

Sequences producing significant alignments:	Score (bits)	E Value
emb|AL035538|ATF20D10 Arabidopsis thaliana DNA chromosome 4, BAC...	4084	0.0
emb|AL161592|ATCHRIV88 Arabidopsis thaliana DNA chromosome 4, co...	4084	0.0
emb|AX508236|AX508236 Sequence 2931 from Patent WO0216655.	1643	0.0
emb|AX505536|AX505536 Sequence 231 from Patent WO0216655.	1326	0.0
emb|AY129478|AY129478 Arabidopsis thaliana AT4g37990/F20D10_110 ...	1326	0.0
emb|X67815|ATELI32 A.thaliana mRNA for Eli3-2	1326	0.0
emb|AV806808|AV806808 Arabidopsis thaliana cDNA clone:RAFL09-48-...	638	e-179
emb|AQ967371|AQ967371 LERIR45TF LERG Arabidopsis thaliana genomi...	611	e-170
emb|AQ967372|AQ967372 LERIR45TR LERG Arabidopsis thaliana genomi...	561	e-155
emb|AY050931|AY050931 Arabidopsis thaliana putative cinnamyl-alc...	478	e-130
:		
emb|AY028929|AY028929 Lotus corniculatus cinnamyl alcohol dehydr...	48	0.45
emb|AL512662|AL512662 Human DNA sequence from clone RP11-479O17 ...	48	0.45
emb|AC092598|AC092598 Homo sapiens BAC clone RP11-114B11 from 2,...	48	0.45
emb|AL929356|PFA929356 Plasmodium falciparum strain 3D7, chromos...	46	1.8
emb|AL929356|PFA929356 Plasmodium falciparum strain 3D7, chromos...	46	1.8
emb|AP005760|AP005760 Oryza sativa (japonica cultivar-group) gen...	46	1.8
emb|AC133897|AC133897 Mus musculus clone RP24-357G2, LOW-PASS SE...	46	1.8
emb|AC133897|AC133897 Mus musculus clone RP24-357G2, LOW-PASS SE...	46	1.8

banque ***AC*** ***description*** ***score*** ***E-value***

BLAST out: schéma



INSTITUT PASTEUR

BLAST out: alignements

```
>emb|AX505536|AX505536 Sequence 231 from Patent WO0216655.
    Length = 1080

 Score = 1326 bits (669), Expect = 0.0
 Identities = 669/669 (100%)
 Strand = Plus / Plus

Query: 2402 ggcatgagatcgtgggcgtggtgactgaagtcggagccaaagtgactaaattcaaaaccg 2461
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 200  ggcatgagatcgtgggcgtggtgactgaagtcggagccaaagtgactaaattcaaaaccg 259

   :
Query: 3062 tcatctttg 3070
            |||||||||
Sbjct: 860  tcatctttg 868


 Score =  422 bits (213), Expect = e-114
 Identities = 213/213 (100%)
 Strand = Plus / Plus


Query: 3194 gagaggaagatggtaatgggaagtatgataggagggataaaagagacccaggaaatgata 3253
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 868  gagaggaagatggtaatgggaagtatgataggagggataaaagagacccaggaaatgata 927

:

Query: 3374 gccaacacattgaagcctaatcctaatttataa 3406
            ||||||||||||||||||||||||||||||||
Sbjct: 1048 gccaacacattgaagcctaatcctaatttataa 1080
```

