BIOINFORMATIQUE

POLYCOPIE DE COURS

MASTER : BIOCHIMIE / SEMESTRE : S1
INTITULE DE L'UE : UEM1
CREDITS : 5 / COEFFICIENTS : 3

MASTER : TOXICOLOGIE / SEMESTRE : S1
INTITULE DE L'UE : UEM1
CREDITS : 4 / COEFFICIENTS : 2

MASTER : MICROBIOLOGIE APPLIQUEE / SEMESTRE : S2
INTITULE DE L'UE : UEM1
CREDITS : 4 / COEFFICIENTS : 2

Année universitaire 2019/2020

Ver. MARS/2020

Graphique page de garde issue du site web de la 6ème école de bioinformatique AVIESAN-IFB. http://www.france-bioinformatique.fr/fr/evenements/EBA2017

Préparé par : M. AMARA KORBA R.
Université Mohamed El Bachir El Ibrahimi - Bordj Bou Arréridj -
Faculté SNV-STU
Département des Sciences Biologiques
Ver. 03/2020

Table des matières

Liste des abréviations 5
Chapitre I. Introduction à la Bioinformatique 6
I.1. Qu`est-ce que la bioinformatique 7
I.2. Les différentes facettes de la bioinformatique 8
I.2.1. Compilation et organisation des données 8
I.2.2. Traitements systématiques des séquences (l'annotation des séquences) 8
I.2.3. Bioinformatiques et logiciels 8
I.2.4.1. Les outils lignes de commandes 9
I.2.4.2. Les Outils Web (Web-Based Software) 9
I.3. Domaines de la bioinformatique 10
I.4. Historique de la bioinformatique 11
I.5. Projet de recherche en bioinformatique et en génomique 11
I.6. Rappel en biologie moléculaire du gène 13
I.6.1. L'ADN 13
I.6.2. Génomique et bioinformatique 13
I.6.3. Le génome 15
I.6.3.1. Le génome des procaryotes 15
I.6.3.2. Le génome des eucaryotes 16
I.6.3.3. Composition des chromosomes 16
I.7. Le travail en bioinformatique 17
Chapitre II. Techniques d'acquisition des données biologiques 18
II.1. Les techniques de séquençage 19
II.1.1. Les premières techniques de séquençage 19
II.1.1.1. La technique de Maxam et Gilbert 19
II.1.1.2. La technique de Sanger 19
II.1.1.3. Automatisation de la méthode de Sanger 22
II.1.2. Les nouvelles techniques de séquençage (NGS) 24
II.1.2.1. La technique de pyroséquençage 24
II.1.2.2. La technique de séquençage Illumina/Solexa 27
II.1.3. Les techniques de séquençage de 3ème génération 29
II.1.3.1. La technologie SMRT sequencing 29
II.1.4. Comparaison des techniques 29
Chapitre III. Les bases de données biologiques 32
III.1. Définition 33
III.2. Le rôle des banques et bases de données biologiques 33
III.3. Contenues des bases de données biologiques 33
III.4. Les types de banques de données 33
III.4.1. Les banques de données généralistes 35
III.4.2. Les banques de données spécialisées 35
III.4.3. Les banques de séquences nucléiques 35
III.4.4. Les banques de séquences protéiques 35
III.5. Structuration et organisation 36
III.5.1. Fichiers et formats 36
III.5.1.1. Le format FASTA 37
III.5.1.2. Le format EMBL 38
III.5.1.3. Le format Genbank 41
III.5.2. La qualité des données 42

III.6. Interrogation des bases de données....43
Chapitre IV. Exploitation et Analyse des données (Annotation)....47
IV.1. Introduction....48
IV.2. Les algorithmes....48
IV.2.1. Définition....48
IV.2.2. Algorithme, Programme, Logiciel, structures de données....48
VI.2.3. Les propriétés d'un algorithme....49
VI.2.4. L'écriture d'un algorithme....49
VI.2.4.1. La déclaration des variables....50
VI.2.4.2. Initialisation et affectation de valeurs....50
VI.2.4.3. Les instructions de contrôle....50
VI.2.4.4. Les instructions d'affichage....51
VI.2.5. L'algorithme DNA Walk....53
IV.3. L'annotation des séquences....57
IV.3.1. Annotation syntaxique : la recherche d'objets génétiques....58
IV.3.1.1. La recherche de signaux de séquence codante chez les procaryotes....58
IV.3.1.2. La recherche de signaux de séquence codante chez les eucaryotes....61
IV.3.1.3. Analyse du contenu en base des séquences codantes....64
IV.3.1.4. Programmes bioinformatiques d'annotation syntaxique....65
IV.3.2. Annotation fonctionnelle : la recherche de fonctions potentielles....66
IV.3.2.1. Les outils bioinformatique de comparaison de séquence....66
Références bibliographiques....68

Liste des abréviations

BLAST : Basic Local Alignment Search Tool
CCD : Charge Coupled Device
CDS : Coding Sequence
CNRS : Centre National de la Recherche Scientifique
EBI : The European Bioinformatics Institute
EMBL : The European Molecular Biology Laboratory
HMM : Hidden Markov Model
INRA : Institut national de la recherche agronomique
IP : Institut Pasteur de Paris
Kb : Kilobases
Mb : Mégabases
NCBI : The National Center for Biotechnology Information
NGS : Next Generation Sequencing
ORF : Open Reading Frame
Pb : Paires de Bases
PCR : Polymerase Chain Reaction
PDB : Protein Data Bank
PIR : Protein Information Resource
RBS : Ribosome Binding Site
SMRT : Single Molecule Real Time sequencing
SMS : Single Molecule Sequencing
SOLiD : Sequencing by Oligo Ligation and Detection
UCSC : University of California Santa Cruise

Chapitre I. Introduction à la Bioinformatique

I.1. Qu`est-ce que la bioinformatique

À l'interface entre biologie, informatique et mathématiques, la bioinformatique analyse et interprète, au moyen de méthodes informatiques, les données biologiques que sont les séquences des gènes et des protéines cellulaires, et apporte ainsi de nouvelles connaissances sur le fonctionnement des cellules et des organismes vivants [1].

La bioinformatique est définie comme l'utilisation de bases de données et d'algorithmes informatiques pour analyser, les gènes, les protéines, et la collection complète d'acide désoxyribonucléique (ADN) d'un organisme vivant (le génome) [2].

Un défi majeur en biologie consiste à comprendre les énormes quantités de données de séquence et de données structurelles générées par les expériences biologiques (ex : les projets de séquençage) [2].

Les outils de la bioinformatique comprennent des programmes informatiques qui aident à révéler les mécanismes fondamentaux à la base des problèmes biologiques liés à la structure et fonction des macromolécules, des voies biochimiques, des processus pathologiques et évolutive [2].

Selon l'Institut national de recherche sur le génome humain (***NHGRI : N****ational* ***H****uman* ***G****enome* ***R****esearch* ***I****nstitute*), « la bioinformatique est la branche de la biologie qui se préoccupe de l'acquisition, du stockage, de l'affichage et de l'analyse des informations contenues dans les données sur les séquences d'acides nucléiques et de protéines » [2].

La bioinformatique est l'approche **« *in silico* »** de la biologie qui consiste en une analyse informatisée des données biologiques en utilisant un ensemble de moyens :

- 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.

C'est une discipline complémentaire aux approches classiques de la biologie :

- *In vivo* (tests au sein des organismes vivants) ;
- *In situ* (tests dans les milieux naturels) ;
- *In vitro* (tests dans des tubes).

I.2. Les différentes facettes de la bioinformatique

Pour l'analyse des données expérimentales que représentent les séquences biologiques, cet apport informatique concerne principalement trois aspects [3] :

I.2.1. Compilation et organisation des données

Cet aspect concerne essentiellement **la création de bases de données**. Certaines ont pour vocation de réunir le plus d'informations possible **(bases de données généralistes)** sans expertise particulière de l'information déposée alors que d'autres sont spécialisées dans un domaine considéré avec l'intervention d'experts **(bases de données spécialisées)** [3].

Les banques de données spécialisées sont généralement construites autour de thèmes précis comme l'ensemble des séquences d'une même espèce ou les facteurs de transcription. Incontestablement, toutes ces banques de données constituent une source de connaissance d'une grande richesse que l'on peut exploiter dans le développement de méthodes d'analyse ou de prédiction [1].

I.2.2. Traitements systématiques des séquences (l'annotation des séquences)

L'objectif principal est de **repérer ou de caractériser une fonctionnalité** ou **un élément biologique intéressant**. Ces programmes représentent les traitements couramment utilisés dans l'analyse des séquences comme l'identification de phases codantes (CDS) sur une molécule d'ADN ou la recherche de similitudes d'une séquence avec l'ensemble des séquences d'une base de données [3].

I.2.3. Bioinformatiques et logiciels

Il est maintenant facile et courant d'effectuer certaines opérations plus ou moins complexes à **l'aide de logiciels** plutôt que manuellement. Pourtant, ces pratiques ne sont pas toujours systématiques car il est souvent difficile pour certains utilisateurs de savoir quel programme utiliser en fonction d'une situation biologique déterminée ou d'exploiter les résultats fournis par une méthode [2].

C'est pourquoi ce cours contient la présentation d'un certain nombre d'outils ou de méthodes couramment utilisés et reconnus dans l'analyse informatique des séquences.

Remarque : Ce cours ne constitue en aucun cas un exposé exhaustif de tout ce qui existe dans le domaine de la bioinformatique.

Il existe deux approches totalement différentes de la bioinformatique : l'utilisation d'outils Web (WEBicielles) et de lignes de commande (Figure 1).

I.2.4.1. Les outils lignes de commandes

Ces outils peuvent être difficile a utilisé pour la plupart des biologistes, mais offrent presque toujours plus d'options pour l'exécution des programmes. Ils sont plus appropriés pour analyser des ensembles de données à grande échelle qui sont rencontrés actuellement en bioinformatique.

I.2.4.2. Les Outils Web (Web-Based Software)

Les outils Web, parfois appelés « point-and-click », ne nécessitent pas de connaissances en programmation et sont immédiatement accessibles à la communauté scientifique.

Le domaine de la bioinformatique s'appuie fortement sur Internet pour accéder aux données de séquence, aux logiciels utiles pour analyser les données moléculaires et pour intégrer différents types de ressources et d'informations relatives à la biologie. Nous allons décrire une variété de sites Web. Dans un premier temps, nous nous concentrerons sur les principales bases de données accessibles au public qui servent de référentiels pour les données sur l'ADN et les protéines. Ceux-ci comprennent :

- Le Centre national d'information sur la biotechnologie (**NCBI**), qui héberge la GenBank et d'autres ressources ;
- L'Institut européen de bioinformatique (**EBI**) ;
- Ensembl, qui comprend un navigateur génomique et des ressources pour étudier des dizaines de génomes ;
- Le site de bioinformatique du génome de l'Université de Californie à Santa Cruz (**UCSC**), comprenant un navigateur Web et un navigateur de tableaux pour diverses espèces.

À travers les chapitres de ce cours, nous présentons plusieurs sites Web supplémentaires concernant la bioinformatique. Les principaux avantages offerts par les sites Web sont un accès facile, des mises à jour rapides, une bonne visibilité pour la communauté scientifique et une facilité d'utilisation (étant donné que les compétences de d'algorithmique et de programmation ne sont pas nécessaires).

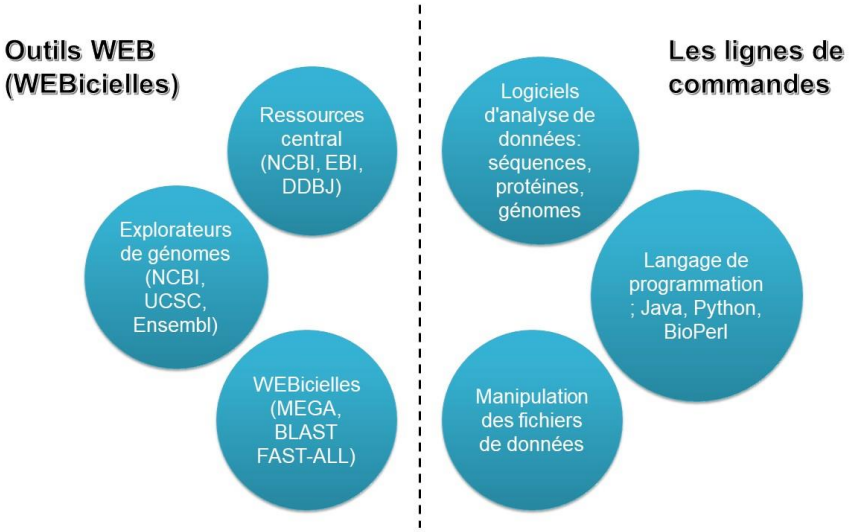


Figure 1. Ressources en bioinformatique. Les ressources **"Web"** ou **"point-and-click"** apparaissent à gauche, notamment les principaux portails (NCBI, EBI, …), les principaux navigateurs génomiques (Ensembl, UCSC), des bases de données et des sites Web spécialisés. Les ressources **"lignes de commandes"** sont affichées à droite. Celles-ci incluent les langages de programmation utilisant des algorithmes appropriée (tels que R, C++, Python, BioPerl et Java) et les logiciels de ligne de commande. (*Modifiée à partir de Pevsner, 2015*).

I.3. Domaines de la bioinformatique

- Stockage et Gestion des données : Banques de données généralistes et spécialisées.
- Structures moléculaires : Visualisation, analyse, classification, prédiction.
- Analyse de séquences : Alignements, recherches de similarités, détection de motifs.
- Génomique structurale : Annotation des génomes, génomique comparative.
- Génomique fonctionnelle : Transcriptome, protéome, interactome.
- Phylogénie : Relations évolutives entre gènes, entre génomes, entre organismes ; Inférence de scénarios évolutifs.
- Analyse des réseaux biomoléculaires : Réseaux métaboliques, d'interactions protéiques, de régulation génétique, …

I.4. Historique de la bioinformatique

Le terme de « *Bioinformatique* » n'est apparu dans la littérature scientifique qu'au tout début des années 90. Cependant, ce domaine de recherche ne vient pas d'émerger. Le tableau 1 retrace les grandes étapes de la bioinformatique, et montre à quel point cette discipline a accompagné et souvent précédé le développement des concepts biologiques et des outils informatiques sur laquelle elle est fondée [3].

I.5. Projet de recherche en bioinformatique et en génomique

- Cancer Genome Atlas : Cartographier le génome pour plus de 25 types de cancers a généré 1 petabyte de données (à ce jour), représentant 7 000 cas de cancer. Les scientifiques n'attendent pas moins de 2,5 petabytes (1petabyte= 1015 bytes = 1000 terabytes).

- Encyclopedia of DNA Elements (ENCODE) : Le catalogue des éléments fonctionnels du génome humain : 15 terabytes de données brutes (1 terabytes = 1000 gigabytes).

- Human Microbiome Project : l'un des projets visant à caractériser le microbiome à différents endroits du corps : 18 terabytes - environ 5 000 fois plus de données que le premier projet « génome humain ».

- Earth Microbiome Project : Caractérisation des communautés microbienne sur la terre : 340 gigabytes (17109 séquences, ~ 20,000 échantillons, 42 biomes). 15 terabytes attendus.

- Genome10K : Volume de données brutes pour le projet de séquençage de 10,000 espèces de vertébrés devrait atteindre 1 petabyte (1 petabyte = 1024 Terabytes).

Tableau 1. Étapes-clés dans l'histoire de la bioinformatique [3].

