République algérienne démocratique & populaire
Ministère de l'enseignement supérieur & de la recherche scientifique

جامعة غليزان
RELIZANE UNIVERSITY

Université de Relizane
Faculté des Sciences et Technologies
Département des Sciences Biologiques

Polycopié

Présenté par : Dr MELLALI Sarah

Intitulé

Cours de Bio-informatique

Ce polycopié est destiné aux étudiants de :
1ère année Master Biochimie Appliquée

Année universitaire : 2022/2023

Chapitre 1 :

I. Initiation à la bioinformatique :

I.1. **La bioinformation**

La bioinformation est l'information liée aux molécules biologiques : leurs structures, leurs fonctions, leurs liens de "parenté", leurs interactions et leur intégration dans la cellule.

Il y a deux types de bioinformation : la séquence des nucléotides et la séquence des acides aminés. Les séquences constituent l'un des principaux types de bioinformation qu'analyse la bioinformatique.

Divers domaines d'études permettent d'obtenir cette bioinformation : la génomique structurale, la génomique fonctionnelle, la protéomique, la détermination de la structure spatiale des molécules biologiques, la modélisation moléculaire ...

I.**2. La bioinformatique :**

La bioinformatique est une discipline récente et un champ de recherche multidisciplinaire où travaillent en concert biologistes, médecins, informaticiens mathématiciens et des physiciens dans le but de résoudre un problème scientifique posé par la biologie.

C'est une discipline qui permet l'analyse et l'interprétation des informations biologiques contenues soit dans génome (séquences ADN, ARN) soit dans le protéome (l'ensemble des protéines bio synthétisées) , soit dans transcriptome (ARNm transcrits).

On peut également la définir comme étant la discipline de l'analyse" *in silico "* de l'information biologique contenue dans les séquences nucléiques et protéiques.

I.3. Les applications de la bioinformatique :

La bioinformatique a différents objectifs et différentes applications :

A-Collecter et stocker des informations dans des bases de données, accessibles en ligne.

B-Fournir des outils de comparaison de séquences (protéiques ou nucléotidiques) afin de:

-identifier une séquence par rapport à une base de données

-déterminer le degré de similitudes entre deux séquences (intérêt en taxonomie)

-repérer des motifs structuraux :

-gènes, promoteurs, etc. pour un nucléotide.

-zone de repliement, site actif, etc. pour un polypeptide.

C-Fournir des outils de traduction de séquences afin de :

-simplifier les taches de traduction

-proposer plusieurs possibilités de protéines pour une même séquence

-repérer exons / introns

D-Fournir des outils de prédiction :

- ❖ Prédiction physiologique et fonctionnelle afin de :

-repérer un opéron

-repérer un gène ou une protéine anormale

-prévoir la structure 3D d'une protéine

-repérer des mutations

-prédire une pathologie…

- ❖ Prédiction expérimentale afin de :

-repérer des sites de restriction

-prévoir la digestion d'un nucléotide

-prévoir / simuler la migration de fragments nucléotidiques ou protéiques lors d'une électrophorèse…

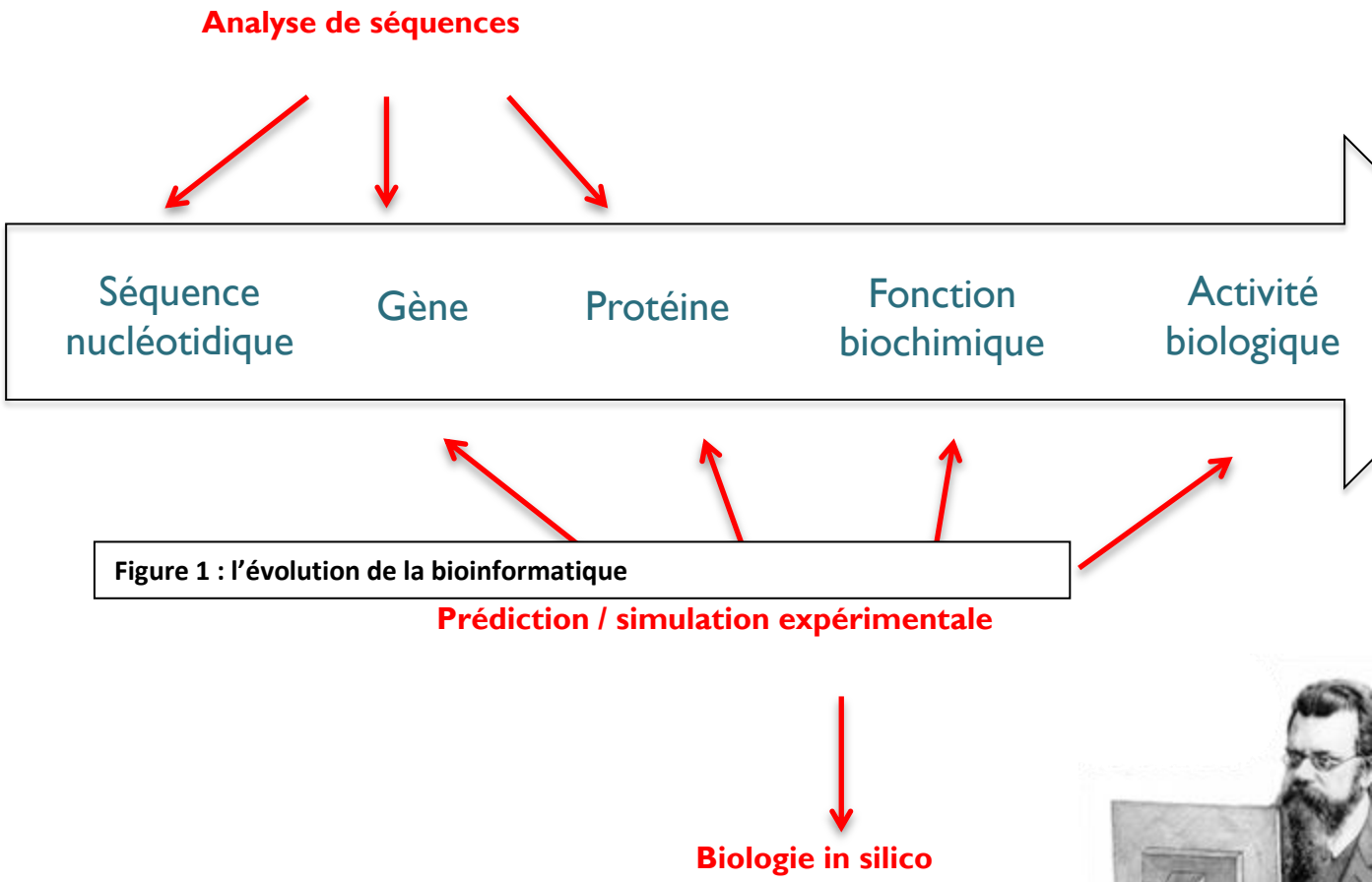


Figure 1 : l'évolution de la bioinformatique

I.4. Histoire de la bioinformatique :

˜ **Années 80** : - Début de la micro-informatique

- Création des premières bases de données (EMBL, GenBank, PIR (Accès téléphonique à la base PIR)

Années 90 : - Développement de l'internet et des réseaux

- Apparition des logiciels de comparaison (d'alignement) de séquences (FASTA, BLAST)

Années 2000 :

- Consultation libre en ligne des bases de données
- Mutualisation des données avec les projets de séquençages de génomes

I.5. Comment ça marche ?

La bioinformatique fournit des bases de données centrales, accessibles mondialement, qui permettent aux scientifiques de présenter, rechercher et analyser de l'information. Elle propose des logiciels d'analyse de données pour les études de données et les comparaisons et fournit des outils pour la modélisation, la visualisation, l'exploration et l'interprétation des données.

