

Les alignements de séquences

Pourquoi: Comparaison de séquences?

Le « dogme central » de la bioinformatique: La déduction par homologie

L'évolution des gènes laisse une trace parfaitement visible lorsque l'on compare leur séquence:

-Les régions **fonctionnelles** des gènes (sites catalytique, de fixation, etc.) sont soumises à sélection. Elles sont relativement préservées par l'évolution car des mutations trop radicales sont désavantageuses.

-Les régions **non fonctionnelles** ne subissent aucune sélection et divergent rapidement à mesure que s'accumulent les mutations.

-Les nouveaux gènes apparaissent surtout par remaniement de gènes ancestraux:

→→→ La fonction des gènes peut être déduit par comparaison avec les gènes « homologues » d'autres espèces

(Evolution des gènes= mutations, insertions, délétions, recombinaisons)

Comparaison de séquences

La comparaison de séquences comme méthode de prédiction
Activité principale en bioinformatique

Alignement: processus de comparaison de séquences permettant d'obtenir le maximum de correspondances entre les lettres qui les composent. Il est quantifié par un score de similarité

Similarité: mesure du degré de ressemblance entre séquences, quantifié par un score, calculé à l'aide d'une matrice de score.

Homologie: parenté évolutive. Inférence déduite à partir du degré de similitude. Mais deux séquences similaires ne sont pas forcément dérivées d'un ancêtre commun.

Pourquoi: Comparaison de séquences?

Une ressemblance entre séquences peut indiquer:

- ✓ **une fonction biologique proche**
- ✓ **une structure tridimensionnelle semblable**
- ✓ **une origine et/ou une histoire évolutive commune**

Terminologie

Identité

Proportion des paires de résidus **identiques** entre deux séquences alignées. (Exprimé généralement en %).

Similitude

Mesure de la ressemblance entre deux séquences. Le degré de similitude est quantifié par un score basé sur le % de similarité (% **identité** + % **substitutions conservatives**) des séquences.

Gaps / Indels

Proportion d'**Indels** entre deux séquences alignées. (Exprimé en %).

Homologie

Deux séquences sont homologues si elles ont un **ancêtre commun**.
Il n'y a pas de degré d'homologie (oui ou non !).
On ne dit pas: tres homologue, faible homologie, etc....

- De 100% à quelques nucléotides/aminoacides en commun
- Il n'y a pas vraiment de limite, mais en dessous de 20-25% (*twilight zone*), il devient très difficile de distinguer une homologie d'une ressemblance fortuite
- Des séquence sans ressemblance apparente peuvent parfaitement être homologues (on le retrouve par ex. au niveau 3D). **Des séquences homologues ne sont pas nécessairement similaires**