1951	Première séquence protéique (Insuline, Sanger)
1960	Lien entre séquence & structure (Globines, Perutz)
1965	Premiers Ordinateurs IBM/360
1965	Evolutionary divergence and convergence in Proteins (Zuckerkandl & Pauling)
1967	"Construction of Phylogenetic Trees" Fitch & Margoliash.
1968	Atlas of Protein Sequences (M. Dayhoff, Georgetown)
1968	Mini-ordinateur DEC PDP-8
1970	A general method applicable to the search for similaries in sequences of two proteins (Needleman & Wunsch).
1971	Premiers travaux sur le repliement des ARNs (J. Ninio)
1972	Premier microprocesseur Intel 8008
1973	"Génie Génétique" (Cohen et al.)
1974	"Prediction of Protein Conformation" (Chou & Fasman)
1975	Intel 8080, kit Altair
1977	Mini-ordinateur DEC-VAX.
1977	Micro-ordinateurs (Apple, Commodore, Radioshack
1977	Séquençage d'ADN (Sanger, Maxam, Gilbert)
1977	Premier "package" Bioinformatique (Staden)
1978	Banques de données : EMBL, GenBank, ACNUC, PIR
1980	Accès téléphonique à la base de données PIR
1981	IBM-PC (8088), 16-32kb
1981	Los Alamos-GenBank : 270 séquences, 370.000 nucléotides
1981	Programme d'alignement local (Smith-Waterman)
1983	IBM-XT Disque DUR (10 Mbytes)
1984	MacIntosh : interface graphique & souris
1985-88	Programme "Fasta" (Pearson-Lipman)
1989	INTERNET succède à ARPANET et BITNET
1990	Programme "Blast" (Altschul et al.)
1990	Clonage positionnel et séquençage de NF-1
1991	Grail, programme performant pour localiser les gènes (Mural et al.)
1991	Étiquettes d'ADNc "EST" (Venter et al., Matsubara et al.)
1992	Séquençage complet du chromosome III de levure
1995	Première séquence complète d'un micro-organisme (Venter et al.; H. influenza)
1996	Séquence complète de la levure (consortium européen)
1997	Programme "Gapped Blast" (Alschul et al.)
1997	11 génomes bactériens disponibles
1998	2 Mbase/jour de nouvelles séquences publiques
2001	Séquence ("premier jet") complète du génome humain.

I.6. Rappel en biologie moléculaire du gène

I.6.1. L'ADN

- L'ADN est le **support de l'information génétique**.
- L'ADN est une **longue molécule**, faite de **deux brins** s'enroulant en une **double hélice**.
- Les deux brins de la double hélice suggèrent un mécanisme de réplication de l'ADN
- Chaque brin est le support d'une **succession de nucléotides**
- **Quatre types de nucléotides :** (Adénine, Cytosine, Guanine, Thymine).
- Le texte génomique est écrit dans un alphabet de 4 lettres : A, C, G, T

La structure de l'ADN représenté par la figure 2, les deux brins sont orientés, on parle de l'orientation 5' → 3' par des considérations biochimiques.

- L'extrémité 5' se termine par un groupement phosphate.
- L'extrémité 3' se termine par un groupement hydroxyle.

Il faut retenir qu'il y a un brin sens orienté 5' → 3' et un brin anti-sens orienté 3' → 5', et généralement lorsqu'on veut connaitre la séquence des nucléotides le long d'un brin, on va la lire traditionnellement dans le sens 5' → 3'.

Figure 2. Représentation schématique d'une succession de nucléotides d'un fragment d'ADN double brins.

I.6.2. Génomique et bioinformatique

La génomique au sens large est la science qui étudie les génomes et l'ensemble de leurs gènes :

- La structure
- Le fonctionnement
- L'évolution
- Le polymorphisme, …

Les études en génomique nécessitent des outils bioinformatiques

Au sens strict, on sépare la génomique (**Génomique structurale**) de la post-génomique **(Génomique fonctionnelle)** [4].

Dans ce cas, la génomique *sensu stricto* concerne les informations sur le contenu des génomes, tandis que les analyses post-génomiques concernent la comparaison génomique (variation du contenu et structure des génomes) et l'analyse fonctionnelle des gènes [4].

L'analyse fonctionnelle des gènes peut porter sur [4] :

- Le **transcriptome** (l'ensemble des transcrits ou ARNm) ;
- Le **protéome** (l'ensemble des protéines).

Par la suite, on a défini d'autres notions qui découlent des **analyses post-génomiques** comme le **sécrétome** (l'ensemble des protéines sécrétées), le **métabolome** (l 'ensemble des métabolites produits par une cellule), l'**interactome** (l'ensemble des protéines et/ou acides nucléiques interagissant entre eux), etc… [4].

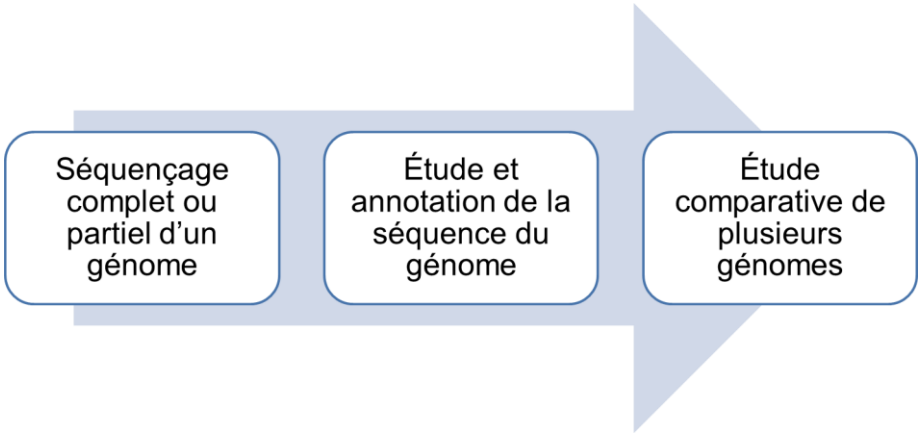


Figure 3. Les différentes étapes du travail effectuée en génomique

Les programmes ou logiciels de bioinformatique sont utilisés :

- À différentes étapes du séquençage des génomes
- Pour la lecture des séquences à la sortie des séquenceurs
- Assemblage des génomes à partir des fragments séquencés
- Pour la recherche des répétitions pour corriger les mauvais assemblages
- Pour la localisation des gènes sur les génomes
- Pour le regroupement des séquences appartenant à un même gène
- Pour comparer les séquences obtenues
- Pour la comparaison 2 à 2, multiple, une séquence contre une banque
- Dans la création et la gestion des banques de données
- Dans la collecte puis stockage des séquences et bien plus …

I.6.3. Le génome

Le génome, c'est-à-dire l'ensemble des gènes d'un organisme, est structurellement défini par les chromosomes au sein de chaque cellule, selon la théorie chromosomique de l'hérédité. Chaque chromosome est constitué d'une molécule d'ADN (Figure 5), support physique des gènes, et de protéines associées [4].

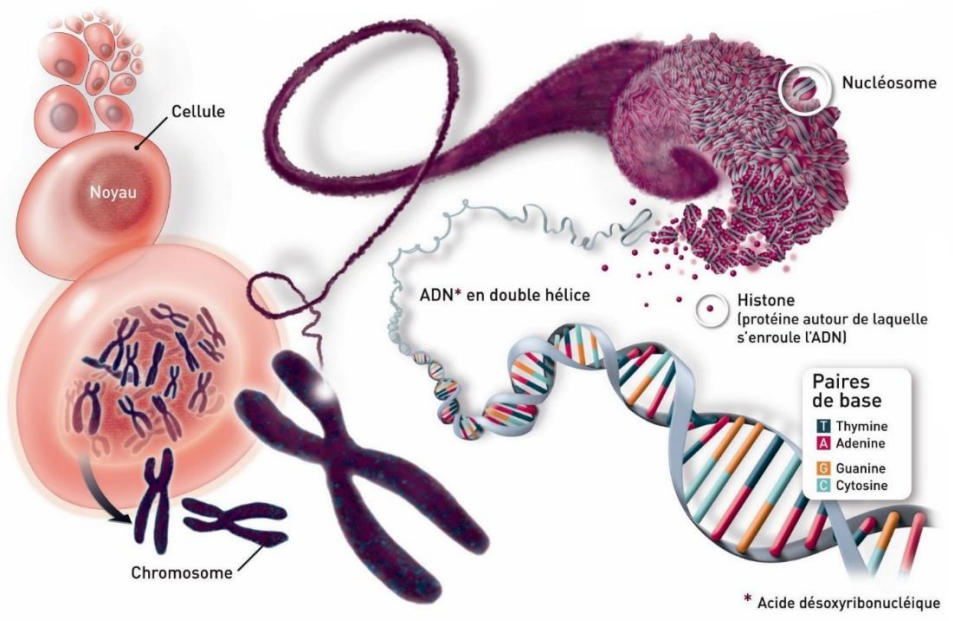


Figure 5. Schéma du génome humain, avec la structure en double hélice de l'ADN. (Source : Premamag.wordpress.com)

I.6.3.1. Le génome des procaryotes

Le matériel génétique des bactéries n'est pas organisé de la même façon que chez les eucaryotes. Toutefois, les gènes bactériens sont disposés linéairement sur le chromosome [4].

Les procaryotes possèdent en général un seul chromosome circulaire, en exemplaire souvent unique. Néanmoins, avec l'utilisation de l'électrophorèse en champ pulsé permettant de séparer les molécules d'ADN de grande taille, il a été montré que certaines bactéries ont un chromosome linéaire ou plusieurs chromosomes circulaires et linéaires [4].

En plus les procaryotes possèdent de l'ADN sous forme extra-génomique. Il s'agit d'une petite molécule circulaire d'ADN, appelée plasmide, capable de se répliquer indépendamment du chromosome [4].

I.6.3.2. Le génome des eucaryotes

Chez les eucaryotes, les génomes sont en fait visualisés comme des structures filamenteuses, non circulaires, situées majoritairement dans le noyau, et qui peuvent présenter des configurations variables suivant le cycle cellulaire (cf. mitose). Il existe également des chromosomes mitochondriaux et chloroplastiques qui sont pour la plupart circulaires. Ceux-ci sont plus petits que les chromosomes nucléaires et ne présentent pas d'aspect filamenteux. Les gènes présents sur ces chromosomes extra-nucléaires ne suivent pas les lois de la transmission mendélienne [4].

I.6.3.3. Composition des chromosomes

Avec l'avènement de l'électrophorèse en champ pulsé permettant de séparer de grandes tailles de molécules d'ADN, il a été admis que chaque chromosome eucaryote est constitué d'une molécule unique et linéaire d'ADN, repliée sur elle-même (voir figure x) [4].

I.6.3.4. Caractéristiques des chromosomes

La différence entre la taille et la formes des chromosomes [4] :

- Les chromosomes sont variables en taille : ainsi chez l'humain, le chromosome 1 est environ 3 à 4 fois plus grand que le chromosome 21 ;
- Concernant la forme, les chromosomes apparaissent avec un « étranglement » au niveau de la région centromérique dont la position est variable ;
- La longueur des 2 bras de chaque chromosome par rapport au centromère est propre à chaque chromosome.

De plus, chaque espèce contient un nombre de chromosomes qui lui est propre. On note n, le nombre de chromosomes dans un ensemble génomique élémentaire ; c'est le nombre **haploïde**. Ce nombre, spécifique à chaque espèce, permet de constituer **le caryotype**, c'est-à-dire la **« carte d'identité »** de l'espèce étudiée [4].

Lorsque les espèces ne présentent qu'un seul jeu de chromosomes, les cellules sont dites **haploïdes** ; c'est le cas de la plupart des champignons et des algues. Lorsque les espèces présentent deux jeux de chromosomes, c'est-à-dire des paires de chromosomes, les cellules sont dites **diploïdes** et sont représentées par 2n : c'est le cas des cellules somatiques de la plupart des animaux et végétaux. Dans le tableau 2 sont représentés des exemples d'espèces en fonction du nombre n de chromosomes [4].

Tableau 2. Exemples du nombre n de chromosomes chez différentes espèces

Espèce	Nombre *n*
Moustique commun (*Culex pipiens*)	3
Riz (*Oryza sativa*)	12
Souris commune (*Mus musculus*)	20
Homme (*Homo sapiens*)	23
Pomme de terre (*Solanum tuberosum*)	24
Chien (*Canis familiaris*)	39

I.7. Le travail en bioinformatique

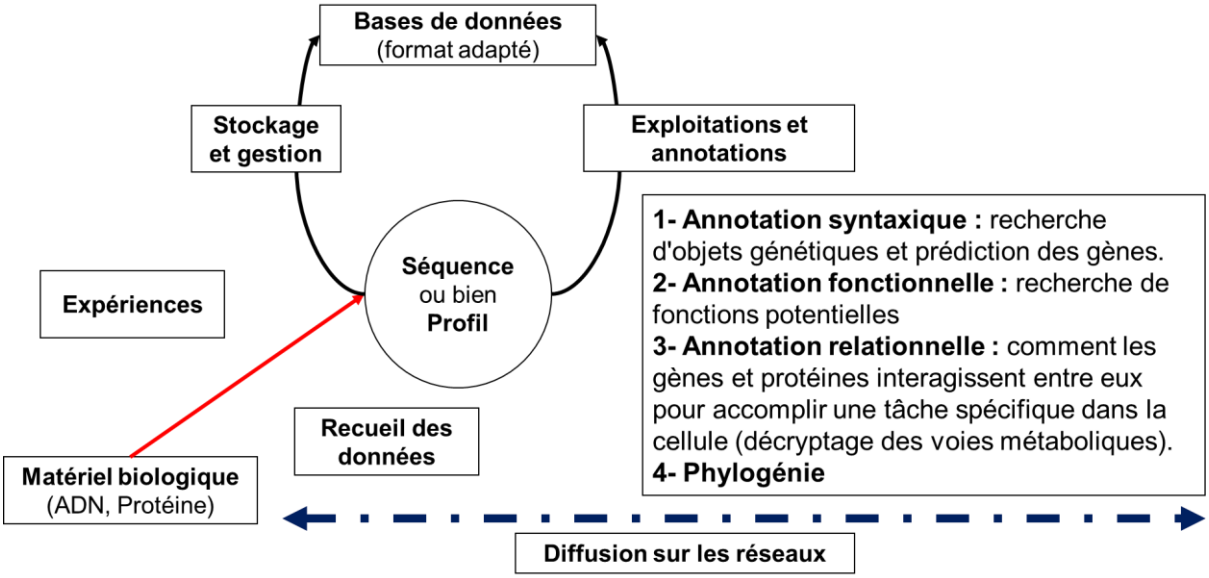


Figure 6. Représentation schématique des procédés utilisées par la bioinformatique

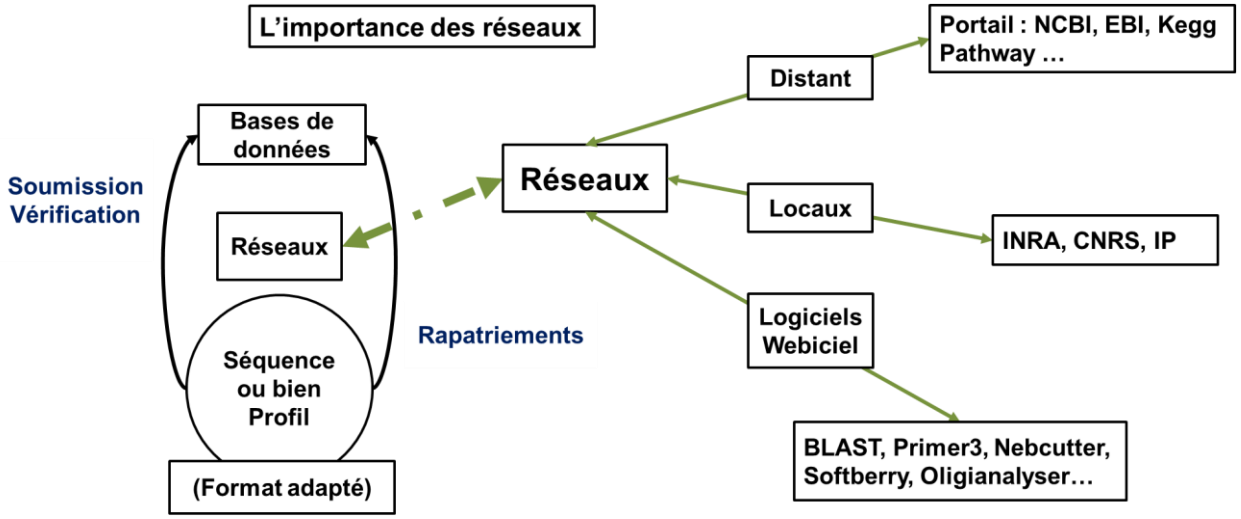


Figure 7. Représentation schématique de l'importance des réseaux en bioinformatique

Chapitre II. Techniques d'acquisition des données biologiques

II.1. Les techniques de séquençage

II.1.1. Les premières techniques de séquençage

Les premières techniques de séquençage portent le nom de leurs inventeurs : la technique de **Maxam et Gilbert** et la technique de **Sanger**. Mises au point à la fin des années 1970, ces deux techniques utilisent **un principe commun**. La molécule d'ADN est **découpée** progressivement en **fragments plus petits**. La séquence de l'ADN est **reconstituée** suite à la **séparation par électrophorèse** sur gel de polyacrylamide de fragments d'ADN simple brin. Ces techniques, qui allaient bouleverser la biologie de la fin du 20ème siècle, ont valu à Gilbert et Sanger le prix Nobel de chimie en 1980 [4].

