Université de Jijel,

Faculté des Sciences de la Nature et de la Vie

Département des Sciences de l'Environnement et sciences Agronomiques

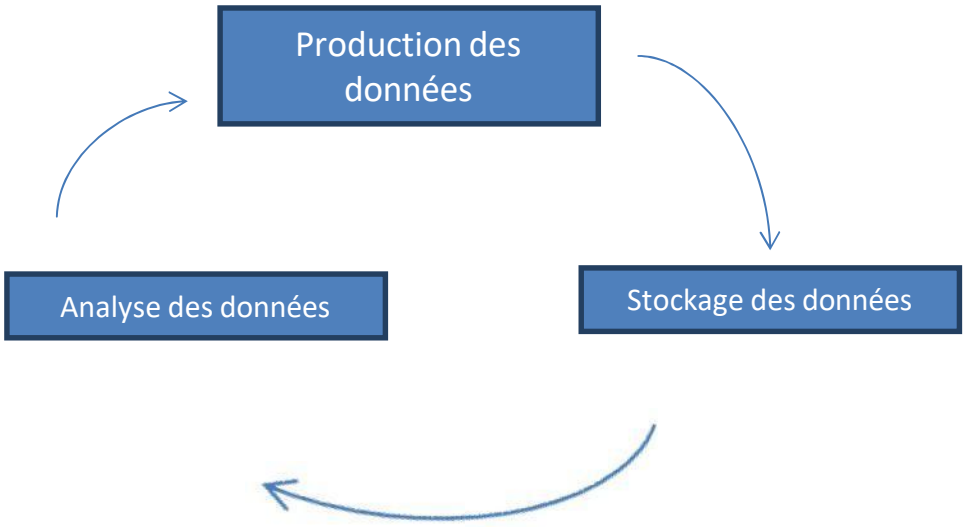
Matière : Bioinformatique

I- Introduction

Qu'est-ce que la bioinformatique ?
L'approche *in silico de* la biologie
Discipline relativement nouvelle, qui évolue en fonction des nouveaux problèmes posés par la biologie moléculaire. Un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie.
La bioinformatique est la discipline de l'analyse de l'information biologique, essentiellement sous la forme de séquences nucléotidiques, de séquences d'acides aminés et de structures de protéines. La bioinformatique est donc une branche théorique de la Biologie.

Trois activités principales :

- Acquisition et organisation des données biologiques
- Conception de logiciels pour l'analyse, la comparaison et la modélisation des données
- Analyse des résultats produits par les logiciels



Son but est

- d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories),
- d'énoncer des hypothèses généralisatrices (ex.: comment les protéines se replient ou comment les espèces évoluent ?),
- de formuler des prédictions (ex.: localiser ou prédire la fonction d'un gène).

II. Les bases de données

II.1. Définition

Une base de données (BD) est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise à jour, recherche et éventuellement analyse dans les systèmes les plus évolués).

Rôle des banques/bases de données

-Collecter les informations (séquences, cartographie physique, génétique…, données structurales, relationnelles…, - *auprès de*: biologistes, littératures, autres bases de données)
-Stocker et organiser
-Distribuer l'information

-Faciliter l'exploitation

Il existe un grand nombre de bases de données d'intérêt biologique. D'une façon générale, on distingue:

- les bases de données généralistes
- les bases de données spécialisées.

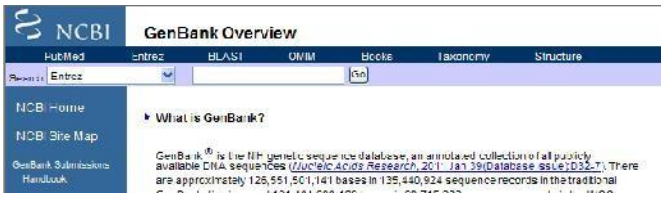
II.2. Bases généralistes

Les bases de données généralistes de séquences nucléotidiques et protéiques couvrent tous les secteurs de la biologie et toutes les espèces. Les grandes bases de séquences généralistes: Genbank, l'EMBL et DDBJ. Elles sont maintenant devenues indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique. Leur principale mission est de rendre publiques les séquences qui ont été déterminées, ainsi un des premiers intérêts de ces banques est la masse de séquences qu'elles contiennent.

Les qualités des BD généralistes:

- Enorme richesse de séquences en un seul ensemble ;
- Grande diversité d'organismes ;
- Nombreuses informations qui accompagnent les séquences (annotations, expertise, bibliographie);
- Présence de lien vers d'autres bases de données (spécialisées), soit nucléique, soit (encore mieux) protéique.

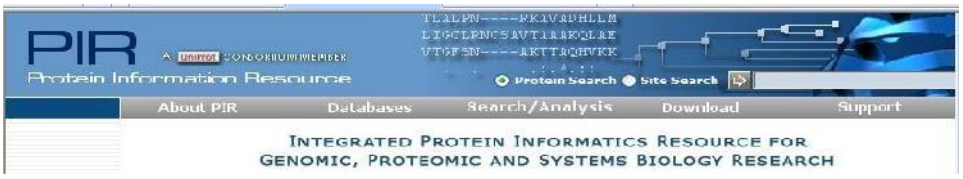
GenBank créée en 1982 au Los Alamos National Laboratory, est maintenue au NCBI (National Center for Biotechnology information), qui dépend du NIH (*National Institute of Health*) américain.



EMBL créée en 1980 par l'EMBL (European Molecular Biology Laboratory), est maintenue à l'EBI (European Bioinformatics Institute);

DDBJ (DNA Data Bank of Japan) maintenue par le Centre d'Information Biologique de l'Institut National de Génétique, créée en 1986.

PIR-NBRF : D'abord, elle fut créée par la NBRF (National Biomedical Researche Foundation) en 1984. Actuellement, elle constitut un ensemble dû à la fusion de MIPS (Martinsried Institute for Protein sequences, Munich Allemagne) et de JIPID Japan International Protein Information Database).



SwissProt : Créée par le biochimiste Amos BAIROCH en 1986 à l'Université de Genève. Actuellement développée en collaboration entre l'Institut Suisse de BioInformatique (ISB-SIB) et l'EBI. Elle contient la séquence de quasiment toutes les protéines découvertes jusque-là.

II.3. Bases spécialisées

Pour des **besoins spécifiques** liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques, de nombreuses bases de données spécifiques ont été créées au sein des laboratoires.
Elles ont pour but:
- de recenser des familles de séquences autour de caractéristiques biologiques précises comme les signaux de régulation, les promoteurs de gènes, les signatures peptidiques ou les gènes identiques issus d'espèces différentes.
- de regrouper des classes spécifiques de séquences comme les vecteurs de clonage, les enzymes de restriction, ou toutes les séquences d'un même génome.

Quelques exemples de bases de données spécialisées:
PDB, protein DataBank (PDB): Exemple: regroupe les molécules pour lesquelles on a obtenu les coordonnées 3D par résonance magnétique ou diffraction aux rayons X. Ces structures peuvent être facilement visualisées à l'aide de logiciels de visualisation 3D.