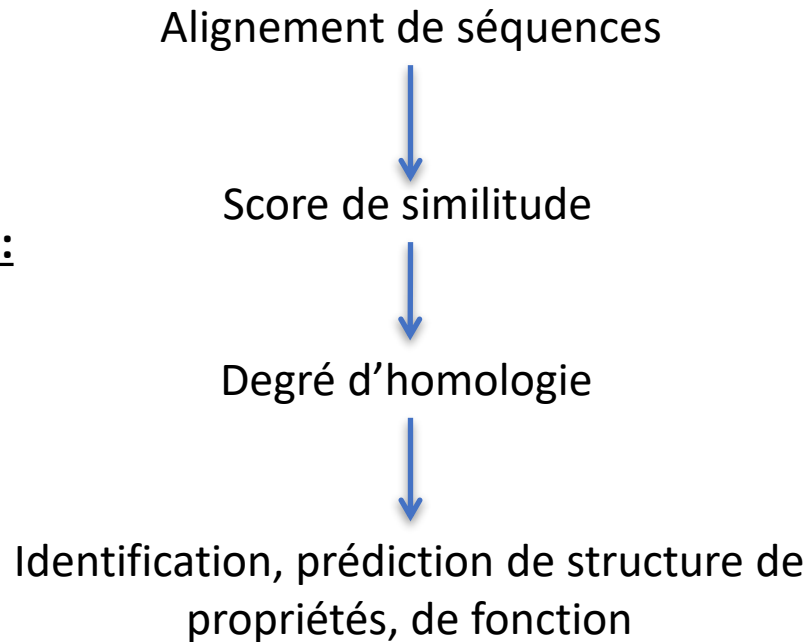
Principe de la comparaison de séquences

La comparaison de séquences est **l'outil central** en bioinformatique :

→Repose sur des calculs matriciels ou des algorithmes complexes qui rendent des résultats sous forme de données statistiques (% match, **score**, e-value...)

→Logiciel d'alignement le plus connu = **BLAST** (Basic Local Alignment Search Tool)

Démarche globale :



Comment effectuer une recherche?

Utiliser des algorithmes « heuristiques »

➤ **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 deux programmes heuristiques les plus utilisés par les biologistes sont:

- **FASTA** – FAST Alignment (Pearson et Lipman, 1988)
- **BLAST** - Basic Local Alignment Search Tool (Altschul et al., 1990, 1997)

- Format commun de manipulation des données:
le format FASTA (Fast – alignment)

Objectif: manipuler facilement des séquences dans les bases de données, à l'aide d'un **format universel**, compatibles avec les traitements de texte (sous forme de fichier texte), ou par copier – coller.

