Traitement informatique des données

Présentation du module:

Cette matière contribue au développement des acquis de l'étudiant en ce
Qui concerne les nouvelles bases de la bioinformatique

Connaissances préalables recommandées
Génétique
Biologie moléculaire
Microbiologie
Biochimie

Cours 01 : Initiation à la bioinformatique

Chargée de cours : ALIANE S

« Les biologistes se noient dans les données, la bioinformatique leur apprend à nager »

1. Généralités

Récemment, nous avons commencé à entendre le mot « Bioinformatique », cette discipline a fait son apparition dans les années quatre vingt, parallèlement à la création des premières banques de données, prenant l'exemple de GenBank et EMBL.

Cette matière s'intéresse à l'étude et l'analyse "*in silico* " (au moyen d'ordinateur) de l'information biologique contenue dans les séquences nucléotidiques et protéiques. A partir des années quatre vingt-dix, cette discipline devient indispensable avec l'accumulation des données de séquençage des protéines, des gènes d'identification comme ceux de l'ARN16S, des gènes de résistance, virulence et même de génomes complets.

Pour un microbiologiste, il est important de comprendre les méthodes utilisées en bioinformatique an fin de réussir l'application des outils disponibles et également pour une interprétation correcte des résultats (Ex. L'identification, la caractérisation des mutations, les ORF, l'évolution, l'homologie...etc.). Parmi ces outils, celles appliqués à la comparaison des génomes bactériens comme l'alignement (Ex. BLAST), sont d'une importance cruciale.

Globalement, la bioinformatique propose des méthodes et des logiciels qui permettent:

- Le recueil, le stockage et la gestion des données biologiques et leur distribution à travers les réseaux.
- Les études phylogénétiques et l'évolution moléculaire des êtres vivants.
- Le développement des outils pour analyser les problèmes de biologie moléculaire.
- L'analyse, la comparaison et la prédiction de la structure des gènes.
- La modélisation et la prédiction de la structure et de la fonction des protéines.

2- Historique

1955-1965: Premiers langages informatiques, premier ordinateur commercial

1970: 1er programme pour la comparaison de séquences protéiques, Alignement optimal entre deux séquences (Needleman & Wunsh)

1971: PDB - Protein Data Bank (structures 3D macromolécules)

1978: Matrice de substitution (PAM) (Dayhoff *et al.*)

1981: Similarités de séquences dans les banques (Smith & Waterman)

1980/1986: Création de des banques de données: EMBL (1980), GenBank (1982), DDBJ (1986), SwissProt (1986)

1989: L'internet diffusé pour le publique

1990: BLAST – Simulation de séquence dans les banques.

2000: - Séquençage du 1er génome de plante, *Arabidopsis thaliana*

- Publication du "draft" de la première carte complète du génome Humain (3000 MB).

2001: Publication des travaux de séquençage du génome humain presque complet.

3. Définitions

3.1. Bioinformatique

Domaine interdisciplinaire, situé au carrefour de l'informatique, des mathématiques et de la biologie, qui traite de l'application de l'informatique aux sciences biologiques.

« *Domaine jeune en effervescence, novateur, hybride et dynamique* »

La bioinformatique est une discipline **récente** et **hydride**, permettant **l'analyse** et l'**interprétation** de la bioinformation par des moyens informatiques.

Selon **NCBI (2001)** « *la bioinformatique est un axe de recherche dans le quel la biologie, l'informatique, la technologie de l'information sont fusionnées en une seule discipline* ».

Selon OLF (2001) « *la bioinformatique est un vaste domaine qui recouvre l'ensemble des utilisations de l'informatique pour la gestion, l'entreposage, le traitement, l'organisation, la comparaison et la diffusion de données relatives à l'ensemble des sciences biologiques (Physiologie, écologie, biochimie, biologie moléculaire et, dans une large mesure génétique et génomique*) »

3.2. Bioinformation

La bioinformation est l'information liée aux biomolécules : leur séquence, leur fonction, leur lien de parenté, leurs interactions et leur intégration cellulaire.