ECD, base sur les séquences nucléiques d'*Escherichia coli*.
NRL3D, base de séquences protéiques dont la structure tridimensionnelle a été déterminée. TFD, base de facteurs de transcription.
Prosite, bases de motifs protéiques. Elle peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.
CATH, base sur les classifications hiérarchiques (ordonnées) des structures protéiques.
IMGT, base de séquences des immunoglobulines et des récepteurs T.
GENATLAS, base d'informations issues de la cartographie des gènes humains.
KEGG, bases de voies métaboliques.

Les bases de motifs
On sait que certains segments d'ADN ou de protéines sont déterminants dans l'analyse des séquences car ils correspondent à des sites précis d'activité biologique comme par exemple les éléments de régulation des gènes ou les signatures peptidiques. C'est pourquoi des bases spécialisées se sont naturellement constituées autour de ces séquences. L'utilisation des bases spécialisées comme les bases de motifs, est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée.

- TFD ou IMD : sont employées pour des séquences promotrices des gènes
- Prosite ou BLOCKS : sont utilisées pour des protéines inconnues ou bien des séquences protéiques traduites à partir de cDNA ou de séquences génomiques.

Pour détecter une fonctionnalité sur une séquence, il suffit d'exécuter un programme qui s'appliquera à repérer la présence de certains motifs recensés dans ces bases et ainsi à prédire l'appartenance de la séquence testée à un groupe de séquences ayant une signature commune.

II.4. Bases de données Bibliographiques
Les bases de données bibliographiques répertorient toute catégorie d'objets bibliographiques : livres, journaux scientifiques, articles …

Ex : PubMed est une base de données bibliographiques en sciences biologiques et sciences biomédicales dont la couverture débute en 1946 et qui contient plus de 30 millions de références. En plus des articles indexés dans MEDLINE, PubMed contient aussi des références additionnelles, incluant les articles en accès libre de PubMed Central et les livres du NCBI. Il a été développé par le (NCBI), et est hébergé par la Bibliothèque nationale de médecine (NLM) américaine du National Institutes of Health

II.5-Organisation de l'information
Ces bases offrent des fiches descriptives de séquences nucléiques ou protéiques (AND, ADNc, ARN, protéines) ; ces fiches sont appelées des entrées.
Une "entrée" (entry en anglais) contient principalement : Une séquence, qui l'a déposée dans la base, La date de dépôt, voire de mise à jour, des informations sur l'organisme qui contient cette séquence, des informations sur le génome (par exemple nucléaire ou mitochondrial).
Eventuellement (souvent) : une ou des références à des articles scientifiques, la description de la composition de la séquences (les annotations : ou se situe le gène, la CDS, les introns, exons...), des hyperliens vers d'autres bases de données ou séquences.

Numéros d'Accession : Pour identifier ces séquences, les différentes banques de données leur assignent des Numéros d'Accession (Accession number) uniques au sein de leurs collections respectives. Un numéro d'accession en bio-informatique est un identifiant unique donné à toute séquence d'ADN ou de protéine enregistrée dans une base de données. Lors de la soumission d'une séquence à une base de données primaire on lui attribue un numéro d'accession nouveau et unique. Cet identifiant sera utilisé pour citer la séquence dans les articles scientifiques ou sert à retrouver rapidement cette information car cet identifiant ne changera jamais.
Les numéros d'accession sont de la forme : anciennes séquences : U43555 ... nouvelles séquences: AP007209

Exemple d'une entrée dans la base de donné genbank

Le numéro de version: chaque entrée possède un numéro de version. Par exemple lors de sa première soumission une séquence sera attribuée le numéro d'accession AB334763 et le numéro de version 1. Si bien que son identifiant complet sera: AB334763.1
Si on change quelque chose dans l'entrée, le numéro de version deviendra 2, et donc l'identifiant complet sera AB334763.2 etc...

II.6- Utilisation des bases de données

Avant 1990 : envoi postal de bandes magnétiques puis de CD-Rom aux laboratoires abonnés. A partir de 1990: Utilisation des réseaux informatiques (Internet) avec mise à disposition gratuite et mise à jour quotidienne des données. Depuis le début des années 90, avec l'installation massive des réseaux informatiques (Internet) à hauts débits, beaucoup de laboratoires consultent les bases de données via ces réseaux à partir de serveurs publics. Ces réseaux informatiques rapides et les services qui en découlent permettent une large diffusion des bases. Beaucoup de serveurs mettent gratuitement à disposition de nombreuses bases (pour la plupart des banques), dont les grandes banques de séquences généralistes comme l'EMBL ou Swissprot avec une mise à jour quotidienne des données, mais également un grand nombre d'autres bases dont la diffusion était auparavant plus restreinte.
L'utilisation de l'Internet pour la recherche de l'information biologique est d'actualité.

Il existe des règles spécifiques pour la formulation des requêtes d'interrogation de bases de séquences biologiques, en particulier: le nom d'espèce à laquelle appartiennent une séquence et les mots-clés à utiliser pour décrire les séquences recherchées

Interrogation des bases de données

On peut interroger une BD pour plusieurs raisons :
• Pour connaître la séquence d'un gène ou d'une portion de ce gène
• Pour connaître la structure primaire d'une protéine
• Pour comparer deux séquences, …

Le résultat de l'interrogation des BD est une fiche descriptive de la molécule. On parlera alors d'une **entrée** (ou fiche descriptive de la séquence recherchée). La structure d'une entrée est presque la même quelque soit la BD interrogée.

Systèmes d'interrogation des bases de données

Toutes les banques de données possèdent leurs systèmes (ou outils, ou logiciels) d'interrogation.
Chaque banque de séquences à son propre système d'interrogation, avec quelquefois des versions différentes proposées par certains serveurs. Des outils d'interrogation qui permettent des interrogations dans de nombreuses banques de séquences, généralistes ou spécialisées, ont été développés, les plus connus et utilisés sont : **SRS (Sequence Retrieval System), ENTREZ, ACNUC, DBGET**

Bases de données bibliographiques

- Ce sont des bases de connaissances scientifiques ou on trouve seulement les articles et revues scientifiques.
- Il s agit d'un ensemble structuré de données sous formes de références accessible au moyen d'un logiciel
- La consultation dune seul base bibliographique n'est en un aucun cas suffisante pour faire une bibliographie exhaustive. Il faut faire plusieurs recherches dans les différentes bases bibliographiques pour obtenir un résulta complet
- Il existe de nombreuses bases de données bibliographiques , certains sont gratuits et d'autres sont payants

Les bases de données sont des produits documentaires qui rassemblent :

- Soit des documents immédiatement utilisables (articles , photos) , dans ce cas on parle d'information primaire : consultation direct de l'article)
- Soit des informations sur ces articles (auteurs, résumé, titre..) l'information est alors qualifiée d'information secondaire

Pour assurer la diffusion du journal , il faut l'indexer dans les bases de données

Un journal indexé dans pubmed prend les résumé de chaque article , résumé , titre, volume , page ,auteurs…

Les bases de données proposent de plus en plus fréquemment des liens entre la notice et l'article et le texte intégral

Les services offerts :

- Faciliter la recherche
- Générer une liste d'articles sur le sujet
- Déterminer la recherche par date , type de document(pdf, word, html) , ou encore par langue
- Rechercher tout les articles publiés dans une revue
- Identifier une référence dont on possède pas touts les éléments pour la localiser (par exemple :le titre de la revue , l'auteur , la date…)
- Ouvrir un compte dans la base de données et sélectionner la discipline et recevoir directement dans la boite de réception électronique les références dés leur parution .