II.1.1.1. La technique de Maxam et Gilbert

La technique de Maxam et Gilbert repose sur un **procédé chimique** qui coupe une molécule d'ADN marquée radioactivement à son extrémité 5' ou 3' au niveau d'une base ou d'une famille de base spécifique. Les conditions utilisées sont adaptées pour conduire à une coupure partielle. Par conséquent, la longueur des fragments marqués identifie la position de la base. Les réactions chimiques effectuées clivent préférentiellement l'ADN aux guanines, aux adénines, aux cytosines et thymines et aux cytosines. Les produits des quatre réactions sont résolus par électrophorèse. La séquence d'ADN peut être lue directement à partir du profil des bandes radioactives [4].

La technique de Maxam et Gilbert s'est peu développée car elle nécessite des réactifs chimiques toxiques, de plus elle n'est pas facile à automatiser et reste limitée quant à la taille des fragments d'ADN qu'elle permet d'analyser (< 250 nucléotides) [4].

II.1.1.2. La technique de Sanger

La technique de Sanger utilise le principe de la synthèse enzymatique de l'ADN à séquencer en présence d'inhibiteurs d'élongation de l'ADN polymérase, les di-désoxynucléotides (ddNTP). La technique originelle va nous permettre de décrire le principe de cette méthode [4].

La réaction de séquençage s'effectue grâce à quatre réactions enzymatiques menées parallèlement. Dans chaque tube, sont placés [4] :

- L'ADN matrice (ADN à séquencer) : Si l'ADN à séquencer est de petite quantité, il peut être préalablement amplifié par une réaction de (**PCR**). Sinon, une amplification de la matrice par multiplication clonale, le plus fréquemment, dans un vecteur bactérien, est nécessaire ;
- Une amorce capable de s'hybrider à un des brins de la matrice et qui permet le démarrage de la polymérisation d'ADN ;
- Un mélange équimolaire de dCTP, dGTP et dTTP ;
- Du dATP marqué radioactivement ;
- Un ddNTP correspondant à une des quatre bases (figure 8).

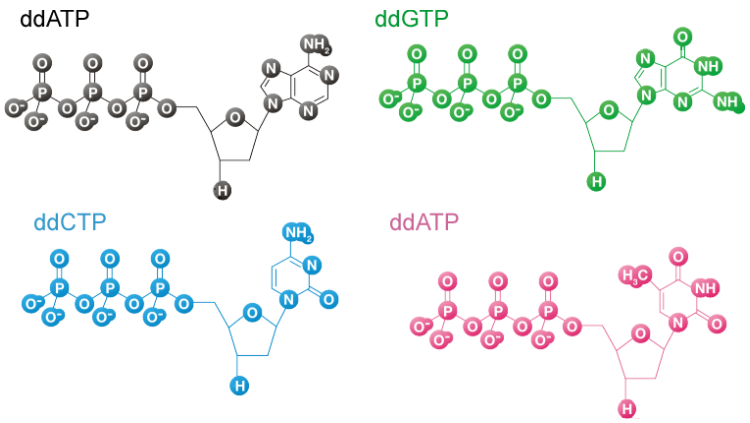


Figure 8. L'utilisation des ddNTP permet d'obtenir des fragments d'ADN de différentes tailles (Voir figure 10)

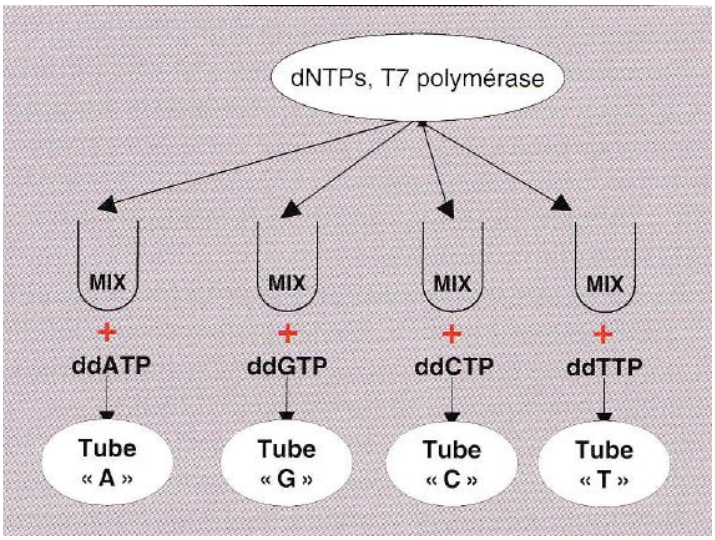


Figure 9. Ajout du mélange réactionnel contenant l'ADN polymérase, les désoxynucléotides, et d'un di-désoxynucléotide différent dans chaque tube (modifiée à partir d'Ameziane et al. 2005).

Comme les ddNTP ne présentent pas de groupement hydroxyle sur le carbone 3' (Figure 8), la liaison phosphodiester entre cet atome de carbone et l'atome de carbone 5' d'un autre nucléotide ne peut pas s'établir. Par conséquent, l'incorporation d'un ddNTP à la place d'un dNTP interrompt la polymérisation. Le ratio entre la concentration des ddNTP et les dNTP est tel que statistiquement il y a l'incorporation d'un ddNTP à chaque position. Ainsi, la réaction enzymatique va générer un mélange de fragments d'acides nucléiques ayant tous la même extrémité 5' et avec des résidus ddNTP à l'extrémité 3' (figure 10) [4].

Les différents fragments de chaque réaction enzymatique séparés par électrophorèse sur un gel de polyacrylamide coulé entre deux plaques de verre peuvent donc être visualisés par autoradiographie. La longueur des fragments identifiera la position de la base. La résolution électrophorétique de ces quatre réactions produit un profil de bandes radioactives qui reflète la séquence complémentaire de l'ADN matrice (Figure 11) [4].

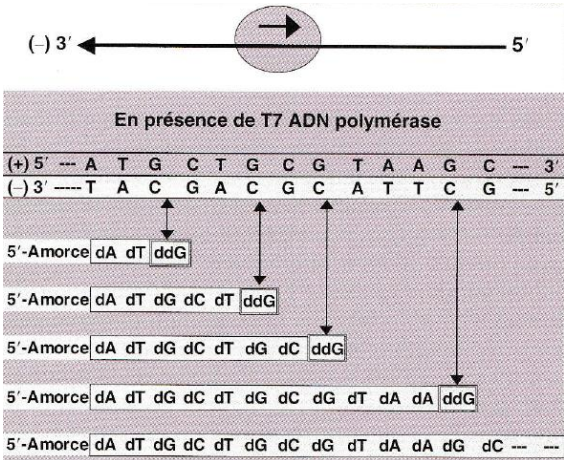


Figure 10. Action de l'ADN polymérase. (Source Ameziane et al. 2005).

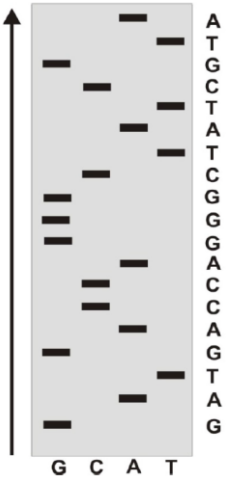


Figure 11. Séparation sur gel électrophorèse des fragments obtenues

Contrairement à la technique de Maxam et Gilbert, la technique de Sanger s'est développée avec succès, **car les produits utilisés sont moins toxiques et la longueur d'une séquence est plus importante** (jusqu'à 700 nucléotides avec la technique originelle). Par conséquent, jusqu'à ces dernières années, elle constituait la technique courante de séquençage [4].

II.1.1.3. Automatisation de la méthode de Sanger

Dans un souci d'automatisation, la technique de Sanger a évolué. Elle est désormais mise en œuvre sur des plateformes automatisées. Le principe est toujours le même, mais plusieurs modifications ont été apportées. Le marquage des fragments synthétisés ne se fait plus avec de la dATP radioactive mais avec des **ddNTP marqués avec des fluorochromes**. L'émission de fluorescence est mesurée à 4 longueurs d'onde correspondant aux 4 fluorophores. **Il est donc possible de repérer individuellement les quatre types de marquages dans un mélange**. Comme chaque ddNTP a un signal spécifique qui permet de l'identifier, **les quatre réactions enzymatiques sont effectuées dans un même et seul tube** dont le contenu est soumis à électrophorèse [4].

Les fragments d'ADN sont soumis à électrophorèse en gel de polyacrylamide coulé non plus entre deux plaques de verre mais dans des **capillaires en verre** (diamètre : # 250µm). Le gain de place permet d'avoir des robots, appelés **séquenceurs**, qui sont capables d'analyser **jusqu'à 96 séquences en parallèle** (96 capillaires). À l'extrémité du capillaire, un laser excite les fluorochromes et une caméra réceptionne les émissions aux différentes longueurs d'onde. La fluorescence émise permet de déterminer la nature du ddNTP incorporé à l'extrémité 3' du fragment sortant du capillaire [4].

Après traitement informatique, les signaux de fluorescence sont présentés sous forme d'un chromatogramme qui permet une lecture directe de la séquence du brin d'ADN complémentaire du brin séquencé (figure 12) [4].

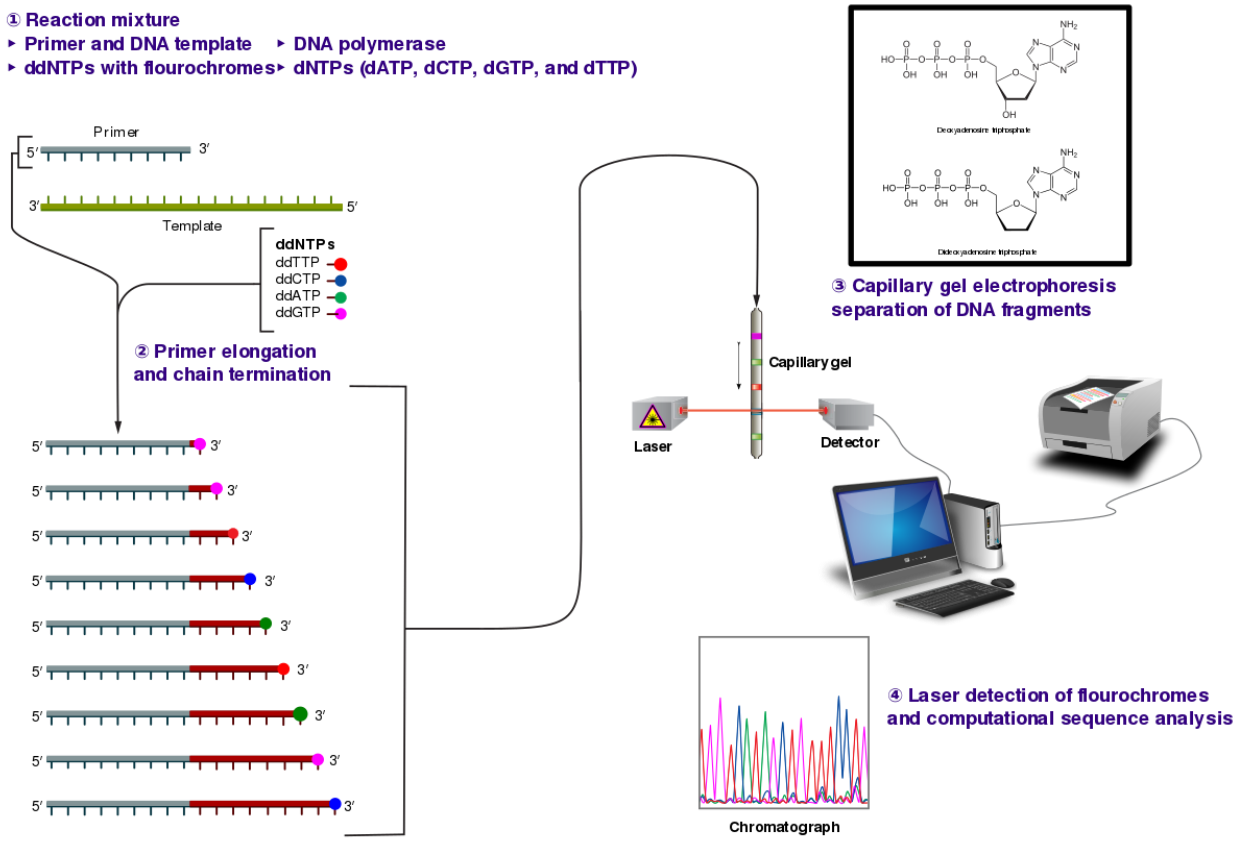


Figure 12. Automatisation de la technique de Sanger grâce à l'utilisation des ddNTP marqués avec des fluorochromes (source Wikipédia).

II.1.2. Les nouvelles techniques de séquençage (NGS)

Depuis 2004, de nouvelles techniques de séquençage sont disponibles sur le marché. Par contraste avec les techniques traditionnelles, elles ont été développées par des industriels qui commercialisent les plateformes automatisées permettant d'utiliser ces techniques. Un autre point commun très important à toutes ces nouvelles technologies est que l'amplification des banques d'ADN matrice ne passe plus par la multiplication clonale, mais par des réactions de PCR [4].

Les deux types de réactions PCR utilisées sont :

- La PCR en émulsion (Emulsion PCR ou EmPCR) ; les fragments d'ADN sont liés à des billes d'agarose ;
- La PCR par pontage (bridge amplification) ; les fragments d'ADN sont fixés sur une lame de verre appelée flow-cell. Ces amplifications par PCR évitent tout biais de représentation des fragments.

En effet, lors d'un clonage bactérien, certains fragments d'ADN peuvent présenter une toxicité pour les cellules bactériennes. Ces nouvelles techniques reposent sur [4] :

- La synthèse d'ADN (**Pyroséquençage, Solexa/Illumina et Ion Torrent**).
- L'hybridation sur des puces à ADN (**SOLiD** développé par Applied Biosystems®) ;
- La détection en temps réel de molécules (non encore mis en œuvre du point de vue commercial).

Le pyroséquençage et la technique Solexa sont couramment utilisées en combinaison avec la technique de Sanger pour **le séquençage de novo**. Les techniques Solexa et SOLiD sont utilisées pour du reséquençage. Comme ce sont les plus couramment utilisées, nous décrirons ici les techniques reposant sur la synthèse d'ADN [4].

II.1.2.1. La technique de pyroséquençage

La banque d'ADN matrice est amplifiée par PCR en émulsion (voir figure 13). Chaque bille d'agarose portant un exemplaire d'ADN de la banque est déposée dans le puits d'une lame à alvéole analysable par fibre optique **(Picotiter plate)**. Dans chaque alvéole se déroule la réaction de séquence (figure 14). Les dNTP sont ajoutés en flux successif (à l'inverse de la technique de Sanger). Quand le dNTP ajouté est complémentaire du nucléotide du brin matrice, il est incorporé dans le brin en cours de synthèse et un pyrophosphate inorganique (PPi) est libéré [4].