Cette bioinformation est issue de diverses disciplines, citant la biochimie, la génétique, la génomique structurale, la génomique fonctionnelle, la transcriptomique, la protéomique, la biologie structurale…etc.

4. Les données biologiques

Depuis les années soixante-dix, l'extraction et la collecte des données biologiques ont été progressivement développées, grâces aux différentes découvertes comme la méthode de Sanger (séquençage de l'ADN), qui ont permis d'automatiser une partie de l'extraction de connaissances des entités biologiques et ont entrainés la croissance exponentielle des banques des données qui y sont consacrées.

5. Les banques de données biologiques

Les banques de données ou bases de données biologiques sont définies comme étant des bibliothèques contenant des données de séquences biologiques collectées grâce à des expériences scientifiques, largement diffusées par le réseau internet. Elles sont généralement reliées entre elles par des liens.

L'origine des banques de données biologiques remonte à l'utilisation des premiers ordinateurs par des cristallographes ou des biochimistes. Parmi ceux, Margaret Dayhoff, biochimiste américaine, fut la première à voir l'intérêt de rassembler toutes les données sur les séquences des protéines en vue d'étudier leurs relations évolutives et de classer en familles.

En 1965, elle publia le premier Atlas de protéines contenant la séquence et la structure de 65 d'entre elles (*Atlas Of Protein Sequence And Structure*). Et en 1984, ce volumineux Atlas est devenu la banque des données P.I.R. (*Protein Information Resource*) de National Biomedical Research Foundation (N.B.R.F.). C'est la première banque des protéines et qui reste une référence pour leur analyse. En 2004, elle contenait quelque 283 000 séquences protéiques qui totalisent 96 millions d'acides aminés.

Des banques de données des séquences nucléiques ont fait leur apparition en 1982. Aux Etat Unis, GenBank est une banque de données qu'a prises corps au L.A.N.L (Los Alamos Nuclear Laboratory) avec Doug Brutlag et Temple Smith. Cette derniere est gérée et distribuée par le N.C.B.I (National Center for Biotechnology Information). Alors que la banque nucléique européenne a pris le nom de laboratoire au sein duquel elle a été développée à Heidelberg (Allemagne) E.M.B.L. (European Molecular Biology Laboratory).

Il est important de signaler qu'une antenne spéciale pour l'informatique a été créée à Cambridge en 1997, E.B.I. (European Bioinformatics Institute), afin de poursuivre le développement de la banque E.M.B.L.

Tableau 01: Exemples de quelques bases de données biologiques

Base de données Description Site web (ou adresse internet)	**Base de données Description Site web (ou adresse internet)**	**Base de données Description Site web (ou adresse internet)**
Genbank	Banques de séquences nucléotidiques publiquement disponibles et leur traduction en protéines.	*www.ncbi.nlm.nih.gov/enter (Acides nucléiques)*
EMBL (European Molecular Biology Librery).		*www.srs.ebi.ac.jp (Acides nucléiques)*
NCBI (national center for biotechnology Information)	Portail américain de bioinformatique: BLAST, banques de séquences nucléotidiques et génomiques, PubMed.	*www.ncbi.nlm.nih.gov (Portail)*
Pubmed-Medline	Articles, publications en biologie et en médecine	*www.ncbi.nlm.nih.gov (Bibliographie)*
Google Scholar	Publication scientifiques avec google	*http://scholar.google.com (Bibliographie)*
Connectsciences	Articles, thèses	*http://connectsciences.inist.fr (Bibliographie*

6. Principes de bases de l'alignement des séquences

Durant les dernières décennies, des progrès énormes ont été réalisés dans trois domaines qui ont profondément affecté la classification des microorganismes. Il s'agit d'une part de l'étude détaillée de la structure cellulaire par microscopie électronique, d'autre part de la caractérisation physiologique et biochimique de nombreux microorganismes et enfin de la comparaison des séquences d'acides nucléiques et de protéines d'une grande variété de microorganismes.