Format FASTA

Séquence
simplifiée
(pas de chiffre,
pas d'espace)

```
>gi|31341071|ref|NM_181006.2| Bos taurus chromogranin B (secretogranin 1)
(CHGB), mRNA
GGCACGTGAGGATACAAGGTTGTTTTCCACAGCATCTTCATCTCGCCTTCCGTCCCCTTCTCATTACC
CTTCCGACTGCTATCCTCTTCTCCGCGCAGATTTGGACGAGCGAGGCCATGCAGCCGGCCGCCCTTCTCG
GCCTTCTGGGAGCCACGGTGGTTGCAGCCGTGAGCTCTATGCCAGTGGACATCAGGAACCACAATGAAGA
AGTGGTGACTCACTGCATCATCGAGGTCCTCTCAAATGCCCTATTGAAGTCCAGCGCTCCACCCATCACC
CCTGAGTGCCGACAAGTCCTTAAGAAGAATGGAAAAGAGCTCAAAGATGAAGAGAAAAGTAAAAATGAAA
ACACAAGGTTTGAAGTGAAGTGTGAGAGACCCAGCTGACACCTCAGAAGCCCCCGGGCTCTCCAGTAG
GGAGGACTCAGGGGAGGGGGATGCCCAAGTCCCAACAGTAGCAGACACGGAGAGCGGTGGGCATAGCCGA
GAGCGGGCAGGTGAGCCCCGGGAAGTCAAGTGGCCAAAGAAGCAAAGACACGCTATTCTAAGAGCGAGG
GACAGAACAGGGAGGAAGAAATGGTAAAATACCAGAAAAGGGAACGTGGGGAAGTTGGCAGTGAAGAGAG
ACTGTCTGAAGGGCCGGGAAAGGCACAAATGGCTTTTCTCAACCAAAGAAACCAGACTCCGGCTAAGAAA
GAGGAGTTAGTGTCCAGATATGATACACAGTCTGCCAGGGCCTTGAGAAGTGCACAGCCGGGAAAGGA
GCAGCCAGGAGAGTGGAGAGGAGACCAAGAGCCAGGAGAACTGGCCCCAAGAGCTGCAACGCCATCCGGA
GGGCCAGGAAGCACCCGGAGAAAAGTGAAGAGGATGCCAGCCCCGAGGTGGACAAAACGGCGCTCGAGGCCA
AGACACCACCAGGGAGGAGCAGGCCCGACAGGTCCTCCCAGGAGGGGAATCCTCCCCTCGAGGAGGAGT
CACACGTGGGCACGGGCAACTCAGACGAAGAGAAAAGCCCCGCCATCCAGCCCACCTTTAGGGCTTTGGAGGA
GGGAGCCGAATATGGGGAGGAAGTGAAGAGACTCAGCTGCCAGGCTCCTGGAGACTTGCAGGGGGCA
CGATTCCGGGGCAGAGGACGTGGAGAGCACCAGGCTCTAAGGCGTCCCAGCGAGGAGAGCCTAGAGCAGG
AAAACAAGAGACATGGCCTCAGCCCGGATCTAAACATGGCGCAGGGATACAGCGAGGAAAGCGAGGAAGA
GAGGGTCCGGCCCCGGGACCCAGCTACAGAGCCCCGGGGAGGGGAGGGCGGCCTACTCCACACTAGGC
CAAACAGATGAGAAACGGTCTTGGGTGAAACGCACCACCGTGTTCAGGAAAGCCAGAGGGACAAGGCGA
GGCGCCGCCTACCAGGCGAGCTGAGAAATTACCTCGACTATGGTGAAGAAAAGGGTGAAGAACAGCCAG
AGGGAAGTGGCAGCCGCAGGGAGACCCGCGAGACGCTGACGAGAACAGGGAAGAGGCTAGGCTTCGAGGC
AAACAGTATGCTCCCCATCACATCACTGAAAAGAGATTAGGGGAGCTACTCAATCCATTCTACGACCCTT
CCCAGTGAAGAGCAGCCGTTTTGAGAGAAAAGACCCCATGGATGACAGTTTTCTTGAGGGTGAAGAGGA
AAACGGGCTGACCTTGAATGAGAAAAATTTCTTCCAGAATACAACATGACTGGTGGGAGAAAAAGCCC
TTTGAAGAGGATGTAAACTGGGGTATGAGAAGAGAAACCCGGTCCCCAAACTGGATCTAAAAAGGCAGT
ATGACCGAGTGGCCGAACGGACCAGCTCCTTCACTACAGGAAGAAGTCAAGTGGTTCAGACTTCTA
TACTCCGAGGAGCAGGTGAGCCACAACACACAGCAGAAAATGAAGAGGAGAAGGCTGGCCAAGGAGTT
CTGACGGAGGAAGAGGAAAAAGAACTTAAAACTTGGCTGCGATGGATTTGGAACACAGAAAATAGCTG
AGAAGTTCAGTGGTACCCGAAGGGGCTAATGGTCATTAGAGTAAAGGGCAGATTTTAAAGACAGCCTTAC
ATGATCTGTTATCCACCCTTCACTGAAAGACCCATTTATTTATCCAAAGGCAGAAAAGTAGAATTTACT
CATCCAATGTTTGACACAATTGAAATGTCTTTGATTTTTGCCAGAGTGTATTGGAATATAAATAGCA
TGACTTGTAGTATACTCTTTATAAAAAAGTAGATATATTAACATGCTTGTGACAATGACTGTGCTACTGT
CCTTGAAAAAATGTCTTAGTTTTGAAGTAATAAAAGATTACCTGAGGCCAAAAGCGTCATGTTGCGAGCT
TCCCTTGGTCTAATAGTCTGACTTTCAGATCCATTCTTCAAATAAATTCTAAAATACAGC
```

➤ Format commun de manipulation des données:
le format FASTA (Fast – alignment)

Remarques:

-Les bases nucléotidiques ne référencient que des **monobrans d'ADN**
(même si la séquence soumise est de l'ADN bicaténaire ou de l'ARN)

→ la séquence est toujours dans le sens 5'P – 3'OH

-Les séquences nucléotidiques quel que soit leur nature (ADN, ARN...) seront écrites le plus souvent avec **A, T, C et G**

