REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE des Sciences

DEPARTEMENT des sciences de la Nature et de la Vie

N°……..

DOMAINE : Sciences de la Nature et de la Vie

FILIERE : Biotechnologies

OPTION : Biotechnologie Végétale

Mémoire présenté pour l'obtention
Du diplôme de Master Académique

Par

MENICHE Anis / SAIT Massil / BARKA Achraf

Intitulé :

Outil Bio-informatique utiles en Biologie : Banque et Base de données

Soutenu le 6.6.2022 devant le jury composé de :

Dr. GHADBANE Mouloud	Pr.	Université de Msila	Président
Dr. BENDIF Hamdi	MCA	Université de Msila	Encadreur
Mr. HARRAR Abdenasseur	MAA	Université de Msila	Examinateur

Année universitaire : 2021 / 2022

REMERCIEMENTS

Nous remercions Dieu, le tout puissant de nous avoir donné la volonté et la patience nécessaire pour accomplir ce travail.

La réalisation d ce mémoire a été possible grâce au concours de plusieurs personnes à qui nous voudrions témoigner toute ma gratitude.

Nous voudrions dans un premier temps remercier, notre directeur de mémoire ***, Dr. BENDIF Hamdi****, Maitre de conférences à Université de M'Sila, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter nos réflexions. Nos remercîments vont également au membres jury ;* ***Pr. GHADBANE Mouloud*** *et Mr.* ***HARRAR Abdenasseur*** *pour accepter de juger notre modeste travail.*

Nous remercions également toute l'équipe pédagogique de l'université de M'Sila et les intervenants professionnels responsables de notre formation durant tout notre cursus universitaire.

Notre reconnaissance va également à

Nous chers parents pour leur soutien constant et leurs encouragements,

Anis / Massil / Achraf

Résumé

Avec le développement de la génétique et des nouvelles technologies à très haut débit, nous faisons actuellement face à la production de données à un niveau encore jamais atteint. En effet, il est aujourd'hui démontré que les données produites par les technologies de séquençage à haut débit seront plus importantes que tout ce qui a jamais été produit dans le passé y compris le web lui-même ! Nous faisons donc face à de multiples challenges tant pour le stockage de ces données (les nouvelles plateformes de séquençage peuvent produire jusqu'à 0,1 téraoctets de données par heure) que pour leur analyse. L'objectif de ce travail est donc de faire le point sur les apports de la bioinformatique notamment par les différentes bases de données et outils bioinformatiques qu'elle a permis de créer ces dernières années et qui sont aujourd'hui autant d'outils incontournables pour les généticiens.

Mots clé : Bioinformatique, Bases de données, Banque de données

ملخص

مع تطور علم الوراثة والتقنيات الجديدة عالية السرعة، نواجه حاليًا إنتاج البيانات بمستوى لم يتحقق من قبل. في الواقع، لقد ثبت الآن أن البيانات التي تنتجها تقنيات التسلسل عالية الإنتاجية ستكون أكبر من أي شيء تم إنتاجه في الماضي بما في ذلك الويب! لذلك نواجه تحديات متعددة لتخزين هذه البيانات (يمكن أن تنتج منصات التسلسل الجديدة ما يصل إلى 0.1 تيرابايت من البيانات في الساعة) ولتحليلها. لذلك فإن الهدف من هذا العمل هو تقييم مساهمات المعلوماتية الحيوية، لا سيما من خلال قواعد البيانات المختلفة وأدوات المعلوماتية الحيوية التي أتاح إنشاؤها في السنوات الأخيرة، التي أصبحت متاحة اليوم لعلماء الوراثة.

الكلمات المفتاحية: المعلوماتية الحيوية، قواعد البيانات، بنك البيانات

Sommaire

Liste des abréviations

ENSEMBL	European Bioinformatics Institute / Wellcome Trust Sanger Institute
ADN	Acide Désoxyribonucléique
ARN	Acide Ribonucléique
BED	"Browser Extensible Data", un format de fichier basé sur le texte
BS	BASE DE DONNEES
F ASTQ	Format basé sur le texte pour la représentation des séquences nucléotidiques et leurs
	Scores de qualité.
FASTA	Format basé sur le texte pour la représentation des séquences nucléotidiques.
NCBI	(National Cancer for Biology Information)
NGS	Séquençage de nouvelle génération ("Next Generation Sequencing")

Liste des figures

Liste des tableaux

Introduction

Introduction générale

« La création, c'est l'art de déborder le sujet ». Marie-Catherine Dupuy

À l'interface entre biologie, informatique et mathématique, la bio-informatique analyse et interprète au moyen de méthodes informatique, les données biologiques que sont les séquences de gène et des protéines cellulaires, et apporte ainsi de nouvelles connaissances sur le fonctionnement des cellules et des organismes vivants (**Rechenman, 2004)**.Dans ce mémoire, nous présentons une étude détaillée sur relation entre la biologie et la bio-informatique, et comment ça passe le traitement des quantités énormes de données générées par cette technologie qui nécessite des moyens informatiques puissants et efficaces. A la lumière de ce qui a été développé précédemment, notre travail de recherche est centré sur l'outil bio-informatique utiles en biologie : banque et base de données afin de répondre à la problématique suivante :

Face aux quantités énormes des données générées, quelles sont les outils de la bio-informatique auxquels la biologie fait recours ?

Pour mieux comprendre ce sujet, diverses interrogations découlent de notre problématique, à savoir :Comment se définit la bio-informatique ? Quelles-sont ses démarches ?

- Qu'est-ce que les acides nucléiques ?
- Comment se passe le stockage de la bio-informatique ?

Ce sont ces questions, qui vont orienter notre problématique et nous permettre de rédiger notre mémoire. Pour tenter d'avoir les éléments de réponses à notre problématique, notre présent travail s'est principalement basé sur une démarche de recherche bibliographique, à travers des lectures d'ouvrages, des rapports de stage et des mémoires.

Ce document se compose de deux chapitres principaux, une conclusion. Le premier chapitre intitulé « bibliographie » décrit les principes de base de bio-informatique, suivi d'une étude détaillée des acides nucléiques et les protéines. Ensuite, le deuxième chapitre qui présente le stockage de la bio-informatique. La description sera subdivisée en deux grandes parties. La première décrira les banques de données. La deuxième partie, quant à lui, concernera la description des bases de données.La section conclusion présente une synthèse basée sur les résultats obtenus par l'étude et les perspectives de recherche qui en découlent.

Chapitre I :
Bibliographie

INTRODUCTON

La bio-informatique est un domaine de recherche qui analyse et interprète des données biologiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie.

Tout au long de ce chapitre, nous essaierons de répondre à ces questions : Qu'est-ce-que la bio-informatique et quelles sont ses descriptions et ses démarches ? Quelles sont les acides nucléiques et les protéines?

Ce chapitre sera l'occasion pour nous de comprendre la notion de la bio-informatique à travers sa définition, ses descriptions et ses démarches, et pour conclure ce chapitre, on abordera les acides nucléiques et les protéines.

I.1. La bio-informatique :

Dans cette section nous présentons des notions de biologie qui sont en principe nécessaires pour bien mettre en contexte la bio-informatique.

La bio-informatique (bioinformatics) ou biologie computationnelle (computational biology) traite de la gestion et de l'analyse de données biologiques ; elle constitue l'un des principaux piliers des branches modernes des sciences de la vie et des industries connexes. Parmi les domaines d'application de la bio-informatique figurent la recherche pharmaceutique, la science du diagnostic médical, la production d'aliments pour animaux et la recherche de nouvelles énergies. Il convient aussi de faire mention des applications faisant appel aux nanotechnologies. De nos jours, il n'est plus possible de faire de la recherche de pointe en biologie sans avoir recours à la bio-informatique.

- C'est une science interdisciplinaire qui utilise des méthodes théoriques assistées par ordinateur pour résoudre des problèmes scientifiques. Elle a contribué aux grandes avancées de la biologie et de la médecine modernes. La bio-informatique s'est fait connaître dans les médias en 2001 grâce au rôle primordial qu'elle a joué dans le séquençage et le décryptage du génome humain (ADN).
- La bio-informatique est une discipline qui est fortement compartimentée tant eu égard aux problèmes posés que concernant les méthodes appliquées. Les principales branches de la bio-informatique sont la gestion et l'intégration de données biologiques, l'analyse de séquences, la bio-informatique structurale et l'analyse de données issues des technologies à haut débit. Comme la bio-informatique est indispensable à l'analyse de

grandes quantités de données, elle constitue l'un des piliers essentiels de la biologie systémique.

Dans le monde anglophone, bioinformatics est souvent opposé à computational biology, qui couvre un domaine plus large que la bio-informatique classique ; le plus souvent, ces deux termes sont toutefois utilisés comme synonymes. Aujourd'hui, la bio-informatique est une discipline à part entière et bien établie qui compte parmi les sciences de base de la biologie et de la médecine et qui, à ce titre, est enseignée dans plusieurs villes de Suisse.

Pour la Suisse, le potentiel de la bio-informatique réside dans le fait que cette discipline est axée à la fois sur les applications pratiques et sur la recherche pure. La priorité est donnée au soutien d'industries clés comme l'industrie pharmaceutique, la science du diagnostic ou la biochimie spécialisée et d'entreprises comme Novartis, Roche Pharma, Merck Serono, Actelion, Roche Diagnostics, DSM ou Lonza. Ces industries et entreprises, pour rester compétitives, ont des besoins importants en bio-informatique, raison pour laquelle elles investissent dans cette discipline. Par ailleurs, il existe aussi, dans le monde académique (et en particulier au sein de l'Institut suisse de bioinformatique, SIB), un grand savoir-faire et une concentration de compétences en bio-informatique.

Avec deux des plus grands fournisseurs du monde (GeneData et GeneBio) et des succursales d'entreprises informatiques proches des milieux de la recherche, la Suisse est, à l'instar des Etats-Unis, de l'Allemagne et du Royaume-Uni, l'un des leaders de la discipline. Il convient de maintenir et de renforcer cette position afin de parvenir, à moyen ou à long terme, à garantir des emplois hautement qualifiés fondés sur le savoir et à établir de nouveaux domaines d'application offrant d'importants débouchés sur le plan international. D'ici à quelques années, les chercheurs espèrent faire bénéficier les patients des progrès du diagnostic moléculaire, objectif pour lequel la bio-informatique est fortement mise à contribution (médecine personnalisée).

Les enjeux liés à la bio-informatique résident dans la sécurité et la qualité des données, la gestion des données (massives et complexes), l'évaluation et l'intégration des données et, indirectement, dans la protection de la propriété intellectuelle (brevetage des logiciels, p. ex.), la validation des systèmes informatiques (certification des logiciels, p. ex.), la transparence relative aux questions d'ordre juridique et éthique et aux procédures étatiques (procédures d'approbation, etc.).

I.1.1. Définition de la bio-informatique :

On trouve un grand nombre de définitions selon l'acception du terme et selon la prépondérance de "bio" sur "informatique" ou l'inverse

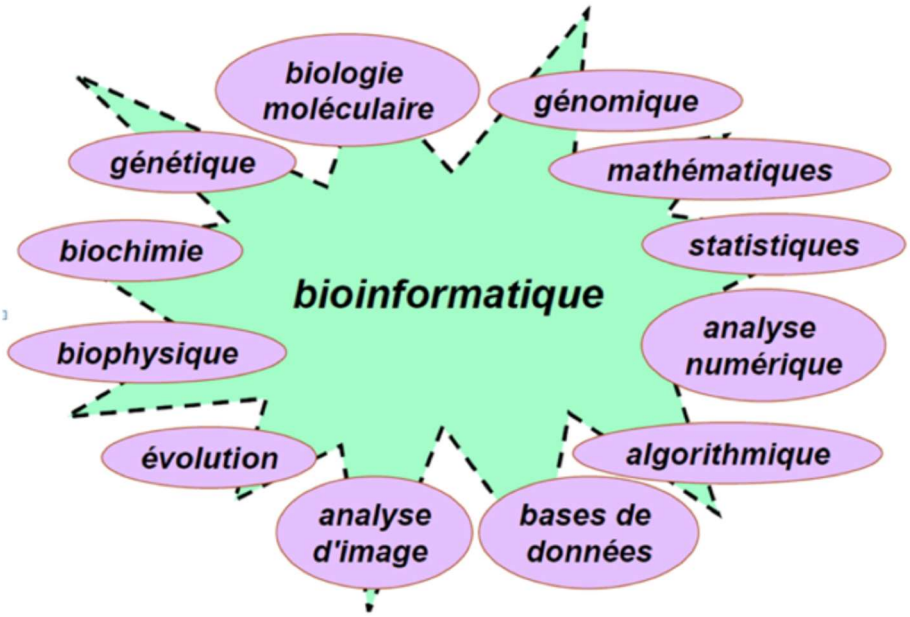
- La bio information est l'information liée aux molécules biologiques : leur séquence, leur nombre, leur(s) structure(s), leur(s) fonction(s), leurs liens de "parenté", leurs interactions et leur intégration dans la cellule ...
- Cette bio information est issue de diverses disciplines : la biochimie, la génétique, la génomique structurale, la génomique fonctionnelle, la transcriptomique, la protéomique, la biologie structurale (structure spatiale des molécules
- Une définition de la bioinformatique : analyse de la bio information par des moyens informatiques

La définition du NCBI (2001) est : "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline."