Chapitre 2 : Banques et bases de données biologiques

Banque de données

Ensemble de données relatif à un domaine ; Ensemble de fichiers manuels ou informatiques sans relation entre eux (*fichier plat*)

⇩

Des données exhaustives, donc offre un ensemble plutôt hétérogène d'informations

Base de données

Ensemble de relations entre les données, gérées à l'aide d'un système de gestion de base de données

⇩

Des données plus homogènes établies autour d'une thématique

Pour éviter toute confusion sémantique nous parlerons

⇩

Bases de données généralistes

⇩

Bases de données spécialisées

I. Les banques de séquences généralistes

C'est au début des années 80 que les premières banques de séquences sont apparues sous l'initiative de quelques équipes dont la première à l'initiative de Grantham et C. Gautier à Lyon. Elles couvrent tous les secteurs de la biologie, toutes les espèces. Ainsi, plusieurs organismes ont pris en charge la production de telles bases de données.

I.1.Les banques de séquences nucléiques

- **EMBL** (European Molecular Biology Laboratory ou Laboratoire Européen de Biologie Moléculaire) : banque européenne créée en 1980 et financée par l'EMBO (European Molecular Biology Organization), elle est aujourd'hui diffusée par l'EBI (European Bioinformatics Institute, Cambridge, UK) ;
- **GenBank** : créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information, Los Alamos, US) ;
- **DDBJ (DNA Data Bank of Japan)** : créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon) ;

I.2.Les banques protéiques :

- **PIR-NBRF (Protein Information Resource-National Biomedical Research Foundation) :** créée en 1984 par la NBRF (National Biomedical Research Foundation). Elle est maintenant un ensemble de données issues du MIPS (Martinsried Institute for Protein Sequences, Munich, Allemagne) et de la banque japonaise JIPID (Japan International Protein Information Database) ;
- **SwissProt** : créée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes, traduites de l'EMBL.

Elles contiennent la protéine obtenue de plusieurs manières différentes :

- *in silico* : déduite à partir de la séquence nucléique, par simple traduction du ou des exons la codant
- isolée à partir de la cellule
- ou encore par génie génétique

I.3. Avantages et inconvénients :

- ***Avantages :***

- Ces banques sont d'une importance majeure car elles offrent des informations qui ne sont plus reproduites dans la littérature scientifique (livres ou articles)

- Ces informations sont gratuites

- On y trouve une bibliographie et une expertise directement liées aux séquences traitées

❖ ***Inconvénients***

- Manques de vérification de séquences soumises

- Le temps d'insertion des séquences (retard … !)

II. Les bases de données de séquences spécialisées :

Elles couvrent un secteur défini de la biologie. Pour des besoins spécifiques, de nombreuses bases de données spécialisées ont été créées, Certaines sont pérennes et continuent d'être développées et mises à jour, d'autres sont laissées à l'abandon et enfin d'autres ont disparu. On en dénombre à cette date un peu plus d'un millier, accessibles directement par le Web. La nature ainsi que la quantité d'informations sont très variables.

II.1. Organisme :

Ces banques regroupent les données pour **un organisme particulier**, ou un groupe, contenant tout ou partie des informations suivantes :

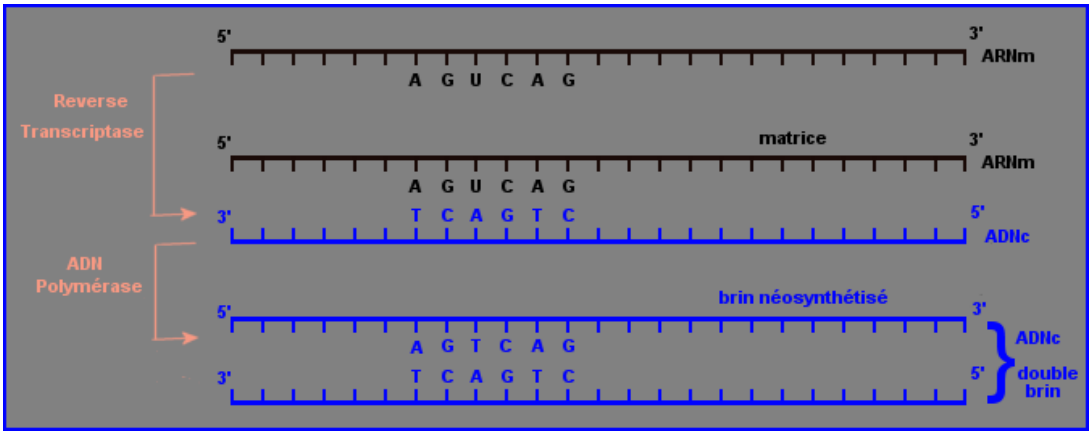
- **carte physique chromosomique** : la cartographie *physique* est de localiser les gènes sur les *chromosomes*.

- **carte génétique et liaison** :

- clonage positionnel pour les gènes :

- **EST (marqueurs de séquences exprimées) :** Un **marqueur de séquence exprimée**, ou ***expressed sequence tag*** (**EST**), est une courte portion séquencée d'un ADN complémentaire (ADNc), utilisée comme marqueur pour différencier les gènes entre eux dans une séquence ADN et identifier les gènes homologues dans d'autres espèces.
Parce qu'il est généralement assez facile de récupérer des brins d'ARNm des cellules, les biologistes récupèrent ces séquences et les convertissent en ADNc, qui est bien plus stable. Un ARNm étant forcément l'expression d'un gène du génome, cet ADNc n'est pas une copie exacte de la séquence ADN qui a généré l'ARN car, à la suite de l'épissage, l'ARN ne garde pas les régions non codantes de l'ADN (introns).

- **Banque d'ADNc :** L'**ADN complémentaire** (ou **ADNc**, **Acide désoxyribonucléique complémentaire**) est un simple brin artificiellement synthétisé à partir d'un ARNm, représentant ainsi la partie codante de la région du génome ayant été transcrit en cet ARNm. Il est obtenu après une réaction de transcription inverse d'un ARNm mature et équivaut donc à la copie ADN de l'ARNm qui a été extrait dans une cellule donnée à un moment donné.



- **Banque de vecteurs de clonage** : On appelle vecteur l'ADN dans lequel on insère le fragment d'ADN à étudier. L'ADN inséré est appelé insert ou ADN étranger ou ADN exogène. Cette séquence nucléotidique est capable de s'auto-répliquer.

Les vecteurs sont donc des petits ADN dans lesquels on insère un fragment d'ADN que l'on veut étudier. Ces petits ADN sont généralement des plasmides ou des bactériophages. Il est nécessaire d'introduire ces bactériophages ou plasmides dans les bactéries pour réaliser une multiplication de ceux-ci.

Les plasmides : Les plasmides sont des petits fragments d'ADN circulaire présents dans la cellule bactérienne et indépendants du génome bactérien.

Les phages : Les phages sont des virus qui infectent les bactéries.

- Gène et expression

- Cytogénétique et anomalies chromosomiques

- Gène et maladie

- Oncogènes

II. 2. Banques nucléiques spécialisée :

Elles sont spécialisées dans les informations suivantes :

- EST, ADNc
- ARN
- Structure secondaire d'ARN
- Signaux et éléments de régulation
- Sondes, amorces
- Alignements
- Famille de gènes

II.3.Banques protéiques spécialisées :

Elles sont spécialisées dans les informations suivantes :

- Alignement

- Classification structurale
- Familles de protéines
- Interactions
- Enzymes
- Modifications protéiques post-traductionnelles
- Pathologies
- Gels bidimensionnels
- Bases protéiques sur l'interaction et la thermodynamique des protéines

II.4. Banques immunologiques :

Elles sont spécialisées dans les informations suivantes :

- Séquences
- Récepteur (cellule T, par exemple)
- **Complex MHC** (Major Histocompatibility Complex) : un système de reconnaissance du soi présent chez la plupart des vertébrés. Les molécules du CMH sont à la surface de toutes les cellules nucluéées pour le CMH de classe I et les cellules présentatrices de l'antigène pour le CMH de classe II qui assurent la présentation de l'antigène aux lymphocytes T afin de les activer.

