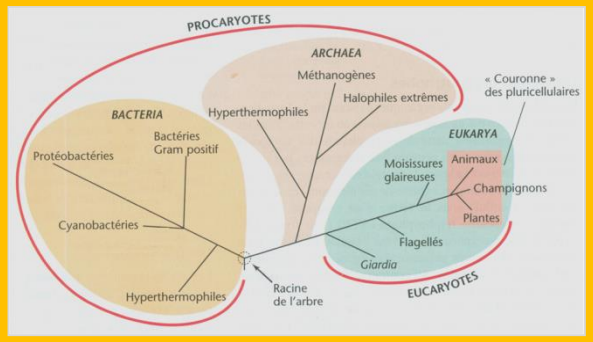
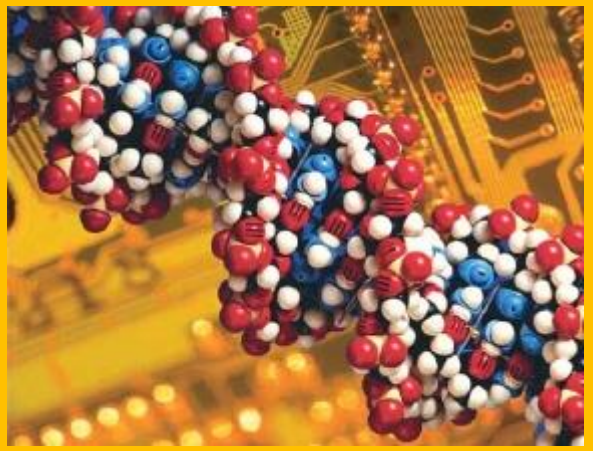
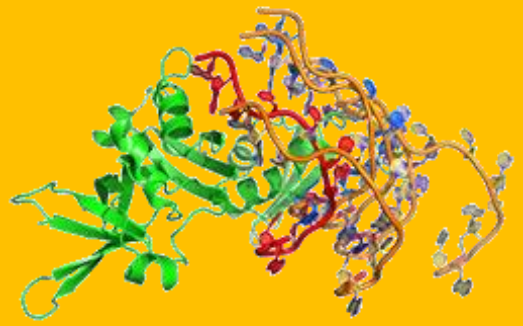
Université Frères Mentouri Constantine 1
Institut de la Nutrition, de l'Alimentation et des Technologies Agro-alimentaires (INATAA)
1e année Master Biotechnologie alimentaire
2019-2020