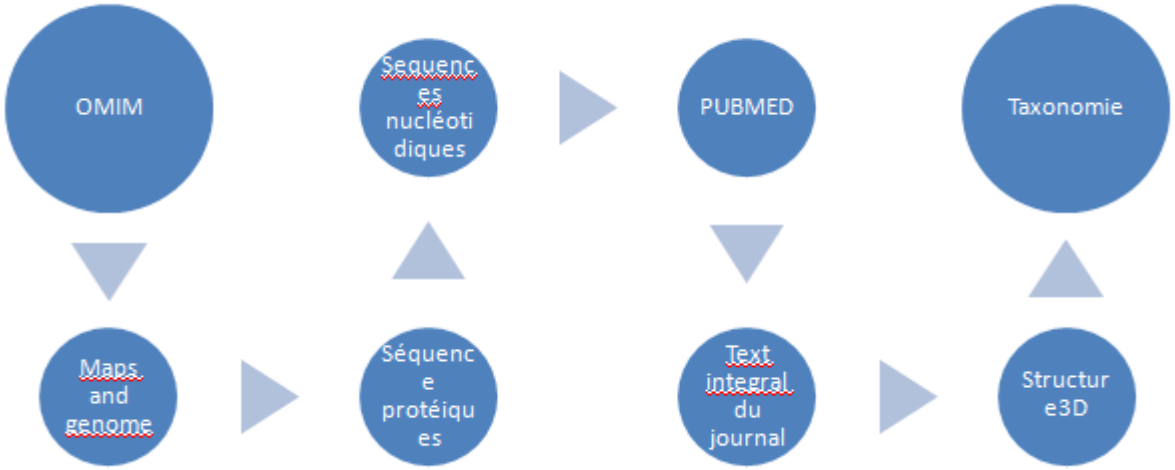
Springer: est un éditeur de journales scientifiques

Current Microbiologie est un journal publié par Springer

Scopus : base de données pour les sciences appliquées (sciences biologiques, physiques , chimiques…)

PubMed: base centré dans le domaine des sciences biologiques , il contient des liens ver les textes intégrales sur les sites des différents éditeurs comme Springer

Les liens croisées



Sur la base PubMed

Les résultats sont classés dans des pages qui contiennent 20 résultats

Les services:

- **Journals databases**
- **MeSH database**
- **Single citation matcher**
- **Clinical Queries**
- **Linkout**
- **My NCBI**

Journals databases : Base de données périodique qui donne des informations sur la revue scientifique (la fréquence de la revue , la période …

Interrogée par le titre complet , titre abrégé , numéro d'identification ISSN

MeSH database : Contient le vocabulaire scientifique (comme un dictionnaire) , donne une définition et suggestion pour chaque terme scientifique

- **Single Citation matcher** : est un système de recherche dune seul référence , permet de trouver une référence complète dont vous connaissez que quelques éléments: auteur , année….

Clinical Queries : Conçu pour les cliniciens, donne des information pour les médecins , et permet de filtrer la recherche . Diagnostic , thérapie , éthologie et maladies

Linkout : Conçu surtout pour les éditeurs proposer pour Pub Med , le contact avec l'éditeur se fait via pub Med

My NCBI : Permet de stocker des stratégies de recherche

Pour cela il faut créer un compte

OMIM Online :

- Donne des informations sur les maladies héréditaires
- Rescence les maladies génétiques
- Donne des références bibliographiques

- Les mutations
- Les maladies liées a la mutation
- La carte cytogénique

III. Alignement

Introduction

En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines…) repose essentiellement sur la notion de l'alignement1 , 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 œuvre de procédures de calcul (algorithmes) et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences.

Alignement : opération de base en bioinformatique qui a pour but d'identifier des zones conservées entre séquences:

- **Identifier** des sites fonctionnels
- **Prédire** la ou les fonctions d'une protéine
- **Prédire** la structure secondaire (voir tertiaire ou quaternaire) d'une protéine
- **Établir** une phylogénie (évolution : parenté entre les organismes).

Les étapes pour faire l'identification d'une espèce

- Extraction de l ADN
- Amplification par PCR du gène 16 S , on utilise des amorces pour déterminer le géne , les amorces sont universelles
- Purification de l ADN
- Vérification par électrophorèse si l ADN est périmé
- Le produit de PCR doit être 1,5 Kb
- On fait le séquençage dans les deux sens
- Comparaison des séquences sur Genbank avec le programme BLAST

- **Alignement :** processus par lequel deux (ou n) séquences sont comparées afin d'obtenir le plus de correspondances (identités ou substitutions conservatives) possibles entre les lettres qui les composent.
- **Alignement local :** alignement des séquences sur une partie de leur longueur
- **Alignement global :** alignement des séquences sur toute leur longueur
- **Alignement optimal :** alignement des séquences qui produit le plus haut score possible
- **Alignement multiple :** alignement global de trois séquences ou plus

- Brèches ou "gap" : espace artificiel introduit dans une séquence pour contre-balancer et matérialiser une insertion dans une autre séquence. Il permet d'optimiser l'alignement entre les séquences
- indel :
- "in" = insertion "del" = délétion

- **Similarité** : c'est le pourcentage d'identités et/ou de substitutions conservatives entre des séquences**.**
- **Le degré de similarité est quantifié par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquences.**
- **Homologie** : 2 séquences sont homologues si elles ont un ancêtre commun.
- **mésappariement : non correspondance entre deux lettres.**
- **Un mésappariement peut être** : soit la substitution d'un caractère par un autre, c'est-à-dire une mutation soi l'introduction d'un "gap"
- **Score** : un score global permet de quantifier l'homologie. Il résulte de la somme des scores élémentaires calculés sur chacune des positions en vis à vis des deux séquences dans leur appariement optimal. C'est le nombre total de "bons appariements" pénalisé par le nombre de mésappariements

- **Motif :** petite séquence commune entre les séquence de la même famille
  On obtient ses motifs par la comparaison (l'alignement entre les séquences)

Le rôle de la comparaison (alignement)
• Trouver les motifs de la même famille
• Faire l'arbre phylogénétique
• Identifier les espèces

• **Séquence similaire :** il y a un degré de similitude
• **Séquence identique:** l'ordre et les bases sont les mêmes pour les deux séquences
• **Identité :** Estimation de la fraction de résidus identiques entre deux **séquences**
• **CATGCATA**
GATGCATT

Les fractions de résidus sont considérées similaires lorsque leur score de substitution est supérieur à 0

TRAITEMENT DES SEQUENCES NUCLEIQUES (ADN ou ARN)

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)… Au 10ème point de comparaison, les deux séquences contiennent le même nucléotide T donc un score élémentaire de 1. 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 est de la forme :

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

Il faut savoir qu'il existe une matrice (matrice d'identité) qui 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