De manière générale :

- discipline récente (quelques dizaines d'années).
- discipline hybride : elle est fondée sur des concepts et des formalismes issus de la biologie, de l'informatique, des mathématiques et de la physique, de la chimie (techniques de séquençage, ...).
- discipline qui utilise tout le potentiel de traitement de l'informatique : modèles théoriques, algorithmes et programmes, bases de données, ordinateurs, réseau Internet, protocoles de communication, langages,... **(COURS BIO INFORMATIQUE, Faculté des Sciences Exactes et Appliquées , , / M1- CTC 2019/2020)**

Figure I. 1: la bio-informatique



I.1.2. Description de la bio-informatique :

- Discipline récente (quelques dizaines d'années).
- Discipline hybride : elle est fondée sur des concepts et des formalismes issus de la biologie, de l'informatique, des mathématiques et de la physique, de la chimie (techniques de séquençage, ...).
- Discipline qui utilise tout le potentiel de traitement de l'informatique : modèles théoriques, algorithmes et programmes, bases de données, ordinateurs, réseau Internet, protocoles de communication, langages, ... (www.aquaportail.com/definition-531-arn.html consulté le 02/02/2022)

I.1.3 Démarche de la bio-informatique :

La démarche de la bio informatique peut être résumée selon les étapes suivantes :

1)- Compilation et organisation des données biologiques dans des bases de données :

- bases de données généralistes (elles contiennent le plus d'information possible sans expertise très poussée de l'information déposée)
- bases de données spécialisées autour de thèmes précis

2)- Traitements systématiques des données : l'un des objectifs est de repérer et de caractériser une fonction et/ou une structure biologique importante. Les résultats de ces traitements constituent de nouvelles données biologiques obtenues "in silico".

3)- Elaboration de stratégies :

- apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "in silico".
- ces connaissances permettent, à leur tour, de développer de nouveaux concepts en biologie.
- concepts qui, pour être validés, peuvent nécessiter le développement de nouvelles théories et outils en mathématiques et en informatique.

Quelques étapes clé en biologie moléculaire, en informatique et en bioinformatique (la liste ne peut évidemment pas être exhaustive), sont données dans le tableau[1] suivant, pour nous permettre de comprendre l'évolution :

Tableau I. 1: l'évolution de la bio-informatique

1965	Margaret Dayhoff et al. : Première compilation de protéines ("Atlas of Protein Sequences"). Matrices de substitution
1967	Article : "Construction of Phylogenetic Trees" - Fitch & Margoliash
1970	Algorithme pour l'alignement global de séquences : Saul Needleman & Christian Wunsch
1971	Premier microprocesseur Intel 4004
1972	Clonage de fragments d'ADN dans un virus, l'ADN recombiné : Paul Berg, David Jackson, Robert Symons
1973	Découverte des enzymes de restriction qui coupe spécifiquement l'ADN. Méthode de transfection (introduction d'un ADN étranger) des cellules eucaryotes grâce à un virus (vecteur).
1974	Programme de prédiction de structures secondaires des protéines : "Prediction of Protein Conformation" - Chou & Fasman. Vint Cerf et Robert Khan développent le concept des réseaux reliant des

[1] Etablis par l'étudiante

	ordinateurs au sein d'un « internet » et développent deux protocoles fondamentaux "Transmission Control Protocol" (TCP) et "Internet Protocol" (IP).
1977	Développement des micro-ordinateurs accessibles à tous Techniques de séquençage d'ADN : Frederick Sanger / Maxam & Gilbert
1978 - 1980	Mutagénèse dirigée : Michael Smith Séquençage du 1er génome à ADN, le bactériophage phiX174 : Frederick Sanger Premières bases de données : EMBL, GenBank, PIR Accès téléphonique à la base de données PIR
1981 : 370.000 nucléotides GenBank : 270 séquences	Micro-ordinateur IBM-PC 8088 Programme d'alignement local de séquences : Temple Smith & Michael Waterman
1983	IBM-XT disque dur (10 Mb)
1984	Amplification de l'ADN : réaction de polymérisation en chaîne (PCR - Karry Mullis) MacIntosh : interface graphique & souris
1985	"FASTA" : Programme d'alignement local de séquences - David Lipman & William Pearson
1987	Nouveau vecteur permettant de cloner des fragments d'ADN 20 fois plus grands : le YAC (Yeast Artificial Chromosome) qui rend possible le séquençage de grands génomes.
1988	Taq polymérase, enzyme thermostable pour la PCR. Création du "National Centre for Biotechnology Information" (NCBI).
1989	INTERNET succède à ARPANET
1990	Clonage positionnel et premier essai de thérapie génique. "BLAST" : Programme d'alignement local de séquences - Altschul et al.
1991	"Expressed Sequences Tags" (EST) : méthode rapide d'identification des gènes (C. Venter).
1992	Séquençage complet du chromosome III de levure
1993	"European Bioinformatics Institute" (EMBL). Création à terme du "European Bioinformatics Institute" (EMBL - EBI).
1995	Analyse du transcriptome : début des puces à ADN
1996	Séquençage complet de la levure (consortium européen).

1997	11 génomes bactériens séquencés Evolutions de BLAST : "Gapped BLAST" et "PSI-BLAST"
1998	Séquençage de 2 millions de nucléotides par jour. Interférence ARN
2000	Séquençage
2001	Séquence "premier jet" complète du génome humain
Années 2000	Epigénétique : développement de technologies d'analyse des modifications de l'ADN et des histones. Accès aux revues et journaux scientifiques : développement de l'"open access". Montée en puissance de la biologie synthétique. Détermination de structures de systèmes biologiques de plus en plus complexes (ribosomes, spliceosome, virus, ...) - cryo-microscopie électronique et autres techniques ("femtosecond pulses / X-ray free-electron laser")
2007 - 2008	Avènement des nouvelles technologies de séquençage à très haut débit, dites de seconde génération et maintenant de 3è génération. Prise de conscience du phénomène "big data" (pas seulement en biologie) qui devient peu à peu une discipline scientifique.
Mars 2019 : > 303 milliards de nucléotides > 49 millions séquences d'acides aminés	Plus de 18.900 génomes eucaryotes et procaryotes séquencés et des milliers en cours de séquençage (Genomes OnLine).

(Juliana Silva Bernendes ;Hugue Richards ; introduction à la bio informatique)

I.1.4.Domaines d'applications

1)- L'acquisition des données biologiques

- les séquences nucléotidiques et les séquences polypeptidiques
- les gels bidimensionnels et les différentes méthodes de spectromètrie de masse (protéomique)
- les données de puce à ADN
- les données de structures tridimensionnelles
- l'uniformisation - standardisation des (formats de) données
- la recherche de phase de lecture ouverte (gène) et de signaux de régulation de la transcription et de la traduction, détection de bornes introns/exons

- la recherche de régions transcrites (EST) - profil d'expression des gènes (puces à ADN, analyse d'images)
- la détection de polymorphismes de nucléotide simple ou d'insertion / délétion
- la reconstruction d'arbres phylogèniques
- l'analyse de génomes entiers (génomique structurale, synténie) - réseaux de gènes
- l'ontologie : l'organisation hiérarchique de la connaissance sur un ensemble d'objets par leur regroupement en sous-catégories suivant leurs caractéristiques essentielles

2)- Le séquençage :

La bio-informatique intervient aussi dans le séquençage, avec par exemple l'utilisation de puces à ADN ou biopuce. Le principe d'une telle puce repose sur la particularité de reformer spontanément la double hélice de l'acide désoxyribonucléique face au brin complémentaire. Les quatre molécules de base de l'ADN ont en effet la particularité de s'unir deux à deux. Si un patient est porteur d'une maladie, les brins extraits de l'ADN d'un patient, vont hybrider avec les brins d'ADN synthétiques représentatifs de la maladie.

Depuis l'invention du séquençage de l'ADN par Frederick Sanger dans la deuxième moitié des années 1970, les progrès technologiques dans ce domaine ont été tels que le volume des séquences d'ADN disponibles a progressé de manière exponentielle, avec un temps de doublement de l'ordre de 15 à 18 mois, c'est-à-dire un peu plus rapidement que la puissance des processeurs des ordinateurs (Loi de Moore). Un nombre exponentiellement croissant de séquences de génomes ou d'ADN complémentaires sont disponibles, dont l'annotation (ou interprétation de leur fonction biologique) reste à effectuer.

La première difficulté consiste à organiser cette énorme masse d'information et de la rendre disponible à l'ensemble de la communauté des chercheurs. Cela a été rendu possible grâce à différentes bases de données, accessibles en lignes. À l'échelon mondial, trois grandes institutions sont chargées de l'archivage de ces données : le NCBI aux États-Unis, l'EBI en Europe et le DDBJ (en) au Japon. Ces institutions se coordonnent pour gérer les grandes bases de données de séquences nucléotidiques comme GenBank ou l'EMBL database, ainsi que les bases de données de séquences protéiques comme UniProt ou TrEMBL. Il faut ensuite développer des outils d'analyse de séquences afin de pouvoir déterminer leurs propriétés.

- Recherche de protéines à partir de la traduction de séquences nucléiques connues. Celle-ci passe par la détermination des cadres de lecture ouverts d'une séquence nucléique et de sa ou ses traduction(s) probables

- Recherche de séquences dans une banque de données à partir d'une autre séquence ou d'un fragment de séquence. Les logiciels les plus fréquemment utilisés sont de la famille BLAST (blastn, blastp, blastx, tblastx et leur dérivés).
- Alignement de séquences : pour trouver les ressemblances entre deux séquences et déterminer leurs éventuelles homologies. Les alignements sont à la base de la construction de parentés suivant des critères moléculaires, ou encore de la reconnaissance de motifs particuliers dans une protéine à partir de la séquence de celle-ci.
- Recherche de motifs ou structures consensus pour caractériser les séquences

3)- Modélisation moléculaire :

Les macromolécules biologiques sont en général de dimensions trop petites pour être accessibles à des moyens d'observations directes telles que la microscopie. La biologie structurale est la discipline qui a pour objet de reconstruire des modèles moléculaires, par l'analyse de données indirectes ou composites. L'objectif est d'obtenir une reconstruction tridimensionnelle 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), de résonance magnétique nucléaire, de cryomicroscopie électronique ou de techniques de diffusion aux petits angles (diffusion des rayons X ou diffusion des neutrons). Les données issues de ces expériences constituent des données (ou contraintes) expérimentales qui sont utilisées pour calculer un modèle de la structure 3D. Le modèle moléculaire obtenu peut être est un ensemble de coordonnées cartésiennes des atomes composant la molécule, on parle alors de modèle atomique, ou une "enveloppe", c'est-à-dire une surface 3D décrivant la forme de la molécule, à plus basse résolution. L'informatique intervient dans toutes les étapes conduisant de l'expérimentation au modèle, puis dans l'analyse du modèle par la visualisation moléculaire (voir les protéines en 3D).

Un autre volet de la modélisation moléculaire concerne la prédiction de la structure 3D d'une protéine à partir de sa structure primaire (l'enchaînement des acides aminés qui la composent), en prenant en compte les différentes propriétés physico-chimiques des acides aminés. Cela a un grand intérêt car la fonction, l'activité d'une protéine dépend de sa forme. De même, la modélisation des structures 3D d'acides nucléiques (à partir de leur séquence nucléotidique) revêt la même importance que pour les protéines, en particulier pour les structures d'ARN :

La connaissance de la structure tri-dimensionnelle permet d'étudier les sites actifs d'une enzyme, mettre au point informatiquement une série d'inhibiteurs potentiels pour cette enzyme,

et ne synthétiser et ne tester que ceux qui semblent convenir. Cela permet de réduire les coûts en temps et en argent de ces recherches.

- De même la connaissance de cette structure permet de faciliter l'alignement de séquences protéiques.
- La visualisation de la structure tridimensionnelle d'acides nucléiques (ARN et ADN) fait également partie de la palette des outils bio-informatique très utilisés.

4)- Construction d'arbres phylogénétiques

On appelle gènes homologues des gènes descendant d'un même gène ancestral. De façon plus spécifique, on dit de ces gènes qu'ils sont orthologues s'ils se retrouvent dans des espèces différentes (spéciation sans duplication), ou qu'ils sont paralogues s'ils se retrouvent chez la même espèce (duplication à l'intérieur du génome).

Il est alors possible de quantifier la distance génétique entre deux espèces en comparant leurs gènes orthologues. Cette distance génétique est représentée par le nombre et le type de mutations qui séparent les deux gènes.

Appliquée à un nombre plus important d'êtres vivants, cette méthode permet d'établir une matrice des distances génétiques entre plusieurs espèces. Les arbres phylogénétiques rapprochent les espèces qui ont la plus grande proximité. Plusieurs algorithmes différents sont utilisés pour tracer des arbres à partir des matrices de distance. Ils reposent chacun sur des modèles de mécanismes évolutifs différents. Les deux méthodes les plus connues sont la méthode UPGMA et la méthode du Neighbour Joining mais il existe d'autres méthodes basées sur le Maximum de Vraisemblance et le Bayésien Naïf.