La diversité génétique est due à des mutations ponctuelles et à des insertions ou (et) délétions apparues au cours de l'évolution. **L'alignement de séquences** (globale, local ou multiple) de deux ou plusieurs microorganismes est une méthode de comparaison qui permet la détermination de leur **degrés de similarité** (deux organismes sont similaires si leur score de substitution est supérieur à 0), **d'homologie** pour montrer s'ils dérivent d'un ancêtre commun, et leur identité par estimation de la fraction de résidus identiques.

6-1- Alignement et détermination du score

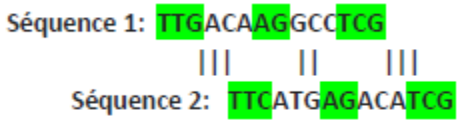
L'alignement de séquences est une manière de représenter deux ou plusieurs séquences (nucléotides, acides aminés) les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires (les zones de concordance). Elle nous permet de révéler trois opérations d'édition: **1**) Substitution **2**) Délétion **3**) Insertion.

Les pénalités des brèches doivent être suffisamment coûteuses pour éviter les alignements sans signification biologique. La recherche de la similarité entre séquences nécessite la détermination d'un score de similarité.

Le score de l'alignement est la somme des scores élémentaires :

Score = Σ scores élémentaires - Σ score pénalité.

Exemple: Déterminer le score des séquences suivantes:



On a 8 appariements et six mésappariement (1 appariement "match": vaut +1 et 1 mésappariement "mismatch" vaut 0): **Score:** 8-0 = 8

On a 12 appariements, 1 mésappariement, 2 brèches (1 brèche vaut -1): **Score:** 12-0-2 = 10

6-2- Alignement global

Les alignements globaux (alignement des séquences sur la totalité de leur longueur en tenant compte de tous les résidus) sont plus souvent utilisés quand les séquences mises en jeu sont similaires et de **tailles comparables**. Une technique générale, appelée algorithme de Needleman-Wunsch et basée sur la programmation dynamique permet de réaliser des alignements globaux de manière optimale même pour les longues séquences.

6-2-1- Algorithme de Needleman et Wunsch

C'est la méthode la plus utilisée est connue qui réalise le meilleur alignement global entre deux séquences. Il permet d'évaluer le degré de similarité entre deux séquences connues (détermine le meilleur alignement évalué par un score). Pour cela il construit le chemin optimal allant du coin supérieur gauche au coin inferieur droit d'un tableau à deux dimensions. Pour déterminer la valeur d'une case on doit respecter les trois règles suivantes: Règle 1: chaque case va contenir un score; le score de l'alignement sera celui de la case en bas à droite. Règle 2: le score d'une case se déduit à partir de celui des cases au-dessus, à gauche ou en diagonale. Règle 3: un pas horizontal/vertical coûte 1 gap un pas diagonal coûte 1 position alignée (match ou mismatch). Ci-dessous un exemple de l'alignement de deux séquences. **$\sigma(S1_{(i)}, S2_{(j)})$:** la similitude entre les deux séquence (match oumismatch).

		i=0	i=1	i=2	i=3	i=4	i=5
		m	**T**	**G**	**G**	**T**	**G**
j=0	**n**	0	-2	-4	-6	-8	-10
j=1	**A**	-2	-1	-3	-5	-7	-9
j=2	**T**	-4	-1	-2	-4	-4	-6
j=3	**C**	-6	-3	-2	-3	-5	-5
j=4	**G**	-8	-5	-2	-1	-3	-4
j=5	**T**	-10	-7	-4	-3	0	-2

$$T(i, j) = \text{Max} \begin{cases} (T_{(i-1, j-1)} + \sigma(S1_{(i)}, S2_{(j)})) \\ (T_{(i-1, j)} + \text{gap penality}) \\ (T_{(i, j-1)} + \text{gap penality}) \end{cases}$$

$$\text{Score} \begin{cases} \text{Match} = 1 \\ \text{Mismatch} = -1 \\ \text{Gap} = -2 \end{cases}$$