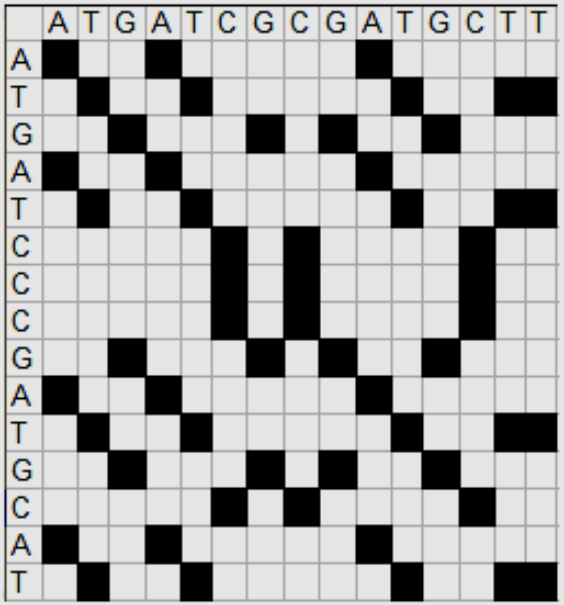


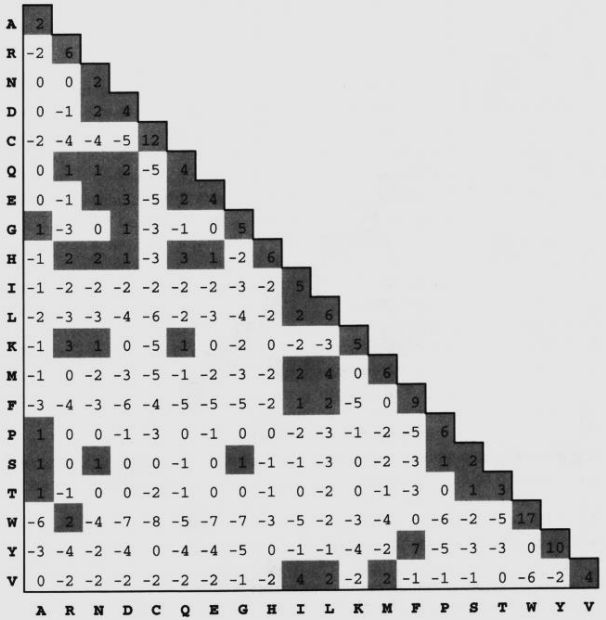
COURS DE BIOINFORMATIQUE

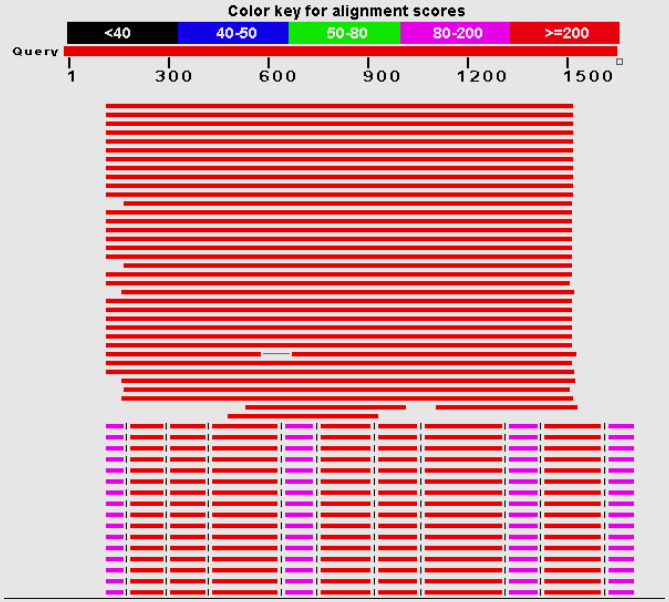


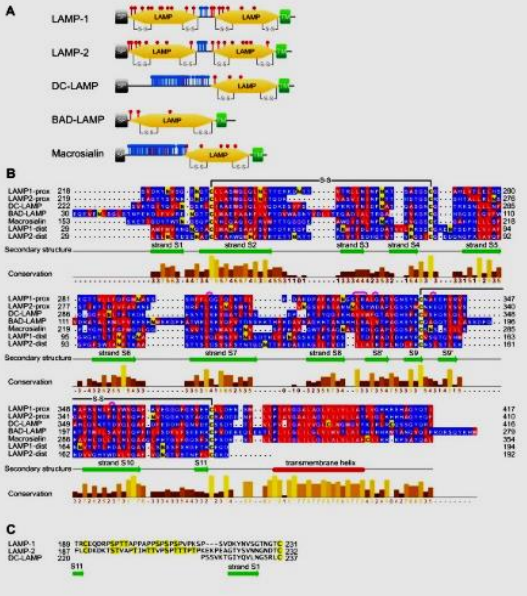


CHAPITRE II
LES ALIGNEMENTS (SUITE)









ALIGNEMENT MULTIPLE

Dans l'alignement multiple, il est question de comparer plusieurs séquences (des dizaines voire des milliers) à la fois, contrairement à l'alignement par paire. Cet alignement multiple va permettre de mettre en évidence les relations structurales et génétiques qui existent entre ces séquences.

ALIGNEMENT MULTIPLE

POURQUOI RÉALISER UN ALIGNEMENT MULTIPLE ?

- ✓ Afin de caractériser les familles de gènes ou de protéines, d'identifier les régions d'homologie partagée, de détecter les homologies en générale ;
- ✓ Identifier des sites fonctionnels, des séquences motifs ;
- ✓ Définir une séquence consensus de plusieurs séquences ;
- ✓ Contribuer à la prédiction des structures secondaires et tertiaires de nouvelles séquences ;
- ✓ Analyser l'évolution moléculaire (analyse phylogénétique) : reconstituer les relations évolutives (ou de parenté) entre séquences ;
- ✓ Choisir des amorces PCR permettant l'amplification de séquences homologues chez des organismes différents, etc.

ALIGNEMENT MULTIPLE

PRINCIPE

- ✓ Il est possible d'utiliser les algorithmes de la programmation dynamique pour un alignement de plus de deux séquences. Cependant, ceci s'avère en pratique impossible en raison de l'importance du temps et de la mémoire virtuelle nécessaires en fonction du nombre de séquences considérées ;

Par exemple, pour aligner 10 séquences d'une longueur d'environ 300 résidus, il faudra un espace mémoire de 515 Giga-Octets.