1er HSP

« alignement » local

2me HSP

Donner un sens aux scores: E-value

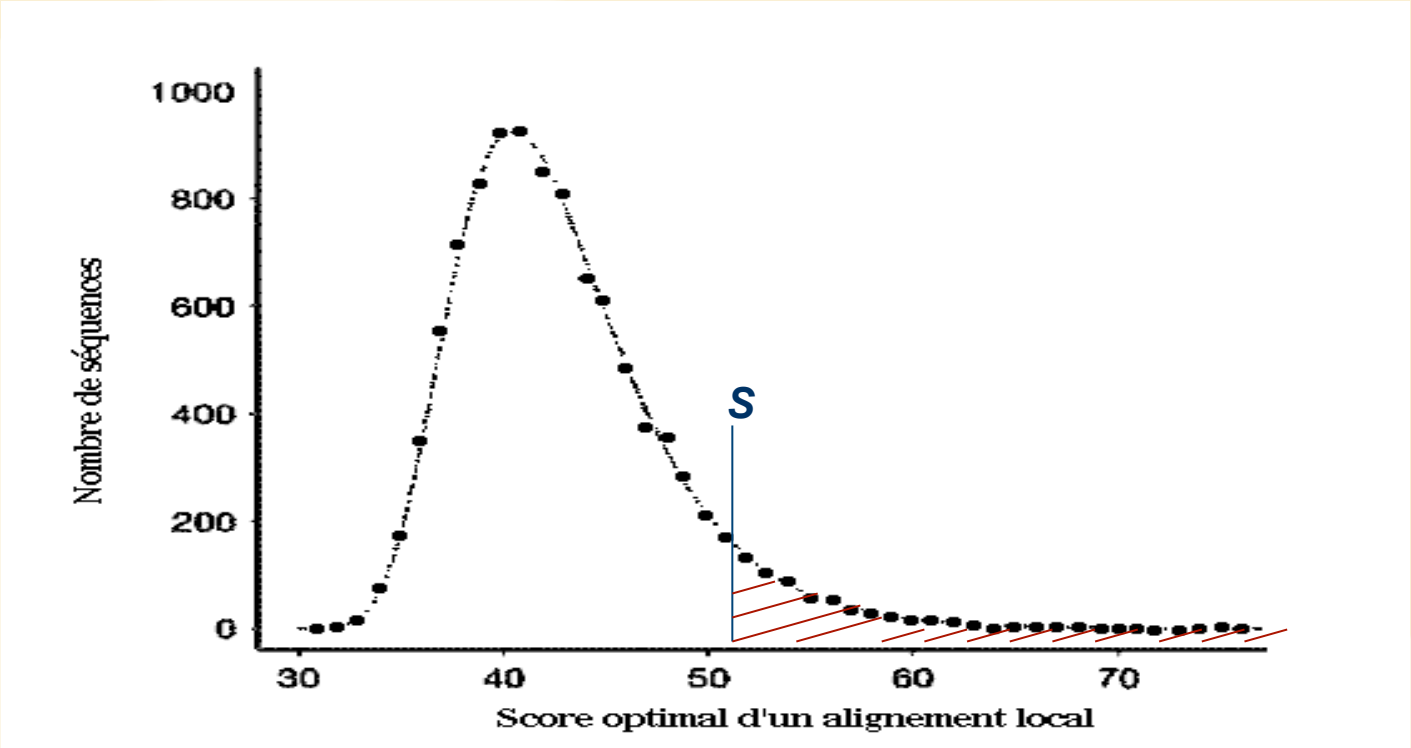
• **Question:** Quelle est la **probabilité** qu'un alignement, ayant un score **supérieur** ou égal à *S*, soit obtenu par **hasard** en cherchant les meilleurs alignements dans une banque de données.

➥ Il faut connaître la **distribution statistique** (distribution des valeurs extrêmes) des scores des alignements sous un model aléatoire.

• Un **model aléatoire** est défini par:

- ✓ des séquences aléatoires composée de résidus indépendants avec des probabilité p_i dans la banque
- ✓ le score *S* du MSP (sans gap) de 2 séquences de longueur n et m (n et m doivent être grands)
- ✓ un système de scores S_{ij}

Distribution des valeurs extrêmes



Surface sous la courbe = probabilité d'obtenir un score > *S*
= p-value

Transformer en E-Value

INSTITUT PASTEUR

Théorie statistique de Karlin-Altschul

- **A l 'aide du modèle aléatoire, on calcule une « p-value » et une « E-value » pour un MSP et un score *S* donnés**

- **p-value = probabilité d'obtenir pour un MSP donné un score supérieur ou égal à S par hasard .**