On distingue les complexes majeurs d'histocompatibilité de classe I et de classe II. Chez l'être humain, on parle d'antigène HLA.

- **Système HLA**

II.5. Banques Structure 2D ou 3D :

Elles sont spécialisées dans les informations suivantes :

- Coordonnées 3D de protéines
- Structure secondaire des protéines
- Domaines structuraux : est une partie d'une protéine capable d'adopter une structure de manière autonome ou partiellement autonome du reste de la molécule
- Centre actif des enzymes
- Complexes récepteurs-ligands
- Atlas de topologie structurale des protéines

❖ **Avantages**

- Elles fournissent des informations détaillées, spécifiques du domaine biologique qui n'existent pas dans les systèmes généralistes

- Les données sont en général contrôlées, donc plus fiables et de meilleure qualité que dans les bases généralistes

- Elles évoluent en fonction des progrès scientifiques dans le domaine plus facilement

- **Inconvénients :** ne cible pas toujours ce que l'on veut; toutes les banques possibles n'existent pas

Chapitre 3 : Alignement des séquences

I.Introduction

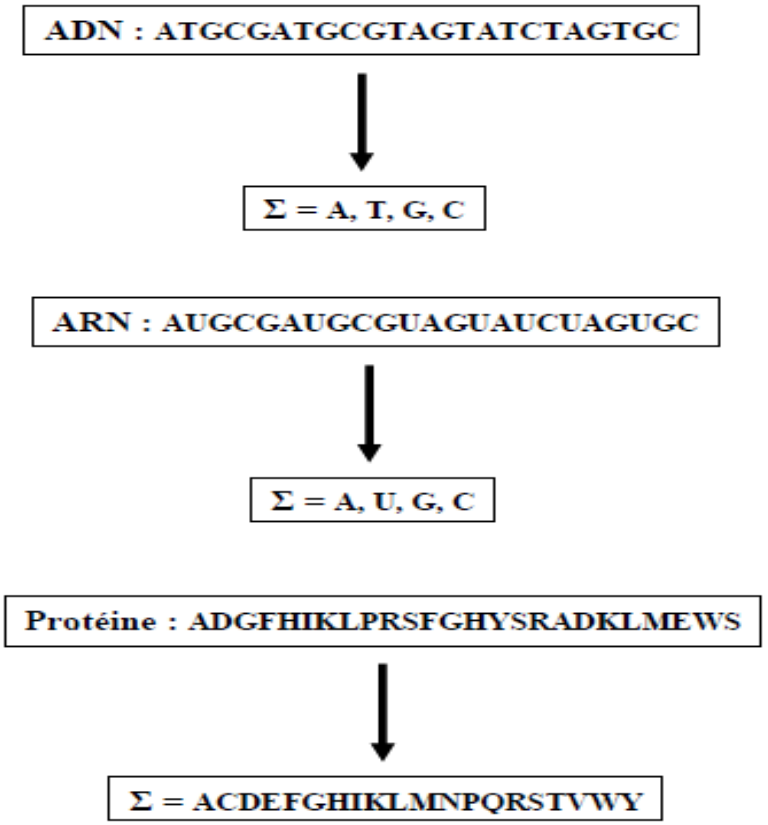
En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines : ARNm, régions 5'UTR, les EST, des clones, …) repose essentiellement sur la notion de l'**alignement**, et permet de déterminer le **degré de ressemblance** entre celles-ci (similitude ou identité en révélant des régions proches dans leurs séquences primaires). Cela peut alors indiquer que :

- La structure (primaire, secondaire ou tertiaire) des deux séquences est semblable,
- La fonction biologique est proche ou différente (dans le cas de la dissémilarité),
- L'origine des séquences alignées est commune ou éloignée (notion d'homologie), …

Cependant, la comparaison pour l'obtention d'un alignement optimal entre deux séquences biologiques, nécessite néanmoins la mise en oeuvre de procédures de calcul (algorithmes) et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences.

I.1.Qu'est-ce qu'une séquence ?

- Du point de vue d'un bio-informaticien, une séquence biologique est un **MOT**.
- Un **MOT** est une collection ordonnée de symboles choisis dans un alphabet (Σ).



Amino acid codes	
A	Ala Alanine
R	Arg Arginine
N	Asn Asparagine
D	Asp Aspartic acid
C	Cys Cysteine
Q	Gln Glutamine
E	Glu Glutamic acid
G	Gly Glycine
H	His Histidine
I	Ile Isoleucine
L	Leu Leucine
K	Lys Lysine
M	Met Methionine
F	Phe Phenylalanine
P	Pro Proline
S	Ser Serine
T	Thr Threonine
W	Trp Tryptophan
Y	Tyr Tyrosine
V	Val Valine
B	Asx Aspartic acid or Asparagine
Z	Glx Glutamine or Glutamic acid
X	Xaa Any amino acid

Nucleic acid codes	
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base

On ne considère que la structure primaire des séquences

- La séquence est représentée sous un format donné

I.2. Les différents types de formats :

➢ Format d'une entrée :

- 3 parties :
  - Description générale de la séquence
  - « Features » Description des objets biologiques présents sur la séquence
  - La séquence

```
ctccggcagc ccgaggtcat cctgctagac tcagacctgg atgaacccat agacttgcgc      60
tcggtcaaga gccgcagcga ggccggggag ccgcccagct ccctccaggt gaagcccgag     120
acaccggcgt cggcggcggt ggcggtggcg gcggcagcgg cacccaccac gacggcggag     180
```

- Chaque ligne commence par un mot-clé
  - Deux lettres pour EMBL
  - Maximum 12 lettres pour Genbank et DDBJ
- Fin d'une entrée : //

➢ Informations d'une entrée de la banque EMBL

ID : identificateur

C'est le nom de l'entrée contenant

```
--------------------------------------------------------------
ID   SCCBP6      standard; DNA; FUN; 937 BP.
XX
AC   M10154;
XX
DT   19-SEP-1987 (Rel. 13, Created)
DT   22-APR-1990 (Rel. 23, Last updated, Version 1)
XX
DE   Yeast (S.cerevisiae) nuclear gene CBP6 for cytochrome b,
DE   complete cds.
XX
KW   cytochrome; cytochrome b.
XX
OS   Saccharomyces cerevisiae (yeast)
OC   Eukaryota; Plantae; Thallobionta; Eumycota; Hemiascomyce
OC   Endomycetales; Saccharomycetaceae.
XX
RN   [1]
RP   1-937
RX   MEDLINE; 85105014.
RA   Dieckmann C.L., Tzagoloff A.;
RT   "Assembly of the mitochondrial membrane system";
RL   J. Biol. Chem. 260:1513-1520(1985).
XX
DR   SWISS-PROT; P07253; CBP6_YEAST.
XX
CC   There is a putative 'tata' box at position 215 to 219.
```