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 : il permet ainsi de guider l'alignement total. **(Bernendes, Richards, & informatique)**

I.2.Les acides nucléiques et les protéines

I.2.1. Généralités sur les acides nucléiques

Les acides nucléiques sont des polymères formés par l'association de plusieurs nucléotides. Chaque nucléotide est constitué lui-même d'une base hétérocyclique purique (A et G) ou pyrimidique (C, U, T), d'un pentose et d'un acide phosphorique. Selon la nature du pentose, on distingue deux types d'acide nucléiques : les acides ribonucléiques ou ARN (contenant du ribose) et les acides désoxyribonucléiques ou ADN (contenant du désoxyribose).

(https://www.etudier.com/G%C3%A9n%C3%A9tique%20mol%C3%A9culaire consulté le 02/02/2022)

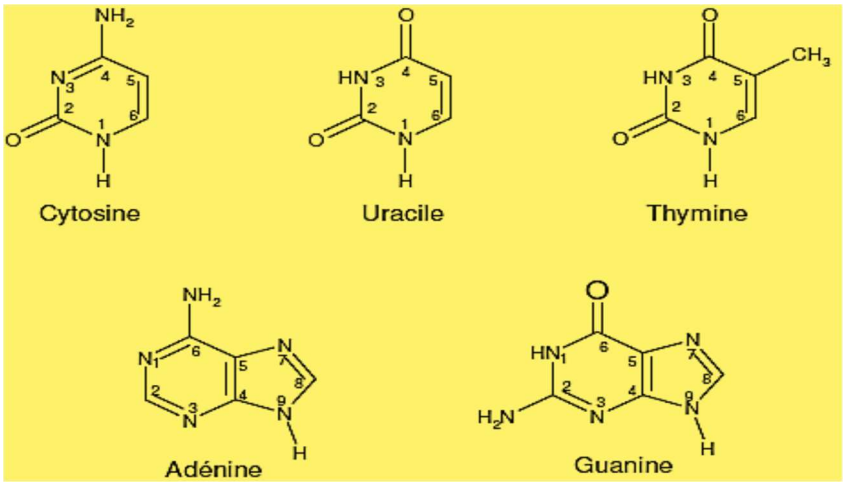


Figure I. 2 : Structure chimique des principales bases nucléiques

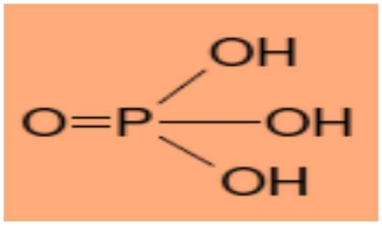


Figure I. 3 : l'acide phosphorique :

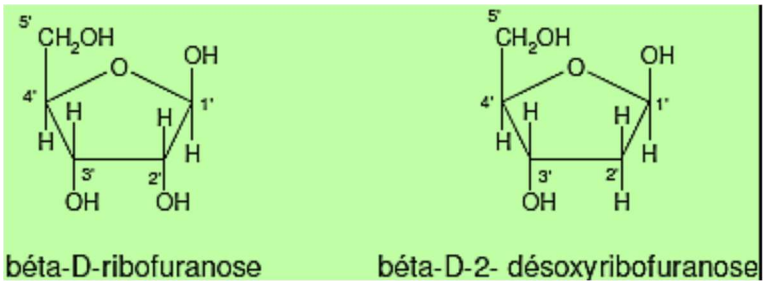


Figure1. 4 : structure chimique du ribose et du désoxyribose :

I.2.2. L'ADN : L'acide Désoxyribonucléique

La biologie moléculaire est une discipline scientifique au croisement de la génétique, de la biochimie et de la physique, dont l'objet est la compréhension des mécanismes de fonctionnement de la cellule au niveau moléculaire. Le terme « biologie moléculaire », utilisé la première fois en 1938 par Warren Weaver, désigne également l'ensemble des techniques de manipulation d'acides nucléiques (ADN, ARN), appelées aussi techniques de génie génétique.

Après la découverte de la structure en double hélice de l'ADN en 1953 par James Watson (1928-), Francis Crick (1916-2004), Maurice Wilkins (1916-2004) et Rosalind Franklin (1920-1958), la biologie moléculaire a connu d'importants développements pour devenir un outil incontournable de la biologie moderne à partir des années 1970.

Depuis la fin des années 1950 et le début des années 1960, les biologistes moléculaires ont appris à caractériser, isoler et manipuler les composants moléculaires des cellules et des organismes. Ces composants incluent l'ADN, support de l'information génétique, l'ARN, et les protéines, molécules structurelles et enzymatiques les plus importantes des cellules. **(Introduction à l'analyse génétique Griffith S. Wessler. 4eme édition de boeck)**

1. Définition de l'ADN :

ADN est l'acide désoxyribonucléique, un acide nucléique composé de désoxyribose, de phosphate, d'adénine, de cytosine, de guanine et de thymine. L'ADN contient les instructions génétiques utilisées dans le développement et le fonctionnement de tous les organismes vivants et de certains virus, et qui est responsable de sa transmission héréditaire.

Cette macromolécule constitue le support des informations génétiques de tous les êtres vivants excepté les virus à ARN. Elle est formée d'une double chaîne hélicoïdale de désoxyribonucléotides, chaque chaîne ou brin étant complémentaire de l'autre.

2. Structure de l'ADN:

- Eléments constitutifs de l'ADN

L'ADN est constitué de 03 éléments des nucléotides monophosphates.

- ✓ Un groupe phosphate (phosphoryle)
- ✓ Un sucre, le désoxyribose : un pentose (5 carbones)
- ✓ Une base azotée qui peut être l'une des 4 bases azotées suivantes : (figure 05)

Adénine : A, Guanine : G, Cytosine : C, Thymine : T C

Les bases **A** et **G** sont des purines ou bases puriques.

Les bases **C** et **C** sont des pyrimidiques ou bases pyrimidiques.

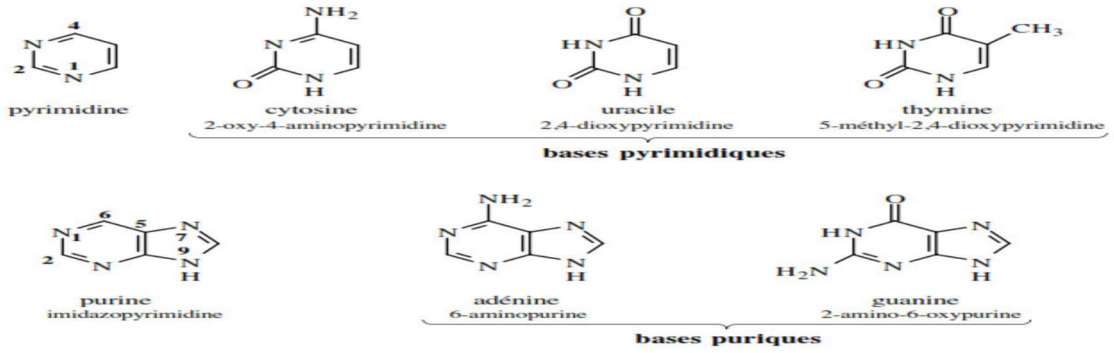


Figure I.5: les bases azotées

L'association de chaque base avec une molécule de sucre constitue un nucléoside, et l'ajout d'un groupe phosphate donne un nucléotide monophosphate.

- **Structure primaire de l'ADN :**

ADN est constitué par l'enchainement linéaire de sous unité de base, les nucléotides qui forment un filament non ramifié. Un nucléotide est composé d'un groupement phosphate, d'un sucre, le D-désoxyribose (qui constituent le squelette de l'ADN) et d'une base purique ou pyrimidique (figure 06). **(Cours de biologie moléculaire faculté de médecine de Batna : 2015/2016)**

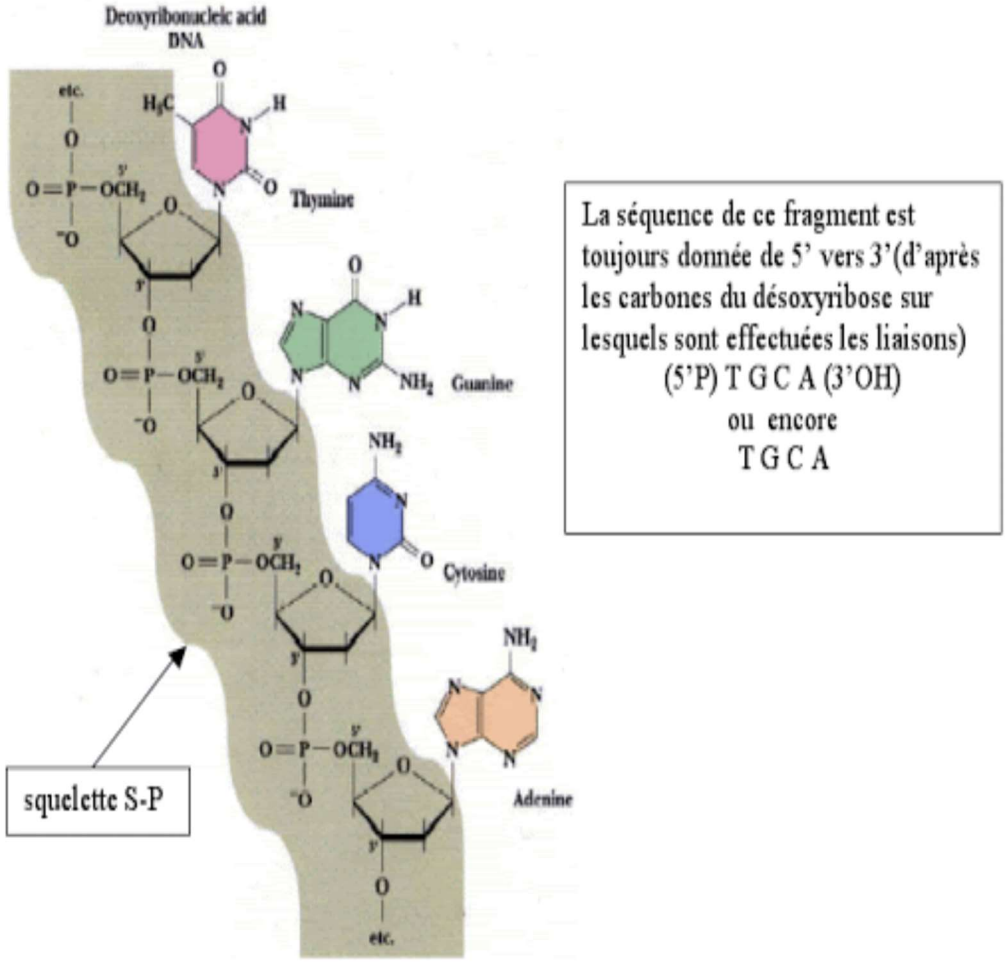


Figure I. 6 : Structure primaire de l'ADN

- **Structure secondaire de l'ADN**:

Selon Watson et Crick en 1953, l'ADN est formé de deux chaines (polymère) enroulés l'une autour de l'autre en une double hélice.

Les bases azotées des deux brins sont reliées par des liaisons faibles c'est des liaisons d'hydrogènes. Les bases sont associées par paires. Dans chaque paire il y a toujours une purine associée à une pyrimidine.

- ✓ Les bases A sont associées aux bases T par 2 liaisons Hydrogène.
- ✓ Les bases G sont associées aux bases C par 3 liaisons Hydrogène.

Les deux brins sont complémentaires l'un de l'autre mais ne sont pas identiques. Chaque brin d'ADN possède une extrémité 5′-phosphate et une extrémité 3′-hydroxyle.par convention, une orientation est définie de l'extrémité 5′P 3′OH. Les deux brins sont orientés en sens opposés, on dit qu'ils sont anti parallèles (figure 07).