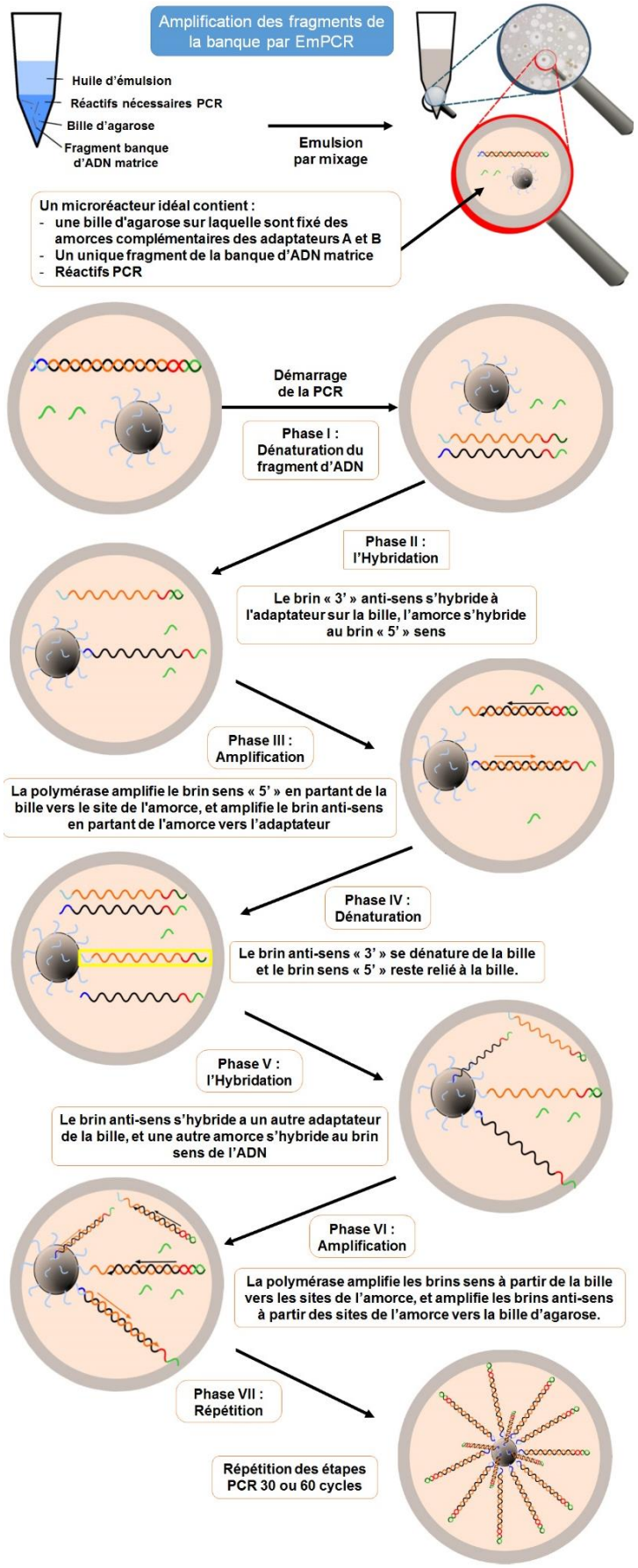


Figure 13. Les différentes étapes de la PCR en émulsion ou EmPCR

La libération du PPi est suivie par chimioluminescence suite à une cascade de réactions enzymatiques : en présence d'adénosine 5'-phosphosulfate (APS), l'ATP sulfurylase transforme le PPi en ATP qui est utilisé par une luciférase pour transformer la luciférine en oxyluciférine, molécule générant un signal lumineux dans le visible. Puis, l'apyrase dégrade les nucléotides non incorporés et l'excès d'ATP avant qu'un nouveau dNTP soit ajouté (Figure 17) [4].

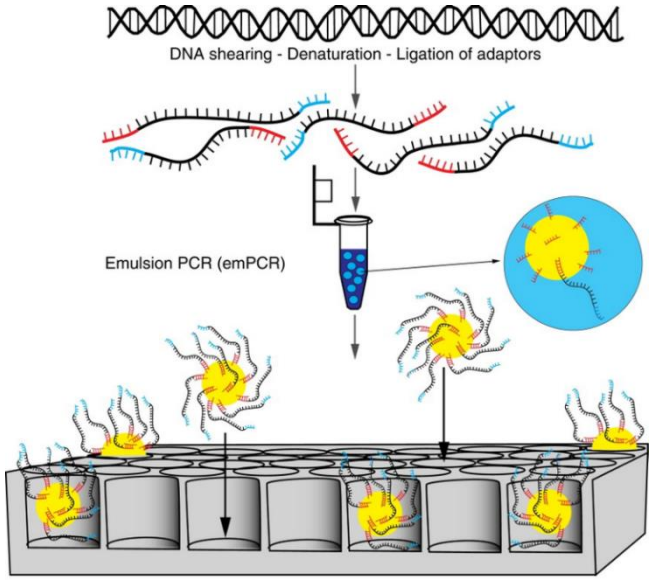


Figure 14. Chaque bille d'agarose portant un exemplaire d'ADN de la banque est déposée dans le puits d'une lame à alvéole analysable par fibre optique **(Picotiter plate).**

Une caméra **CCD** capte le signal lumineux émis et le transforme en un signal électrique qui se traduit par un pic sur le pyrogramme. La hauteur du pic est proportionnelle à l'intensité du signal lumineux, elle-même proportionnelle au nombre de nucléotides incorporés au même moment. On déduit la séquence à partir de la taille des pics obtenus (figure 17) [4].

Figure 15. Fonctionnement du Genome Sequencer FLX System.

(a) : Les dNTP sont ajoutés en flux successif.

(b) : Quand le dNTP ajouté est complémentaire du nucléotide du brin matrice, il est incorporé dans le brin en cours de synthèse et un pyrophosphate inorganique (PPi) est libéré. La libération du PPi est suivie par chimioluminescence suite à une cascade de réactions enzymatiques.

(c) : Une caméra CCD capte le signal lumineux émis.

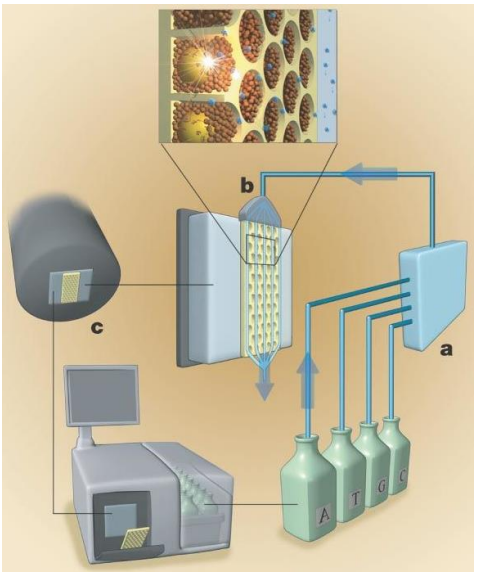


Figure 16. Principe du pyroséquençage

Quand le dNTP ajouté est complémentaire du nucléotide du brin matrice, il est incorporé dans le brin en cours de synthèse et un pyrophosphate inorganique (PPi) est libéré.

En présence d'adénosine 5'-phosphosulfate (APS), l'ATP sulfurylase transforme le PPi en ATP qui est utilisé par une luciférase pour transformer la luciférine en oxyluciférine, molécule générant un signal lumineux dans le visible.

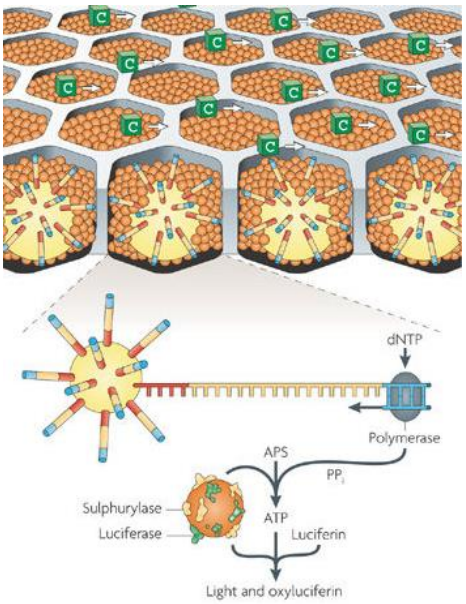
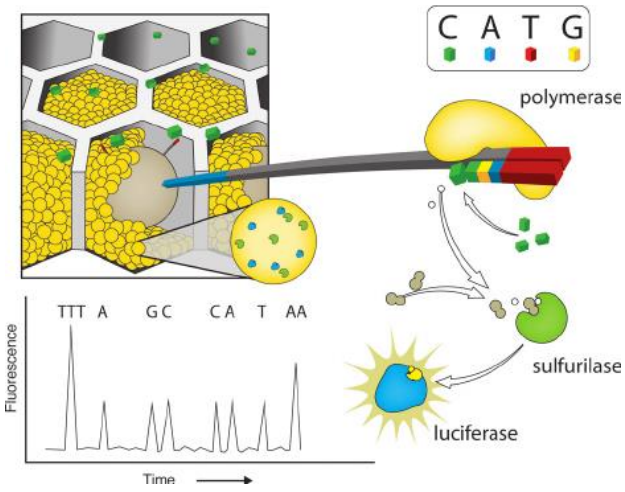


Figure 17. Principe du pyroséquençage

Une caméra **CCD** capte le signal lumineux émis et le transforme en un signal électrique qui se traduit par un pic sur le pyrogramme.

La hauteur du pic est proportionnelle à l'intensité du signal lumineux, elle-même proportionnelle au nombre de nucléotides incorporés au même moment.

On déduit la séquence à partir de la taille des pics obtenus.



La technique de pyroséquençage a été développée par la société ***454 Life Science***. Les plateformes de séquençage correspondantes « ***Genome Sequencer FLX System*** » (*GS20 ; première génération*), « ***Genome Sequencer FLX System*** » (GS FLX et GS FLX Titanium) sont commercialisées par les laboratoires **Roche® Diagnostic** [4].

II.1.2.2. La technique de séquençage Illumina/Solexa

Les fragments du brin d'ADN sont amplifiés par « *bridge amplification* » (figure x). Chaque exemplaire d'ADN de la banque est représenté par un groupe de fragments clonaux, appelé « cluster ». Les clusters sont attachés à la surface d'une lame constitué de 8 canaux et analysable par fibre optique qu'on nomme *flow-cell.* Lors d'un premier cycle de synthèse, l'ADN polymérase, l'amorce d'initiation de la synthèse d'ADN et les 4 dNTP sont ajoutés [4].

Les dNTP ont deux caractéristiques [4] :

- Chacun est marqué avec un fluorochrome différent ;
- Tous sont modifiés chimiquement de façon à bloquer l'extrémité 3'-OH et donc inhiber l'élongation.

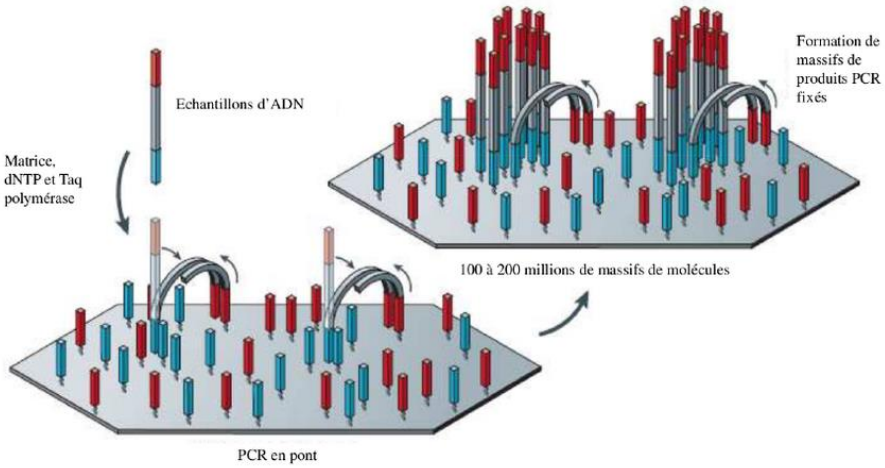


Figure 18. Principe de la PCR par pontage ou Bridge PCR

La fluorescence du dernier dNTP incorporé est mesurée avant qu'une réaction chimique n'enlève le fluorochrome et débloque l'extrémité 3'-OH afin de permettre la poursuite de la synthèse d'ADN (Figure 19). La plateforme de séquençage correspondante, le « ***Genome Analyzer sequencing system*** », est commercialisée par **Illumina®** [4].

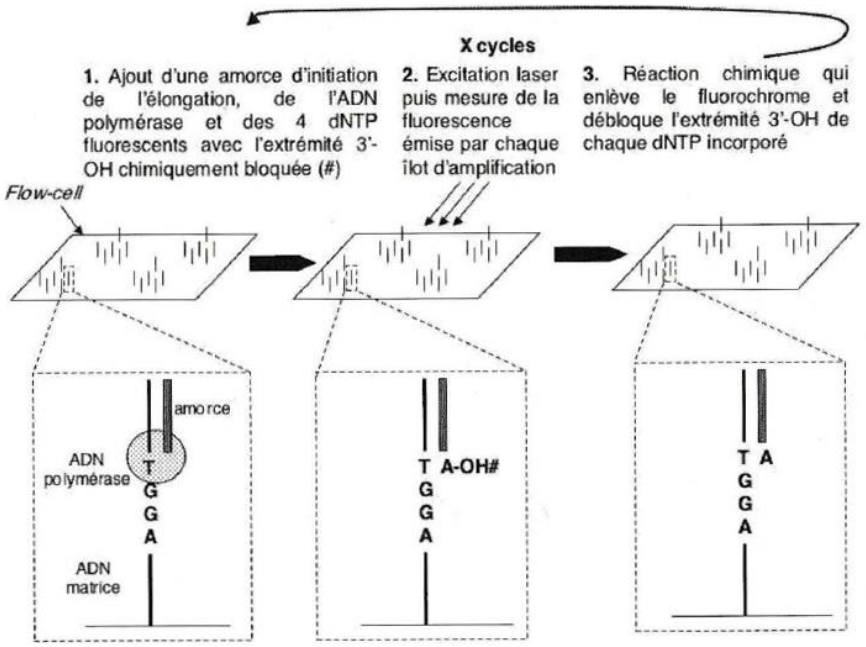


Figure 19. Principe de la technique de séquençage Illumina/Solexa [4].

II.1.3. Les techniques de séquençage de 3ème génération

II.1.3.1. La technologie SMRT sequencing

Cette technologie utilise le marquage par fluorescence couleur des nucléotides ajoutées aux brins d'ADN transcrits par polymérase. Leur ajout est détecté en temps réel au fur et à mesure de leur ajout au brin d'ADN à séquencer [5].

Son bénéfice principal est de permettre de lire d'une seul fois des séquences **allant jusqu'à 3000 bases**. Cela contribue à diminuer le nombre d'erreurs et à réduire le niveau de taux de couverture (le nombre de lecture, c-à-d le nombre de bases à détecter par redondance / nombre de bases de l'ADN à séquencer) [5].

II.1.4. Comparaison des techniques

Ces nouvelles techniques permettent d'effectuer un séquençage rapide et à très haut débit. De plus. Une fois l'acquisition de la plateforme automatisée effectuée, le séquençage est beaucoup moins coûteux qu'un séquençage par la technique de Sanger. Néanmoins, Pour les étapes d'assemblage de génome et particulièrement dans des régions avec de nombreuses répétitions, la technique de Sanger reste encore parfois nécessaire [4].

Tableau 3. Comparaison des 4 principales techniques de séquençage [4].

Technique	Banque	Amplification de la banque	Principe de séquençage	Longueur de la lecture (Nucléotides)	Nombre de nucléotides lus par expérimentation (en Mb)	Prix approximatif par Mb (en euros)
Technique de Sanger sur un séquenceur automatisé (96 réactions)	Fragments d'ADN double brin dans un vecteur réplicatif	Multiplication bactérienne	Synthèse enzymatique en présence d'inhibiteurs d'élongation, les ddNTP et électrophorèse	Jusqu'à 800	0,096	5000
Pyroséquençage sur une plateforme FLX	Fragments d'ADN simple brin ligaturés avec des adaptateurs	PCR en émulsion	Synthèse enzymatique et suivi du relargage du pyrophosphate généré lors de l'incorporation d'un nucléotide	200-300	80-120	75
Technique Solexa/Illumina	Fragments d'ADN simple brin ligaturés avec des adaptateurs	PCR par pontage	Synthèse enzymatique, inhibition réversible de l'élongation et suivi de la fluorescence du nucléotide incorporé	30-40	1000	5
Technique SOLiD	Fragments d'ADN simple brin ligaturés avec des adaptateurs	PCR en émulsion	Hybridation / ligature d'amorces et suivi de la fluorescence des oligonucléotides hybridés	35	1000-3000	5

Chapitre III. Les bases de données biologiques

III.1. Définition

Les bases de données biologiques sont 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.