Outils d'Alignement de séquence: BLAST


NCBI: Blast

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information


All Databases Search


COVID-19 Information ✕
[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)


 **UNITE**
A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.
LEARN MORE


NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases


Download
Transfer NCBI data to your computer


Learn
Find help documents, attend a class or watch a tutorial


Popular Resources
[PubMed](#)
[Bookshelf](#)
[PubMed Central](#)
[BLAST](#) 
[Nucleotide](#)
[Genome](#)
[SNP](#)
[Gene](#)
[Protein](#)
[PubChem](#)

NCBI: Blast



U.S. National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

NEWS

A new feature was added to the NCBI IgBLAST webpage

IgBLAST is now able to determine Ig isotypes

Mon, 01 Nov 2021 12:00:00 EST

[More BLAST news...](#)


Web BLAST



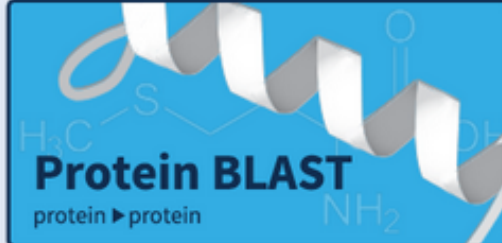
Nucleotide BLAST
nucleotide ▶ nucleotide