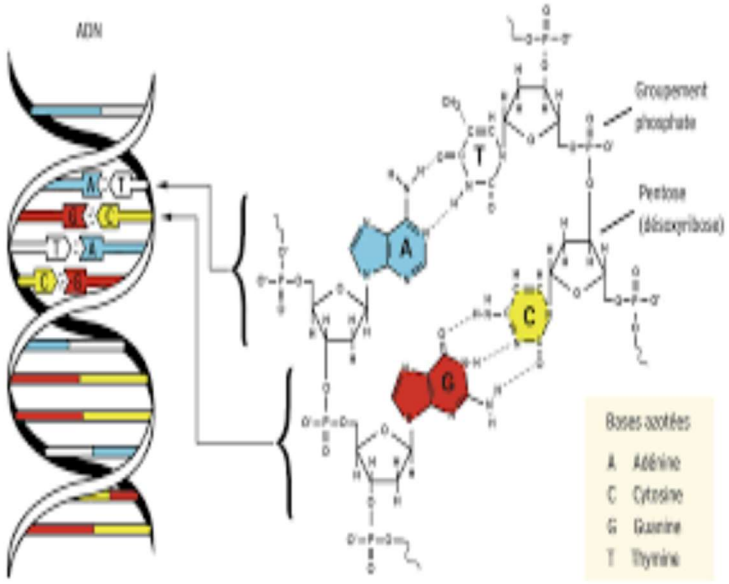


Figure 7 : Structure secondaire

- ➢ Structure Tertiaire de l'ADN (double Hélice d'ADN) :

Chaque brin d'ADN ressemble à une hélice, les deux brins forment une structure en double hélice. La double hélice présente deux types de sillons : les grands sillons et les petits sillons. C'est au niveau de ces deux types de sillons que se fixent les protéines nécessaires aux processus de transcriptions, de réplication ou de réparation de l'ADN. **(Biologie moléculaire Houali K. Lahcen S.2005.)**

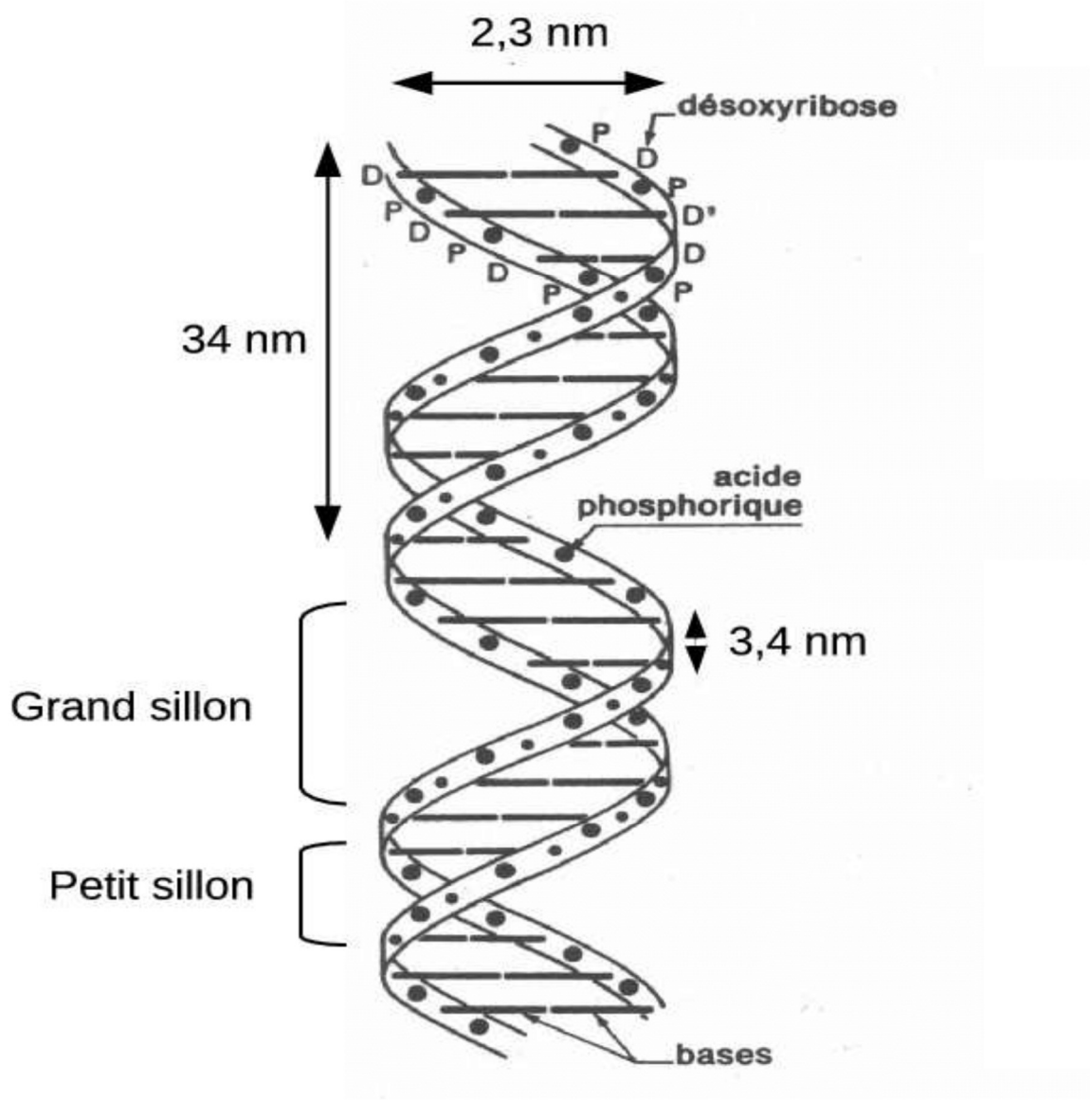


Figure I. 8 : Structure Tertiaire de l'ADN (double Hélice d'ADN)

3. Propriétés physicochimique de l'ADN:

- **La taille** : les acides nucléiques (ADN et ARN) sont les plus grandes macromolécules naturelles. 03 caractéristiques peuvent être utilisées pour exprimer la taille :
  - ✓ La longueur : (en µm ou en nm).
  - ✓ La masse moléculaire en Dalton : 1 dalton= la masse d'un atome d'hydrogène. 1 kilo dalton (Kda) = 100 Da.
  - ✓ Le nombre de nucléotides ou de bases (noté b) pour les molécules monocaténaires (simple brin) ou de paires de bases (noté pb) pour les bicaténaires (doubles brins) : 1 kilo paires de bases (Kpb) = 1000 pb.

Spectre d'absorption de l'ADN : l'ADN présente une absorption de la lumière (densité optique) qui est maximale dans l'ultraviolet UV à une longueur d'onde = 260nm.

Température de fusion : Si une solution d'ADN est chauffée, à une certaine température, les liaisons hydrogène qui assurent la cohésion des 2 brins appariés se rompent figure 4. On parle de fusion de l'ADN caractérisée par la température de fusion (Tm : melting temperature) (La dénaturation et la renaturation des brins d'ADN en solution sont des reconstitutions critiques pour diverses fonctions biologiques normales (réplication; transcription …etc.)

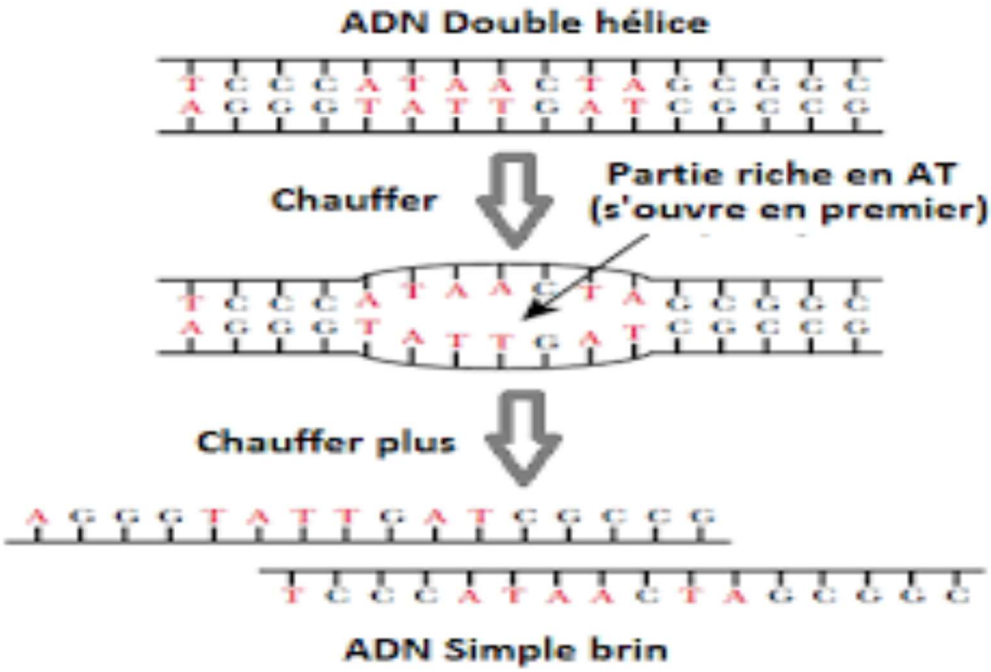


Figure I. 9 : Séparation de brin d'ADN en fonction de température

Renaturation de l'ADN : la fusion est un phénomène réversible : en baissant lentement la température, il y a une réassociation des 2 brins, c'est la renaturation de l'ADN. Deux simples brins provenant de 2 ADN différents chauffés peuvent aussi réassocier elle s'appelle Hybridation.

Solubilité de l'ADN : l'ADN est soluble dans l'eau mais pas dans l'éthanol.

(www.aquaportail.com/definition-530-adn.html)

I.2.3. L'ARN : L'acide ribonucléique

ARN l'acide ribonucléique est un acide nucléique formé par une chaîne de ribonucléotides, composé de ribose, de phosphate, d'adénine, de cytosine, de guanine et

d'uracile. Il est présent dans les cellules procaryotes et les eucaryotes, et est le seul matériel génétique de certains virus (virus à ARN). Cet acide nucléique résulte de la transcription de l'ADN. L'ARN est constitué d'une chaîne de monomères répétitifs appelés nucléotides. Les nucléotides sont joints les uns après les autres par des liaisons phosphodiester chargées négativement. **(Introduction à l'analyse génétique Griffith S. Wessler. 4eme édition de boeck)**

- **Structure de l'ARN:**
- La première différence principale dans la structure de l'ARN est la fonction hydroxyle en 2' du ribose qui permet à l'ARN de faire une liaison phosphodiester intramoléculaire en milieu basique avant de faire la liaison 3'-5'. De ce fait les ARN ont une demi-vie très courte figure 10.
- La deuxième différence principale est le remplacement de la thymine par l'uracile.
- Les molécules d'ARN sont simple brin et linéaire, et les seuls appariements de paires se font intramoléculaires par des liaisons hydrogènes sous forme de structures en tiges-boucles aux extrémités de l'ARN et des structures en épingles à cheveux à l'intérieur de l'ARN.
- Les hélices d'ARN formées lors des appariements sont de type A et sont plus courtes et plus trapues que l'hélice de type B de l'ADN. L'effet hypochrome est également présent et dû aux appariements intramoléculaires. La courbe d'absorption UV en fonction de la température est cette fois-ci par pallier correspondant aux différentes boucles à épingles à cheveux dénaturées. **(https://www.etudier.com/ consulté le 02/02/2022)**

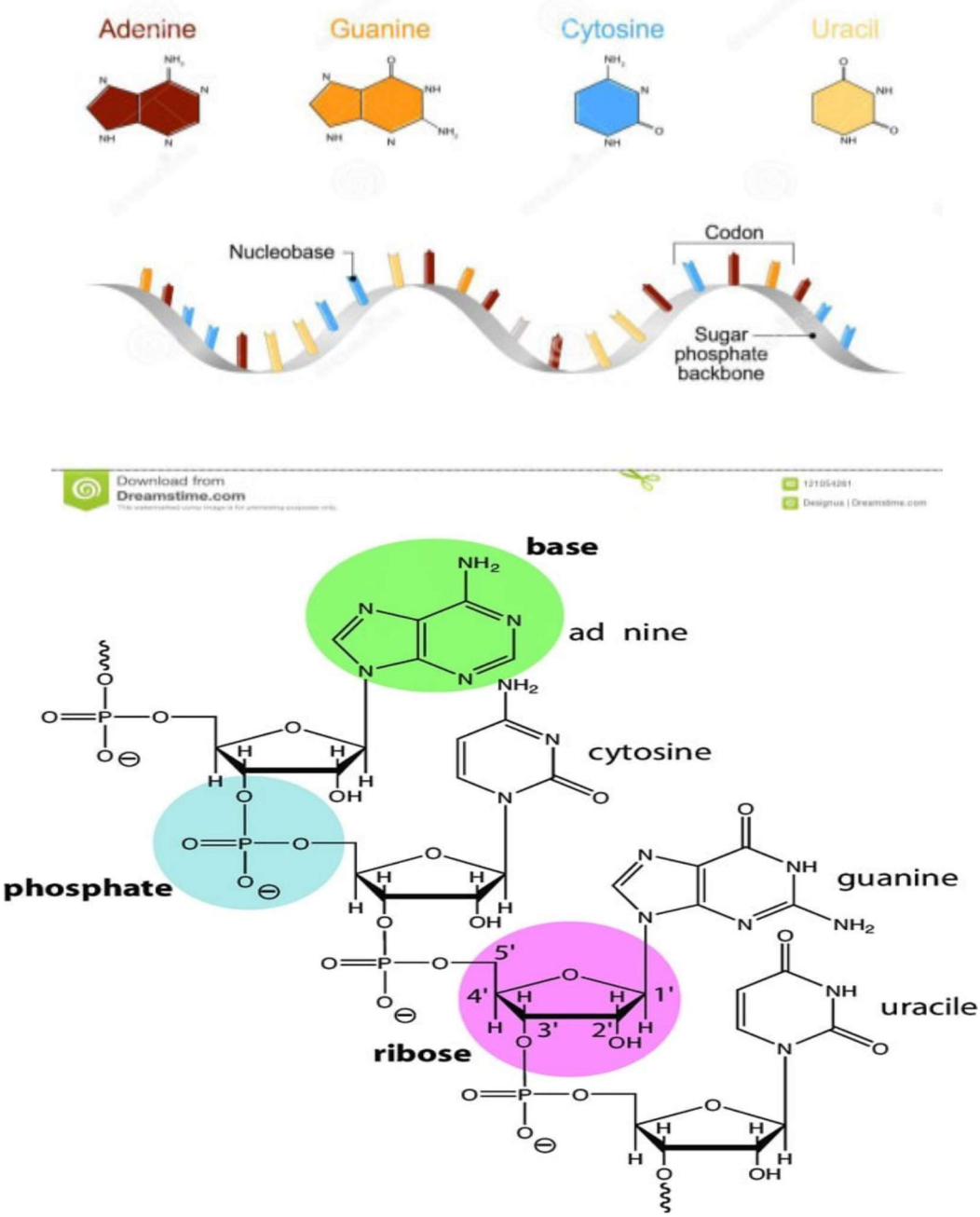


Figure I. 10 Structure de l'ARN

- **Les différents types de l'ARN :**

Il existe trois grands types d'ARN qui sont impliqués dans différents aspects de la synthèse des protéines et sont par conséquent nécessaires à l'expression de l'information génétique sont :

- **L'ARN messager (ARNm) :** qui représente 5% des ARN totaux, est un support temporaire de l'information génétique. Il est utilisé par la cellule pour transmettre l'information correspondant à un gène donné, il provient de la transcription de l'ADN et sert de matrice pour la traduction en protéines figure2.la séquence nucléotidique de l'ARNm est une séquence linéaire, complémentaire et antiparallèle à la séquence matrice de l'ADN dont elle est issue.