III.2. Le rôle des banques et bases de données biologiques

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 [3].

III.3. Contenues des bases de 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.

III.4. 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 [3].

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 [3].

Tableau 4. Quelques banques de données généralistes

→ **Banques de séquences nucléiques généralistes**			
Nom	Lien	Date de création	Description
EMBL	http://www.ebi.ac.uk/embl/	1980	Banque européenne (European Moleculary Biology Laboratory) diffusée par l'EBI (European Bioinformatics Institute, Cambridge)
GenBank	http://www.ncbi.nlm.nih.gov/	1982	Banque américaine diffusée par NCBI (National Center for Biotechnology Information, Los Alamos)
DDBJ	http://www.ddbj.nig.ac.jp/	1986	DNA Data Bank of Japan diffusée par le NIG (National Institute of Genetics)
→ **Banques de séquences protéiques généralistes**			
UniProt	https://www.uniprot.org/	1986	Séquences annotées & séquences codantes traduite de l'EMBL

Tableau 5. Quelques banques de données spécialisées

→ **Banques de donnés spécialisées**		
Ensembl	https://www.ensembl.org/index.html	Banque intégrative génomique
Prosite	http://prosite.expasy.org/	Recense les motifs protéiques ayant une signification biologique
Reactome	https://reactome.org/PathwayBrowser/	Banque intégrative métabolique
Kegg Pathway	http://www.genome.jp/kegg/pathway.html	Interactions moléculaires et réactions
PFAM	http://xfam.org/	Domaines protéiques
Interpro	http://www.ebi.ac.uk/interpro/	Regroupe plusieurs banques existantes
PDB	http://www.rcsb.org/pdb/home/home.do	Structure 3D de protéines, acides aminés et molécules biologiques
PubMed	https://www.ncbi.nlm.nih.gov/pubmed	Citations, résumés et articles (recherche bibliographique)

III.4.1. 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

III.4.2. 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, …

III.4.3. 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'.

III.4.4. 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

III.4.5. Les bases de données bioinformatiques 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 of Proteins*

III.5. Structuration et organisation

Les grandes banques de séquences généralistes telles que **GenBank** ou **l'EMBL** sont des projets internationaux qui constituent des leaders dans le domaine. 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 [6].

III.5.1. Fichiers et formats

Les séquences sont stockées en général sous forme de **fichiers texte** qui peuvent être soit des fichiers personnels (présents dans un espace personnel), soit des fichiers publics (séquences des banques) accessibles par des outils Web.

Le format correspond à l'ensemble des règles (contraintes) de présentation auxquelles sont soumises la ou les séquences dans un fichier donné. Le format permet :

- Une mise en forme automatisée

- Le stockage homogène de l'information
- Le traitement informatique ultérieur de l'information.

Une seule pièce d'informations dans une base de données est nommée **"entrée"**

Pour que l'utilisateur puisse se repérer, toutes ces informations sont mises à la disposition de la collectivité scientifique selon une organisation en **rubriques ou en champs** [6].

III.5.1.1. Le format FASTA

Il existe plusieurs formats dont le plus courant est le format FASTA :

Appelé aussi format (Pearson) est un format de fichier texte utilisé pour stocker des séquences biologiques de nature nucléique ou protéique.

La séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">". Plusieurs séquences peuvent être ainsi mises dans un même fichier.

La simplicité du format FASTA rend la manipulation et la lecture (ou analyse syntaxique) des séquences aisées par l'utilisation d'outils de traitement de texte et de langages de programmations tels que C++, Java, Python, R, Matlab ou Perl.

Ainsi un fichier FASTA se présente sous la forme suivante (les X représentant acides nucléiques ou aminés) :

```
> Identifiant|Commentaire

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Exemples types :

Voici un exemple de séquence nucléique :

```
>gi|373251181|ref|NG_001742.2| Mus musculus olfactory receptor
GA_x5J8B7W2GLP-600-794 (LOC257854) pseudogène on chromosome 2

AGCCTGCCAAGCAAACTTCACTGGAGTGTGCGTAGCATGCTAGTAACTGCATCTGAATCTTTCAGC
TGCTTGTTGGGCCTCTCACAAGGCAGAGTGTCTTCATGGGACTTTGATATTTATTTTTGTACAACC
TAAGAGGAACAAATCCTTTGACACTGACAAATTGGCTTCCATATTTTATACCTTAATCATCTCCAT
GTTGAATTCATTGATCAACAGTTTAAGAAAAAAAGATGTAAAAATGCTTTTAGAAAGAGAGGCAAA
GTTATGCACAATAACTTCTCATGAAGTCACAGTTTGTTAAAAGTTGCCTTAGTTCACAATAAATAA
TTATGTATGC
```

III.5.1.2. Le format EMBL

https://www.ebi.ac.uk/ena : L'exemple d'une séquence d'ADN génomique d'un micro-organisme *Saccharomyces cerevisiae*

```
ID   M10154; SV 1; linear; genomic DNA; STD; FUN; 937 BP.
XX
AC   M10154;
XX
DT   19-SEP-1987 (Rel. 13, Created)
DT   22-APR-1990 (Rel. 23, Last updated, Version 1)
XX
DE   Yeast (S.cerevisiae) nuclear gene CBP6 for cytochrome b,
DE   complete cds.
XX
KW   cytochrome; cytochrome b.
XX
OS   Saccharomyces cerevisiae (yeast)
OC   Eukaryota; Plantae; Thallobionta; Eumycota; Hemiascomycetes;
OC   Endomycetales; Saccharomycetaceae.
XX
RN   [1]
RP   1-937
RX   MEDLINE; 85105014.
RA   Dieckmann C.L., Tzagoloff A.;
RT   "Assembly of the mitochondrial membrane system";
RL   J. Biol. Chem. 260:1513-1520(1985).
XX
DR   SWISS-PROT; P07253; CBP6_YEAST.
XX
CC   There is a putative 'tata' box at position 215 to 219.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..937
FT                   /organism="Saccharomyces cerevisiae"
FT   CDS             301..789
FT                   /note="CBP6 protein"
FT                   /note="pid:g171173"
XX
SQ   Sequence 937 BP; 345 A; 159 C; 166 G; 267 T; 0 other;
ATACGATTAT TTTGGAAGTT TATAAAAGAA GTGCGGAAAT CACATCTGCT GTTTATTTAG      60
CCATTCCTCA CACTAATAGT TAAAGTACTT TCATAGCAGC TCTGCGCATG GTCGGACATG     120
CGAAAAATTC TGATATCAAG AAAAAGCGAA ATATTTCCGG CCTTGTAGGG GCCAAAACAT     180
TAACGTATAT CAAGATTTCC TGTGGTAGCA ACATTATAAG AAAAAAAGGT AGCCTTCATT     240
GAAACATTCT CTCTATCAGC TTACCAAGTT AAACTCCGTA TTCCACAAGC AAGTGCCAAA     300
ATGTCTTCTT CCCAGGTCGT CAGGGATTCT GCCAAAAAAT TAGTTAATTT ACTGGAAAAA     360
TATCCAAAGG ATCGTATACA CCACTTGGTC TCATTCAGGG ATGTACAAAT AGCAAGATTT     420
AGACGTGTAG CGGGTCTGCC AAATGTAGAT GACAAAGGAA AATCTATAAA AGAGAAAAAA     480
CCCTCATTAG ATGAAATAAA AAGTATAATT AACAGAACTT CCGGTCCATT AGGACTGAAT     540
AAGGAGATGT TAACCAAAAT TCAAAATAAA ATGGTAGATG AGAAATTCAC GGAAGAAAGC     600
ATCAACGAGC AAATTCGTGC CTTGAGCACT ATAATGAATA ATAAATTCAG AAACTATTAC     660
GATATTGGCG ATAAGCTCTA TAAACCTGCA GGAAATCCCC AATATTATCA ACGGTTAATA     720
AATGCCGTTG ACGGTAAGAA AAAGGAAAGC TTATTTACTG CAATGAGAAC TGTATTATTT     780
GGTAAATAAA GAGCACATTA TTTTCTAAGC TTGTAAATAC ATATTTATTC ATAATGGAGA     840
ACGTTATTCA AATTTATCTG TGAATTTCTT TACTCGAGGT ATACTTCCGC AAAGGAAATT     900
CTACTTAGCA AATCCTATGG TAACGTCATT GTTTTGT                              937
//
```

Une explication de l'organisation du format EMBL est donnée ci-dessous :

ID : Identificateur, c'est le nom de l'entrée contenant la séquence. Cette ligne a la structure suivante : nom de l'entrée ; classe de la donnée ; molécule ; division ; longueur. Le nom est suivi de l'indication de la classe de donnée, puis du type de molécule ADN, ARN ou ADNc (XXX si l'entrée n'a pas été annotée) ; ensuite la division à laquelle l'entrée appartient et enfin la longueur de la séquence en paires de bases (bp).

AC : Numéro d'accession de l'entrée qui ne varie pas au cours des versions successives de la banque. Il peut y avoir plusieurs numéros d'accessions pour une même entrée. En effet lorsque deux entrées sont fusionnées en une seule, un nouveau numéro peut être attribué à la nouvelle entrée et ceux provenant des ex-entrées indépendantes sont conservés.

DT : Donne la date d'incorporation dans la base (1ère ligne) et la date de la dernière mise à jour de l'entrée (2ème ligne).

DE : Cette ligne contient des informations descriptives sur la séquence comme le nom du gène, la région du génome dont elle est issue etc... C'est en fait le titre de la séquence.

KW : Donne-le(s) mot(s)-clé(s) désignés par les auteurs. Ils peuvent être utilisés pour retrouver l'entrée dans la base. Les mots-clés séparés par des ; sont rangés par ordre alphabétique.

OS : Spécifie l'organisme d'où provient la séquence ; le plus souvent, on donne le nom latin suivi du nom commun anglais entre parenthèses. Dans le cas d'hybrides les lignes OS/OC sont spécifiées pour chaque organisme de l'hybride.

RN : Numéro unique attribué à chaque référence bibliographique de l'entrée. Ce numéro est utilisé pour désigner la référence dans les commentaires (CC comments) et le champ des caractéristiques biologiques (FT features).

RP : Donne la région du gène pour laquelle la référence bibliographique est associée.

RX : Donne la référence MEDLINE associée à la bibliographie. MEDLINE Est une base de données bibliographiques regroupant la littérature relative aux sciences

biologiques et biomédicales. La base est gérée et mise à jour par la Bibliothèque américaine de médecine (NLM).

RA : Indique les auteurs de l'article ou du travail cité. Les auteurs sont cités dans l'ordre donné dans la publication.

RT : Indique le titre de l'article. Si la séquence a été soumise à la base et non publiée, la ligne ne contiendra qu'un ;

RL : Donne d'une manière abrégée les références du journal. Pour un article sous presse le numéro du volume et des pages sera de 0.

DR : Etablit des liaisons avec d'autres bases de données qui contiennent une information en relation avec cette entrée. Par exemple, si la traduction protéique d'une séquence existe dans la banque de données SWISS-PROT, la ligne DR pointera sur l'entrée correspondante dans SWISS-PROT. Cette ligne est composée de plusieurs champs qui sont les suivants :

- Identificateur de la banque de données : L'identificateur de la base de données est le nom abrégé courant que l'on donne à cette base.
- Identificateur primaire : pointe sur l'entrée de cette base et dépend de la base référencée. Il pointe sur le numéro d'accession si la base est SWISS-PROT, sur le champ ID si la base est TFD ou FLYBASE et sur le code d'entrée si la base est EPD (Eucaryotic Promoter Database)
- Identificateur secondaire : complète l'information donnée par l'identificateur primaire et dépend de la base référencée, par exemple c'est le nom de l'entrée pour UniProt.

CC : Donne les commentaires sur la séquence.

FH : Cette ligne sert à améliorer la lecture d'une entrée lorsqu'elle est imprimée ou affichée sur l'écran du terminal : c'est l'en-tête du champ FT (feature)

FT : Caractéristiques de la séquence (features).

SQ : Séquence (60 nucléotides par ligne dans le sens 5'--->3').

CC : Commentaires

// Fin de l'entrée.

III.5.1.3. Le format Genbank

GenBank: M10154.1

FASTA Graphics

Go to:

```
LOCUS       YSCCBP6                   937 bp    DNA     linear   PLN 27-APR-1993
DEFINITION  Yeast (S.cerevisiae) nuclear gene CBP6 for cytochrome b, complete
            cds.
ACCESSION   M10154
VERSION     M10154.1
KEYWORDS    cytochrome; cytochrome b.
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Saccharomyces.
REFERENCE   1  (bases 1 to 937)
  AUTHORS   Dieckmann,C.L. and Tzagoloff,A.
  TITLE     Assembly of the mitochondrial membrane system. CBP6, a yeast
            nuclear gene necessary for synthesis of cytochrome b
  JOURNAL   J. Biol. Chem. 260 (3), 1513-1520 (1985)
   PUBMED   2981859
COMMENT     Original source text: Yeast (S.cerevisiae; strain D273-10B) DNA,
            clone pG154/ST1.
            There is a putative 'tata' box at position 215 to 219.
FEATURES             Location/Qualifiers
     source          1..937
                     /organism="Saccharomyces cerevisiae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:4932"
     CDS             301..789
                     /note="CBP6 protein"
                     /codon_start=1
                     /protein_id="AAA34476.1"
                     /translation="MSSSQVVRDSAKKLVNLLEKYPKDRIHHLVSFRDVQIARFRRVA
                     GLPNVDDKGKSIKEKKPSLDEIKSIINRTSGPLGLNKEMLTKIQNKMVDEKFTEESIN
                     EQIRALSTIMNNKFRNYYDIGDKLYKPAGNPQYYQRLINAVDGKKKESLFTAMRTVLF
                     GK"
ORIGIN      86 bp upstream of RsaI cut site.
        1 atacgattat tttggaagtt tataaaagaa gtgcggaaat cacatctgct gtttatttag
       61 ccattcctca cactaatagt taaagtactt tcatagcagc tctgcgcatg gtcggacatg
      121 cgaaaaattc tgatatcaag aaaaagcgaa atatttccgg ccttgtaggg gccaaaacat
      181 taacgtatat caagatttcc tgtggtagca acattataag aaaaaaaggt agccttcatt
      241 gaaacattct ctctatcagc ttaccaagtt aaactccgta ttccacaagc aagtgccaaa
      301 atgtcttctt cccaggtcgt cagggattct gccaaaaaat tagttaattt actggaaaaa
      361 tatccaaagg atcgtataca ccacttggtc tcattcaggg atgtacaaat agcaagattt
      421 agacgtgtag cgggtctgcc aaatgtagat gacaaaggaa aatctataaa agagaaaaaa
      481 ccctcattag atgaaataaa aagtataatt aacagaactt ccggtccatt aggactgaat
      541 aaggagatgt taaccaaaat tcaaaataaa atggtagatg agaaattcac ggaagaaagc
      601 atcaacgagc aaattcgtgc cttgagcact ataatgaata ataaattcag aaactattac
      661 gatattggcg ataagctcta taaacctgca ggaaatcccc aatattatca acggttaata
      721 aatgccgttg acggtaagaa aaaggaaagc ttatttactg caatgagaac tgtattattt
      781 ggtaaataaa gagcacatta ttttctaagc ttgtaaatac atatttattc ataatggaga
      841 acgttattca aatttatctg tgaatttctt tactcgaggt atacttccgc aaaggaaatt
      901 ctacttagca aatcctatgg taacgtcatt gttttgt
//
```

III.5.2. La qualité des données

On doit savoir que l'information contenue dans ces bases présente un certain nombre de **lacunes**. Une des principales est **le manque de vérifications systématiques des données soumises ou saisies**, surtout pour les séquences anciennes.

Les auteurs des séquences ont parfois du mal à restituer les connaissances qu'ils détiennent à propos de leurs données ou bien n'ont pas fait un certain nombre de vérifications de base sur leurs séquences.

Il arrive par exemple, que l'on retrouve :

- Des segments de vecteurs de clonage dans certaines séquences ou des incohérences dans les caractéristiques biologiques (parties codantes, définition des espèces ou des mots clés...)
- Des informations biologiques incomplètes, voire erronées.

Les organismes responsables de la maintenance de ces banques ont pris conscience de ces problèmes et maintenant de nombreuses **vérifications** sont faites systématiquement **dès la soumission de la séquence**.

Ceci n'élimine pas la totalité des imprécisions comme par exemple **l'existence de doublons** car il s'agit là de séquences extrêmement similaires qui correspondent à des entrées différentes dans la banque et dont il est souvent difficile de savoir s'il s'agit de **polymorphisme**, **de gènes dupliqués** ou **de gènes dupliqués** tout simplement **d'erreurs établies lors de la détermination des séquences**.

Il existe d'ailleurs des boites aux lettres électroniques (e-mail) pour informer les gestionnaires des banques d'éventuelles erreurs ou rectifications que chacun pourrait déceler ou proposer.

Des boîtes aux lettres

- pour l'EMBL : update@ebi.ac.uk ou datasubs@ebi.ac.uk
- pour GenBank : update@ncbi.nlm.nih.gov ou gb_admin@ncbi.nlm.nih.gov

Des pages web

- EMBL update http://www3.ebi.ac.uk/Services/webin/update/update.html
- UniProtKB update https://web.expasy.org/docs/swiss-prot_guideline.html

Un autre problème important est **le retard de l'insertion d'une nouvelle séquence dans une banque**, lié souvent au volume des séquences à traiter qui engendre des priorités ou des choix. Ainsi, il peut y avoir une dizaine de mois de **délai** entre la **détermination expérimentale** d'une séquence et **l'introduction** de celle-ci dans une banque.

Malgré cela, il faut souligner l'énorme **richesse** que représentent ces banques de données, en particulier dans le cadre de **l'analyse des séquences**.

- Le fait que la majorité des séquences connues soit réunie en un seul ensemble est un élément **fondamental** pour **la recherche de similitudes** avec une nouvelle séquence.
- La grande **diversité d'organismes** qui y est représentée permet d'aborder des analyses de type évolutif. Par exemple, on peut extraire les séquences d'un même gène issu de plusieurs espèces.
- Un autre intérêt de ces bases réside dans **l'information qui accompagne les séquences (annotations, expertise, bibliographie)**, même si celles-ci sont souvent de qualité inégale. Ces dernières peuvent parfois constituer les rares annotations disponibles sur certaines séquences.
- Enfin la présence de **références à d'autres bases** permet d'avoir accès à d'autres informations non répertoriées. Ainsi on peut connaître l'entrée dans une base protéique de la protéine qui correspond au gène que l'on a repéré dans une base nucléique.

La banque **UniProt** particulièrement riche en références croisées avec d'autres banques et en annotations (par exemple, la notion de "prouvé ou pas expérimentalement" a été récemment introduite dans la table des caractéristiques biologiques) **est un exemple de la qualité des données** que l'on peut retrouver dans les différentes banques de séquences généralistes de ces dernières années.

III.6. Interrogation des bases de données

On peut interroger une **BD** (**B**anque de **D**onnées) pour plusieurs raisons :

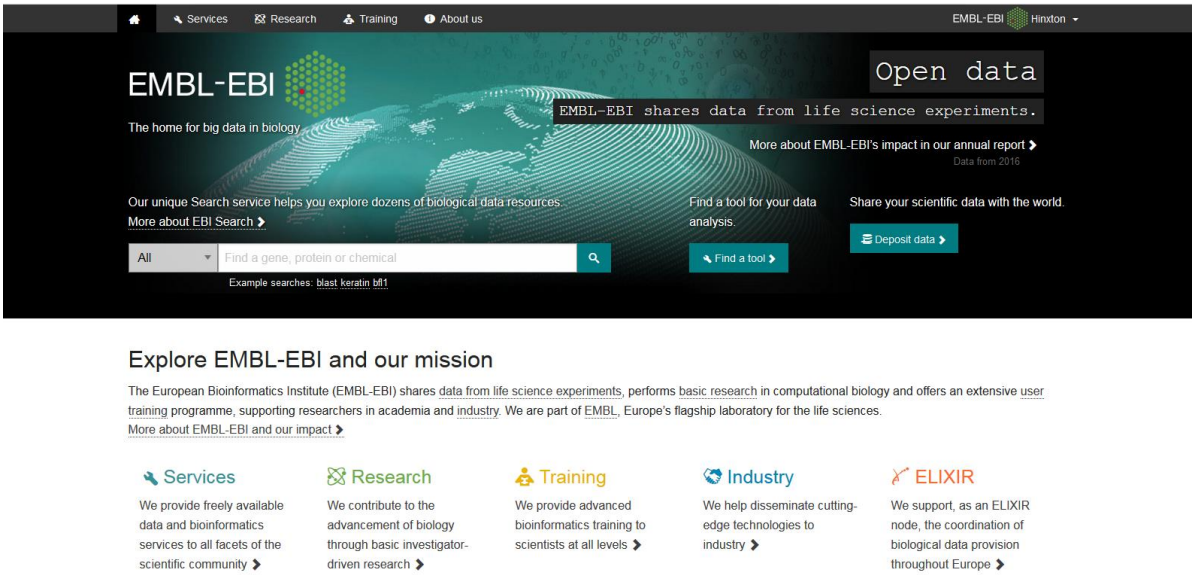
- Pour connaître la séquence d'un gène ou d'une portion de ce gène ;
- Pour connaître la structure primaire d'une protéine ;
- Pour comparer deux séquences, …

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 quel que soit la BD interrogée et la différence se situe beaucoup plus dans l'intitulé des champs dans l'entrée.

Travail Pratique 1 : Par exemple réalisons le TP suivant sur Internet pour connaître la structure primaire d'ARNm du gène de la phosphatase alcaline humaine, en interrogeant la base de données **EMBL**.

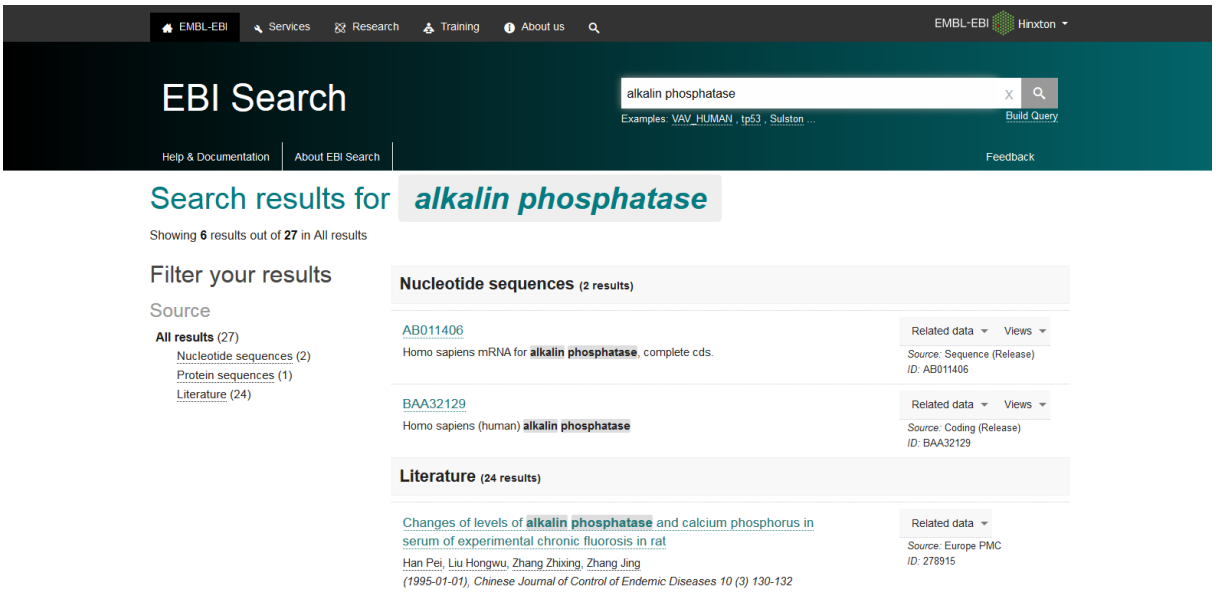
Suivre les étapes suivantes :

1. Allez sur le moteur de recherche google est glisser l'adresse de EMBL : https://www.ebi.ac.uk/



2. Sur l'interface recherche de celle-ci tapez le mot alkalin phosphatase (toujours en anglais svp) puis cliquez sur le bouton Go 🔍

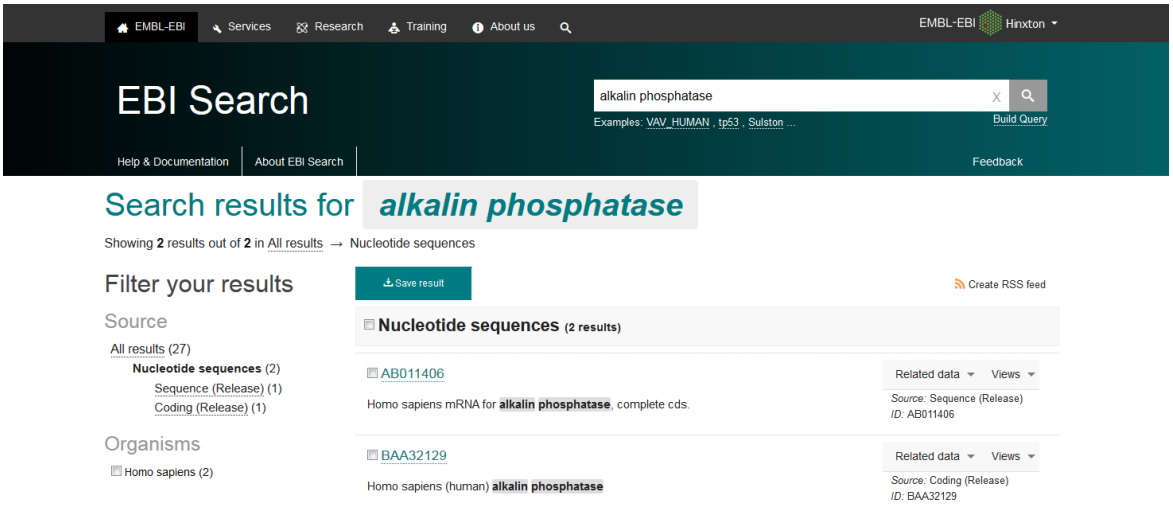
3. Une nouvelle page (interface) apparaît avec une proposition de résultats :



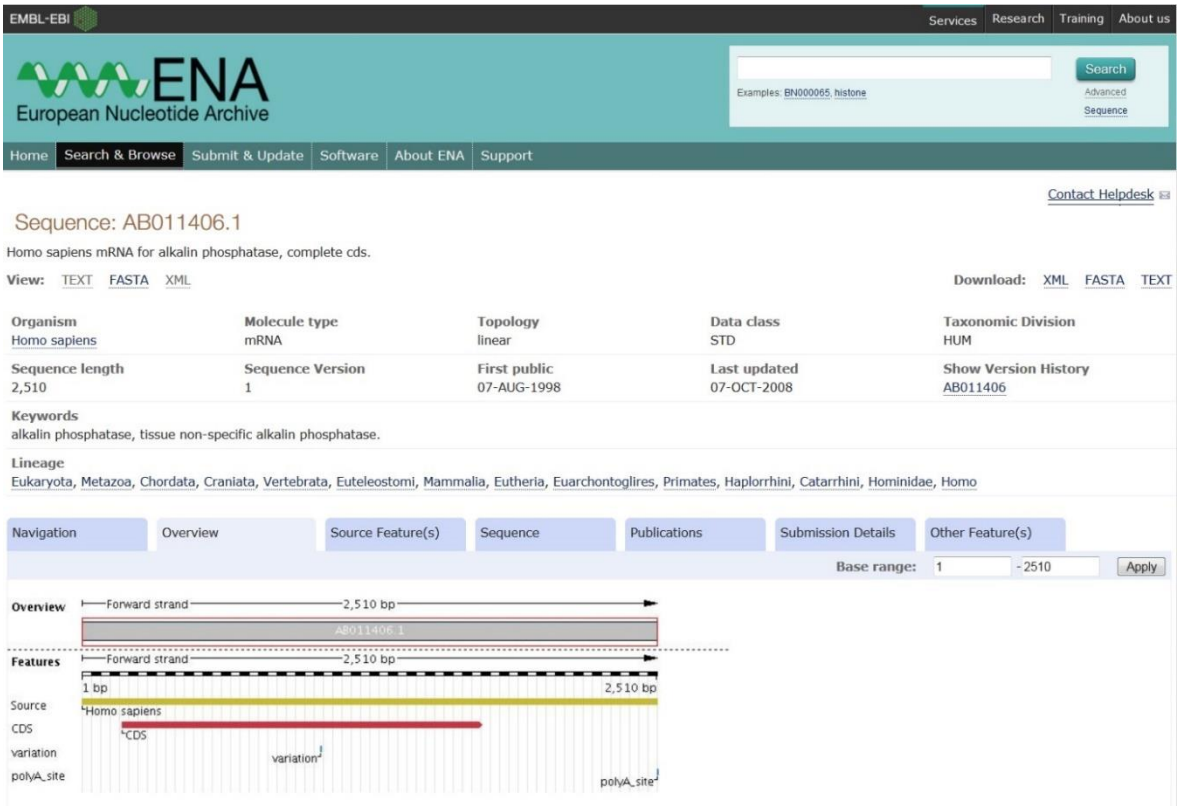
Dans cette interface, EMBL propose les différentes banques de données qui ont été consultées afin de répondre à notre requête (à gauche de la page).

On constate que **" Nucletide Sequences "** propose **deux (2)** résultats et **" Literature "** en propose **24**. **Attention :** ces chiffres peuvent changer à cause des mises à jour !

4. Cliquez sur le bouton de **" Nucletide Sequences "** et vous aurez l'interface suivante :



5. Cliquez sur le numéro d'accession de la séquence : **AB011406** C'est le code d'entrée à cette fiche. Vous obtiendrez la page de l'entrée suivante :



Vous avez le choix entre 3 types de format :

- Le format TEXT (EMBL)
- Le format FASTA
- Le format web XML

3 – Lorsqu'on clique sur FASTA on obtient la page suivante :