Traceback ou le chemin optimal apparu en jaune

Première application

T (1, 0) = Max
- (T (1-1), (0-1))= T(0, -1) et cette position n'existe pas
- T (1-1, (0))+ (-2) = -2 (le bon score)
- T (1, (0-1)= T(1, -1),cette position n'existe pas

Deuxième application

T (2, 0) = Max
- (T (2-1), (0-1))= T(1,-1) cette position n'existe pas
- (T(1,0)= -2) + (-2) = - 4 (le bon score)
- (T (2, (0-1)= T(2, -1) cette position n'existe pas

Troisième application

T (3, 0) = Max
- (T (3-1), (0-1))= T(1,-1) cette position n'existe pas
- (T (2,0)= -4)) + (-2) = -4 -2 = -6 (le bon score)
- (T (3, (0-1)= T (3, -1) cette position n'existe pas

Quatrième application

T (0, 1) = Max
- (T (0-1), (1-1))= T(-1,0) cette position n'existe pas
- (T (-1+1)= -4)) + (-2) = 0 -2 = -2 (le bon score)
- T (3, (0-1)= T(0, 0) cette position est déjà remplie

Cinquième application

T (1, 1) = Max
- T(0,0)+ (-1)= -1 (le bon score)
- (T (0,1)= -2)) + (-2) = -4
- T (1, 0) + (-2)= -2+ (-2)= -4

Étape 1: Création d'un tableau indexé par les deux séquences.

i / j		A	C	G	G	C	T	A	T
	0	-2	-4	-6	-8	-10	-12	-14	-16
A	-2	2	0	-2	-4	-6	-8	-10	-12
C	-4	0	4	2	0	-2	-4	-6	-8
T	-6	-2	2	3	1	-1	0	-2	-4
G	-8	-4	0	4	5	3	1	-1	-3
T	-10	-6	-2	2	3	4	5	3	1
A	-12	-8	-4	0	1	2	3	7	5
G	-14	-10	-6	-2	-1	0	4	5	6

- Case Sim (**i**; **j**) : score entre les i premières bases de U=ACGGCTAT et les j premières bases de V=ACTGTAG.Match =2, Mismatch = -1, GAP (Insertion/Délétion = -2.

- Cas de base (rouge): initialisation

- Remplissage ligne par ligne en gardant en mémoire le mouvement qui donne le meilleur score

Exercice 01:

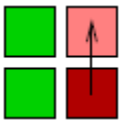
Afin de mieux comprendre comment remplir le tableau ci-dessus, prière de regarder la vidéo intitulé « Introduction to bioinformatics- Needleman Wunsch Algorithhm » https://youtu.be/of3B02hZGS0

Étape 2: Recherche du chemin optimal dans la matrice.

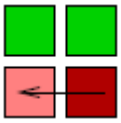
i / j		A	C	G	G	C	T	A	T
	0	-2	-4	-6	-8	-10	-12	-14	-16
A	-2	2	0	-2	-4	-6	-8	-10	-12
C	-4	0	4	2	0	-2	-4	-6	-8
T	-6	-2	2	3	1	-1	0	-2	-4
G	-8	-4	0	4	5	3	1	-1	-3
T	-10	-6	-2	2	3	4	5	3	1
A	-12	-8	-4	0	1	2	3	7	5
G	-14	-10	-6	-2	-1	0	4	5	6

Étape 3: Construction de l'alignement.

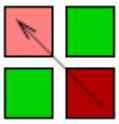
Sur le chemin des scores maximaux, on regarde quelle est l'opération correspondante.