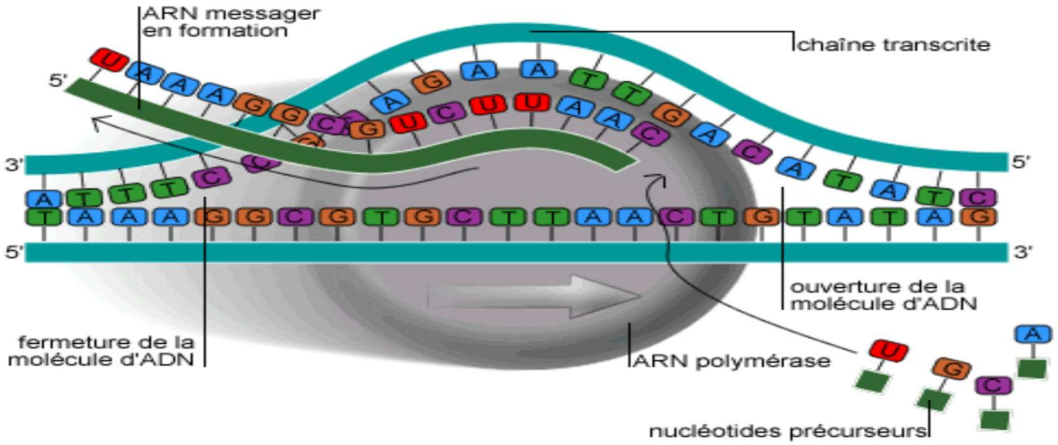


Figure I. 11 : Structure ARN message

- **L'ARN de transfert (ARNt) :** représentant 15% des ARN totaux, est un adaptateur qui reconnait les codons de l'ARNm par appariement de bases complémentaire et insère l'acide aminé adéquat au moment de la traduction. C'est une molécule d'ARN monocaténaire qui peut se replier et former des structures secondaires par appariements entre ses bases complémentaires, donnant une structure en épingle à cheveux résultant d'alternance de zones formant des doubles hélices (les tiges) et les zones non appariées (les boucles) figure12.

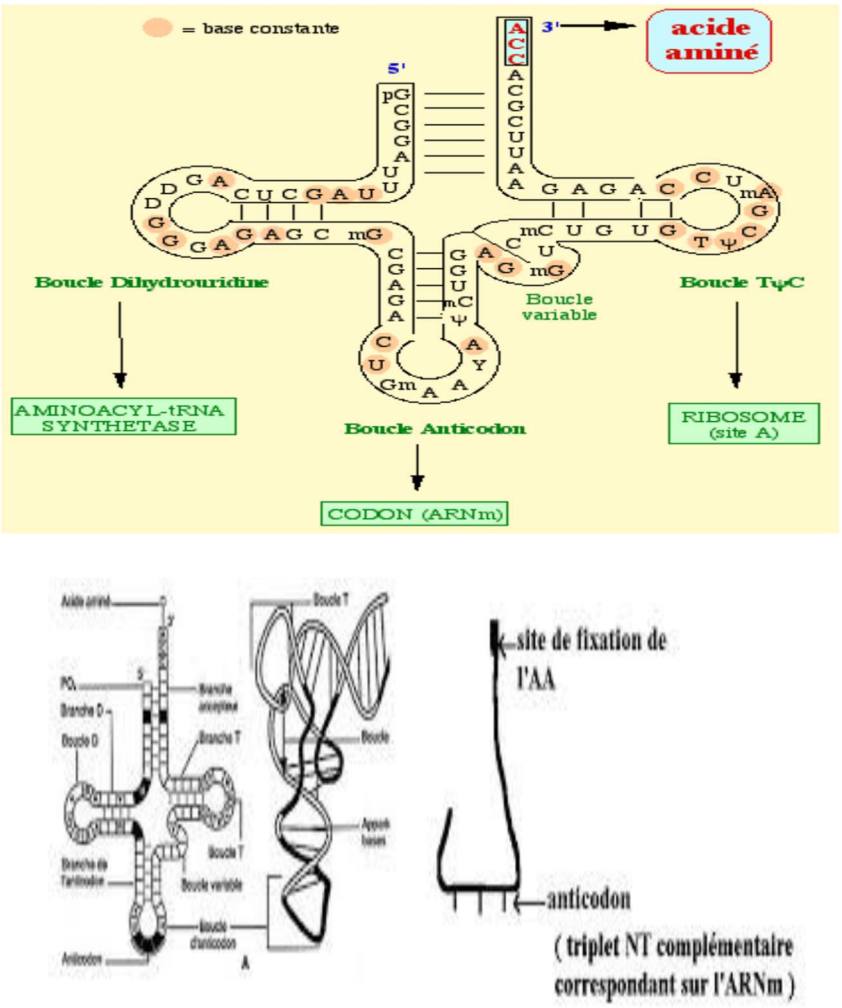


Figure I. 12: Structure ARNt

Les ARNt sont donc des polymères, contenant des régions simple brin et double brin, et composés de ribonucléotides. Leur fonction dans la cellule est d'assurer la correspondance entre l'information génétique portée par l'ARN messager, et les acides aminés contenus dans la protéine codée par cet ARNm. Ils sont les acteurs clés de la traduction de code génétique.

Chaque ARNt porte des 20 acides aminés attaché par une liaison ester à son extrémité 3′-OH et transporte ce dernier au ribosome. Trois des nucléotides de chaque ARNt forment un anticodon spécifique de l'acide aminé. L'anticodon s'apparie au codon sur l'ARNm assurant ainsi la correspondance entre codon et acide-aminé, conformément au code génétique.

L'interaction codon-anticodon s'effectue dans le ribosome qui vérifie la complémentarité des bases de l'ARNm et l'ARNt. Quand celle-ci est réalisée, le ribosome catalyse l'allongement de la chaine protéique encours de synthèse et avance sur l'ARN messager. Une fois que

l'ARNt a été utilisé par le ribosome, il ne porte plus d'acide aminé à son extrémité 3´-OH, il est alors chargé par une enzyme spécifique, nommée aminacyl-ARNt synthétase, qui catalyse l'estérification de l'acide aminé spécifique.

- **ARN ribosomal (ARNr) :** Les ARN ribosomiaux représentent plus de 80% des ARN cellulaires totaux s`associent à des protéines pour former le ribosome qui est le support de la synthèse des protéines. Les ribosomes sont une association de 2 sous unités : 50S et 30S chez les procaryotes et 60S et 40S chez les eucaryotes figure 12. Le coefficient de sédimentation S (Svedberg) est l`unité de mesure de la vitesse de sédimentation. Le coefficient de sédimentation d`une particule dépend non seulement de sa masse mais aussi de sa forme et de sa rigidité. Par définition, la constante de sédimentation S, est la vitesse de sédimentation par unité d'accélération (force G). Comme cette constante est faible, on utilise, comme unité, le Svedberg (un Svedberg = 10-13 seconde). **(https://scholar.google.com/ consulté le 02/02/2022).**

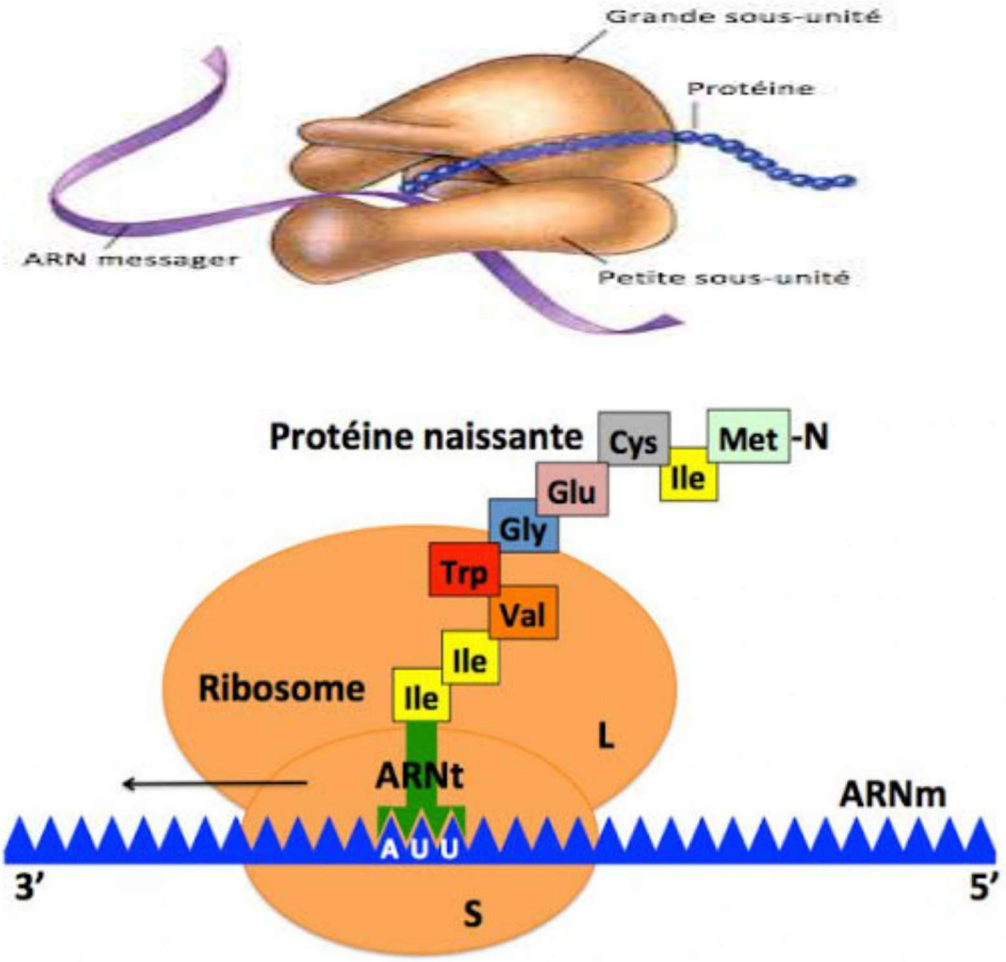
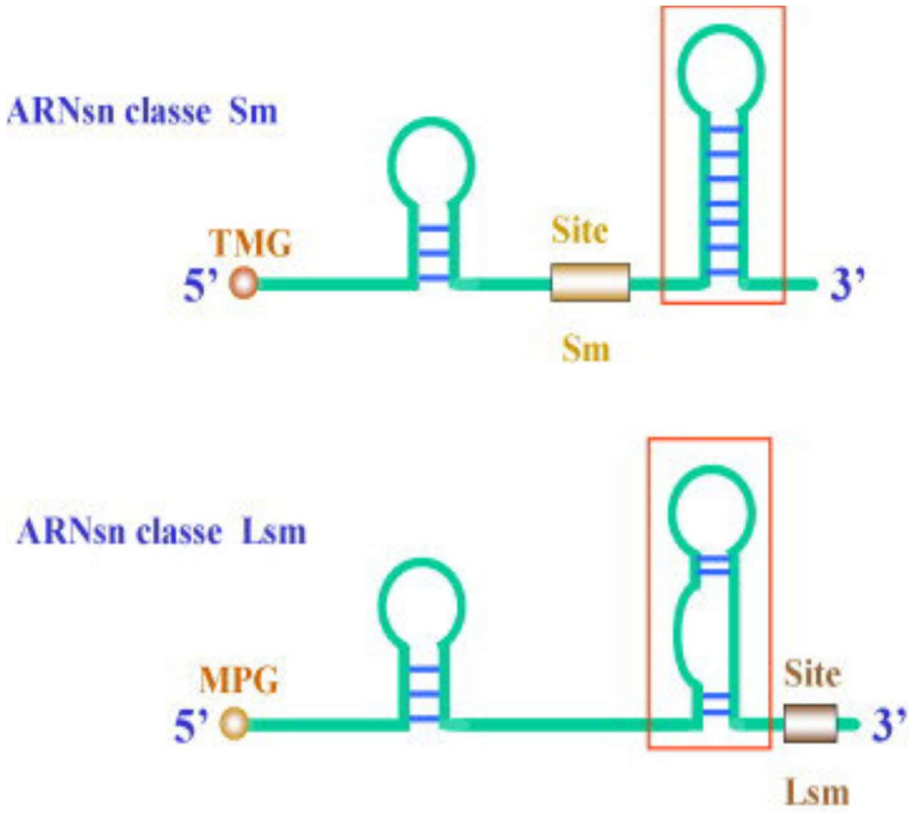


Figure I. 13: Structure de L'ARN ribosomal

- **<u>D'autre type d'ARN</u>:** Au delà du rôle primaire de l'ARN dans la synthèse des protéines, plusieurs variétés d'ARN existent qui sont impliqué dans la modification goujontranscriptionnelle, la réplication de l'ADN, et le règlement de gène. Quelques formes d'ARN sont seulement en particulier les formes trouvées de la durée, comme dans des eucaryotes ou des bactéries.
- **. Petit ARN nucléaire (snRNA) :** le snRNA est impliqué dans transformer des PRÉ-ARN MESSAGERS (pré-ARNm) en ARNm mature. Ils sont très courts, avec une longueur moyenne de seulement 150 nucléotides figure 14**.**



<u>Figure 14 : Structure Petit ARN nucléaire</u>

- **ARNs de réglementation :** Un certain nombre de types d'ARN sont impliqués dans la régulation de l'expression des gènes, y compris l'ARN micro (miARN), le petit ARN de intervention (siARN) et l'ARN antisens (aARN).