AC numéro d'accession de l'entrée correspondant à une soumission

DT date de création et de la mise à iour

DE titre de la séquence

KW mots clés données par les auteurs

OS/OC nom de l'espèce classé dans l'arbre de l'espèce

RN/RL références bibliographiques notés de 1 à n

DR références croisées avec d'autres bases

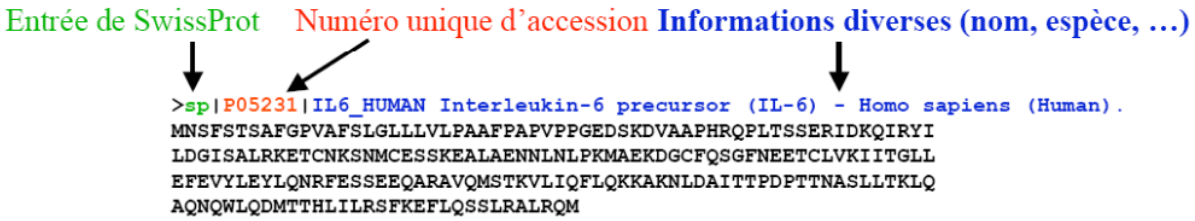
CC commentaires

```
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..937
FT                   /organism="Saccharomyces cerevisiae"
FT   CDS             301..789
FT                   /note="CBP6 protein"
FT                   /note="pid:g171173"
XX
SQ   Sequence 937 BP; 345 A; 159 C; 166 G; 267 T; 0 other;
ATACGATTAT TTTGGAAGTT TATAAAAGAA GTGCGGAAAT CACATCTGCT GTTTATTTAG      60
CCATTCCTCA CACTAATAGT TAAAGTACTT TCATAGCAGC TCTGCGCATG GTCGGACATG     120
CGAAAAATTC TGATATCAAG AAAAAGCGAA ATATTTCCGG CCTTGTAGGG GCCAAAACAT     180
TAACGTATAT CAAGATTTCC TGTGGTAGCA ACATTATAAG AAAAAAAGGT AGCCTTCATT     240
GAAACATTCT CTCTATCAGC TTACCAAGTT AAACTCCGTA TTCCACAAGC AAGTGCCAAA     300
ATGTCTTCTT CCCAGGTCGT CAGGGATTCT GCCAAAAAAT TAGTTAATTT ACTGGAAAAA     360
TATCCAAAGG ATCGTATACA CCACTTGGTC TCATTCAGGG ATGTACAAAT AGCAAGATTT     420
AGACGTGTAG CGGGTCTGCC AAATGTAGAT GACAAAGGAA AATCTATAAA AGAGAAAAAA     480
CCCTCATTAG ATGAAATAAA AAGTATAATT AACAGAACTT CCGGTCCATT AGGACTGAAT     540
AAGGAGATGT TAACCAAAAT TCAAAATAAA ATGGTAGATG AGAAATTCAC GGAAGAAAGC     600
ATCAACGAGC AAATTCGTGC CTTGAGCACT ATAATGAATA ATAAATTCAG AAACTATTAC     660
GATATTGGCG ATAAGCTCTA TAAACCTGCA GGAAATCCCC AATATTATCA ACGGTTAATA     720
AATGCCGTTG ACGGTAAGAA AAAGGAAAGC TTATTTACTG CAATGAGAAC TGTATTATTT     780
GGTAAATAAA GAGCACATTA TTTTCTAAGC TTGTAAATAC ATATTTATTC ATAATGGAGA     840
ACGTTATTCA AATTTATCTG TGAATTTCTT TACTCGAGGT ATACTTCCGC AAAGGAAATT     900
CTACTTAGCA AATCCTATGG TAACGTCATT GTTTTGT                              937
//
```

FH/FT caractéristiques biologiques données par les auteurs, Informations communes aux 3 bases nucléiques selon la

SQ Longueur, statistique des bases et séquence primaire.

- **Entrée de SwissProt:**



Figue 1 : Figure montrant une structure d'une entrée

II. Alignement des séquences :

II.1. Qu'est-ce qu'un alignement :

Est une manière de représenter deux ou plusieurs séquences de macromolécules biologiques (ADN, ARN ou protéines) les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires.
Rechercher le maximum d'appariements entre les résidus des séquences comparées.
L'alignement est d'autant plus parfait qu'il n'y a pas de mésappariements et de brèches (insertions ou délétions).

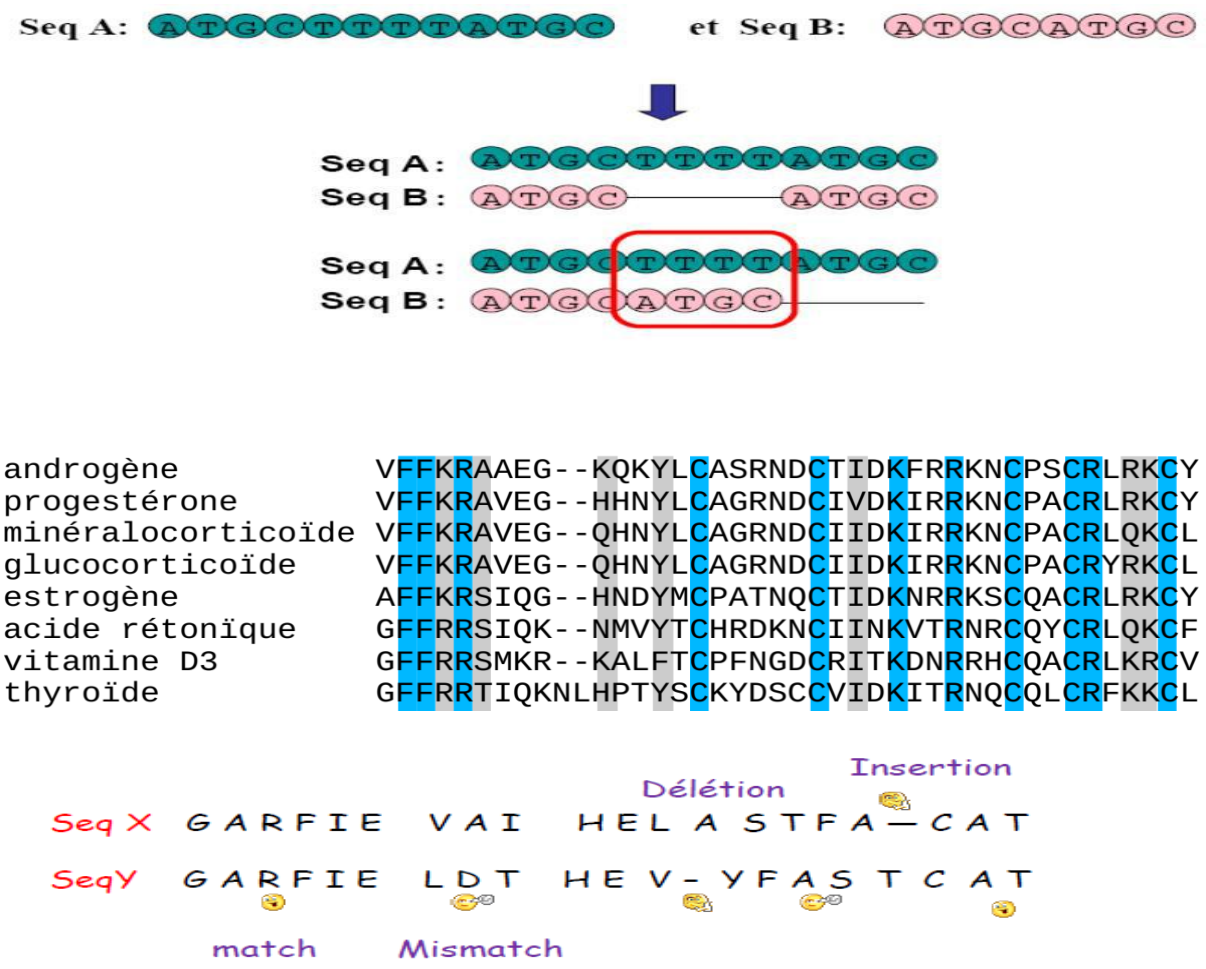
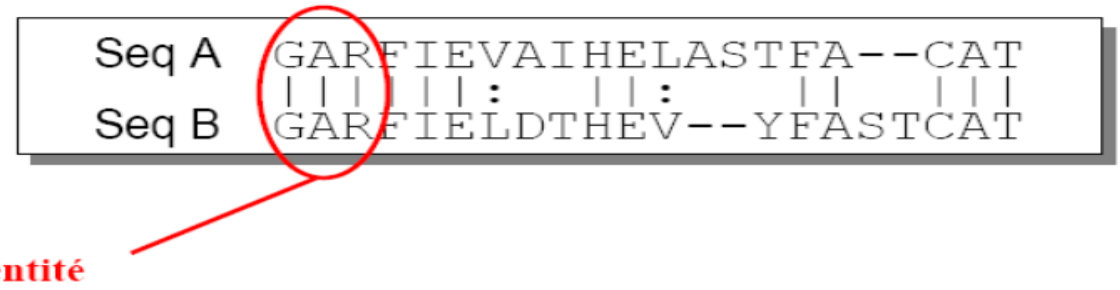
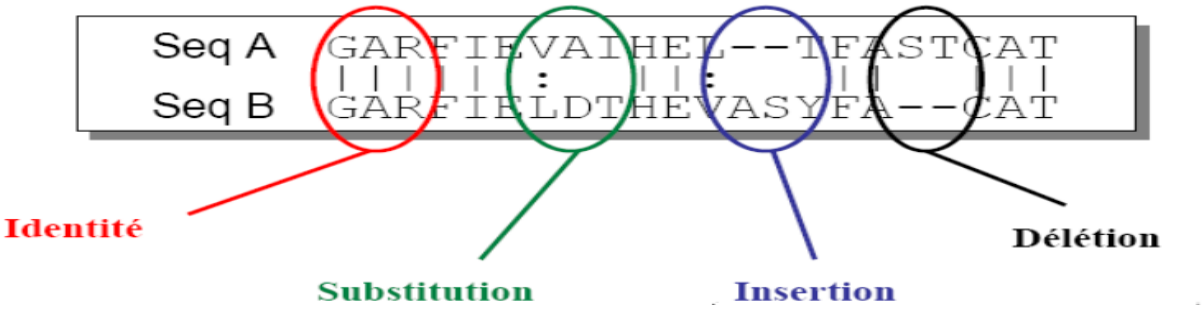