$$\text{p-value} = 1\text{-e}(K\ mn\ e^{-\lambda S})$$

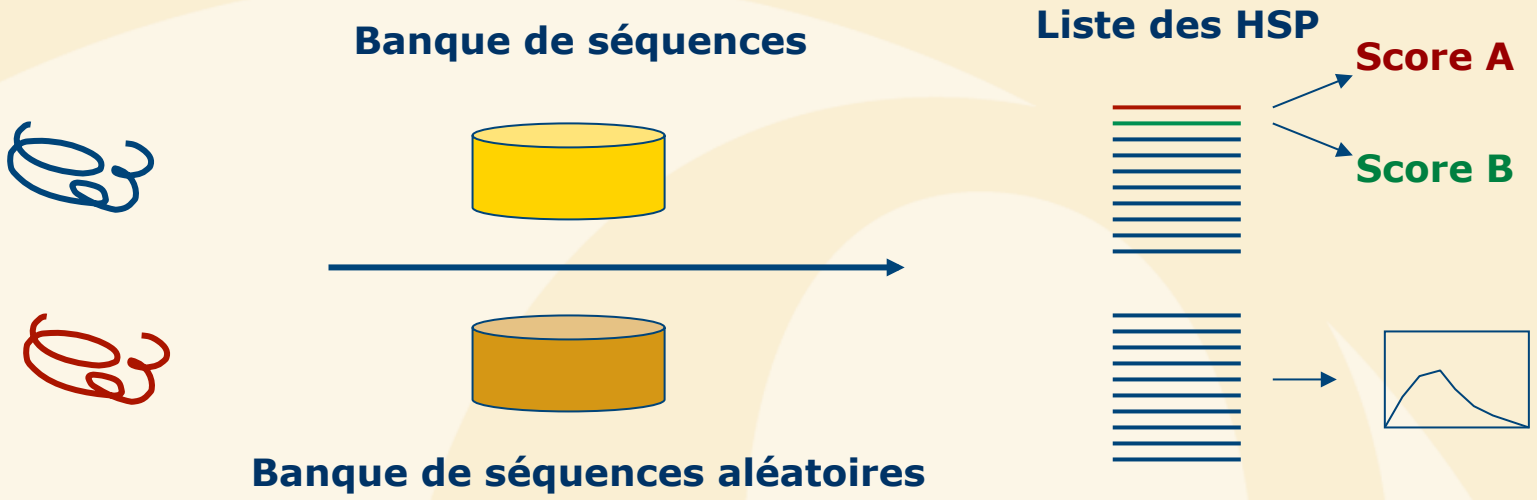
- **E-value = nombre d'alignement attendu par hasard ayant un score supérieur ou égal au score *S*.**

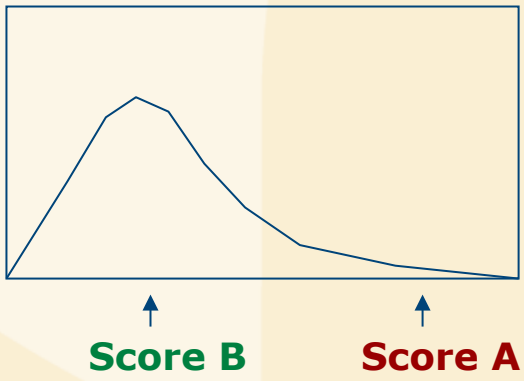
$$\text{E-value} = K\ mn\ e^{-\lambda S}$$

m / n: longueur de la séquence requête et de la séquence de la banque
K / λ : constante dépendant du système de scores ($S_{i,j}$) et de la composition de la banque (p_i)

- **Score convertit en « bits »:** $S_{bits} = \frac{\lambda S - \ln K}{\ln 2}$

Distribution des valeurs extrêmes





Score A: significatif

Score B: non significatif

INSTITUT PASTEUR

E-value

• **La E-value ne représente pas une mesure de la similitude entre 2 séquences:**

Score=247 E-value = 5

• **Par hasard, 5 séquences de la banque ont un score ≥ 247.**

Score=247 E-value = 10^{-3}

• **Par hasard, 10^{-3} séquences de la banque ont un score ≥ 247 <=> il faut qu'il y ait 1000x plus de séquences dans la banque pour trouver une seule séquence par hasard qui ait un score ≥ 247**

Score=247 E-value = 1 db = 1 million

• **Par hasard, 1 séquence de la banque a un score ≥ 247.**
☛ **peut être significatif**