blastx
translated nucleotide ▶ protein



tblastn
protein ▶ translated nucleotide



Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

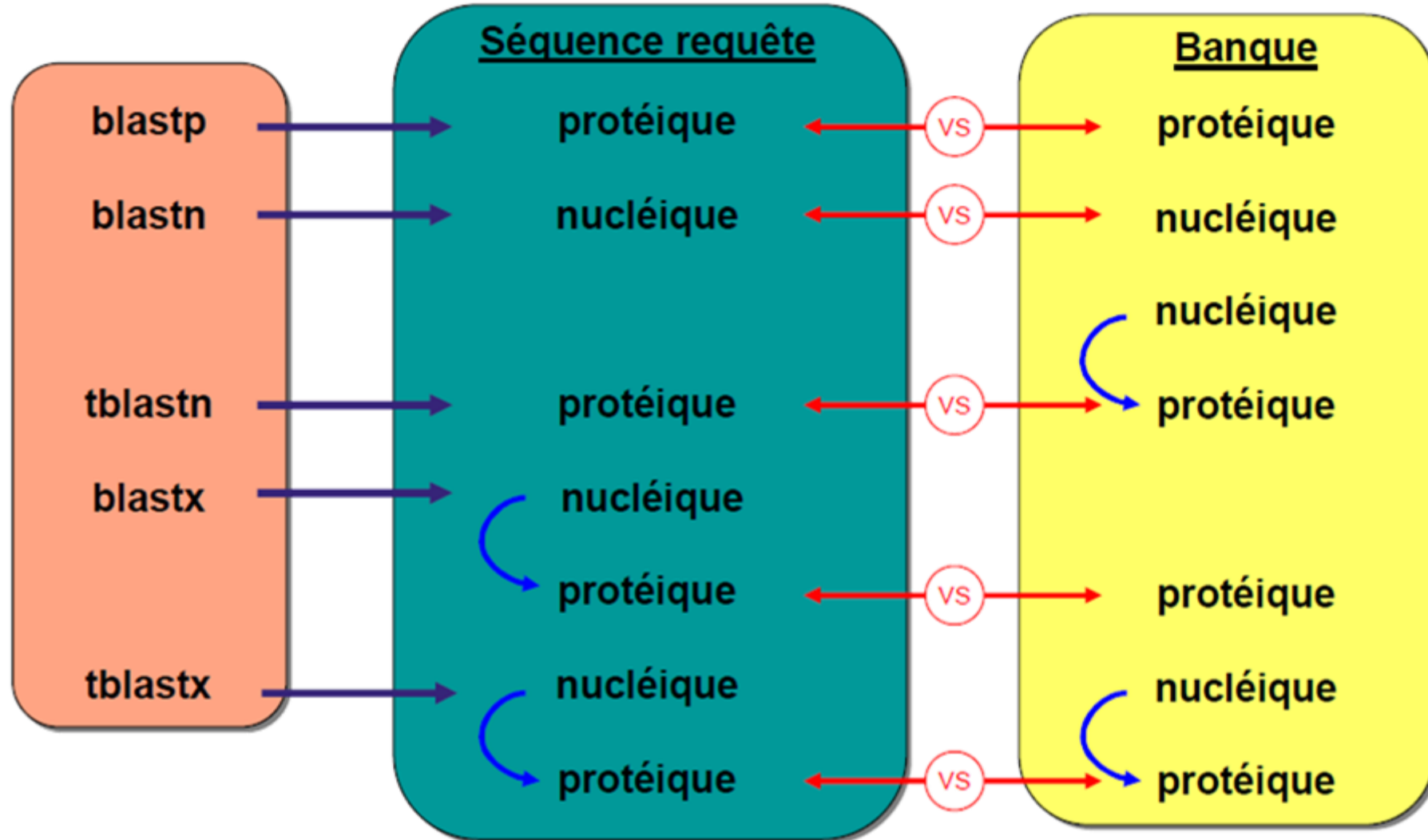
Human

Mouse

Rat

Microbes

BLAST (Basic Local Alignment Search Tool)



- blastn**
- blastp
- blastx
- tblastn
- tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

[Clear](#)

Query subrange

```
TCCAGATATGATACACAGTCTGCCAGGGGCTTGAGAAGTCGCACAGCCG
GAAAGGAAGCCAGGAGAGTGGAGAGGAGACCAAGAGCCAGGAGAACT
GGCCCAAGAGCTGCAACGCCA
```

From

To



Or, upload file

Parcourir...

Aucun fichier sélectionné.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Human genomic + transcript Mouse genomic + transcript Others (nr etc.):

Nucleotide collection (nr/nt)

Organism

Optional

Enter organism name or id--completions will be suggested Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude

Optional

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query

Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window



[Algorithm parameters](#)



[← Edit Search](#)

[Save Search](#)

[Search Summary ▾](#)

[? How to read this report?](#)

[▶ BLAST Help Videos](#)

[↶ Back to Traditional Results Page](#)

Job Title	Nucleotide Sequence
RID	ZXSP0T3001R Search expires on 02-07 03:49 am Download All ▾
Program	BLASTN ? Citation ▾
Database	nt See details ▾
Query ID	lc Query_554479
Description	None
Molecule type	dna
Query Length	309
Other reports	Distance tree of results MSA viewer ?

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity to

E value to

Query Coverage to

[Filter](#) [Reset](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download ▾

[New Select columns ▾](#)

Show

100 ▾

select all 74 sequences selected

[GenBank](#)

[Graphics](#)

[Distance tree of results](#)

[New MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	PREDICTED: Bos indicus x Bos taurus chromogranin B (CHGB), mRNA	Bos indicus x ...	571	571	100%	1e-158	100.00%	2515	XM_027559408.1
<input checked="" type="checkbox"/>	PREDICTED: Bos indicus chromogranin B (CHGB), mRNA	Bos indicus	571	571	100%	1e-158	100.00%	2724	XM_019973370.1
<input checked="" type="checkbox"/>	Bos taurus genome assembly, chromosome: 13	Bos taurus	571	571	100%	1e-158	100.00%	83216874	LR962888.1
<input checked="" type="checkbox"/>	Bos taurus chromogranin B (CHGB), mRNA	Bos taurus	571	571	100%	1e-158	100.00%	2441	NM_181008.2
<input checked="" type="checkbox"/>	Bos mutus isolate yakQH1 chromosome 13	Bos mutus	566	566	100%	6e-157	99.68%	75983851	CP027081.1
<input checked="" type="checkbox"/>	PREDICTED: Bos mutus chromogranin B (CHGB), mRNA	Bos mutus	566	566	100%	6e-157	99.68%	2510	XM_005891057.2
<input checked="" type="checkbox"/>	Bos taurus genome assembly, chromosome: 13	Bos taurus	566	566	100%	6e-157	99.68%	84328414	LR962743.1
<input checked="" type="checkbox"/>	PREDICTED: Bison bison bison chromogranin B (secretogranin 1)(CHGB), mRNA	Bison bison bi...	560	560	100%	3e-155	99.35%	2476	XM_010836447.1
<input checked="" type="checkbox"/>	B.taurus mRNA for chromogranin B	Bos taurus	549	549	100%	6e-152	98.71%	2361	X55489.1