Insertion



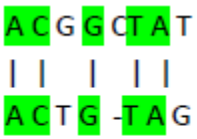
Délétion



Subtitution

(match ou mismatch)

Résultat:



6-3- Alignement local:

Pour obtenir un alignement local optimal, une méthode a été développée par Smith et Waterman. Cette méthode permet l'alignement entre deux séquences (dont leurs longueurs sont différentes) portant sur des régions isolées et permettant de trouver des segments qui ont un haut degré de similitudes (score le plus élevé de la matrice). Cette propriété en fait un outil idéal, rapide et efficace, de recherche dans les bases de données en comparant une séquence inconnue avec les séquences de la banque. Elle permet de trouver des séquences homologues à notre séquence parmi les millions de séquences.

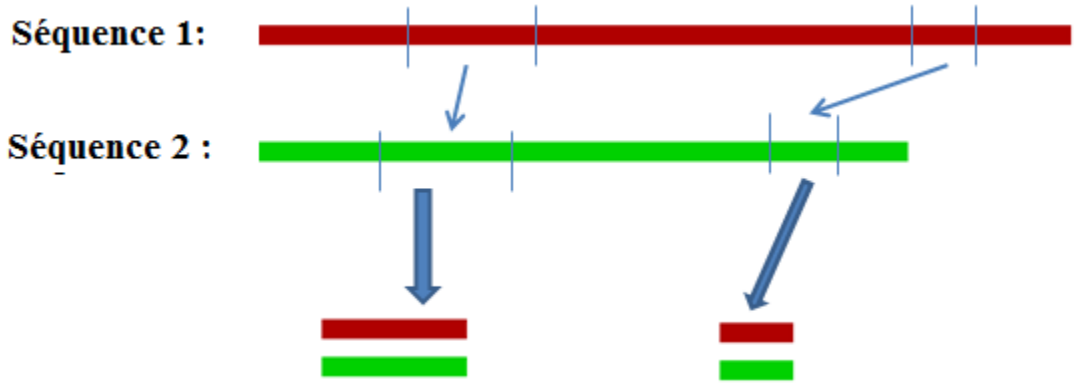


Figure 01: Alignement local.

6-4- BLAST

Le BLAST est acronyme de «Basic Local Alignment Search Tool». Cette méthode de recherche a été développée spécialement pour permettre de trouver les alignements locaux (régions similaires ou homologues entre deux ou plusieurs séquences de nucléotides ou d'acides aminés) statiquement significatifs. Ce programme permet de retrouver rapidement dans des bases de données, les séquences ayant des zones de similitude avec une séquence donnée (introduite par le chercheur).

L'idée de base exploitée par l'algorithme est que les bons alignements doivent contenir quelque part des petites régions très riches en identité. Ces éléments, constituent les points d'ancrage à partir desquels l'alignement est étendu. Ce repérage initial permet de sélectionner rapidement les séquences de la banque potentiellement similaire à la séquence introduite. Le BLAST est utilisé pour trouver des relations fonctionnelles ou évolutives entre les séquences et peut aider à identifier les membres d'une même famille de gènes.

La signification statistique des alignements produits par BLAST est mesurée par E-value "Expected value". Cette valeur indique le nombre d'alignements différents ayant les mêmes degrés de similitude. Si E= 10^{-2} cela signifie que l'alignement sur 100 sera trouvé par hasard. E-value est donné par la formule suivante: $e = k .m .n . e^{-\lambda.S}$

L'E-value dépend donc de la taille de la séquence (m), de la taille de la banque (n), et du *score* d'alignement (S), k et λ sont des paramètres caractérisant la banque de données. Un alignement est d'autant plus significatif que S est élevé et E faible.