- ✓ C'est pourquoi les alignements multiples seront le plus souvent effectués au moyen de méthodes **heuristiques** qui produiront une **approximation** de l'alignement optimal ;
- ✓ Une heuristique est un algorithme qui fournit rapidement une solution réalisable, pas nécessairement optimale, pour un problème d'optimisation difficile.

On distingue, essentiellement, 02 grands types de méthodes :

- **progressives (alignement multiple progressif) :** débutent par l'alignement des deux séquences les plus proches, ensuite les séquences de plus en plus distantes sont **progressivement** ajoutées. Méthodes généralement plus rapides. Exemples : PIMA, PILEUP, Coffee, T-Coffee, Multi-LAGAN, etc.
- **itératives (alignement multiple itératif) :** construisent l'alignement en le calculant en plusieurs répétitions (**itérations**). Méthodes généralement plus précises. Exemples : Dialign, IterAlign, Praline, SAGA, HMMER, Muscle, MAFFT, HMMT, etc.

- ✓ L'alignement progressif est le plus couramment utilisé ;
- ✓ Dans un premier temps cette méthode calcule tous les alignements par paires possibles au moyen d'une procédure semblable à celle de l'algorithme de Needleman et Wunch ;
- ✓ Ensuite, le principe de la méthode consiste à effectuer des regroupements progressifs des séquences. Ce regroupement démarre en utilisant les deux séquences les plus similaires, les autres séquences (ou groupes de séquences alignées) étant ensuite intégrées au moyen d'un algorithme appelé **alignement par profil**.
- ✓ Clustal est l'algorithme progressif le plus populaire.

CLUSTAL (Thompson *et al.*, 1994) : *CLUSTer ALignement*

Principe : reconstruire l'alignement multiple à partir d'un **arbre guide** (constitué de **clusters**).

L'alignement passe par trois principale étapes:

Etape 1: Alignements globaux 2 à 2 par programmation dynamique ;

Etape 2: Regroupements des alignements (clusters), puis construction d'un arbre guide ;

Etape 3: Alignement multiple obtenu par combinaisons des alignements 2 à 2 (profils).

Soient les 4 séquences suivantes :

```
S1 cgatgagtcattgtgactg
S2 cgagccattgtagctactg
S3 cgaccattgtagctacctg
S4 cgatgagtcactgtgactg
```

Etape 1 : Alignements 2 à 2 (match = 1 , mismatch = -1 ; gap = -2) :

```
S1 cgatgagtcattgt-g--actg
   |||    | |||||| |  ||||
S2 cga---gccattgtagctactg
```

```
S2 cgagccattgtagcta-ctg
   ||| |||||||||||| |||
S3 cga-ccattgtagctacctg
```

```
S1 cgatgagtcattg-tgactg
   |||  | |  | | |  |||
S3 cgacca-ttgtagctacctg
```

```
S2 cga---gccattgtagctactg
   |||    | || ||| |  ||||
S4 cgatgagtcactgt-g--actg
```

```
S1 cgatgagtcattgtgactg
   |||||||||| ||||||||
S4 cgatgagtcactgtgactg
```