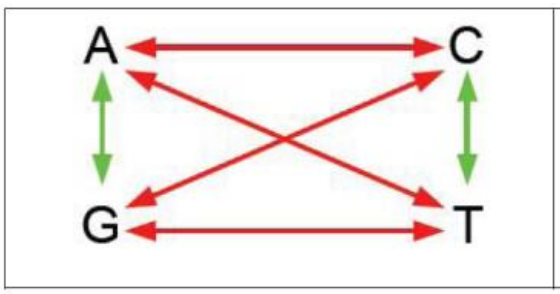
Matrice d'identité nucléique

- Une mutation de délétion peu éliminer 1 ou plusieurs nucléotides .
- L'ouverture de la brèche = élimination dune nucléotide
- Louverture de la bréche est plus couteuse que l'extension
- Exemple :deux alignement :
- Le premier 5 ouvertures , chaque ouverture déléte 1 nucléotide ⟶ 5mutations
- Le deuxième 1 ouverture et délétion de 5 nucléotides ⟶
  1 mutation

  Chaque identité donne 1
  Louverture =mutation donne -10
  Extension est moins couteuse , elle donne -1
  CGATGC AGCAGC AGCATCG
  CGATGC_ _ _ _ _ _AGCATCG
  OUVERTURE EXTENTION
  13(1) + (-10) +6(-1) = -3

  CGATGCAGCAGCAGCATCG
  CG_TG_AGCA_CA_ _ AT_G
  OUVERTURE EXTENTION
  Score : 13(1) +5(-10) +6(-1) =-43
  Le premier alignement est mieux que le deuxième car le score d'alignement du premier est supérieur au deuxième score

- 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 :



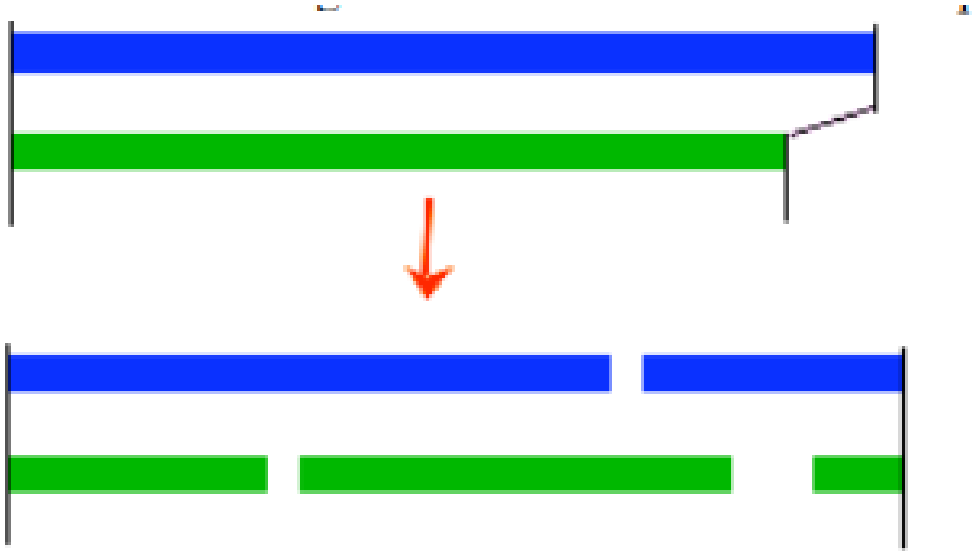
	A	T	G	C
A	3	0	1	0
T	0	3	0	1
G	1	0	3	0
C	0	1	0	3

Matrice de transition/transversion

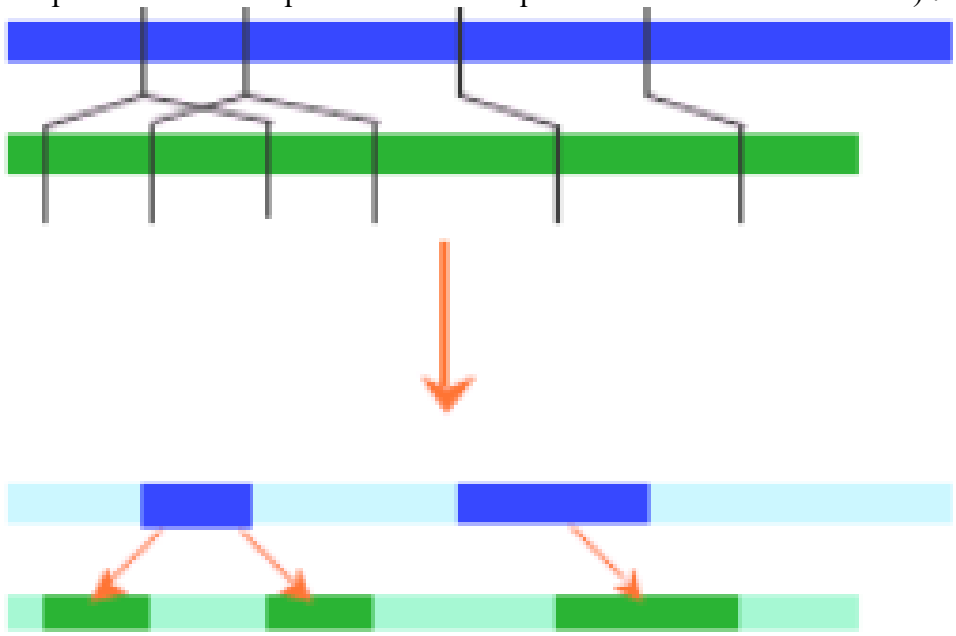
Alignement de deux séquences :

- Les principales méthodes •
- Needleman-Wunch : allignement globale utilisé pour aligner deux séquences
    - *Smith-Waterman* : L'algorithme de Smith et Waterman (1981)
    - Allignement local

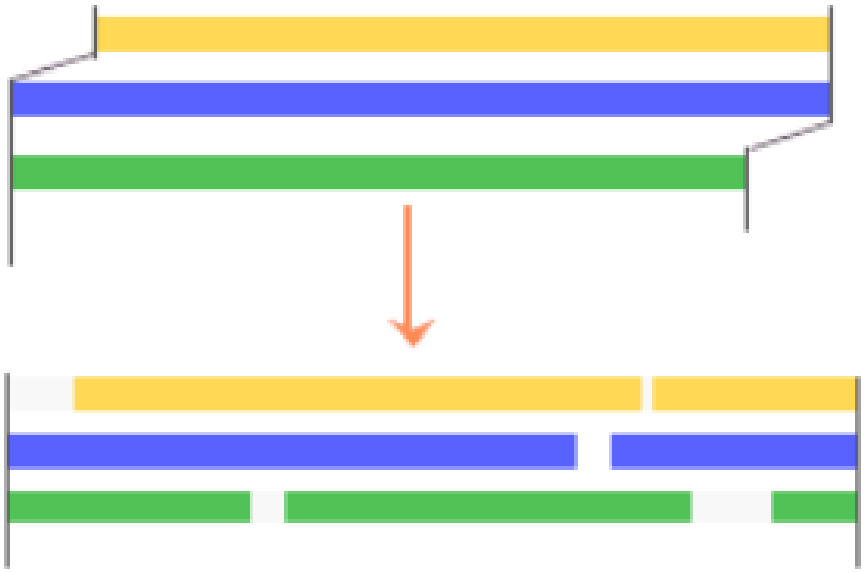
- **Alignement global** : Alignement de deux séquences sur la totalité de leur longueur en tenant compte de tous les résidus. Si les longueurs sont différentes, des insertions / délétions sont introduites pour aligner les deux extrémités des deux séquences. Cet alignement permet de mesurer le degré de similitude entre deux séquences



- **Alignement local:** Alignement de deux séquences sur des régions isolées et permettant de trouver des segments qui ont un haut degré de similitude. Utilisé pour la recherche dans les bases de données (comparaison d'une séquence avec les séquences contenues dans la base) .