[Feedback](#)

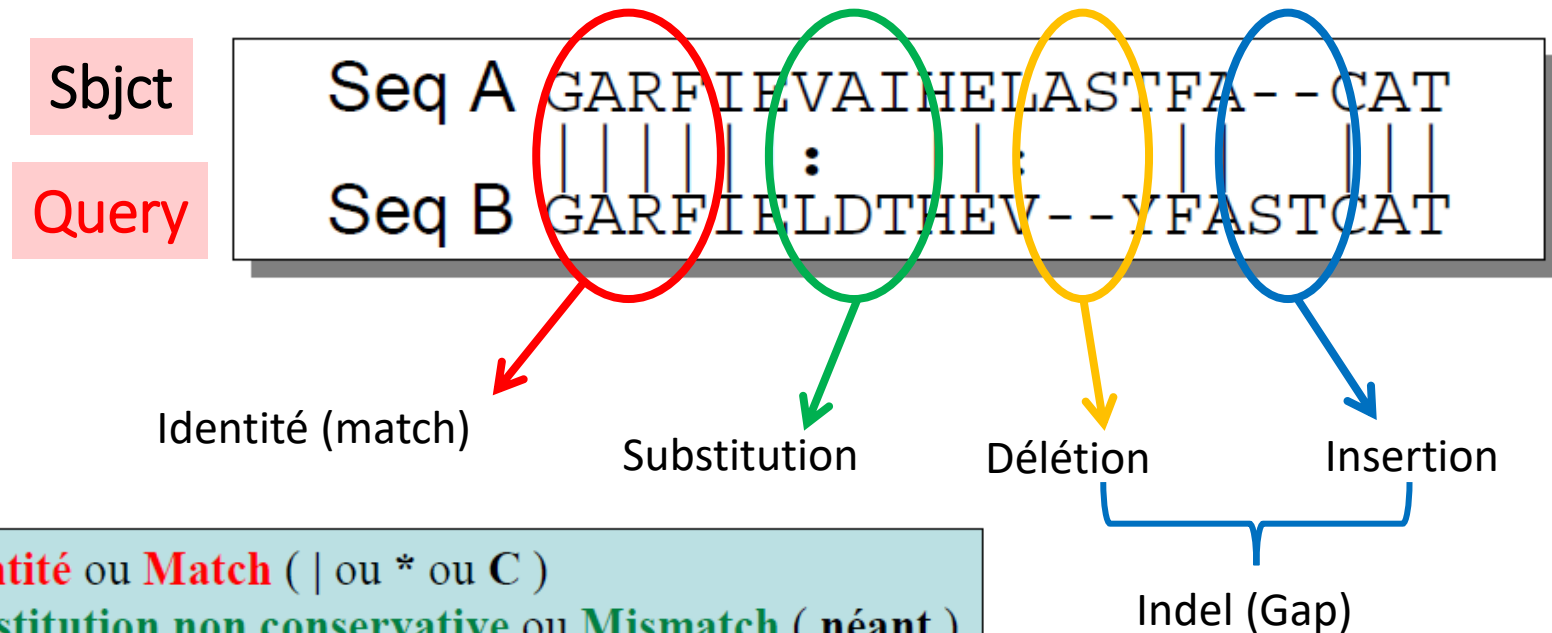
Principe de la comparaison de séquences

Principe du calcul des scores d'alignement :

- En pratique, plus le score d'alignement est élevé, plus les séquences sont similaires et présenteront des propriétés et des fonctions proches.
 - plus de 70% de similarité permettent d'affirmer qu'il y a homologie

Qu'est-ce qu'un alignement?