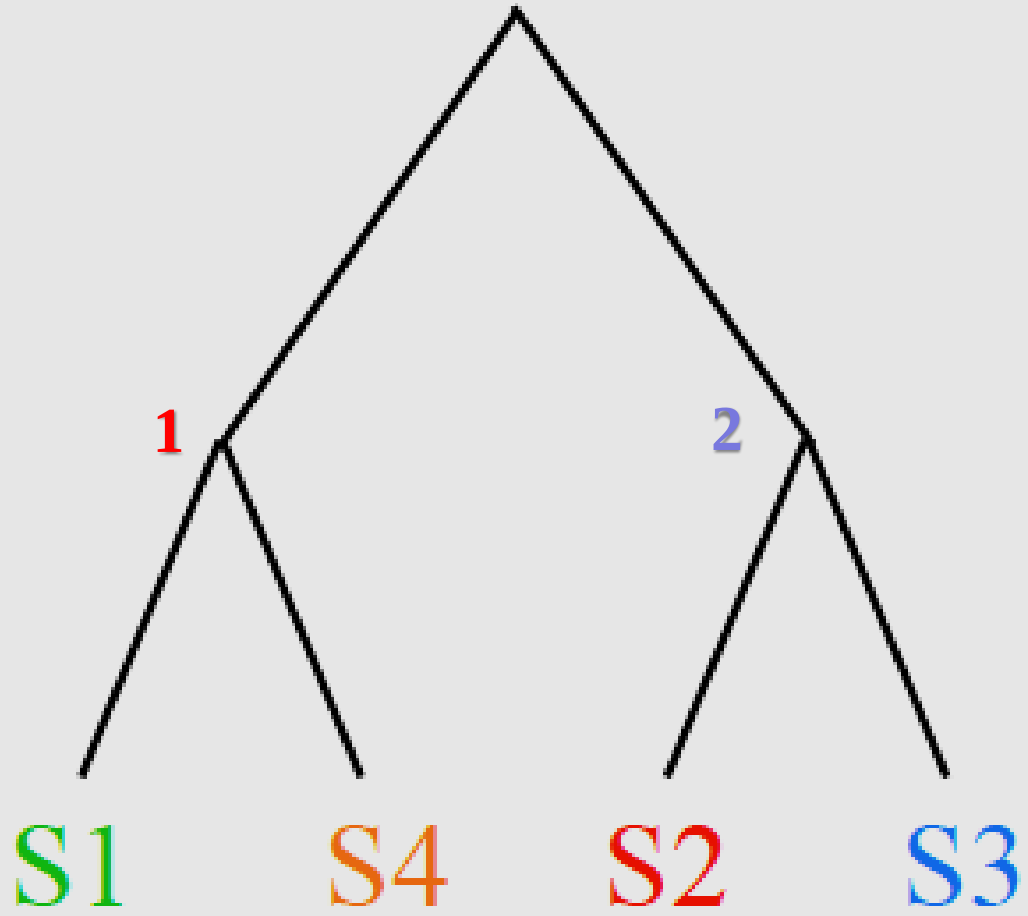
```
S3 cgaccattgtagctacctg
   |||  | |      |  |||
S4 cgatgagtcactgtgactg
```

Pour n séquences :
n(n-1)/2 alignements

Etape 2 : Tracer le tableau de scores de tous les alignements :

	S1	S2	S3	S4
S1	-	2	0	17
S2		-	14	0
S3			-	-1
S4				-

Etape 3 : Construction de l'arbre guide à partir de la table de score:



les alignements sont progressivement ajoutés les uns aux autres en partant des paires ayant les scores les plus élevés :

	S1	S2	S3	S4
S1	-	2	0	17
S2		-	14	0
S3			-	-1
S4				-

Etape 4 : Construction de l'alignement multiple final en suivant l'ordre déterminé par l'arbre guide :

```
S1 cgatgagtcattgt-g--a-ctg
S4 cgatgagtcactgt-g--a-ctg
S2 cga---gccattgtagcta-ctg
S3 cga----ccattgtagctacctg
```

Once a gap, always a gap

```
S1 cgatgagtcattgtgactg
   |||||||||| ||||||||
S4 cgatgagtcactgtgactg
```

```
S2 cgagccattgtagcta-ctg
   ||| ||||||||||||| |||
S3 cga-ccattgtagctacctg
```

S1 cgatgagtcattgtgactg S4 cgatgagtcactgtgactg S2 cgagccattgtagctactg S3 cgaccattgtagctacctg

Pas de méthode universelle. Plus les séquences sont divergentes, moins le résultat est fiable.

Heuristique	Rapidité	Séquences proches	Séquences éloignées	Qualité
Multalin	++	+++	+	++
Clustal	+	++	++	+++
Muscle	+++	+++	+	+++
MAFFT	++	++	+	+++
T-Coffee	+	+	+++	+++
DIALIGN	+	+	+++	+

ALIGNEMENT MULTIPLE

EXEMPLES DE LOGICIEL D'ALIGNEMENT MULTIPLE

L'outil d'alignement multiple **CLUSTAL Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)**.
(Démonstration visionnable ici : **https://youtu.be/Yzo7goqUcVE**)