✓ Le miARN (NT 21-22) est trouvé dans les eucaryotes, et les actes par l'interférence ARN (RNAi). le miRNA peut décomposer l'ARNm qu'il est complémentaire à, à l'aide des enzymes. Ceci peut bloquer l'ARNm de l'traduction, ou accélérez sa dégradation.

✓ Le siARN (NT 20-25) sont souvent produits par la dégradation de l'ARN viral, bien qu'il y ait également des sources endogènes des siARNs. Ils agissent assimilé a miARN.

Un ARNm peut contenir des facteurs de régulation lui-même, tel que des riboswitches, dans le 5' séquence non-traduite ou 3' séquence non-traduite ; ces éléments cis-de réglementation réglementent l'activité de cet ARNm.

- **ARN de Transfert-messager (tmRNA) :** Trouvé dans beaucoup de bactéries et de plastids. La balise de tmRNA les protéines codées par les ARNm qui manquent des codons non-sens pour la dégradation, et empêche le ribosome de caler dû au codon non-sens manquant. **(Biologie moléculaire Houali K. Lahcen S.2005)**

I.3. Extraction et purification des acides nucléiques :

I.3.1. Extraction des acides nucléiques :

L'extraction et la purification des acides nucléiques contenus dans des cellules eucaryotes ou procaryotes sont essentielles pour un grand nombre d'études en biologie moléculaire.

Il existe quatre étapes principales pour extraire les acides nucléiques et les purifier :

- La lyse cellulaire
- La dénaturation des protéines et des complexes nucléoprotéiques
- L'inactivation des nucléases et La purification des acides nucléiques (ADN ou ARN) (séparation de l'acide nucléique souhaité des débris cellulaires)

a. **Lyse cellulaire** : la procédure de lyse idéale doit être suffisamment rigoureuse pour briser le matériau de départ, mais également suffisamment douce pour préserver l'acide nucléique cible. Il existe différentes procédures de lyse cellulaire :

a-1- Lyse mécanique (Broyage ou lyse hypotonique) :

Cette méthode est particulièrement préconisée lors des extractions à partir des cellules sans paroi. Il est également possible d'utiliser cette méthode sur des cellules procaryotes ou des levures en ajoutant des microbilles ou du sable (de fontainebleau, qui contient 95% de silice) pour faciliter la destruction des parois. Néanmoins, ce procédé est de plus en plus délaissé.

a-2- Traitement chimique et enzymatique :

Le traitement chimique et enzymatique implique l'utilisation de détergents, d'agents chaotropiques, la protéinase K…, il se fait en plusieurs étapes :

a-3- Lyse des parois et membranes cellulaires :

Pour les cellules à paroi, on peut utiliser des hydrolases spécifiques comme le lysozyme pour fragiliser la paroi. En effet les lysozymes coupent les liaisons glycosidiques (1-4) de l'acide N-

acetylmuramique (NAM) et la N-actylglycosamine (NAG) du polysaccharide où alterne NAM et NAG constituants les peptidoglycanes de la paroi bactérienne.

Afin de désorganiser les membranes, on utilise le plus souvent des détergents comme le Sodium Dodécyl Sulfate (SDS), le triton X100 et le sarcosyl qui solubilisent les lipides membranaires sous forme de micelles. Cela permet de créer des pores membranaires suffisamment larges pour libérer le contenu du cytoplasme hors des cellules. Suivant leur force, les détergents vont aussi plus ou moins dénaturer les protéines membranaires.

b. Dénaturation des protéines

La déprotéinisation des extraits cellulaires peut se faire par plusieurs procédés :

b-1- Dénaturation par hydrolyse enzymatique :

On utilise le plus souvent une endoprotéase non spécifique comme la protéinase K, active jusqu'à 65°C. Cette digestion est souvent conduite en présence d'un détergent dénaturant comme le SDS, qui facilite l'action de la protéinase K car il déploie les chaînes protéiques.

b-2- Précipitation des protéines en utilisant un agent chaotropique :

Un agent chaotropique est un ion qui modifie la solubilité des molécules (protéines ou acides nucléiques) et qui peut provoquer leur précipitation en neutralisant certaines charges ioniques requises en surface. Il peut également agir en interférant dans les interactions que les protéines établissent avec l'eau ce qui modifie la solubilité des protéines. Ou en dénaturant les protéines par exemple, par rupture des liaisons hydrogènes qui maintiennent leur structure tertiaire entrainant ainsi le démasquage des régions hydrophobes. Les régions hydrophobes ont tendance à s'agréger et les protéines précipitent (défécation). Exemples des agents chaotropique : L'anion chlorate (ClO 3 -), le Thiocyanate (SCN-), I- Li+ , Mg2+, Ca2+ , et Ba 2+ ,le perchlorate de sodium (NaClO4), le thiocyanate de guanidine (TCG), l'iodure de sodium (NaI) et le chlorure de lithium).

c. Autres composants des solutions d'extraction

En fonction des protocoles d'extraction d'autres composés peuvent être rencontrés dans les solutions d'extraction :

c-1- Les thiols :

Les thiols sont des composés soufrés pouvant être considérés en tant qu'analogues des alcools, car ils sont obtenus par remplacement de l'atome d'oxygène du groupe hydroyle, -OH, d'un alcool par un atome de souffre. Le groupe fonctionnel ainsi obtenu est nommé mercapto, -SH. Lorsqu'on utilise les agents chaotropiques pour éliminer les protéines par précipitation, on ajoute quelque fois dans le tampon d'extraction des thiols pour empêcher la reformation de ponts disulfures des protéines qui restent ainsi à l'état dénaturé.

c-2- Les sels :

L'ajout d'une forte concentration de sels (NaCl 0,15 mol/L ; citrate de sodium ; acétate de sodium) dans le milieu d'extraction ne contenant pas d'agents chaotropiques empêche la séparation des deux brins de l'ADN en formant un écran protecteur pour la double hélice.

c-3- EDTA :

L'Ethylène diamine tétra-acétique (EDTA) est un chélateur d'ions divalents comme le magnésium qui est un cofacteur des DNases et RNases. il permet la préservation des acides nucléiques par inhibition des nucléases .

c-4- RNase :

Les extraits acellulaires bruts contiennent les deux types d'acides nucléiques : ADN et ARN. Pour diminuer la concentration en ARN, on utilise une RNase « DNase free », c'est-à-dire dépourvue d'activité DNase ; pour cela les DNases contaminant éventuellement les préparations de RNase du commerce sont dénaturées par chauffage (par exemple 5 min à 100°C). La RNase est une enzyme particulièrement thermostable qui résiste à ce traitement. La RNase peut être ajoutée dès le début de l'extraction-purification car c'est une enzyme très stable.

d. Élimination des débris cellulaires

Après la lyse cellulaire et l'inactivation des nucléases, les débris cellulaires peuvent être aisément retirés par filtration ou centrifugation. **(Biologie moléculaire Houali K. Lahcen S.2005).**

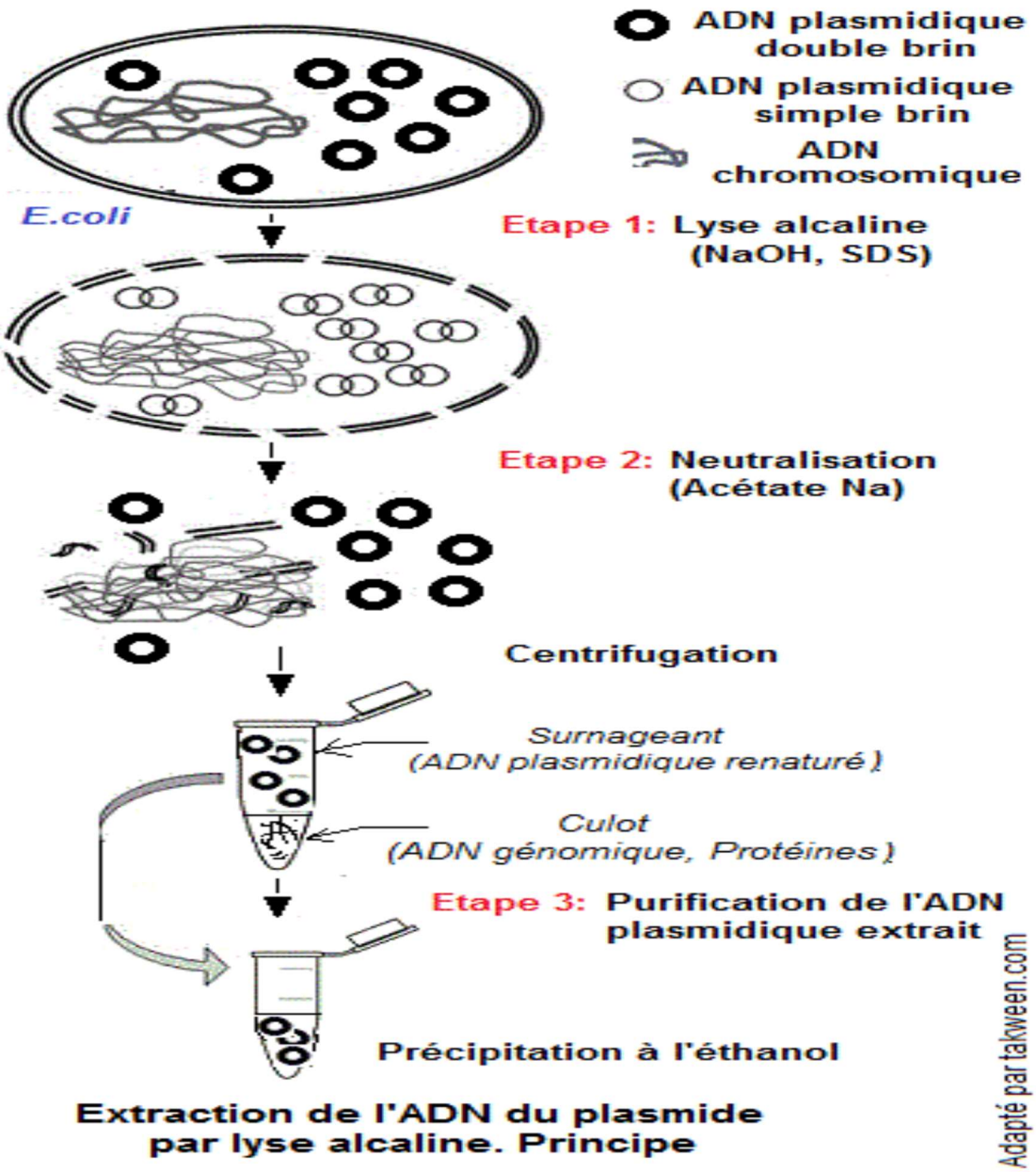


Figure 15 : Extraction de l'ADN du plasmide par lyse alcaline. Principe

I.3.2.Purification des acides nucléiques :

La purification des acides nucléiques à partir d'extraits cellulaires est généralement réalisée par la combinaison entre deux ou plusieurs techniques des techniques suivantes :

1. **Extraction / Précipitation**

L'extraction par solvants est souvent utilisée pour éliminer les contaminants de la préparation d'acides nucléiques. Ainsi plusieurs méthodes de précipitation des acides nucléiques peuvent être utilisées pour les purifier

- **Purification au phénol-chloroforme :** Une combinaison de phénol et de chloroforme sert fréquemment à purifier les acides nucléiques et d'éliminer les protéines. On mélange vigoureusement l'extrait d'acides nucléiques avec une phase hydrophobe. Après centrifugation, on récupère la phase aqueuse supérieure contenant les acides nucléiques.
- **La précipitation par l'isopropanol ou l'éthanol :** Ces techniques sont utilisées pour concentrer les acides nucléiques.
    - **La précipitation par l'éthanol :**

Est réalisée par addition de l'éthanol (un solvant moins polaire que l'eau) a l'extrait d'acides nucléiques (v:v / 2:1). Après refroidissement de l'échantillon, le culot d'ADN est obtenu suite à une centrifugation à une très grande vitesse. Le précipité est lavé avec de l'éthanol à 70 % pour se débarrasser des sels qui puis sécher. Le séchage est obligatoire pour éliminer l'éthanol qui pourrait empêcher la dissolution ultérieure du précipité.