Figure 3 : Exemple d'alignement

❖ Les caractères sont les mêmes : **identité = match** (en anglais)



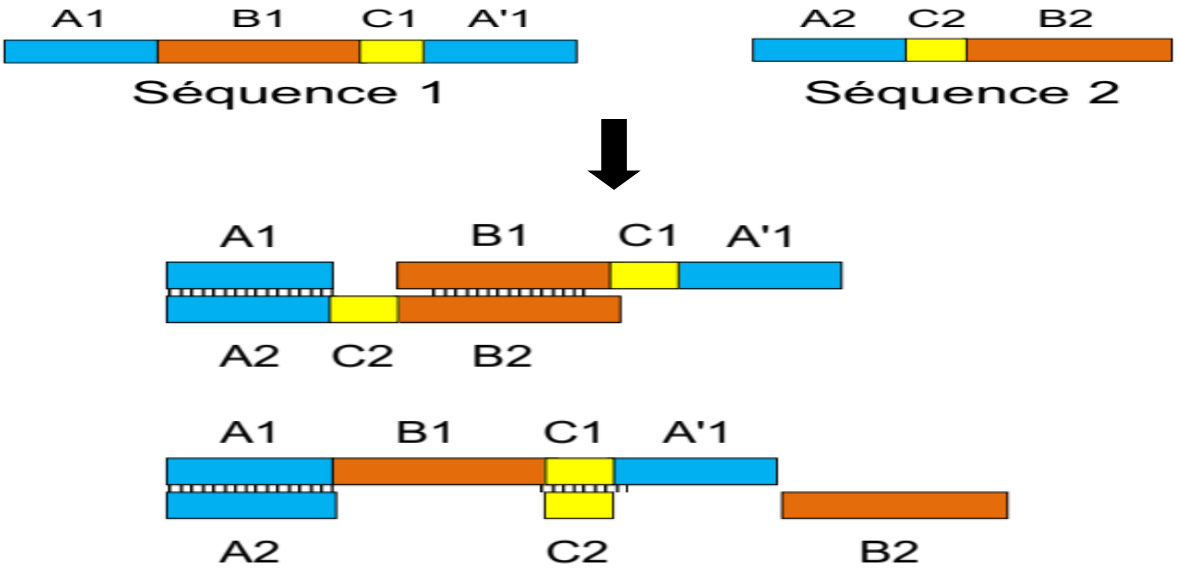
❖ Les caractères ne sont pas les mêmes : **substitution : mismach** (en anglais)



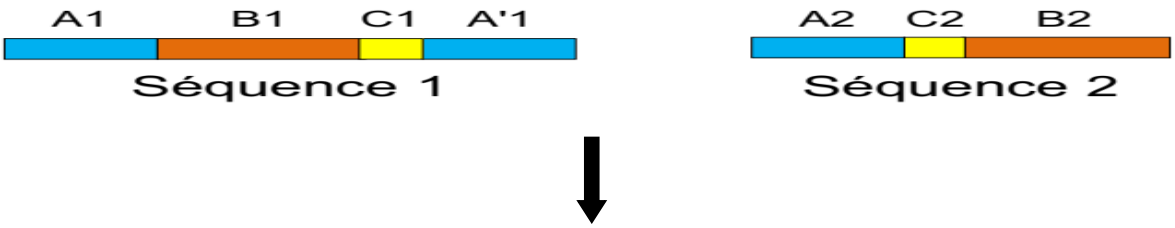
II.2. Les différents types d'alignement :

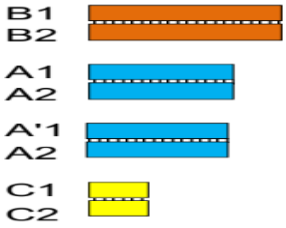
- **Alignement GLOBAL**
- **Alignement LOCAL**
- **Alignement MULTIPLE**

A. Alignement global :

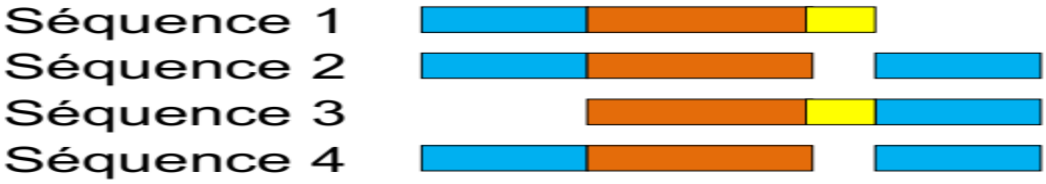


B. Alignement local :





C. Alignement multiple :



II. 3. Les applications de l'alignement :

- ✓ Identifier au sein d'une banque une séquence obtenue en laboratoire de biologie.
- ✓ Localiser une séquence d'acide nucléique au sein du génome d'un organisme
- ✓ Identifier un rôle à une molécule séquencée par comparaison avec des molécules de fonctions similaires déjà répertoriées.
- ✓ Réaliser une étude phylogénétique
- ✓ Prédire la structure secondaire (tertiaire) d'une protéine

II. 3. Les algorithmes utilisés :

- ❖ **Alignement global** : est conçu pour comparer des séquences homologues (apparentées) sur toute leur longueur.
- ✓ Needleman et wunsch
- ✓ Dot plot
- ✓ Stretcher
- ❖ **Alignement local :** est conçu pour rechercher dans la séquence A des régions semblable à la séquence B (ou à des parties de la séquence B).
- ✓ BLAST
- ✓ FASTA
- ✓ Smith et waterman
- ❖ **Alignement Multiple :**