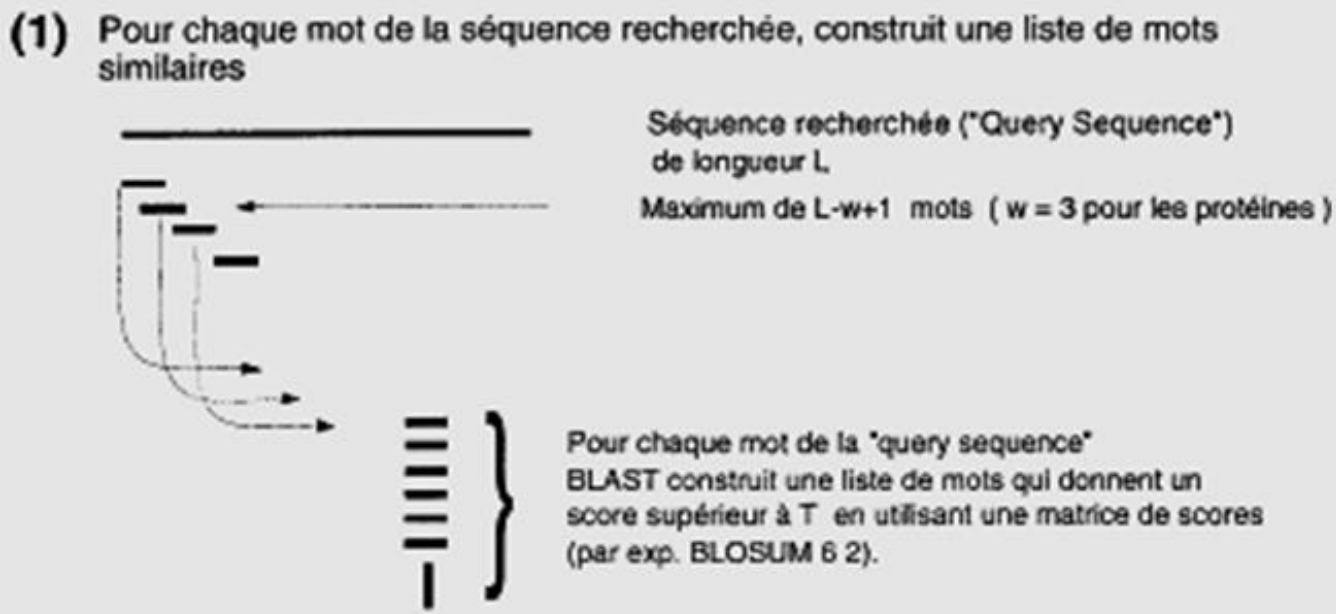
ALIGNEMENT D'UNE SÉQUENCE AVEC UNE BANQUE:

- ✓ Afin de comparer une séquence cible dite « requête » à l'ensemble des séquences d'une banque de données biologiques, des heuristiques permettant des alignements **locaux** ont été développé. Elles réalisent une analyse rapide qui permet de révéler des homologies entre des petits fragments à l'intérieur des séquences mais qui donnent également un résultat traduisant la similarité globale entre les séquences ;
- ✓ Plusieurs programmes utilisent ce type d'heuristiques, les plus connus sont les programmes **FASTA** et **BLAST**. Ce dernier étant le plus employé actuellement.

BLAST (*BASIC LOCAL ALIGNMENT SEARCH TOOL*) (Altschul *et al.*, 1990)

- ✓ BLAST recherche dans une base de données de séquences des segments qui sont localement homologues à une **séquence-requête** fournie par l'utilisateur (*query sequence*) ;
- ✓ Blast utilise une heuristique en trois étapes :

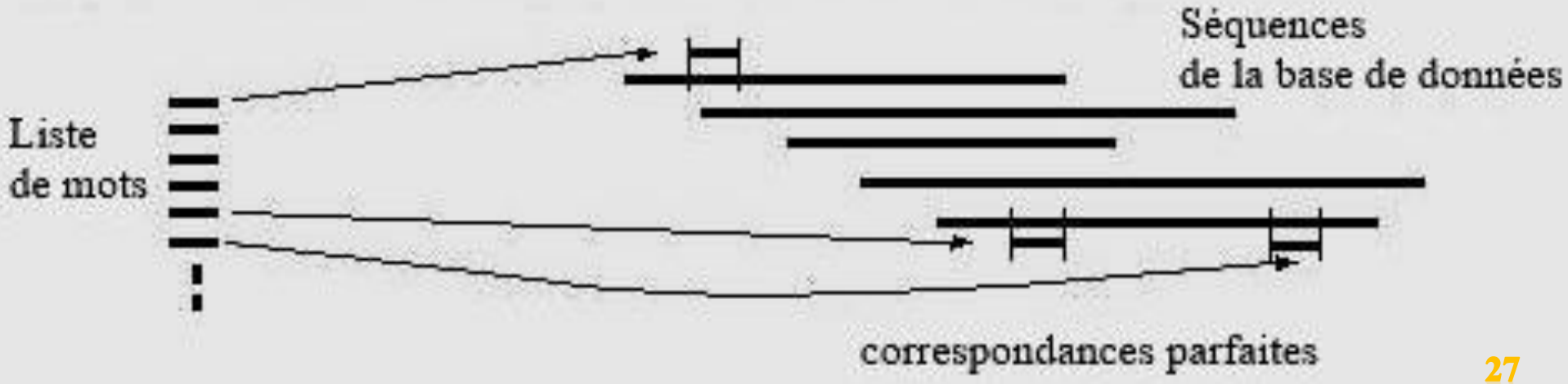
Etape 1: On construit un index de mots de longueur W pour la séquence requête, un automate qui permet de reconnaître ces mots lorsque l'on parcourt un texte est crée ;