- **Alignement Multiple**: Alignement portant sur plusieurs séquences à la fois et dans leur intégralité .Il permet de mettre en évidence les relations entre séquences que l'on ne peut pas visualiser en comparant les séquences deux à deux.



À chaque type d'alignement est associé un programme informatique permettant d'optimiser le traitement

Alignement global :	Alignement local:	Alignement Multiple:
Needle Stretcher	BLAST (Basic Local Alignment Tool) FASTA	T-Coffee

La matrice de points (Dot plot)

- 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 diagonole principal. 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.

		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

- **Matrices protéiques :** Notons tout d'abord que les matrices protéiques utilisées pour réaliser des alignements sont totalement différentes de celles des acides nucléiques (matrice d'identité et matrice de transition/transversion) et ce en raison du nombre des acides aminés (20 acides aminés et non 4 comme le cas des nucléotides) et de la nature physico-chimiques de ceux-ci.
- En effet, le système nucléique basé sur l'identité n'est pas approprié pour le cas des systèmes protéiques. Ceci est du au fait que certains acides aminés peuvent être remplacés par d'autres (à cause de leurs propriétés physicochimiques surtout) sans altérer le rôle et la fonction biologique de la protéine.

- On peut donc classer les acides aminés en familles par rapport à leurs propriétés et obtenir ainsi un système de scores qui rende compte de l'affinité des résidus protéiques entre eux. C'est cette affinité qui permet à un acide aminé d'être substitué par un autre, et les deux structures protéiques ne seront pas identiques à ce point où la substitution a eu lieu mais on dira que les deux séquences sont SIMILAIRES et la fonction de la protéine reste conservée

- Les matrices protéiques peuvent être classées en deux catégories :

❖ Une catégorie qui regroupe les matrices issues d'études montrant le caractère de substitution des acides aminées au cours de l'évolution (matrices liées à l'évolution). Elles représentent les échanges possibles et acceptables d'un acide aminé par un autre lors de l'évolution des protéines.

❖ La deuxième est basée plus particulièrement sur les caractéristiques physicochimiques des acides aminés : caractère hydrophile ou hydrophobe des protéines, la structure secondaire ou tertiaire des protéines. Ce sont les matrices liées à l'évolution qui seront utilisées pour réaliser les alignements des séquences protéiques.

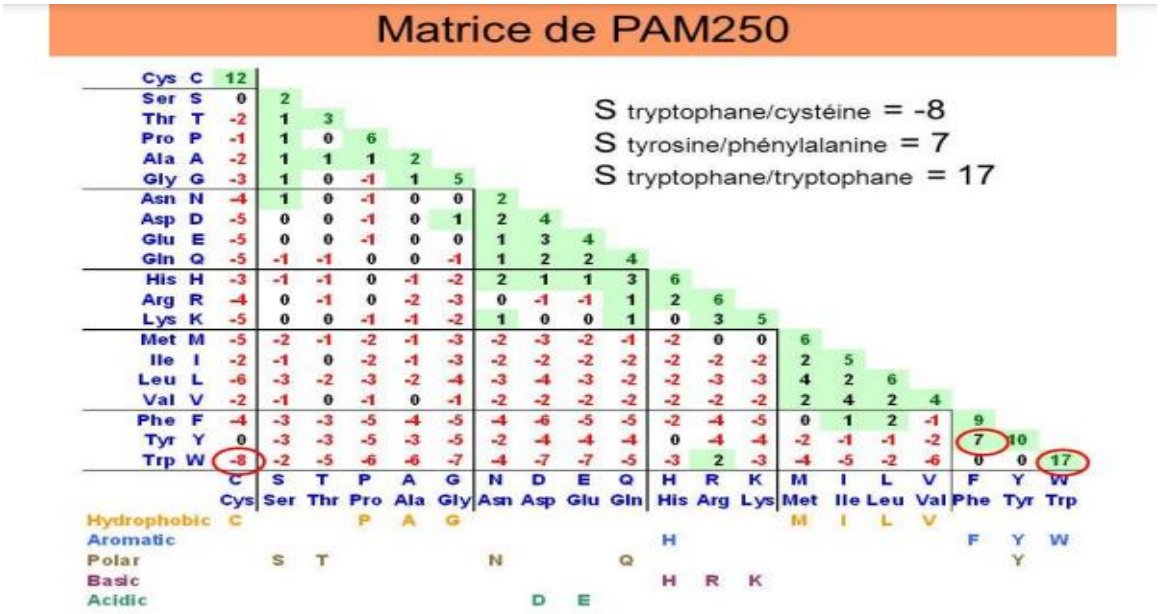
La matrice PAM250 (Purcent Accepted Mutation): La matrice de mutation de Dayhoff.

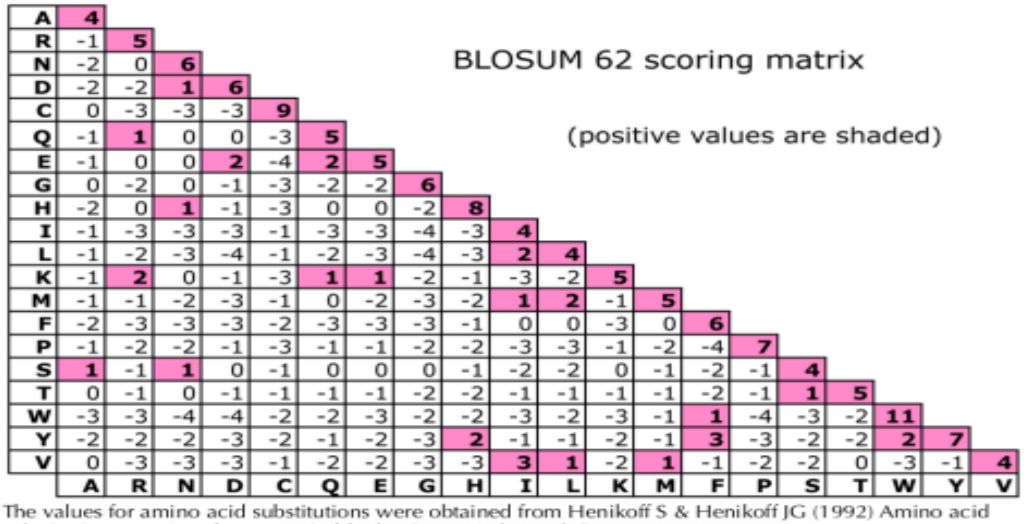
Les matrices de type PAM dérivent d'alignements globaux de protéines très semblables et représentent les échanges possibles et acceptables d'un acide aminé par un autre au cours de l'évolution des protéines : Les acides aminés entrant dans la composition d'une protéine peuvent avoir les mêmes propriétés physicochimiques ou presque et la structure 3d va donc dépendre de ces caractéristiques. Cette similarité des propriétés physico-chimiques est donc suffisante pour permettre la substitution (la mutation) entre ces acides aminés sans pour autant perturber la fonction de la protéine .