✓ T-Coffee

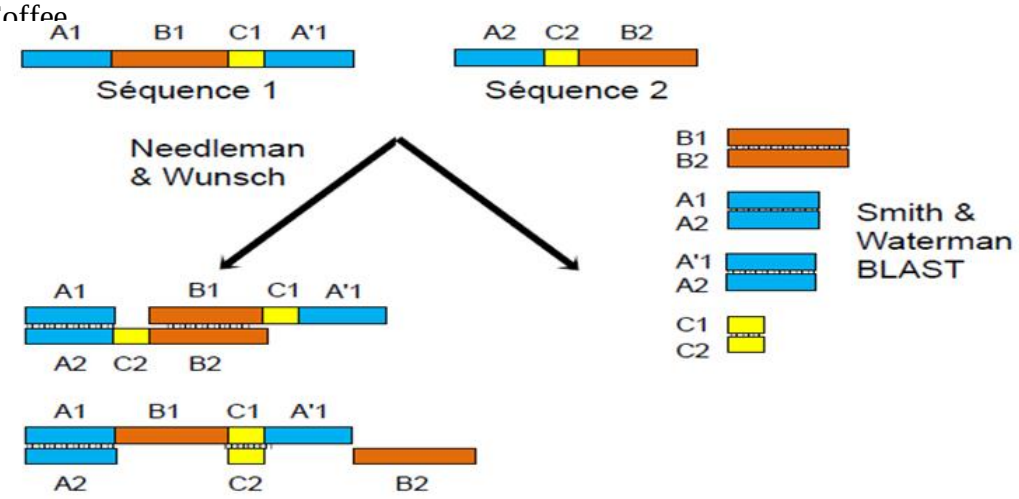


Figure 4 : Alignement global VS alignement local

II. 3.1. BLAST :

BLAST est l'abréviation de « Basic Local Alignment Search Tool » ou, en français, L'outil de recherche basique d'alignement local. Il ressemble à google dans le fonctionnement. On utilise google pour chercher les bases de données d'internet sur les informations d'un mot clé (cancer, par exemple). On cherche les bases de données d'internet sur des sujets qui ressemblent ou qui contiennent le mot clé. BLAST, quand à lui, cherche les bases de données des protéines et ADNs pour des séquences (sujets) qui ressemblent à notre séquence (requête) utilisée comme mot clé.

A. les programmes du BLAST :

La recherche BLAST la plus courante comprend cinq programmes :

- **Protein BLAST (BLASTp) :** cherche dans les bases de données de protéines en utilisant une protéine requête ;

- **BLASTx :** cherche dans les bases de données de protéines en utilisant la traduction d'un ADN ;
- **Nucleotides BLAST (BLASTn) :** cherche dans les bases de données de ADN en utilisant une ADN requête ;
- **tBLASTn :** cherche dans la base de données de l'ADN traduit en utilisant une protéine requête ;
- **tBLASTx :** cherche dans la base de donnée de l'ADN traduit en utilisant une ADN requête traduite

Programme	Base de données (Subject)	Requête (Query)
BLASTN	Nucléotide	Nucléotide
BLASTP	Protéine	Protéine
BLASTX	Protéine	Nucléotide → Protéine
TBLASTN	Nucléotide → Protéine	Protéine
TBLASTX	Nucléotide → Protéine	Nucléotide → Protéine

- ✓ **Query**: séquence requête.
- ✓ **Subject**: séquence sujet de la banque de donnée.

Logiciel	Exemples d'applications
blastn	Comparer les séquences d'ARNm aux séquences génomiques. Aligner un ARN d'interférence (ARNi) sur un génome pour détecter ses cibles potentielles.
blastp	En partant d'une protéine de fonction connue, collecter les protéines similaires dans la base de données Uniprot afin de constituer la famille de protéine supposées homologues.
blastx	Après avoir séquencé un morceau d'ADN, chercher des fragments potentiellement codants (susceptibles de produire un polypeptide similaire à des protéines connues) dans cette séquence même si on ne connaît pas la position des gènes.
tblastn	Identifier une région génomique susceptible de coder pour un homologue d'une protéine d'intéret. Identifier dans un génome les pseudo-gènes (gènes défectifs, qui peuvent contenir un ou plusieurs codons stop) correspondant à une protéine d'intérêt.

B. La recherche des séquences similaires :

Pour trouver des séquences similaires

- **Approche 1 :** naïve (Alignements 2 à 2 requête et cibles)
- ✓ **Avantage** : méthode optimale

- ✓ **Inconvénient :** méthode lente
- ❖ **Approche 2 :** heuristique (accélérer en prenant des raccourcis)
- ✓ **Avantage :** méthode rapide

C. Score :

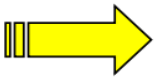
Il est calculé par BLAST pour déterminer le degré de ressemblance entre deux séquences.

Le plus le score augmente le plus les deux séquences se ressemblent.

Exemple 1 :

```
G T T A A G G C G - G G A A A
G T T - - - G C G A G G A C A
* * *       * * *   * * *   *
```

Identité = +1
Substitution = 0
Gap = -1



Score = 10 - 4 = 6

Exemple 2 :

```
Séquence 1   ATGACTGGGCCACT
             ||.  .  .||  .|.|||
Séquence 2   ATACTGGGACAACT
```

8 appariements (« match ») et 6 mésappariements (« mismatch »)

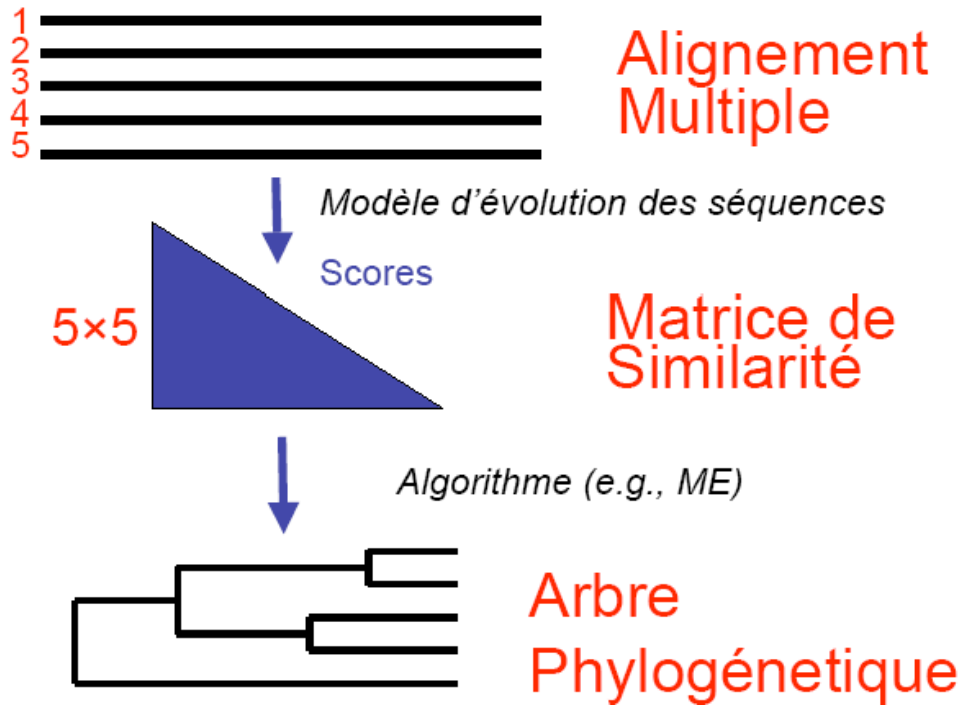
score= 8-0 =8

Exemple 3 :

```
Séquence 1   ATGACTGGGCC—ACT
             ||  || ||||||.|  |||
Séquence 2   AT—ACTGGGACAACT
```

12 appariements et 1 mésappariements et 2 brèches

score= 12-2 =10



D. Exemples des applications de la bioinformatique : (Td)
Exemple 1 : identification d'une souche bactérienne
Exemple 2 : réalisation d'un arbre phylogénétique
Exemple 3 : choix des amorces pour une séquence à amplifier

II. 4. Ressemblance ou similitude entre séquences nucléiques (ADN ou ARN) :

II. 4.1. Notion de score :

- **Le score élémentaire** (noté "s") est une entité numérique que l'on attribue à chaque couple de nucléotides des deux séquences à comparer. Il prend la **valeur de 1** lorsque les deux nucléotides des deux séquences sont **identiques,** et la valeur de zéro sinon.