NB :

- Avant l'ajout de l'éthanol, il faut ajouter à l'extrait d'acides nucléiques, une quantité importante de cations.
- Si la quantité d'acides nucléiques cibles est faible, un véhicule inerte (le glycogène) peut être ajouté au mélange afin d'accroître l'efficacité de la précipitation.
    - **La précipitation par l'isopropanol**

Le principe est le même que précédemment sauf que le sel n'est pas nécessaire et que les petits fragments d'ADN sont éliminés (non précipités). Dans ce cas, on procède à mélanges volume de l'isopropanol à volume identique de l'extrait nucléiques (v:v / 1:1) . Après refroidissement et centrifugation le précipité est lavé pour éliminer les traces d'isopropanol puis séché. On obtient alors les acides nucléiques sous la forme de fibres solides que l'on récupère par entrifugation

2. **Chromatographie**

Plusieurs méthodes chromatographiques peuvent étre utiliser telles que la filtration sur gel, chromatographie échangeuse d'ions, l'adsorption sélective ou la liaison par affinité. Mais les deux techniques les plus employées pour purifier les acides nucléiques sont la chromatographie sur colonne de silice et la chromatographie sur colonne d'échange d'anions.

3. Centrifugation

La centrifugation sélective est une méthode de purification puissante. À titre d'exemple, l'ultracentrifugation isopycnique en gradients de chlorure de césium (CsCl) à des forces gravitationnelles élevées, a été longtemps utilisée pour la purification de plasmides.

3- Contrôle de la pureté de l'ADN extrait :

Le maximum d'absorption des acides nucléiques se situe à 260 nm. Les protéines, principaux contaminant des préparations absorbent aussi à 260 nm, mais avec maximum d'absorption qui se situe vers 280 nm à cause des acides aminés aromatiques.

Le rapport R= A260nm / A280nm constitue alors un bon moyen pour apprécier une éventuelle contamination de la préparation d'ADN par les protéines ou par les ARN. Une contamination par les ARN se traduit par une augmentation du rapport R.

$$R = A260nm/A280nm$$

$$ADN\ pur:\ 1,8 < R < 2$$

$$ADN\ contaminé\ par\ les\ protéines : R < 1,7$$

$$ADN\ contaminé\ par\ les\ ARN : R > 2$$

En absence d'impuretés l'absorbance de la solution d'ADN à 320 nm doit être autour de zéro.

CONCLUSION

A travers ce chapitre, nous avons mis en lumière les principales notions de la bio-informatique. Tout d'abord, nous avons défini la bio-informatique et présenté ses descriptions et ses démarches dans la première section. Ensuite, dans la deuxième section, nous avons présenté les acides nucléiques et les protéines

Chapitre II :

Le stockage de la bio-informatique

INTRODUCTION

Les besoins des biologistes exigent que les informations soient organisées, structurées en groupe homogène. Aussi grâce au développement de la télématique, les entreprises ne pouvant pas pour des raisons coût et stockées toutes les informations dont elles ont besoin peuvent consulter automatiquement un fond documentaire appelé BANQUE DE DONNEES.

Ainsi, ce chapitre sera donc l'occasion pour présenter les banques de données, ainsi que les bases de données

II.1. Les banques de données :

II.1.1. Définition :

- **Dans le règlement :**

À moins d'une disposition expresse contraire, ou à moins que clairement le contexte ne le veuille autrement, dans le présent règlement le terme ou l'expression :

« Banques de données » :

- « Ensemble structuré ou non d'informations concernant des sujets humains généralement organisé en base de données et recouvrant un domaine particulier des connaissances. Pour les fins du présent Règlement, l'expression Banque de données inclut les banques de matériel biologique, de sujets et de données personnelles, médicales ou génétiques ».

« Banque de matériel biologique » :

- « Ensemble structuré ou non de toutes parties du corps, qu'il s'agisse d'organes, de tissus, de liquides organiques, d'ADN, de cellules, de gamètes, d'embryons ou de fœtus ».

« Banque de données de sujets » :

- « Ensemble structuré ou non d'informations constitué à partir de dossier de patients pour des fins de recrutement éventuel pour des projets de recherche futurs.

« Banque de données médicales » :

- « Ensemble structuré ou non d'informations médicales et, le cas échéant, des informations sociales ou administratives connexes relatives à des individus identifiés ou identifiables ».

« Banque de données personnelles » :

- « Ensemble structuré ou non d'informations personnelles (ex. équilibre, force, nutrition, etc.) relatives à des individus identifiés ou identifiables.

« Banque de données génétiques » :

« Ensemble structuré ou non de spécimens humains (ADN, cellules ou tissus) ou d'informations personnelles à caractère génétique ou protéomique issus de sources diverses et auxquels peut s'ajouter l'information provenant des dossiers médicaux et autres dossiers de santé, de l'information généalogique, socio-économique ou environnementale, qui existent de façon autonome ou en relation avec d'autres sources d'information ». **(Règlement adopté par le conseil d'administration du Centre de santé et de services sociaux –)**

- **Autres définitions :**

Les banques de données et de matériel biologique constituées à des fins de recherche sont une ressource inestimable pour la recherche en santé; leur nombre grandissant en témoigne. Aux fins du présent document, une banque de données et de matériel biologique nécessitant l'élaboration d'un cadre de gestion suppose qu'il y ait collecte, conservation et distribution de son contenu et que ce contenu servira à plusieurs recherches dans une perspective de pérennité. Par opposition, un chercheur qui recueille du matériel biologique ou qui met sur support informatique des données aux fins de la gestion ou de l'organisation d'une seule recherche ne constitue pas une banque.

Il s'agit toutefois d'un domaine qui soulève de nombreux enjeux en raison des différentes valeurs qui sont en cause. D'une part, il y a la protection, le respect et la dignité des participants et, d'autre part, la promotion de la recherche et l'utilisation des ressources disponibles pour faire avancer les connaissances

Considérant l'importance de ces valeurs, il est essentiel que la banque de donnée et de matériel biologique soit dotée d'un cadre de gestion. Ce dernier est nécessaire pour assurer la bonne gouvernance de la banque. Il précise la mise en oeuvre et la logistique de la banque ainsi que les considérations éthiques faites par le ou les chercheurs responsables de la banque. Pour ces derniers, le cadre de gestion constitue un document de référence indispensable pour gérer la banque suivant les considérations éthiques et les objectifs déterminés au départ.

Donc une banque de données représente l'ensemble des informations mémorisées par un ordinateur concernant un domaine scientifique économique ou culturel donné et cela d'une façon aussi exhaustive que possible. **(Guide d'élaboration des cadres de gestion des banques de données et de matériel biologique constituées à des fins de recherche ;Unité de l'Éthique ;, Octobre 2012)**

II.2. Caractéristique d'une base de données :

- **Décrire des bases de données :**

Il est possible de décrire une base de données selon les règles de description des documents d'archives. Cependant, du fait de sa complexité, elle demandera davantage de précision. Il existe en effet plusieurs niveaux de description : de son utilisation générale jusqu'à la description technique de chaque élément (table, champ, relations...).

Pour cela, nous nous inspirons de la méthode Merise2 qui distingue 3 niveaux de description:

- ✓ Le modèle conceptuel des données (MCD) :

Il s'agit de décrire des entités (ensemble d'objets ayant des attributs identiques) et des relations (association ou actions entre les entités). Chaque objet doit être identifiable.

- ✓ Le modèle logique des données (MLD) :

Une fois le MCD établi, on peut le traduire en différents systèmes logiques, comme un ensemble de fichiers binaires, un fichier XML particulier ou au sein d'un SGBD.

Il est relativement simple de transposer un modèle conceptuel de données en un modèle logique dédié au SGBD (ou MLBD) : les entités correspondent en fait à des tables, les attributs des entités à des attributs de tables et les relations à des associations entre tables.

- ✓ Le modèle physique des données (MPD) :

Le MPD est une implémentation particulière du MLD pour un SGBD particulier. Il s'exprime en SQL avec des déclinaisons spécifiques au SGBD. On s'appliquera alors à choisir les bons types d'attributs pour les données identifiées dans le MLD, les bonnes relations, etc.…

L'intérêt du modèle conceptuel des données est de voir quels sont les concepts de haut niveau auxquels la base doit répondre. La séparation entre le MLD et le MPD permet entre autres le portage d'une base de données d'un SGBD vers un autre. Par exemple, on peut traduire un MPD MySQL en un MLDR puis traduire ce MLDR en un MPD postgres.

- Les caractéristiques d'une base de données

Dans un contexte d'archivage, il est important de pouvoir caractériser l'usage qui est fait de la base de données. Ces caractéristiques seront de précieux indicateurs pour permettre par la suite d'identifier la meilleure stratégie d'archivage. Dans la mesure où il n'existe pas réellement de vocabulaire standard, nous utilisons ici des termes qui nous sont propres.

Une base de données est dite « vivante » si les éléments qui la constituent sont modifiés ou que de nouveaux éléments sont ajoutés. On parlera de base de données « figée » si aucune modification, ajout ou effacement n'ont été effectués récemment.

Une base de données est fortement « consultée » si un grand nombre de consultations est fait sur les données qu'elle contient.

Une base de données est dite « cumulative » si on ne fait qu'ajouter de nouveaux éléments sans en modifier et sans en effacer. De manière inverse, on parlera de base de données « dynamique» si l'ajout et la modification sont autorisés et utilisés.

Au delà des ces notions, il est important d'avoir conscience de l'environnement d'exploitation de la base de données. Une base de données n'a lieu d'être que si elle est liée à des applicatifs, à des établissements et au final à des personnes. Il sera nécessaire d'analyser et de cerner l'impact de cet environnement pour en intégrer tous les éléments nécessaires à l'exploitation future de la donnée. Ces éléments constitueront alors des informations de représentation (au sens du modèle OAIS) qui aideront le futur utilisateur à comprendre et exploiter la base.

II.3.Rôle de basses de données :

- ✓ Assurer le stockage informatisé organisation de l'enregistrement sur la mémoire secondaire (disques) garantie de pérennité des données même en cas de panne technique
- ✓ Prendre en compte la structure des données stockées avec et selon leur schéma de structuration garantie de cohérence des données
- ✓ Permettre des utilisations simultanées et autorisées contrôle d'accès et gestion de la concurrence des opérations garantie de confidentialité et d'intégrité des données

II.4. Classification d'une base de données bibliographique :

Les bases de données bibliographiques sont utilisées par les chercheurs pour effectuer des recherches documentaires, faire un état de l'art, etc. Mais d'autres applications sont possibles afin de faire émerger de la connaissance sur l'état de la production scientifique d'un pays ou

d'un continent dans un domaine donné. L'accès aux résultats de la recherche et à l'information scientifique et technique est en train de changer de paradigmes et ce grâce au " libre accès " et plus particulièrement au phénomène d'auto-archivage des travaux de recherche dans des archives ouvertes. Face à ce constat, la présente communication situe l'INIST-CNRS et la base PASCAL avec sa composante en sciences médicales comme un outil de valorisation de la production scientifiques, issue des pays Africains et d'offrir des services gratuits concernant l'information sanitaire. Nous montrerons également comment un outil d'interrogation et d'analyse de l'information comme la plateforme Stanalyst permet d'analyser finement les publications scientifiques et d'être un outil pour un large public allant du documentaliste au décideur en passant par le veilleur et le scientifique (ou le praticien) **(https://www.researchgate.net/ consulté le 06/02/2022)**

- **Bases de données bibliographiques**

Les bases de données bibliographiques sont des outils structurés, complets, performants, en accès libre ou payant.

Les références bibliographiques décrivent de façon détaillée (auteurs, titre, résumé, mots-clés, source, etc.), les publications qui ont été sélectionnées en fonction du domaine et de la ligne éditoriale de la base de données. Ces publications peuvent être : des articles, des ouvrages ou des chapitres d'ouvrages, des actes de congrès ou des communications, des thèses, des rapports, des fiches techniques, des cartes, etc.

Les bases de données bibliographiques sont indispensables pour mener une recherche documentaire de qualité.

Elles peuvent être :

- Scientifiques généralistes : Web of Science, Scopus (payantes), Pascal-Archives, SciELO (gratuites), …
- axées sur l'agriculture et le développement rural: Cab Abstracts (payante), Agritrop (gratuite)
- spécialisés : PubMed, PLOS pour les sciences du vivant (gratuites), EconLit pour l'économie (payante), Isidore (gratuite), EconLit (payante) pour les sciences humaines et sociales, …
- ou encore plus spécialisées : FSTA pour l'agro-alimentaire, ZentralblattMath pour les mathématiques (payantes) **(https://coop-ist.cirad.fr/trouver-l-information) …**

Ainsi, les bases de données bibliographiques sont constituées d'un ensemble structuré de références bibliographiques sur un sujet, un domaine, un type de document, etc. Elles peuvent contenir une analyse, un résumé et de plus en plus souvent l'accès au texte intégral du document lui-même.