Les matrices BLOSUM ("BLOcks SUbstitution Matrix") :

Les matrices BLOSUM sont construites à partir de 2000 BLOCKS provenant de plus de 500 familles de protéines. Le degré de substitution des acides aminés a été mesuré en observant des blocs d'acides aminés issus de protéines plus éloignées. Chaque bloc est obtenu par l'alignement multiple sans insertion/délétion de courtes régions très conservées. Ces blocs sont utilisés pour regrouper tous les segments de séquences ayant un pourcentage d'identité minimum au sein de leur bloc. On en déduit des fréquences de substitution pour chaque paire d'acides aminés et l'on calcule ensuite une matrice logarithmique de probabilité dénommée BLOSUM

A chaque pourcentage d'identité correspond une matrice particulière. Ainsi la matrice BLOSUM50 est obtenue en utilisant un seuil d'identité de 50%.





The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

Programmes de comparaison avec les banques :

- Recherches de similitudes dans les banques de séquences, Pourquoi ?
    - Savoir si ma séquence ressemble à d'autres déjà connues
    - Trouver toutes les séquences d'une même famille
    - Rechercher toutes les séquences qui contiennent un motif donné

- La taille sans cesse croissante des banques de séquences a nécessité **l'élaboration d'algorithmes spécifiques** pour effectuer **la comparaison d'une séquence avec une banque de données** car les algorithmes standards de comparaison entre deux séquences sont généralement trop longs sur des machines classiques.

- La plupart de ces programmes constituent des méthodes heuristiques. Leur but est de filtrer les données de la banque en étapes successives car peu de séquences vont avoir des similitudes avec la séquence comparée. Ces méthodes heuristiques utilisent certaines 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. Ces programmes permettent de calculer un **score** pour mettre en évidence les **meilleures similitudes locales** qu'ils ont observées.

- Les deux types de programme les plus utilisés par les biologistes qui sont les logiciels :

- **FASTA**
- Le logiciel regroupe en fait plusieurs programmes de recherche avec les banques de données:
    - Le programme FASTA qui compare respectivement une séquence nucléique avec une base nucléique ou une séquence protéique avec une base protéique.
    - Les programmes TFASTA ou TFASTX qui comparent une séquence protéique avec des bases nucléiques traduites.
    - Les programmes FASTX ou FASTY qui comparent une séquence nucléique traduite avec des bases protéiques.

- **BLAST: Basic Local Alignment Search Tool**
- BLAST est l'abréviation de « Basic Local Alignment Search Tool » ou, en français, L'outil de recherche basique d'alignement local. 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.
- Ce logiciel possède en fait plusieurs programmes de comparaison avec les bases de données :
- **BLASTN** (pour comparer une séquence nucléique contre base nucléique),
- **BLASTP** (Pour comparer une séquence protéique contre base protéique),
- **BLASTX** (comparaison de séquence nucléique (traduite en 6 phases) contre base protéique),
- **TBLASTN** (comparaison de séquence protéique contre base nucléique (traduite en 6 phases)),
- **TBLASTX** (comparaison de séquence nucléique (traduite dans les 6 phases) contre base nucléique (traduite dans les 6 phases)).

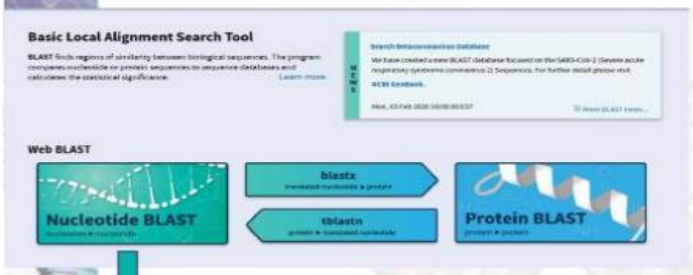
BLAST (Basic Local Alignment Search Tool) :

. Programme pour la recherche de similarités dans les bases de données
. Utilise un algorithme heuristique linéaire pour l'alignement local
. Séquences nucléiques et protéiques
. Disponible sur le Web
. Connecte aux principales banques de données

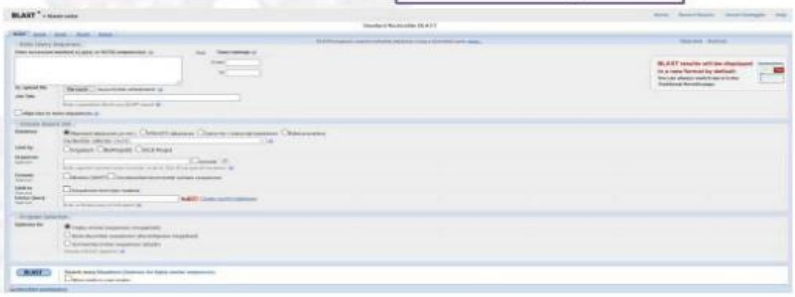
FASTA (FAST All):

L'algorithme est basé sur l'identification rapide des zones d'identité entre la séquence recherchée et les séquences de la banque de données. Cette reconnaissance est essentielle car elle permet de considérer uniquement les **séquences présentant une région de forte similitude** avec la séquence recherchée.

https://blast.ncbi.nlm.nih.gov/Blast.cgi

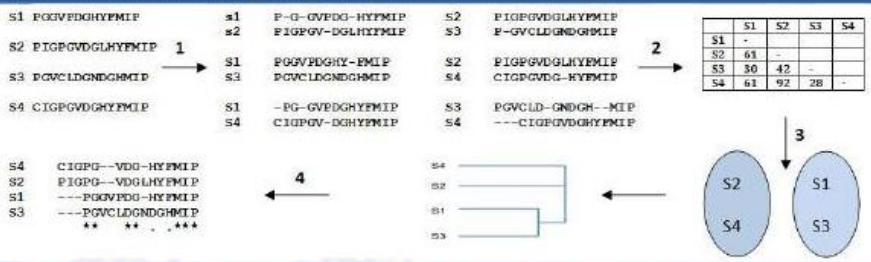


Interface de l'algorithme



CLUSTALW (Cluster Alignment). (Thompson, Higgins et Gibson, 1995)

Principe : CLUSTALW 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'arbre phylogénétique.

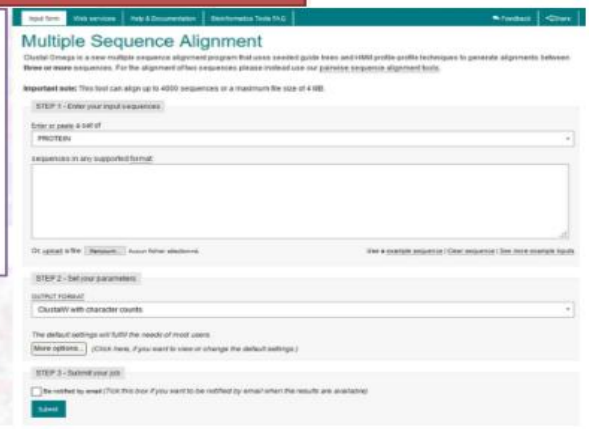


Étape schématique de l'alignement multiple avec CLUSTALW avec 4 séquences d'acides aminés :
1-Alignement de toutes les séquences 2 à 2 et détermination des scores des alignements
2-Construction d'une matrice de score (BLOSUM62) pour l'ensemble des séquences
3-construction d'un arbre guide à partir de la matrice traduisant les relations globales entre les séquences
4-Alignement progressif à partir de l'alignement des 2 séquences les plus proches. les séquences voisines sont alignées de proche en proche jusqu'à l'alignement multiple final.
Légende : « * » : résidus conservés « . » substitution conservatives

CLUSTAL sur internet

sur internet vous trouvez la version recente CLUSTAL Omiga
Comme montré ci-dessous suivant ce lien :
https://www.ebi.ac.uk/Tools/msa/clustalo/

Il est aussi intégré dans plusieurs logiciels d'analyse de séquences comme MEGA X, Bioiedit, UGENE packages...