- 3 situations sont possibles pour une position donnée de l'alignement:
 - les caractères sont les mêmes: **Identité ou match**
 - les caractères ne sont pas les mêmes: **Substitution**
 - l'une des positions est un espace: **Insertion / Délétion**



Identité ou **Match** (| ou * ou C)
Substitution non conservative ou **Mismatch** (néant)
Substitution conservative (+ ou : ou .)
Indel ou **Gap** (- ou .)

Principe de la comparaison de séquences

Principe du calcul des scores d'alignement :

Exemple : Séquence de référence : AAA TTT GGG CCC

Séquence 1 à analyser :

AAA CCC GGG CCC



Alignement :

AAA TTT GGG CCC

AAA CCC GGG CCC



Non identité (mismatch)

Séquence 2 à analyser :

AAA TTT CCC



Alignement :

AAA TTT GGG CCC

AAA TTT - - - CCC



Non correspondance (gap)

Score d'alignement = Somme des scores individuels

(avec identité (match) = +2, non identité = -1 et gap -8)

Score de la séquence 1:

$2+2+2-1-1-1+2+2+2+2+2+2 = 15$

Score de la séquence 2:

$2+2+2-8+2+2+2+2+2+2 = 10$

Principe de la comparaison de séquences

Principe de la E-value:

- The **expect** value E is a parameter that describes the number of hits one can expect to see by chance when searching a database of a particular size. It decreases exponentially as the score (S) of the match increases.
- For example an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance

Quelques règles d'interprétations des résultats de Blast

- **Comparer une séquence à toutes les séquences d'une bases de données:**
 - **Obtention d'une liste classée par score**
- **But : Trouver des séquences similaires avec une signification biologique:**
 - **Les meilleurs scores signifient-ils une parenté fonctionnelle?**
 - **Les mauvais scores signifient-ils une absence de similarité biologique?**
- **Nécessité d'une analyse statistique pour :**
 - **savoir quelle confiance accordée aux résultats.**
 - **estimer la probabilité d'obtenir un score donné par chance** (Ne pas sur-estimer le résultat obtenu par chance, ne pas sous-estimer le résultat au sens biologique)

Quelques règles d'interprétations des résultats de Blast

➤ En générale, lorsque l'alignement est fait sur au moins 70% de la séquence:

- $30 < \text{ID} \leq 50$ ➔ Séquences **faiblement similaires**
- $50 < \text{ID} \leq 70$ ➔ Séquences **similaires**
- $70 < \text{ID} \leq 100$ ➔ Séquences **fortement similaires**

-On peut déjà parler de séquences homologues au delà de 70% de similarité, mais cela reste à confirmer par d'autres hypothèses: présence de motifs communs, etc....



Quelques conseils

- **Méfiez-vous des résultats donnés par les logiciels :**
 - La qualité des résultats est parfois diminuée au profit de la rapidité
 - Certains problèmes admettent un ensemble infini de possibilités
 - Ce n'est pas toujours la solution la meilleure qui est trouvée
 - Beaucoup de logiciels ne font que de la prédiction

- **Méfiez-vous des banques de données :**
 - Les données se sont pas toujours fiables
 - La mise à jour n'est pas toujours récente

La réalité mathématique n'est pas la réalité biologique :

Les ordinateurs ne font pas de biologie, ils calculent ... vite !