```
>ENA|AB011406|AB011406.1 Homo sapiens mRNA for alkalin phosphatase, complete cds.
CGCGCCCGCTATCCTGGCTCCGTGCTCCCACGCGCTTGTGCCTGGACGGACCCTCGCCAG
TGCTCTGCGCAGGATTGGAACATCAGTTAACATCTGACCACTGCCAGCCCACCCCCTCCC
ACCCACGTCGATTGCATCTCTGGGCTCCAGGGATAAAGCAGGTCTTGGGGTGCACCATGA
TTTCACCATTCTTAGTACTGGCCATTGGCACCTGCCTTACTAACTCCTTAGTGCCAGAGA
AAGAGAAAGACCCCAAGTACTGGCGAGACCAAGCGCAAGAGACACTGAAATATGCCCTGG
AGCTTCAGAAGCTCAACACCAACGTGGCTAAGAATGTCATCATGTTCCTGGGAGATGGGA
TGGGTGTCTCCACAGTGACGGCTGCCCGCATCCTCAAGGGTCAGCTCCACCACAACCCTG
GGGAGGAGACCAGGCTGGAGATGGACAAGTTCCCCTTCGTGGCCCTCTCCAAGACGTACA
ACACCAATGCCCAGGTCCCTGACAGCGCCGGCACCGCCACCGCCTACCTGTGTGGGGTGA
AGGCCAATGAGGGCACCGTGGGGGTAAGCGCAGCCACTGAGCGTTCCCGGTGCAACACCA
CCCAGGGGAACGAGGTCACCTCCATCCTGCGCTGGGCCAAGGACGCTGGGAAATCTGTGG
GCATTGTGACCACCACGAGAGTGAACCATGCCACCCCCAGCGCCGCCTACGCCCACTCGG
CTGACCGGGACTGGTACTCAGACAACGAGATGCCCCCTGAGGCCTTGAGCCAGGGCTGTA
AGGACATCGCCTACCAGCTCATGCATAACATCAGGGACATTGACGTGATCATGGGGGGTG
GCCGGAAATACATGTACCCCAAGAATAAAACTGATGTGGAGTATGAGAGTGACGAGAAAG
CCAGGGGCACGAGGCTGGACGGCCTGGACCTCGTTGACACCTGGAAGAGCTTCAAACCGA
GATACAAGCACTCCCACTTCATCTGGAACCGCACGGAACTCCTGACCCTTGACCCCCACA
ATGTGGACTACCTATTGGGTTTCTTCGAGCCAGGGGACATGCAGTACGAGCTGAACAGGA
ACAACGTGACGGACCCGTCACTCTCCGAGATGGTGGTGGTGGCCATCCAGATCCTGCGGA
AGAACCCCAAAGGCTTCTTCTTGCTGGTGGAAGGAGGCAGAATTGACCACGGGCACCATG
AAGGAAAAGCCAAGCAGGCCCTGCATGAGGCGGTGGAGATGGACCGGGCCATCGGCCACG
CAGGCAGCTTGACCTCCTCGGAAGACACTCTGACCGTGGTCACTGCGGACCATTCCCACG
TCTTCACATTTGGTGGATACACCCCCCGTGGCAACTCTATCTTTGGTCTGGCCCCCATGC
TGAGTGACACAGACAAGAAGCCCTTCACTGCCATCCTGTATGGCAATGGGCCTGGCTACA
AGGTGGTGGGCGGTGAACGAGAGAATGTCTCCATGGTGGACTATGCTCACAACAACTACC
AGGCGCAGTCTCCTGTGCCCCTGCGCCACGAGACCCACGGCGGGGAGGACGTGGCCGTCT
TCTCCAAGGGCCCCATGGCGCACCTGCTGCACGGCGTCCACGAGCAGAACTACGTCCCCC
ACGTGATGGCGTATGCAGCCTGCATCGGGGCCAACCTCGGCCACTGTGCTCCTGCCAGCT
CGGCAGGCAGCCTTGCTGCAGGCCCCCTGCTGCTCGCTCTGGCCCTCTACCCCCTGAGCG
TCCTGTTCTGAGGGCCCAGGGCCCGGGCACCCACAAGCCCGTGACAGATGCCAACTTCCC
ACACGGCAGCCCCCCCCTCAAGGGGCAGGGAGGTGGGGGCCTCCTCAGCCTCTGCAACTG
CAAGAAAGGGGACCCAGGAAACCAAAGTCTGCCGCCCACCTCGCTCCCCTCTGGAATCTT
CCCCAAGGGCCAAACCCACTTCTGGCCTCCAGCCTTTGCTCCCTCCCCGCTGCCCTTTGG
CCACCAGGGTAGATTTCTCTTGGGCAGGCAGAGAGTACAGACTGCAGACATTCTCAAAGC
CTCTTATTTTTCTAGCGAACGTATTTCTCCAGACCCAGAGGCCCTGAAGCCTCCGTGGAA
CATTGTGGATCTGACCCTCCCAGTCTCATCTCCTGACCCTCCCACTCCCATCTCCTTACC
TCTGGAACCCCCCAGGCCCTACAATGCTCATGTCCCTGTCCCCAGGCGAGCCCTCCTTCA
GGGGAGTTGAGGTCTTTCTCCTCAGGACAAGGCCTTGCTCACTCACTCACTCCAAGACCA
CCAGGGTCCCAGGAAGCCGGTGCCTGGGTGGCCATCCTACCCAGCGTGCCCAGGCCGGGA
AGAGCCACCTGGCAGGGCTCACACTCCTGGGCTCTGAACACACACGCCAGCTCCTCTCTG
AAGCGACTCTCCTGTTTGGAACGGCAAAAAAAAATTTTTTTTTCTCTTTTTGGTGGTGGT
TAAAAGGGAACACAAAACATTTAAATAAAACTTTCCAAATATTTCCGAGG
```

Chapitre IV. Exploitation et Analyse des données (Annotation)

IV.1. Introduction

Une succession brute de nucléotides n'a aucun sens. **L'annotation** est le travail d'analyse qui permet à **l'aide d'outils informatiques** d'expliquer ou de proposer des hypothèses pour les propriétés biologiques d'un génome.

Pour cela, il faut rechercher les objets génétiques présents dans le génome puis essayer de leur attribuer des fonctions. Ainsi, l'annotation est l'antichambre de l'expérimentation ; elle conduit à élaborer des protocoles expérimentaux qui valident ou invalident la fonction supposée de l'objet biologique [4].

IV.2. Les algorithmes

Les génomes peuvent être considéré comme une longue suite de lettre écrite dans l'alphabet A, C, G, T. Comment interprété ces textes ?

Cela va être le sujet de la bioinformatique à l'aide d'algorithme appropriée.

IV.2.1. Définition

Un **algorithme** est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre un problème ou d'obtenir un résultat.

Le mot algorithme vient du nom arabe "الخوارزمي" du mathématicien perse du 9ème siècle Al-Khwârizmî. Le domaine qui étudie les algorithmes est appelé l'algorithmique.

On retrouve aujourd'hui des algorithmes dans de nombreuses applications telles que le fonctionnement des ordinateurs, la cryptographie, la planification, le traitement d'images, le traitement de texte, **la bioinformatique**, etc.

IV.2.2. Algorithme, Programme, Logiciel, structures de données

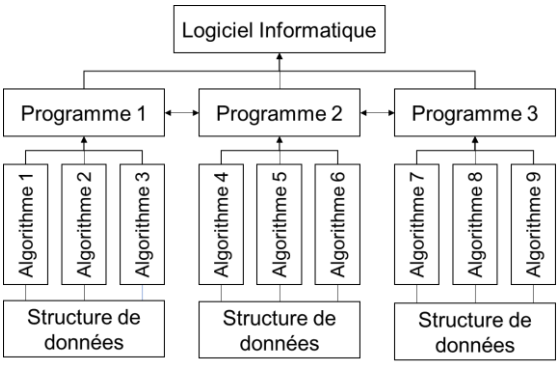


Figure 20. Algorithmes + structures de données = programmes [7]

VI.2.3. Les propriétés d'un algorithme

Une **suite d'opérations à exécuter** pour résoudre un problème (ou une classe de problèmes), Qu'il faudra exprimer d'une manière **formelle**, **explicite** et **non ambigu** [7].

De plus on attend d'un algorithme qu'il possède certaines propriétés [7] :

- Qu'il se termine
- Qu'il soit pertinent (*que son exécution conduise à la résolution du problème posé, ou un diagnostic que ce problème ne peut être résolu*)
- Qu'il soit efficace (*qu'il s'exécute dans un temps raisonnable*)

S'il est destiné à être exécuté par un ordinateur, l'algorithme doit être écrit dans **un langage de programmation** [7]**.**

Il existe de nombreux langages de programmation : R, Java, C++, Perl, Python, Matlab, …

Dans ce cous les algorithmes présentés seront écrits en « pseudo - langage » [7].

VI.2.4. L'écriture d'un algorithme

Ex : Compter les nucléotides et calculer leur fréquence

Notre premier algorithme va être très simple, il va avoir comme objectifs de compter les nucléotides et de calculer leurs fréquences dans une séquence génomique, cette séquence pour un informaticien va avoir une succession de caractères [7].

Donc notre idée c'est de calculer le Nombre de A : nbA, Nombre de C : nbC, Nombre de G : nbG, Nombre de T : nbT, simultanément nous allons aussi calculer TotalNb : le nombre total de nucléotides [7].

Nous calculerons aussi les fréquences qui est le rapport de chacune de ces lettres sur le nombre total de nucléotides [7].

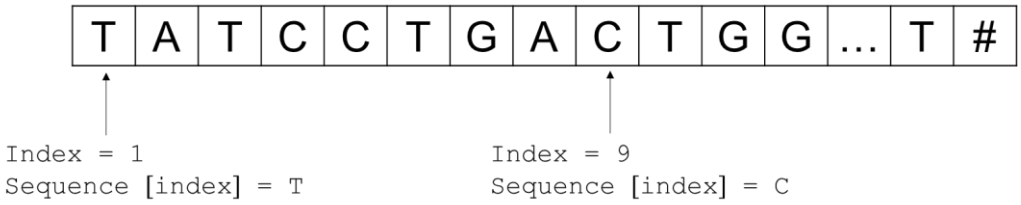
La donnée entrée de notre algorithme sera une séquence de caractère (une séquence génomique), pour marquer sa fin (pour savoir qu'il faut arrêter le comptage nous placerons le caractère (#) [7].

```
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC#
```

VI.2.4.1. La déclaration des variables

L'écriture d'un algorithme débute par la déclaration de ce qu'on appelle des variables. Dans notre exemple, nous avons plusieurs variables : le nombre de A "`nbA`", le nombre de C "`nbC`", le nombre de G "`nbG`" et de T "`nbT`" qu'il faudra calculer, le nombre total de lettres dans la séquence "`TotalNb`". Cette séquence elle-même est définie, déclarée, comme étant une chaîne de caractères débutant à l'indice 1 et de longueur non précisée, ce qu'indique la notation (*). Nous allons ajouter une variable supplémentaire qui va nous aider a progressé dans la séquence nommé "`index`" [7].

```
L 01     | nbA, nbC, nbG, nbT, TotalNb, index: integer
L 02     | sequence: character string [1: *]
```



On peut voir une séquence de caractères sous la forme d'un tableau uni-dimensionnel et chaque cellule du tableau a "`index`". Ici, si je réfère à la séquence de index alors que index vaut 1, je fais référence au contenu : la lettre T. Si l'index vaut 9 : la lettre est C ... [7].

VI.2.4.2. Initialisation et affectation de valeurs

L'étape suivante après avoir déclaré les variables c'est de les initialisés c-à-d de leur donner une valeur initiale. Avant même d'avoir commencé à compter le nombre de A, de C, de T et de G et le nombre total de lettre = 0 et la variable index la valeur de 1 [7].

```
L 01     ¦ nbA, nbC, nbG, nbT, TotalNb, index: integer
L 02     ¦ sequence: character string [1: *]
         ¦
L 03     ¦ nbA, nbC, nbG, nbT, TotalNb ← 0
L 04     ¦ Index ← 1
```

VI.2.4.3. Les instructions de contrôle

Et puis maintenant quelques instructions de contrôles qui vont dire comment les instructions de cet algorithme vont s'exécuté [7].

```
L 01  ¦ nbA, nbC, nbG, nbT, TotalNb, index: integer
L 02  ¦ sequence: character string [1: *]
      ¦
L 03  ¦ nbA, nbC, nbG, nbT, TotalNb ← 0
L 04  ¦ Index ← 1
      ¦
L 05  ¦ repeat
L 06  ¦       case sequence [index] of
L 07  ¦          "A": nbA ← nbA + 1
L 08  ¦          "C": nbC ← nbC + 1
L 09  ¦          "G": nbG ← nbG + 1
L 10  ¦          "T": nbT ← nbT + 1
L 11  ¦       endcase
L 12  ¦       TotalNb ← TotalNb + 1
L 13  ¦       index ← index + 1
L 14  ¦ until sequence [index] = "#"
```

Exécution de l'instruction [7] :

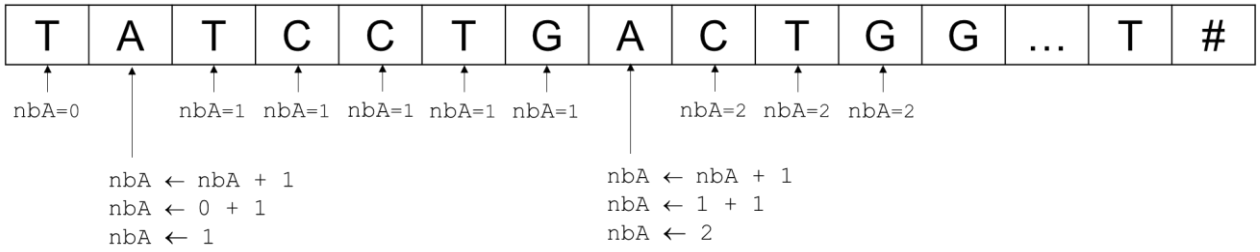
```
L 05  ¦ repeat
L 06  ¦       case sequence [index] of
L 07  ¦          "A": nbA ← nbA + 1
L 14  ¦ until sequence [index] = "#"
```

Lorsque tu trouves dans la **case** la lettre A

Le nombre de A ← prend la nouvelle valeur (nbA + 1)

N'oublions pas que nous avons initialisé les valeurs au départ !

C'est-à-dire (nbA = 0)



VI.2.4.4. Les instructions d'affichage

Cet algorithme-là, va parcourir la séquence, il va calculer le nombre de A, C, G, T, le nombre total de lettre, mais encore faut-il qu'il nous donne les résultats, pour cela il faut rajouter des instructions d'affichages [7] :

- Donc on va afficher les résultats de la longueur de la séquence dont la valeur est (TotalNb) ;
- Ensuite on va afficher la fréquence de chaque lettre dans la séquence.

```
L 01 ¦ nbA, nbC, nbG, nbT, TotalNb, index: integer
L 02 ¦ sequence: character string [1: *]
     ¦
L 03 ¦ nbA, nbC, nbG, nbT, TotalNb ← 0
L 04 ¦ Index ← 1
     ¦
L 05 ¦ repeat
L 06 ¦       case sequence [index] of
L 07 ¦          "A": nbA ← nbA + 1
L 08 ¦          "C": nbC ← nbC + 1
L 09 ¦          "G": nbG ← nbG + 1
L 10 ¦          "T": nbT ← nbT + 1
L 11 ¦       endcase
L 12 ¦       TotalNb ← TotalNb + 1
L 13 ¦       index ← index + 1
L 14 ¦ until sequence [index] = "#"
     ¦
L 15 ¦ display
L 16 ¦        "character string length": TotalNb
L 17 ¦        "%A" =(nbA/TotalNb) x 100
L 18 ¦        "%C" =(nbC/TotalNb) x 100
L 19 ¦        "%G" =(nbG/TotalNb) x 100
L 20 ¦        "%T" =(nbT/TotalNb) x 100
```

Si on donne à notre algorithme une séquence d'entré représenté ci-dessus se terminant par le caractère # marquant la fin de la séquence [7].

```
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC#
```

Il va afficher les résultats suivants :

```
character string length: 70
%A = 28.57
%C = 17.14
%G = 24.28
%T = 30.00
```

Pour que les résultats aient une signification biologique il faut réécrire les instructions d'affichage des résultats : au lieu d'affiché les fréquences de chaque nucléotide individuellement, on va afficher la fréquence "%AT" et "%GC" [7].

```
L 15 ¦ display
L 16 ¦        "character string length": TotalNb
L 17 ¦        "%AT" =(nbA+nbT/TotalNb) x 100
L 18 ¦        "%CG" =(nbC+nbG/TotalNb) x 100
```

Notre algorithme va afficher les résultats suivants [7] :

```
character string length: 70
%GC = 41.43
%TA = 58.57
```

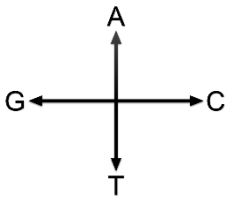
L'explication biologique est que :

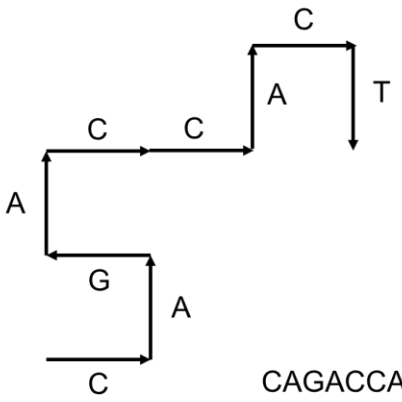
- Les régions d'ADN riches en GC sont beaucoup **plus liées** que celle riches en AT
- Les régions d'ADN riches en GC sont en moyenne **enrichies en gènes**
- Les gènes sont alors **plus compacts** (c'est-à-dire que la proportion d'introns par rapport aux exons est plus faible)

VI.2.5. L'algorithme DNA Walk

La démarche est la suivante :

- On a 4 lettres, et il y a 4 directions dans un plan.
- Le haut, le bas, la gauche, la droite.
- Plusieurs choix sont possibles





CAGACCACTCAGACCTCAAGGACCCAGAAGTGAACACC…

Quand quelque chose est long, fastidieux, sujet à erreurs et répétitifs et peut être formalisé ? On écrit un algorithme ! [7].