T-COFFEE (Noterdame, Higgins, Hering, 2000)

T-COFFEE est fondé également sur un alignement progressif. En plus de réaliser un alignement global entre chacune des paires de séquences, il procède à un alignement local afin d'optimiser l'alignement entre les séquences très divergentes.

VI. Les applications de la bioinformatique

La bioinformatique, véritable clé de la génomique, en est un outil indispensable: à partir d'une séquence d'ADN nouvellement identifiée, elle permet de retrouver les séquences similaires déjà décrites dans des banques de données, déduire quels sont les gènes associés et leur distribution au niveau d'un organe ou d'un tissu, établir un lien entre des gènes présents dans une pathologie, prédire la structure, et même la fonction de cette protéine, cible potentielle pour un futur médicament. On peut classer ces applications en différentes catégories compte-tenu de la diversité des domaines d'action de la bioinformatique.

- l'analyse de séquences qui peut aller de l'identification de gènes aux comparaisons de séquences en passant par la prédiction de motifs ou l'établissement de signatures.
- La structure des protéines nécessite l'usage de la bioinformatique, que ce soit pour la visualisation ou la prédiction de leur repliement.
- En phylogénie de façon à comparer les espèces à l'échelle moléculaire et obtenir ainsi un classement évolutif qui soit plus fiable.
- Des liaisons génétiques peuvent êtres établies grâce à la bioinformatique pour permettre de détecter des gènes candidats de maladies génétiques par exemple.
- l'utilisation de la bioinformatique en génomique fonctionnelle permet de travailler sur le transcriptome, le protéome, l'intéractome...

VI. 1. Construction d'arbres phylogénétiques

-La phylogénie correspond à l'étude des relations d'évolution entre des groupes d'organismes (espèces, populations). Ce qui permet retracer les principales étapes de l'évolution des organismes depuis un ancêtre commun et ainsi de classifier plus précisément les relations de parentés entre les êtres vivants. -La taxonomie c'est la discipline qui consiste à classifier, identifier et nommer les organismes. Ceci est Basé sur les caractéristiques communes des espèces.

Les données de la phylogénie :

Correspondent aux caractères observables (aux différents états : morphologiques, biochimiques et physiologiques) patterns binaires (de type présence d'un caractère donné / absence de ce même caractère) . Exemple:

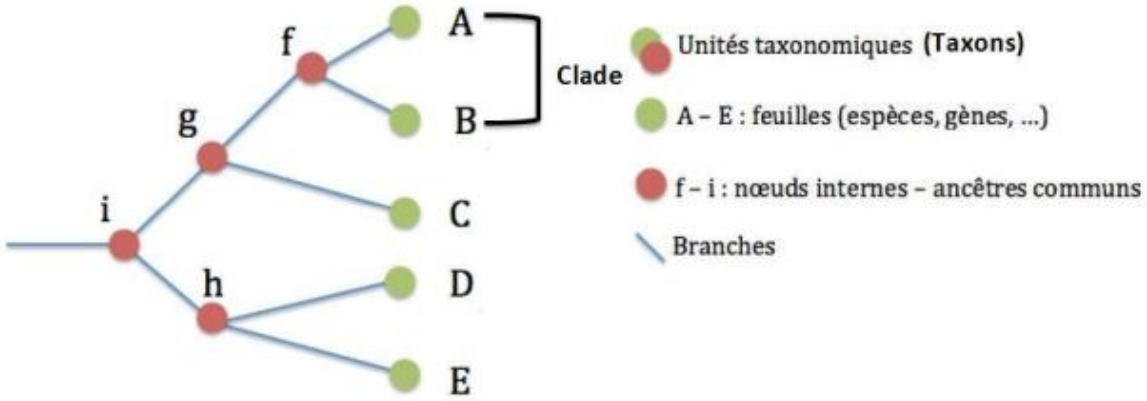
matrice de caractères

	1. Sacs aériens	2. Appendices pairs	3. Mandibule	4. Glandes mammaires	5. Ailes	6. Dents	7. Colonne vertébrale
Truite	0	0	0	0	0	0	0
Chauve-souris	1	1	1	1	1	0	0
Homme	1	1	1	1	0	0	0
Pigeon	1	1	0	0	1	1	0

- La construction d'arbres phylogénétiques est utilisée par les programmes d'alignements multiples de séquences afin d'éliminer une grande partie des alignements possibles et de limiter ainsi les temps de calcul.

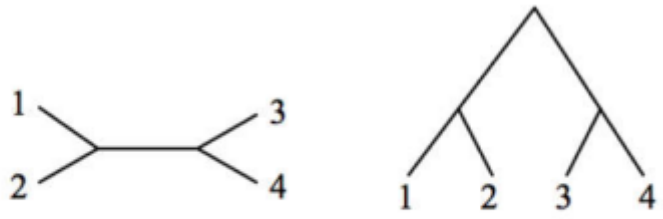
. Arbre de Phylogénie • Premier objectif des études phylogénétiques: Reconstruire l'arbre de vie de toutes les espèces vivantes à partir des données génétiques observées. • Un graphe connexe acyclique; Ensemble de nœuds (ou sommets) connectés par des arêtes (ou branches) branches) de telle sorte que toute paire de nœuds est reliée par exactement un chemin.

Structure d'un arbre phylogénétique



Arbre enraciné et Arbre non-enraciné

Dans l'arbre enraciné, la racine représente l'ancêtre commun le plus récent de tous les taxons considérés. Un arbre enraciné est donc dirigé et prend un « chemin évolutif » de l'ancêtre commun aux taxons actuels.
Un arbre non-enraciné ne représente que les relations entre les taxons.



(a) Arbre non raciné (b) Arbre raciné

Les molécules utilisées

1. Les séquences des gènes des d'ARN ribosomiques (ARNr).
2. Les séquences IGS et ITS des ARNr.
3. Les mêmes séquences dans les mitochondries.
4. Des régions hyper variables du génome mitochondrial.
5. Les séquences de cytochrome C
6. Les séquences de la ribulose 1,5-bisphosphate carboxylase
7. Les séquences du facteur d'élongation alpha (tuf).
8. ……

VI. 2. Prédiction et modélisation moléculaire

VI.2.1. Prédiction de structures secondaires

Les principales caractéristiques physico-chimiques d'un acide aminé sont le volume, l'accessibilité au solvant, l'hydrophobicité, la taille (petit, gros), et l'aromaticité. Le profil hydrophylique (ou hydropathique)

d'une séquence protéique permet d'obtenir quelques indices concernant la structure secondaire de la protéine correspondante. Par exemple, les acides aminés hydrophobiques tendent à se trouver à l'intérieur des protéines globulaires. Dans les protéines transmembranaires, ceux-ci ont tendance à se trouver au niveau de la membrane.