BLAST (*BASIC LOCAL ALIGNMENT SEARCH TOOL*) (Altschul *et al.*, 1990**)**

Etape 2 : A l'aide de cet automate on parcourt le texte complet de la base et à chaque mot identifié on tente de construire un alignement local au voisinage de chaque mot rencontré ;

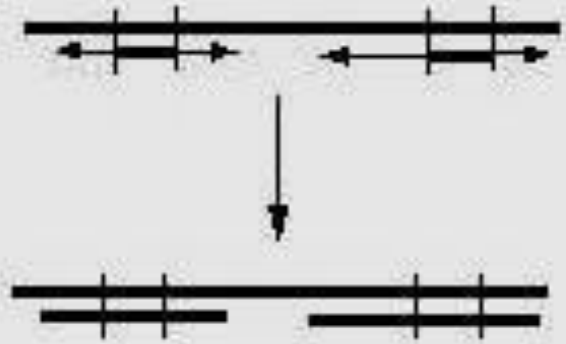
2) Blast compare la liste de mots à la base de donnés et identifie les correspondances parfaites



BLAST (*BASIC LOCAL ALIGNMENT SEARCH TOOL*) (Altschul *et al.*, 1990**)**

Etape 3 : un alignement est fourni si le score obtenu à l'étape précédente est supérieur à un seuil fixé.

3) Pour chacune des correspondances, Blast rallonge l'alignement dans les deux directions pour trouver les alignements ayant un score supérieur à un seuil de valeur S



L'outil d'alignement **BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)**. (Démonstration visionnable ici : **https://youtu.be/iVhVuEruXtQ**)

Les méthodes d'alignement utilisées en bioinformatique peuvent présenter des limites d'utilisations notamment pour l'analyse de grands volumes de données (génomes, métagénomes, séquences issues des méthodes NGS, etc.) :

i. l'homologie ne signifie pas toujours la conservation de régions à l'intérieur des séquences (cas des génomes viraux, de certaines protéines, etc.) ;
ii. la précision de l'alignement baisse significativement lorsque l'identité entre séquences est basse ;
iii. ces méthodes consomment énormément de ressources (mémoire, temps, etc.) ;
iv. il est difficile d'obtenir des résultats précis avec les méthodes d'alignement multiple basées sur des heuristiques ;
v. les paramètres statistiques (matrices de score, pénalités, etc.) employées sont souvent arbitraires.

Des méthodes alternatives aux méthodes basées sur l'alignement des séquences ont été développées. Elles sont classées selon leurs principes :

- méthodes basées sur le calcul des fréquences de sous-séquences d'une longueur définie (**méthodes basées sur la fréquence de mots**). Exemples : *feature frequency profile* (FFP), *composition vector* (CV), *return time distribution*, etc.
- méthodes qui évaluent le contenu informationnel entre des séquences complètes (**méthodes basées sur la théorie de l'information**). Exemples : *base-base correlation* (BBC), *information correlation-partial information correlation* (IC-PIC), *Lempel–Ziv compression*, etc. ;

Des méthodes alternatives aux méthodes basées sur l'alignement des séquences ont été développées. Elles classées selon leurs principes :

- d'**autres méthodes** basées sur la longueur des mots correspondants (exemples : mots communs, mots communs les plus longs, le minimum de mots absents, entre les séquences), ou basées sur la représentation graphique des séquences (cartes itérées).