Filtres avant une recherche FASTA ou BLAST

- **Certaines régions peuvent être filtrées (non prise en compte)**

✓ **Régions de faible complexité: composition limité en acides aminés.**

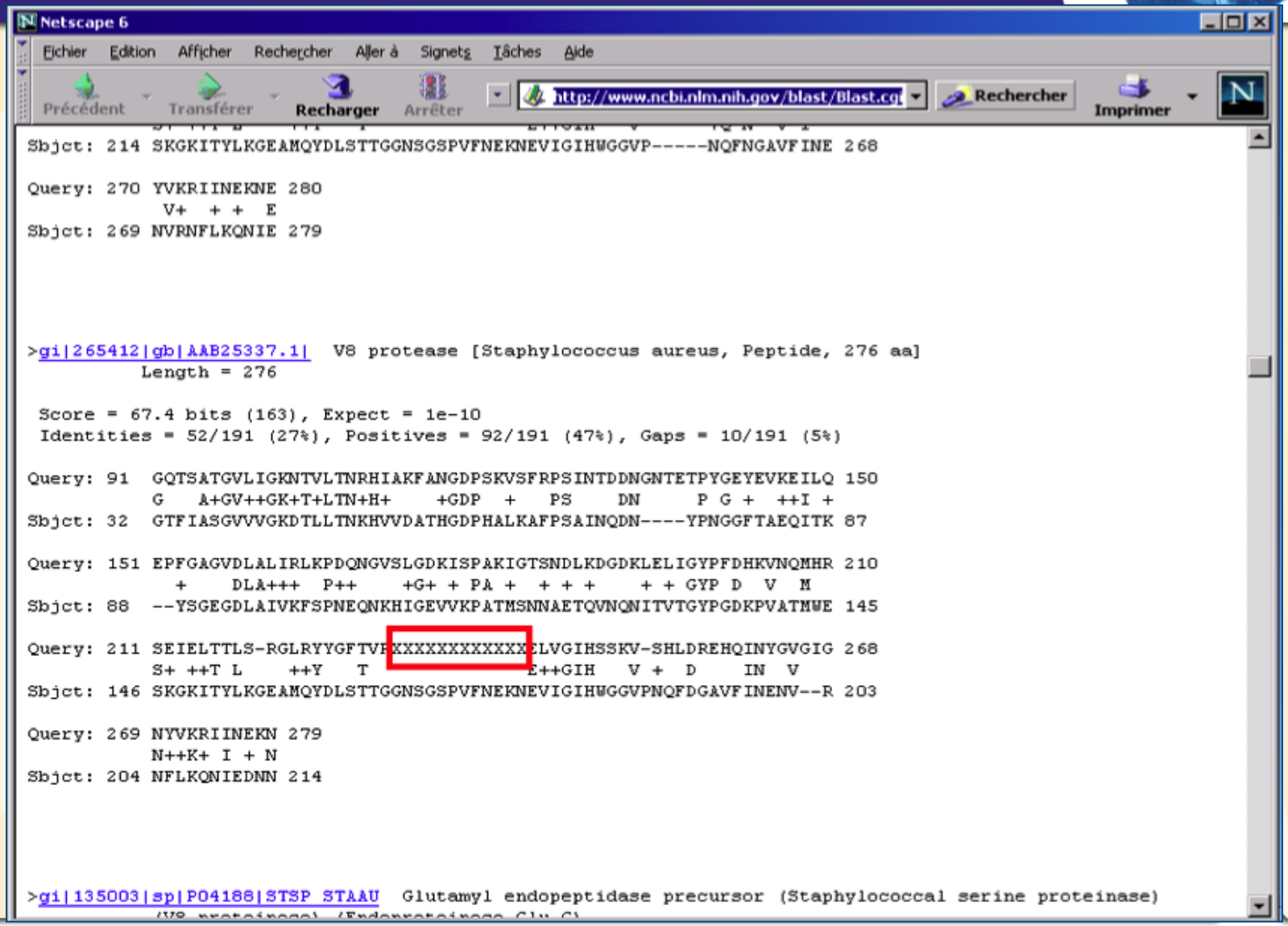
✓ **Souvent une mosaïque d'un nombre limité d'acides aminés, sources de résultats erronés dans la recherche mais parfois fonctionnellement importantes pour certaines protéines.**

✓ **Régions contenant des motifs répétés.**

Human MGT8a protein

Low-complexity segments		High-complexity segments
	1-24	MPDRTEKHSTMPDSPVDVKTQSRL
tpptmpppptt	25-35	
	36-60	QGAPRTSSFTPTTLTNGTSHSPTAL
ngapsppngfsngpsssssslanqqlpp	61-89	
	90-258	ACGARQLSKLKRFLTTLQQFGNDISPEIGE RVRTLVLGLVNSTLTIEEFHSKLQEATNFP LRPFVIPFLKANLPLLQRELLHCARLAKQN PAQYLAQHEQLLLDASTTSPVDSSELLLDV NENGKRRTPDRTKENGFDREPLHSEHPSKR PCTISPGQRYSPNNGLSYQ
pnglphptpppp	259-270	
	271-377	QHYRLDDMAIAHHYRDSYRHPSHRDLRDRN RPMGLHGTRQEEMIDHRLTDREWAEEWKHL DHLLNCIMDMVEKTRRSLTVLRRCQEADRE ELNYWIRRYSDAEDLKK
gggsssshs	378-386	
	387-554	RQQSPVNPDPVALDAHREFLHRPASGYVPE EIWKKAEEAVNEVKRQAMTELQKAVSEAER KAHDMITTERAKMERTVAEAKRQAAEDALA VINQQEDSSESCWNCGRKASETCSGCNTAR YCGSFCQHKDWEKHHHICGQTLQAQQQGDT PAVSSSVTPNSGAGSPMD
tppaatprsttpgtpstiettp	555-576	
	577-577	R

Filtres



INSTITUT PASTEUR

Fin

INSTITUT PASTEUR