Exemple :

Séquence1	A	G	C	T	A	C	C	T	G	T	Score global : Total des scores
Séquence2	A	A	G	T	A	G	C	T	T	T	
Point de comparaison	1	2	3	4	5	6	7	8	9	10	
Score élémentaire (s)	1	0	0	1	1	0	1	1	0	1	**1+0+0+1+1+0+1+1+0+1=6**

Dans cet exemple, constatez qu'au niveau du premier point de comparaison (ou site de comparaison), les deux séquences contiennent le même nucléotide A, donc le score élémentaire (s) à ce point prend la valeur de 1 (s = 1).

Au deuxième point de comparaison, la séquence 1 contient un G et la séquence 2 contient un A elles sont donc différentes en ce point d'où un score élémentaire de zéro (s = 0)…

- **Le score global** : Constatons que la somme des scores élémentaires est égale à six (s = 6).

Donc il y a six points identiques entre les deux séquences ; soit **60% d'identité** entre les deux séquences ([6/10] x100).

On dit alors que **le score global** entre les deux séquences est égal à six. Le score a donc permis de quantifier la ressemblance entre les deux séquences.

- **La relation entre le score global (S) et les scores élémentaires (s) pour deux séquences :**

$$S = \sum_{i=1}^{n} si$$

II. 4.2. Matrice d'identité :

Une **matrice d'identité** donne les valeurs de scores d'identité entre les séquences à comparer. Dans cette matrice, on attribue la valeur de 1 lorsque les deux nucléotides sont identiques et zéro s'ils ne le sont pas.

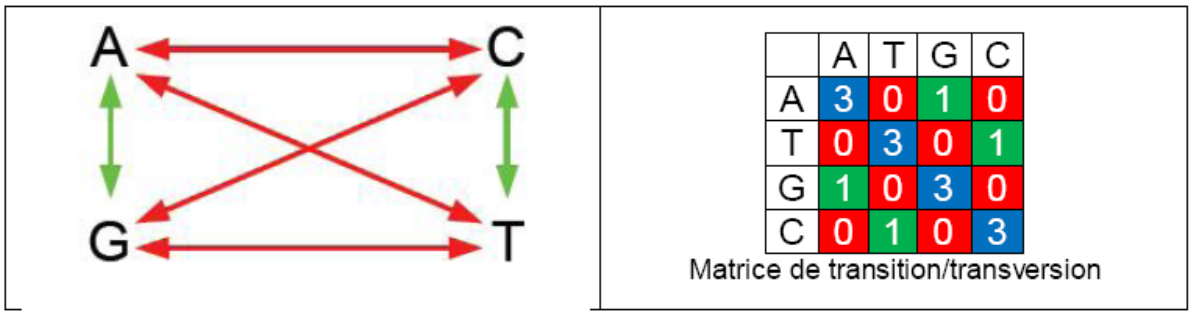
	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

Matrice d'identité nucléique

II. 4.3. La matrice de transition/transversion :

Il existe une autre matrice de score, qui tient compte de l'analogie structurale entre **purines (A et G) et pyrimidines (C, T et U)** et affecte des scores en fonction de cette ressemblance : C'est la matrice de transition/transversion :

La substitution entre purines d'une part, et entre pyrimidines d'autre part est pondérée et n'a pas de score élémentaire nul au moment de la comparaison des séquences :



	A	T	G	C
A	3	0	1	0
T	0	3	0	1
G	1	0	3	0
C	0	1	0	3

Matrice de transition/transversion

- **Quelle matrice utiliser ?**

En bioinformatique, on utilise beaucoup plus **la matrice d'identité.**

II. 4.4. Recherche de segments identiques :

II. 4.4.A. La matrice de points :

Elle permet une vue (méthode visuelle) englobant les similarités entre les régions des séquences à comparer.

- **Exemple de réalisation**: On donne deux séquences **X** et **Y** :

X =**ACTCGGATT** et **Y** =**AGCTCGGT**

Cette méthode consiste à créer une matrice qui va contenir les deux séquences (la séquence x en horizontal et la séquence y en vertical) et de cocher les cases de cette matrice pour le seul cas où les nucléotides sont **identiques (Match**). Quand il n'ya pas identité on parle de **Mismatch**:

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	G					X	X			
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X			
	T			X					X	X

Sur cette matrice, constatons qu'il y a **une diagonale** formée de cinq cases. Donc le segment identique le plus long entre les deux séquences X et Y contient cinq nucléotides identiques et consécutifs qui sont: **CTCGG**

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A									
	G									
	C		X							
	T			X						
	C				X					
	G					X				
	G						X			
	T									

Remarque : Dans le cas où les deux séquences sont complètement identiques, le résultat est une diagonale

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X	X		
	A	X						X		
	T			X					X	X
	T			X					X	X

II. 4.4.B. La méthode du Dot-Plot :

Le dot-plot est utile pour déterminer de combien d'exons est composé un gène en le comparant à son ARNm et pour avoir une idée de la taille des introns et des exons.

Il existe un logiciel de dotplot interactif, Dotlet qui nécessite JAVA. Si JAVA n'est pas installé sur vos machines, vous pouvez utiliser Dottup.

Le principe du dot-plot est basé sur la comparaison de fenêtres de longueur fixe que l'on déplace le long des séquences.

Soit deux séquences A et B à comparer et l la longueur de la fenêtre. On détermine sur la séquence A une première fenêtre de longueur l que l'on va comparer avec toutes les fenêtres possibles de même longueur, obtenues à partir de la séquence B. Un incrément est alors appliqué pour déterminer une deuxième fenêtre sur la séquence A, puis l'on recommence le balayage des comparaisons sur la séquence B. Si l'on choisit un incrément de 1 et que les séquences ont respectivement une longueur de m et n éléments, on effectuera de l'ordre de n x m comparaisons de fenêtres différentes.

Pour chaque comparaison entre deux fenêtres, un score est obtenu et l'on mémorisera uniquement les comparaisons dont les scores sont jugés significatifs, c'est-à-dire supérieurs ou égaux à un seuil que l'on s'est fixé. Par exemple lorsque le score correspond au minimum à 80% d'identité avec l'utilisation d'une matrice unitaire nucléique comme matrice de scores élémentaires.

II.5. des applications d'alignement :

- Considérons, par exemple, les deux séquences A et B suivantes :

Séq A = ATGTAATGCATG et Séq B = TATGTGAATG. La taille du motif (fenêtre) étant choisie égale à 5.

La fenêtre formée des cinq premiers nucléotides de la séquence A est : ATGTA. Il faut la comparer avec toutes les fenêtres possibles de taille égale à cinq retrouvées sur la séquence B. Ces séquences sont :

1. TATGT
2. ATGTG
3. TGTGA
4. GTGAA
5. TGAAT
6. GAATG

Remarque : Au-delà du nucléotide G en 6ème position dans la séquence B, on ne peut plus avoir une fenêtre de taille égale à cinq nucléotides.