Les méthodes de prédiction de structures secondaires ont pour objectif principal de définir la localisation des éléments d'hélice α, de brin β, les coudes et les structures apériodiques. La structure secondaire permet de faire des hypothèses structurales/fonctionnelles en l'absence de données expérimentales. Vouloir prédire la structure à partir de la séquence suppose l'hypothèse vérifiée que l'information nécessaire à l'acquisition de la structure 3D biologiquement active est contenue dans la séquence. Bien que la structure 3D finale soit thermodynamiquement stable, le mécanisme par lequel la chaîne principale se replie est inconnu. Le temps pour que la protéine "retrouve" sa conformation thermodynamiquement stable varie de quelques minutes à plusieurs jours.

Différentes méthodes et approches ont été utilisées pour prédire la structure secondaire. Les méthodes
statistiques, les méthodes de **similarité** est les méthodes
neuronales. La méthode de chou et Fasman (1974-1978)
La méthode de GarNIER Osguthorpe et Robson I (1978)
La méthode de plus proches voisins (Nearest Neighbor method)

La modélisation moléculaire

Les molécules de par leurs dimensions sont invisibles à tout moyen d'investigation direct tel que la microscopie. C'est par l'analyse de données indirectes que les chercheurs peuvent reconstituer un modèle moléculaire, c'est-à-dire une construction intellectuelle présentant la meilleure adéquation avec les résultats expérimentaux. Ces données sont issues principalement d'analyses cristallographiques (étude des figures de diffraction des rayons X par un cristal), ou de Résonance Magnétique Nucléaire. Elles représentent les contraintes expérimentales exercées sur le modèle. Le modèle moléculaire obtenu ensuite est un ensemble de coordonnées atomiques dans l'espace. L'informatique intervient dans toutes les étapes conduisant de l'expérimentation au modèle, puis ensuite dans l'analyse du modèle par la visualisation moléculaire. Elle est utilisée par exemple pour étudier les sites actifs d'une enzyme, mettre au point informatiquement une série d'inhibiteurs possibles pour cette enzyme, et ne synthétiser et ne tester que ceux qui semblent convenir. Cela permet de réduire les coûts de recherche et d'accélérer ces recherches.

La visualisation de la structure tridimensionnelle d'acides nucléiques (ARN et ADN) fait également partie de la palette des outils bio-informatiques très utilisés. Il y a aussi la *dynamique moléculaire*, on essaye de voir le comportement d'une molécule dans son milieu en modélisant les différents champs de force entrant en jeu (force de van der Waals, etc.).

Encore un autre aspect est la prédiction de la structure 3D d'une protéine à partir de sa structure primaire (la liste des acides aminés qui la composent), en modélisant les différentes caractéristiques

des acides aminés. Cela a un grand intérêt car la fonction, l'activité d'une protéine dépendent grandement de sa forme. De même, la modélisation des structures 3D d'acides nucléiques (à partir de leur séquence nucléotidiques) revêt la même importance que pour les protéines.

Un grand nombre de programmes existent: Rasmol, Rastop, Protéine Explorer……

V. La génomique

Pour étudier les génomes il faut d'abor les séquencer.

Séquençage: Le séquençage consiste à déterminer la nature et l'ordre des nucléotides dans un acide nucléique.

Séquençage de l'ADN *(méthode enzymatique selon Sanger)*

La méthode des didesoxyribonucléotides, inventée il y a une vingtaine d'année dans le laboratoire de Fred Sanger à Cambridge en Grande-Bretagne, est aujourd'hui universellement employée pour séquencer l'ADN. Elle repose sur l'allongement par l'ADN polymérase d'un brin à partir d'une amorce, en utilisant un autre brin d'ADN comme matrice. Cet allongement est réalisé en présence des quatre desoxyribonucléotides triphosphate (dATP, dTTP, dGTP, dCTP), monomères utilisés par la polymérase, et d'un analogue didesoxyribonucléotides (ddNTP) qui joue le rôle de terminateur de chaîne.

La très grande majorité des séquences réalisées et publiées aujourd'hui sont réalisées sur des séquenceurs automatiques. Ceux-ci sont capables de réaliser les réactions de séquence, puis de les lire.

Définition de la génomique

La génomique est l'étude des génomes, de leur organisation et de leur évolution, ainsi que de l'expression et de la fonction des gènes. Son objectif est de séquencer l'ADN d'un organisme et de localiser sur celui-ci tous les gènes qu'il porte, puis de caractériser leurs fonctions.

La génomique est la science qui étudie la structure, le contenu et l'évolution des génomes. Elle a pour but de déterminer le plus grand nombre de séquences nucléotidiques, d'analyser l'expression et la fonction des gènes. Les outils bioinformatiques ont été intégrés à ces recherches.

<u>Pourquoi étudier la Génomique?</u>

- Mettre en place des banques de données disponibles sur le Web,
- Établir des cartes du génome (positionner les gènes sur l'ADN),
- Annoter les gènes (trouver la séquence, la fonction et l'expression),
- Étudier les polymorphismes (variations entre individus pour un même locus dans l'ADN),
- Comparer des génomes d'espèces différentes entre eux (*Génomique comparative*) et faire des hypothèses sur l'évolution.

Les applications de la génomique

L'industrie utilise la génomique pour identifier

- Les gènes impliqués dans les pathologies: cibles pharmaceutiques ou marqueurs diagnostiques
- De nouveaux gènes permettant de synthétiser des molécules d'intérêt
- Des gènes responsables de résistances (microbiologie ou agronomie)
- Des gènes permettant de mieux comprendre des mécanismes-clés: cancer, vieillissement, etc.

Génomique fonctionnelle (" post-génomique "): Étude de la fonction des gènes par analyse de leur séquence et de leurs produits d'expression : les ARNm (transcriptome) et les protéines (protéome). Elle s'intéresse à leur mode de régulation, et à leurs interactions (cf. Réseau de régulation). L'analyse des protéines peut aller jusqu'à la détermination de leur structure tridimensionnelle.

Génomique comparative

Génomique comparative: comparaison de génomes entièrement séquencés

Les applications:

- Aide l'annotation en identifiant les régions fonctionnelles (les régions non fonctionnelles sont non conservées)
- Identifier le jeu de gènes de chaque organisme
- Comprendre les solutions trouvées par des organismes différents pour une même fonction
- Étudier des gènes/fonctions particuliers par comparaison de séquence (voir cours de bioinformatique)
- Autres questions spécifiques: adaptation, résistance, pathogénicité, etc.

Transcriptome: Ensemble des ARN messagers transcrits à partir du génome. Comme le protéome, il varie au cours du temps et d'une cellule à l'autre d'un organisme.

Métabolome: l'ensemble des composés organiques (sucres, lipides, amino-acides, …)

Protéome: Le protéome est l'équivalent protéique du génome, c'est à dire l'ensemble des protéines exprimé par le génome d'une espèce donnée. Comme le transcriptome, le protéome n'est pas identique dans toutes les cellules d'un organisme donné.

Protéomique: La protéomique est l'étude du protéome, dans le but de déterminer l'activité, la fonction et les interactions des protéines (cf. Réseau de régulation), et cela dans diverses conditions.