Haemophilus influenzae strain 680 16S ribosomal RNA gene, complete sequence

Sequence ID: ref|NR_044682.2| **Length:** 1486 **Number of Matches:** 1

▷ See 1 more title(s)

Range 1: 432 to 856 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
778 bits(421)	0.0	425/425(100%)	0/425(0%)	Plus/Plus

```
Query  1    GTTCTTTCGGTATTGAGGAAGGTTGATGTGTTAATAGCACATCAAATTGACGTTAAATAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  432  GTTCTTTCGGTATTGAGGAAGGTTGATGTGTTAATAGCACATCAAATTGACGTTAAATAC  491

Query  61   AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGNGTGCGAGCGTT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  492  AGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGNGTGCGAGCGTT  551

Query  121  AATCGGAATAACTGGGCGTAAAGGGCACGCAGGCGGTTATTTAAGTGAGGTGTGAAAGCC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  552  AATCGGAATAACTGGGCGTAAAGGGCACGCAGGCGGTTATTTAAGTGAGGTGTGAAAGCC  611

Query  181  CCGGGCTTAACCTGGGNATTGCATTTCAGACTGGGTAACTAGAGTACTTTAGGGAGGGGT  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  612  CCGGGCTTAACCTGGGNATTGCATTTCAGACTGGGTAACTAGAGTACTTTAGGGAGGGGT  671

Query  241  AGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAATACCGAAGGCGAAGGC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  672  AGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTGGAGGAATACCGAAGGCGAAGGC  731

Query  301  AGCCCCTTGGGAATGTACTGACGCTCATGTGCGAAAGCGTGGGGAGCAAACAGGATTAGA  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  732  AGCCCCTTGGGAATGTACTGACGCTCATGTGCGAAAGCGTGGGGAGCAAACAGGATTAGA  791

Query  361  TACCCTGGTAGTCCACGCTGTAAACGCTGTCGATTTGGGGGTTGGGGTTTAACTCTGGCA  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  792  TACCCTGGTAGTCCACGCTGTAAACGCTGTCGATTTGGGGGTTGGGGTTTAACTCTGGCA  851

Query  421  CCCGT  425
            |||||
Sbjct  852  CCCGT  856
```

Figure 02: Le résultat de la recherche par BLAST, montrant l'identification d'une souche de *Haemophilus influenza* par l'alignement de la séquence d'amplification par PCR du gène codant l'ARN 16S à celle dans la banque de données.

6-5- Alignement de séquences multiples

Un alignement de séquences multiples (MSA: multiple sequence alignment) est un alignement de séquences de trois ou plusieurs séquences biologiques, généralement des

protéines, l'ADN ou de l'ARN. Il permet de mettre en évidence les relations entre séquences que l'on ne peut visualiser en comparant les séquences deux à deux.

Ce type d'alignement est appliqué surtout pour :

- La caractérisation des régions (motifs, domaines) conservées des protéines.

- Prédiction des structures 2D ou 3D par comparaison avec des structures connues.

- Construction des arbres phylogénétiques des séquences homologues.

6-5-1- ClustalW / ClustalX "Cluster alignment"

Ce type de Clustal "Classique " est fondé sur l'utilisation d'un algorithme d'alignement progressif. Les séquences les plus similaires sont alignées en premier puis l'alignement progresse vers les séquences les plus distantes. C'est également un programme de construction d'arbres phylogénétiques.

6-5-1-1 Étapes de l'alignement multiples de type clustal

1) Alignement de toutes les séquences 2 à 2 et détermination des scores des alignements.

2) Construction d'une matrice de score (BLOSUM62) pour l'ensemble des séquences.

3) Construction d'un arbre guide à partir de la matrice traduisant les relations globales entre les séquences. Il exprime les relations entre les séquences dont la longueur des branches est égale (cladogramme). A distinguer du phylogramme dont la longueur des branches est proportionnelle au nombre de changements évolutifs (mutations).

4) Alignement progressif à partir de l'alignement des deux séquences les plus proches. Les séquences voisines sont alignées de proche en proche jusqu'à l'alignement multiple final.