La première comparaison concerne les deux motifs suivants :

Fenêtre de la séquence A = ATGTA

Fenêtre de la séquence B = TATGT

La comparaison de ces deux segments donne un score égal à zéro car il n'ya aucun nucléotide de la séquence A qui soit identique à celui de la séquence B quelque soit le site de comparaison :

Séquence A	A	T	G	T	A	Score global
Séquence B	T	A	T	G	T	
Scores élémentaires	0	0	0	0	0	S = 0

Remarque : Sur la matrice qui contient la totalité des deux séquences A et B, allez au à la case d'intersection qui rejoint le milieu du premier segment de A et celui de B pour insérer la valeur du score global :

	A	T	G	T	A	A	T	G	C	A	T	G
T												
A												
T			0									
G												
T												
G												
A												
A												
T												
G												

En fixant le motif de la séquence A (ATGTA), vous passez au deuxième motif de la séquence B qui est ATGTG :

Séquence A	A	T	G	T	A	Score global
Séquence B	A	T	G	T	G	
Scores élémentaires	1	1	1	1	0	S = 4

La comparaison de la fenêtre de A avec les cinq fenêtres possibles de B donne les résult suivants :

	ATGTA	ATGTA	ATGTA	ATGTA	ATGTA	ATGTA
	TATGT	ATGTG	TGTGA	GTGAA	TGAAT	GAATG
Nucléotides identiques (score)	0	4	1	3	0	1

Ce qui donnera sur la matrice globale :

	A	**T**	**G**	**T**	**A**	**A**	T	G	C	A	T	G
T												
A												
T			**0**									
G			**4**									
T			**1**									
G			**3**									
A			**0**									
A			**1**									
T												
G												

Une fois la comparaison effectuée avec toutes les fenêtres de B, nous incrémentons de un la séquence de A pour avoir la nouvelle fenêtre de cinq autres nucléotides qui sont : **TGTAA.**
C'est cette nouvelle fenêtre de A que nous allons devoir comparer avec les fenêtres de B que nous connaissons toutes maintenant.
Le résultat final est :

Il y a cinq segments formés de cinq nucléotides chacun entre les séquences A et B. Ces segments contiennent tous quatre nucléotides identiques:

1. ATGTA de la séqA et ATGGA de la séqB
2. TAATG de la séqA et GAATG de la séqB
3. TGTAA de la séqA et TGGAA de la séqB
4. TGCAT de la séqA et GGAAT de la séqB
5. GCATG de la séqA et GAATG de la séqB

III. L'alignement des séquences nucléiques: La programmation dynamique

III.A. Pourquoi vouloir réaliser des alignements ?

L'alignement, comme nous allons le voir dans les exemples suivants, permet de mesurer la similitude entre les séquences. S'il ya similitude, cela signifie qu'il est possible que les deux séquences présentent la même fonction biologique, ou du moins les deux séquences

présente une structure fortement similaire. Ce type d'information est nécessaire dans la mesure où, généralement, nous avons à faire à une séquence inconnue. Sa comparaison avec des séquences de structure et de fonction connues permet de tirer un maximum d'informations quant à la structure et la fonction de la séquence inconnue. Dans certains cas, on peut même confirmer si la séquence inconnue est un gène ou une portion de gène après l'avoir aligné avec des séquences de structure génique connue (régions codantes : codons d'initiation et de terminaison, sites d'épissage, zones de fixation des ribosomes).

III.B. L'algorithme de Needleman et Wunsch : Il permet de réaliser un alignement globalentre deux séquences nucléiques. Son expression est de la forme :

$$\boldsymbol{S}(i,j) = Max \begin{cases} \boldsymbol{S}(i+1,j+1) + s(i,j) \\ s(i+1,j) \\ s(i,j+1) \end{cases}$$

- **Exemple :**

Supposons que nous désirons calculer un alignement global des deux séquences suivantes de taille m et n respectivement:
S1 = TAAGTCCG m=8 et S2 = TACGTACG n=8
Remarque : Ici, les deux séquences sont de même longueur (8 résidus chacune). On peutcalculer un alignement entre deux séquences de tailles inégales.
Pour calculer l'alignement entre les deux séquences S1 et S2, quatre étapes sont nécessaires :

Etape 1 : Calcul de la matrice initiale
Il s'agit d'insérer les deux séquences S1 et S2 dans une matrice de sorte que S1 soit à l'horizontal et S2 à la verticale du tableau, puis remplir les cases par 1 (identité des deuxnucléotides de S1 et de S2) ou 0 (sinon) :

	T	A	A	G	T	C	C	G
T	**1**	0	0	0	**1**	0	0	0
A	0	**1**	**1**	0	0	0	0	0
C	0	0	0	0	0	**1**	**1**	0
G	**0**	0	0	**1**	0	0	0	**1**
T	**1**	0	0	0	**1**	0	0	0
A	0	**1**	**1**	0	0	0	0	0
C	0	0	0	0	0	**1**	**1**	0
G	0	0	0	**1**	0	0	0	**1**

Etape 2 : Calcul de la matrice transformée : Initialisation de la matrice
Construisons une nouvelle matrice à m+2 colonnes et n+2 lignes, dans laquelle la 1ère ligne et la 1ère colonne seront initialisées à zéro :

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0								
A	0								
C	0								
G	0								
T	0								
A	0								
C	0								
G	0								

L'application de l'algorithme de Needleman et Wunsh permet de remplir les cases de cette matrice. Le résultat est le suivant :

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2
C	0	1	2	3	3	3	3	3	3
G	0	1	2	3	4	4	4	4	4
T	0	1	2	3	4	5	5	5	5
A	0	1	2	3	4	5	5	5	5
C	0	1	2	3	4	5	6	6	6
G	0	1	2	3	4	5	6	6	7

Etape 3 : Parcours de la matrice transformée
Parcourir la matrice transformée depuis le plus haut score calculé (ici s=6) jusqu'au score le plus petit (ici s=1)

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	**1**	1	1	1	1	1	1	1
A	0	1	**2**	2	2	2	2	2	2
A	0	1	2	**3**	3	3	3	3	3
G	0	1	2	3	**4**	4	4	4	4
T	0	1	2	3	4	**5**	5	5	5
A	0	1	2	3	4	**5**	5	5	5
C	0	1	2	3	4	5	**6**	**6**	6
G	0	1	2	3	4	5	6	6	**7**

Etape 4 : Alignement des deux séquences et calcul de score

Séquence S1	T	A	A	G	T	—	C	C	G
	|	|	|	|	|	*	|	*	|
Séquence S2	T	A	A	G	T	A	C	—	G

Le score global de cet alignement est de 7.
Le pourcentage de l'identité entre les deux séquences S1 et S2 est :
%id = (7/9)*100=77,78%
Le trou retrouvé entre les nucléotides T et C de la séquence S1 est un GAP ou InDel :il signifie qu'à ce point, la séquence S1 a subi une mutation par **DELétion** au cours dela quelle le nucléotide A est perdu par nécessité évolutive et d'adaptation à l'environnement ; en même temps, il est conservé dans la séquence S2 (à la 6ème position face au gap de S1). Comme on peut supposer que c'est la séquence S2 qui asubi une mutation par **INsertion** du nucléotide A par nécessité adaptative. Dans un cas ou dans l'autre une des deux séquences a subi une mutation (**IN**sertion ou **DEL**étion) ; ce point est appelé **INDEL** pour dire INSERTION dans la séquence S2 ou DELETION dans la séquence S1. La même interprétation concerne le deuxième gap retrouvé 8ème position : il s'agit d'une délétion du nucléotide C dans la séquence S2 oude l'insertion de C dans la séquence S1.