- **Bases de données de séquence nucléiques ou protéiques**

Les trois principales bases de données de séquences nucléotidiques sont **GenBank**, **EMBL** (European Molecular Biology Laboratory) et **DDBJ** (DNA Data Bank of Japan). Ces trois bases de données publiques contiennent toutes les séquences nucléiques et protéiques connues, avec annotations bibliographiques et biologiques. Des échanges journaliers permettent d'assurer que les trois bases sont à jour. Jusqu'à récemment, la source principale de données de ces trois bases a été les soumissions directes de séquences par les chercheurs et les échanges journaliers entre bases. Maintenant, la plus grande part de données provient des centres de séquençage, avec une part croissante de génomes complets et d'EST (Expressed Sequence Tags). Après avoir doublé tous les 18 mois, le nombre de séquences double actuellement tous les 15 mois. En août 1998, les 2,5 millions d'entrées représentaient 1,8 milliard de bases. Deux génomes complets avaient été ajoutés en 1996, 6 en 1997 et 10 en 1998, dont le génome de Caenorhabditis elegans, dont la taille est de 100 Mégabases. Environ 20 microorganismes sont actuellement en cours de séquençage et la plupart des résultats sont attendus pour 1999. Plus de 40 000 espèces différentes sont représentées et environ 900 sont ajoutées par mois. Les séquences humaines représentent 54 % des entrées.

Le type d'analyse le plus fréquent effectué sur les bases de séquences nucléotidiques est la recherche de séquences similaires à une séquence donnée. La recherche des meilleurs alignements est réalisée à l'aide de la famille de programmes BLAST. Chaque alignement BLAST est accompagné d'un score et d'une indication de la valeur statistique.

Traditionnellement, les séquences soumises aux bases de données étaient des fragments étudiés par des chercheurs en raison de leur intérêt direct en relation avec un sujet biologique donné : structure de gènes et de familles de gènes, étude de leurs régions régulatrices, phylogénie, étude des séquences répétées. Les soumissions correspondaient à des séquences lues sur les deux brins d'ADN, avec annotations précises par les auteurs

Actuellement, une large part des entrées proviennent de projets systématiques par lesquels une grande quantité de séquences est produite, dont l'analyse du contenu biologique est réalisée

automatiquement (annotation automatique), ou pour lesquelles l'information de séquence ne sert que d'outil pour la construction de cartes. **(https://www.ebi.ac.uk/)**

II.5. Contenues des bases de données biologiques :

Rappelons qu'une base de données ou bien l'ensemble de base de données se définit comme des bibliothèques électronique et informatisé qui contiennent des informations sur les sciences de la vie, collectées grâce à des expériences scientifiques, à la littérature publiée, aux technologies expérimentales à haut débit, et aux analyses informatiques. 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. Entre autres ils ont pour mission l'archivage, le stockage, la diffusion et l'exploitation des données biologiques.

Ces bases de données peuvent contenir des informations : (ADN, protéines, gènes et génomes, taxonomie, autres, …etc.). On y trouve également une bibliographie et une expertise biologique directement liées aux séquences traitées.

II.6. Les types de banques de données :

Il existe un grand nombre de bases de données d'intérêt biologique. Nous nous limiterons dans ce chapitre à une présentation des principales banques de données publiques, basées sur la structure primaire des séquences, qui sont largement utilisées dans l'analyse informatique des séquences

Nous distinguerons deux types de banques, celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations (**banques de données généralistes**) et celles qui correspondent à des données plus homogènes établies autour d'une thématique (**banques de données spécialisées**) et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe de scientifiques

- **Les banques de données généralistes ;**

Ces banques contiennent des données hétérogènes :

- Collecte la plus exhaustive possible
- Banques de séquences nucléiques
- Banques de séquences protéiques

- Avantage : tout est consultable en une fois
- Inconvénients : difficiles à maintenir, difficiles à interroger

➢ **Les banques de données spécialisées ;**

Ces banques contiennent des données homogènes

– Collecte établie autour d'une thématique particulière

- Avantages : facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée, …
- Inconvénients : ne cible pas toujours ce que l'on veut ; toutes les banques possibles n'existent pas
- Exemples : banques spécialisées pour un génome, banques de séquences d'immunologies, banques sur des séquences validées

➢ **Les banques de séquences nucléiques**

- Origine des données

– Séquençage d'ADN et d'ARN

→ Les données stockées : séquences + annotations

– Fragments de génomes
  - Un ou plusieurs gènes, un bout de gène, séquence intergénique, …
  - Génomes complets
  - ARNm, ARNt, ARNr, … (fragments ou entiers)

✓ [Remarque 1] : toutes les séquences (ADN ou ARN) sont écrites avec des T

✓ [Remarque 2] : les séquences sont toujours orientées 5' → 3'.

➢ **Les banques de séquences protéiques**

→ Origine des données

– Traduction de séquences d'ADN

– Séquençage de protéines (Rare car long et coûteux)

– Protéines dont la structure 3D est connue

→ Les données stockées : séquences + annotations

– Protéines entières

– Fragments de protéines

II.7. Les bases de données bio-informatiques les plus utilisées :

- ✓ **Le portail Entrez ou NCBI :**
- GenBank : Séquences d'ADN et d'ARN
- Site officiel de BLAST
- PubMed: Permet la recherche d'articles scientifiques
- COGs: Familles de gènes orthologues …
- ✓ Le **portail EMBL : The European Molecular Biology Laboratory**
- ✓ **ExPASy : Expert Protein Analysis System, Protéomique**
- UniProt : Séquences de protéines
- PROSITE : Domaines et familles de protéines
- SWISS-MODEL : Outil de prédiction 3D de protéines
- Différents outils de recherche
- ✓ **PDB : Protein Data Bank**
  - Base de données de structures 3D de protéines
  - Visualisation et manipulation de structures
- ✓ **SCOP : Structural Classification**

CONCLUSION

A travers ce chapitre, nous avons mis en lumière les principales notions du stockage de la bio-informatique. Tout d'abord, nous avons défini les banques de données et présenté les bases de donnés qui doivent permettre le stockage d'une quantité croissante d'informations dont la nature est hétérogène. Des entités parfois très différentes sont à représenter, ainsi que leurs relations. De plus, les techniques variées utilisées pour générer les données doivent être prises en compte. De ce fait, il existe un foisonnement de bases de données spécialisées, dont nous décrivons ici les principales, ainsi que leurs interrelations.

CONCLUSION GENERALE

CONCLUSION GENERALE

Le bon fonctionnement de la biologie sollicite des moyens informatiques puissants et efficaces. En effet, la bio-informatique est un champ de recherche multidisciplinaire de la biotechnologie où travaillent de concert biologistes, médecins, informaticiens, mathématiciens, physiciens et bio-informaticiens, dans le but de résoudre un problème scientifique posé par la biologie. Plus généralement, la bio-informatique est l'application de la statistique et de l'informatique à la science biologique.

Dans ce sens, notre travail de recherche a permis de mettre l'accent sur l'outil bio-informatique utiles en biologie : banque et base de données. Donc, au cours des trois premiers chapitres consacrés au volet théorique, nous avons mis l'accent sur la bibliographie, la méthodologie, le stockage de la bio6informatique. En revanche, le premier chapitre intitulé « bibliographie » décrit les principes de base de bio-informatique, suivi d'une étude détaillée des acides nucléiques et les protéines. Ensuite, le deuxième chapitre donne un aperçu général sur la méthodologie, et enfin, le dernier chapitre présente le stockage de la bio-informatique. La description sera subdivisée en deux grandes parties. La première décrira les banques de données. La deuxième partie, quant à lui, concernera la description des bases de données.

Pour conclure, nous espérons avoir apporté, tout au long de ce modeste travail, les éléments de réponse attendus à la problématique posée dans ce mémoire et avoir éclairci le mécanisme de l'outil bio-informatique utiles en biologie

BIBLIOGRAPHIE

BIBLIOGRAPHIE

❖ Les ouvrages :

1- Chen, Yangho, Tade Souaiaia, and Ting Chen. 2009. " PerM: Effic ient Mapping of Short Sequencing Reads with Peri odic Full Sensitive Spaced Seeds." Bioinformatics 25 (19): 25 14- 2 1. doi: 1 0.1 093/bioinformatics/btp486.

2- Li, Guoli ang, Metissa J Fullwood, Han Xu, Fabi anus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila uwantha Ariyaratne, et al. 201 O. "ChiA-PET Tool for Comprehensive Chromatin Interaction Analys is with Paired-End Tag Sequencin g." Genome Bio/ogy 11 (2). England: R22. doi: 1 0.1186/gb-20 10-1 l-2-r22.

3- Li, Guoli ang, Metissa J Fullwood, Han Xu, Fabi anus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila uwantha Ariyaratne, et al. 201 O. "ChiA-PET Tool for Comprehensive Chromatin Interaction Analys is with Paired-End Tag Sequencin g." Genome Bio/ogy 11 (2). England: R22. doi: 1 0.1186/gb-20 10-1 l-2-r22.

4- Li, Heng, and Richard Durbin. 2009a. "Fast and Accurate Short Read Alignment with BurrowsWheeler Transform ." Bioinformatics. doi: 10 .1 093/ bioinformatics/btp324. ---. 201 O. "Fast and Accu rate Long-Read Alignment with Burrows-Wheeler Transform ." Bioinformatics (Oxford, Eng/and) 26 (5). England : 589- 95 . doi: 10 .1 093/b ioinfo rm ati cs/btp698.

5- Li, Heng, and Nils Homer. 201 O. "A Survey of Sequence Alignment Algorithms for next-Generation Sequencing." Briefings in Bioinformatics 1 1 (5): 4 73- 83. doi: 1 O. 1 093/bib/bbqO 15 .

6- Nico l, John W, Gregg A Helt, Steven G Blanchard, Archana Raja, and AnnE Loraine. 2009. "The lntegrated Genome Browser: Free Software for Distribution and Exploration ofGenome-Scale Datasets." Bioinformatics. doi: 10 .1 093/bioinformatics/btp472.

7- Pl anet, Evarist, Camill e Stephan-Otto Attolini, Oscar Reina, Oscar Flores, and Dav id Rossel!. 2012. " htSeqTools: High-Throughput Sequencing Quality Control, Processing and Visuali zation in R." Bioinformatics (Oxford, England) 28 (4). England : 589-90. doi: 10 .1 093/bioinformati cs/btr700.

8- Podi cheti , Ram, and Qunfe ng Dong. 2011. "Administering GBrowse Sites with WebGBrowse." Current Protoco/s in Bioinformatics 1 Editora/ Board, Andreas DBaxevanis ... [et Al.} Chapter 9 (March). United States: Unit 9. 14. doi: 10 .1002/0471 25 0953 .bi09 14s33 .

❖ Textes réglementaires :

1. Règlement adopté par le conseil d'administration du Centre de santé et de services sociaux – Institut universitaire de gériatrie de Sherbrooke,

❖ Mémoires et cours :

1. Études scientifiques mandatées par le SECO sur le thème de la cyberéconomie ; Sieber & Partners, Wikipédia
2. Cours Bio Informatique, Faculté des Sciences Exactes et Appliquées / M1- CTC 2019/2020, Université Ahmed Ben Bella 1
3. Juliana Silva Bernendes ;Hugue Richards ; introduction à la bio informatique
4. Cours de biologie moléculaire faculté de médecine de Batna : 2015/2016
5. Introduction à l'analyse génétique Griffith S. Wessler. 4eme édition de boeck
6. Guide d'élaboration des cadres de gestion des banques de données et de matériel biologique constituées à des fins de recherche ;Unité de l'Éthique ;Octobre 2012

❖ La Webographie (consultée entre Janvier- MArs2022)

1. https://www.lalanguefrancaise.com
2. https://www.google.com/
3. https://coop-ist.cirad.fr/trouver-l-information
4. https://www.ebi.ac.uk/
5. www.aquaportail.com/definition-531-arn.html
6. https://www.etudier.com
7. www.aquaportail.com/definition-530-adn.html
8. https://scholar.google.com

ANNEXES :

La **transcriptomique**	La transcriptomique regroupe un ensemble de techniques permettant une analyse quantitative relative des ARN (comparaison des transcriptomes entre différentes conditions expérimentales) et une analyse qualitative par la caractérisation de variant d'épissage, de polymorphismes, de gène fusion etc…
ADN	Macromolécule de poids moléculaire élevé, formée de polymères de nucléotides dont le sucre est le 2-désoxyribose. Il se présente so us forme d'une double chaîne hé licoïdale dont les deux brins sont complémentaires. Il constitue le génome de la plupart des organi smes vivants.
ARN	Macromolécule formée par la polymérisation de nombreux nucléotides dont le sucre est le ribose, présente dans le cytoplasme, les mitochondries ainsi que dans le noyau cellulaire, et servant d'intermédiaire dans la synthèse des protéines.
ARN messager (ARNm)	Copie transitoire d'une portion de l'ADN correspondant à un ou plusieurs gènes utilisé comme intermédiaire par les cellules pour la synthèse des protéines.
BLAST	Méthode de recherche heuristique utilisée en bio-informatique permettant de rechercher les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés.
Nucléotide	Molécule organique composée d'une nucléobase, d'un pentose et de 1 à 3 groupements phosphates. Certains nucléotides forment la base de l'ADN et de I'ARN.
Polymérase	Enzymes qui ont pour rôle la synthèse d'un brin de poly nucléotide (ADN ouARN), le plus souvent en utilisant un brin complémentaire comme matrice et des nucléotides triphosphosphate (NTP ou dNTP) comme monomères
Ribosome	Complexes ribonucléoprotéiques (c'est-à-dire composés de protéines et d'ARN) présents dans les cellules eucaryotes et procaryotes. Leur fonction est de synthétiser les protéines en décodant l'information contenue dans I'ARN messager.