```
L 01  ¦ index: integer
L 02  ¦ sequence: character string [1:*]
L 03  ¦ index ←1
L 04  ¦ repeat
L 05  ¦    case sequence [index] of
L 06  ¦     "A": drawUp
L 07  ¦     "C": drawRight
L 08  ¦     "G": drawLeft
L 09  ¦     "T": drawDown
L 10  ¦    endcase
L 11  ¦    index ← index + 1
L 12  ¦ until sequence [index] = "#"
```

Résolution d'un écran

Le nombre de pixels qui peuvent être affichés dans chacune des deux dimensions

- Par exemple : 3840 x 2160 pixels (l'écran d'un téléviseur 4K)
- Un écran en full HD : 1920 x 1080 pixels

Problème !

Comment « faire rentrer » des suites de plusieurs millions, voire milliards, de segments sur un seul et même écran ?

La solution ! Changer l'échelle du dessin

Au lieu de faire notre tracé à chaque nucléotide, on va faire un tracé à chaque (n) nucléotide.

Les instructions qu'on va donner à notre algorithme [7] :

- Calculer nbA, nbC, nbG, nbT dans **la fenêtre courante de longueur L ;**
- Calculer les coordonnées (x, y) de l'extrémité du nouveau segment ;
- Tracer ce segment ;
- Avancer la fenêtre le long de la séquence.

Pour écrire cet algorithme on va utiliser ce qu'on appelle **le principe de fractionnement de complexité** (c-à-d qu'on va écrire notre algorithme en plusieurs étapes) [7].

1ère étape : On va écrire un sous-algorithme

```
L 01 | L, Index, nbA, nbC, nbG, nbT: integer
L 02 | sequence: character string [1:*]
L 03 | nbA,nbC,nbG,nbT ← 0
L 04 | for Index from 1 to L do
L 05 |     case sequence [Index] of
L 06 |         "A": nbA ← nbA + 1
L 07 |         "C": nbC ← nbC + 1
L 08 |         "G": nbG ← nbG + 1
L 09 |         "T": nbT ← nbT + 1
L 10 |     endcase
L 11 | endfor
```

2ème étape : On va intégrer notre sous-algorithme dans un algorithme plus complexe

```
L 01 | SeqLength, L, Index, InitW, nbA, nbC, nbG, nbT, NbStepsRight,
     | NbStepsUp: integer
L 02 | XEndSegment, YEndSegment, Step: real
L 03 | sequence: character string [1:*]
L 04 | InitW ← 1
L 05 | repeat
L 06 |    for Index from InitW to min (InitW + L – 1, SeqLength) do
L 07 |       nbA, nbC, nbG, nbT ← 0
L 08 |       case sequence [Index] of
L 09 |       "A": nbA ← nbA + 1
L 10 |       "C": nbC ← nbC + 1
L 11 |       "G": nbG ← nbG + 1
L 12 |       "T": nbT ← nbT + 1
L 13 |       endcase
L 14 |    endfor
L 15 |    NbStepsRight ← nbC – NbG
L 16 |    NbStepsUp ← nbA – nbT
L 17 |    XEndSegment ← NbStepsRights x Step
L 18 |    YEndSegment ← NbStepsUp x Step
L 19 |    DrawTill (XEndSegment, YEndSegment)
L 20 |    InitW ← InitW + L
L 21 | until InitW > SeqLength
```

Nous avons appliqué notre algorithme sur le génome de *Borrelia burgdorferi*, cette bactérie possède un génome de plusieurs millions de caractères et nous avons obtenu le tracé représenter dans la figure 21 [7].

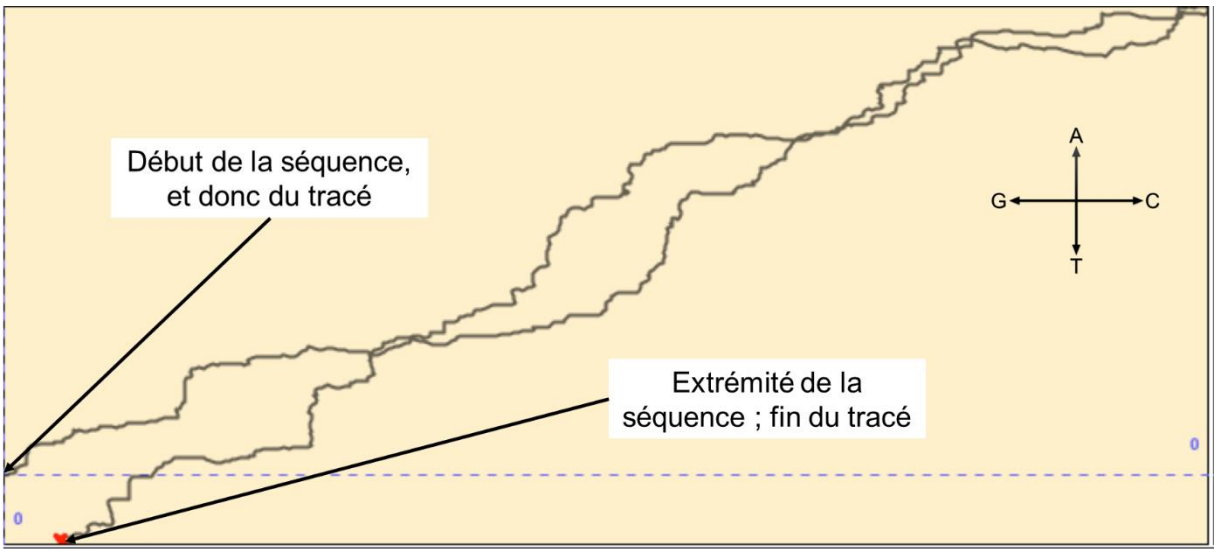


Figure 21. Résultat de l'exécution de l'algorithme (DNA walk) sur le génome de *Borrelia burgdorferi* [7] .

Notre tracé commence au début de la séquence, et on remarque, de façon assez surprenante il va vers le haut à droite [7] .

Explication : c'est qu'il y a plus de A et C dans cette partie de la séquence [7].

Et arrivé un certain point, et de façon symétrique le tracé va partir sur la gauche et vers le bas. On va l'explique de la même manière que la première partie du tracé, c'est qu'il y a plus de G et T [7] .

Donc il y a ce qu'on appelle un biais de A et C d'un côté et G et T de l'autre côté. Ce biais est différent dans les deux parties de la séquence [7].

L'explication est à rechercher dans le mécanisme de réplication de l'ADN. La molécule d'ADN est composée de 2 brins complémentaire. Une des raisons de cette asymétrie dans la molécule d'ADN et que le mécanisme de réplication de l'ADN est asymétrique (Voir le livre [8] page 33)

IV.3. L'annotation des séquences

Classiquement, on distingue trois étapes principales dans le processus d'annotation d'un génome [4] :

- **L'annotation syntaxique :** c'est l'étape qui permet d'identifier les objets génétiques présentant une pertinence biologique (séquences codantes, ARN, séquences répétées, etc.).
- **L'annotation fonctionnelle :** c'est l'étape qui permet de prédire les fonctions potentielles des objets génétiques préalablement identifiés (similitudes de séquences, motifs, structures, et c.) et de collecter d'éventuelles informations expérimentales (littérature, jeux de données à grande échelle) ;
- **L'annotation relationnelle :** c'est l'étape qui permet de déterminer les interactions que les objets biologiques préalablement identifiés sont susceptibles d'entretenir (familles de gènes, réseaux de régulation, réseaux métaboliques, etc.).

L'ensemble de ces informations seront ensuite stockées dans des **bases de données** consultables par l'expérimentateur (voir le chapitre III). L'informatique joue un rôle essentiel dans l'annotation en raison de la puissance de calcul nécessaire pour effectuer les recherches et la masse énorme d'informations générée. La plupart des outils et bases de données présentés ici sont d'accès libres par le biais d'Internet. D'autre part, les outils **bioinformatique** nécessaires aux différentes étapes de l'annotation peuvent être regroupés dans des **plateformes d'annotation.** Ces plateformes ont des interfaces conviviales utilisables par des biologistes non informaticiens (figure 22) [4].

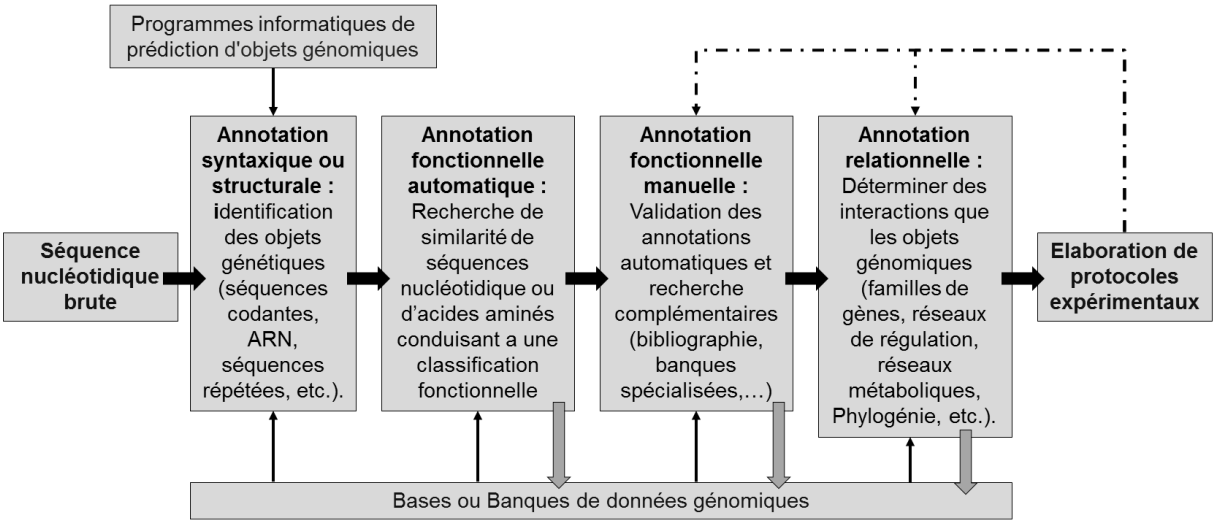


Figure 22. De la séquence nucléotidique brute aux bases de données [4].

IV.3.1. Annotation syntaxique : la recherche d'objets génétiques

Principe : La recherche d'objets génétiques passe principalement par la recherche de gènes au sens large, c'est-à-dire, toute séquence qui, transcrite et/ou traduite, peut avoir un rôle dans le fonctionnement biologique de la cellule. Cela recouvre donc les séquences codantes (Coding Sequence ou **CDS** en anglais), c'est-à-dire séquences traduites en protéines), les ARN non traduits (ARN de transfert ou ARNt, ARN ribosomaux ou ARNr, petits ARN, ARN interférents, etc.) [4].

La recherche de séquences codantes, bien qu'insuffisante pour la bonne compréhension du fonctionnement d'un génome, est néanmoins celle qui est la plus développée et pour laquelle un grand nombre d'outils informatiques existe. C'est ce que nous développerons dans cette partie [4].

IV.3.1.1. La recherche de signaux de séquence codante chez les procaryotes

L'annotation syntaxique des génomes de procaryotes est relativement plus aisée que celle des génomes eucaryotes pour les raisons suivantes [4] :

- Les génomes procaryotes sont plus petits que les génomes eucaryotes et ont surtout une densité de codage bien plus importante, de l'ordre de 80-90 %, tandis qu'elle peut aller de 70% chez la levure à quelques pourcentages chez l'humain ;
- Les gènes procaryotes sont fréquemment organisés en opéron, c'est-à-dire qu'une seule unité de transcription peut contenir plusieurs séquences codantes ;
- Les gènes procaryotes ne sont pas morcelés[1] contrairement à ceux des eucaryotes (figure 23).

[1] Qualifie un gène constitué d'une alternance de séquences codantes (les exons) et non codantes (les introns).

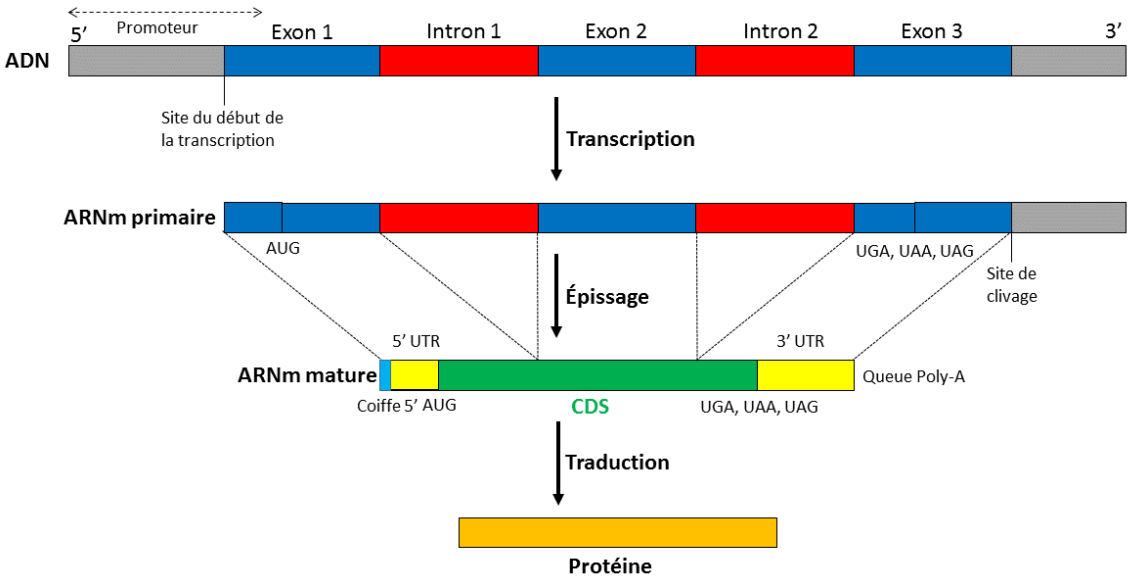


Figure 23. Schéma simplifié du dogme central de la biologie moléculaire. Certaines séquences d'ADN subissent une transcription afin de générer un ARN messager primaire. Cet ARNm subit différentes transformations, notamment **l'épissage**, par lequel les **introns sont enlevés**, pour générer un transcrit mature. Finalement, les ribosomes, à l'aide des ARNt et facteurs de traduction, traduisent la séquence codante en protéine. La séquence codante (CDS) est indiquée en vert (source Wikipédia).

a. ORF et CDS chez les procaryotes

La phase ouverte de lecture **(ORF)** est la région de l'ADN qui sépare deux codons de terminaison de la traduction (donc potentiellement codante). Dans celle-ci, une séquence codante **(CDS)** débute toujours par un **codon d'initiation** de la traduction et se termine toujours par un **codon de terminaison** de la traduction [4].

Le codon universel d'initiation de la traduction ou codon **« Start »** est le codon **ATG**. Néanmoins, chez les procaryotes il existe des codons **« Start »** plus rares tels les codons **GTG** et **TTG**. Les codons de terminaison de la traduction ou codon **« Stop »** sont les codons **TAA**, **TAG** et **TGA**.

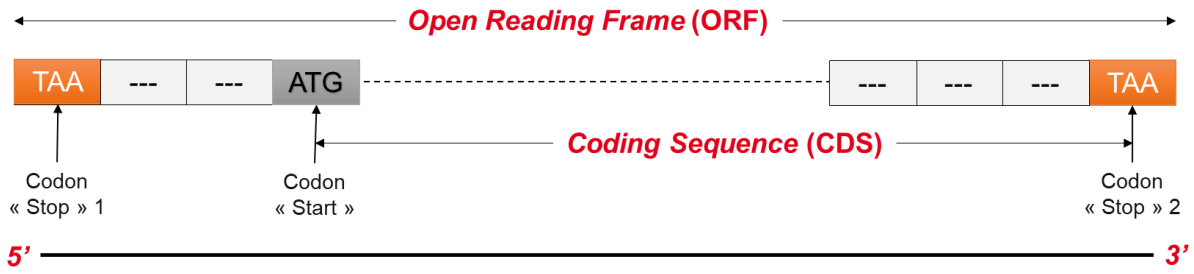


Figure 24. ORF et CDS chez les procaryotes

Chez les procaryotes, chaque séquence codante s'appelle un cistron. Beaucoup d'ARN messagers procaryotes sont **polycistroniques** : ils contiennent plusieurs cistrons ou **CDS** et codent donc pour plusieurs protéines (Figure 25) [4].

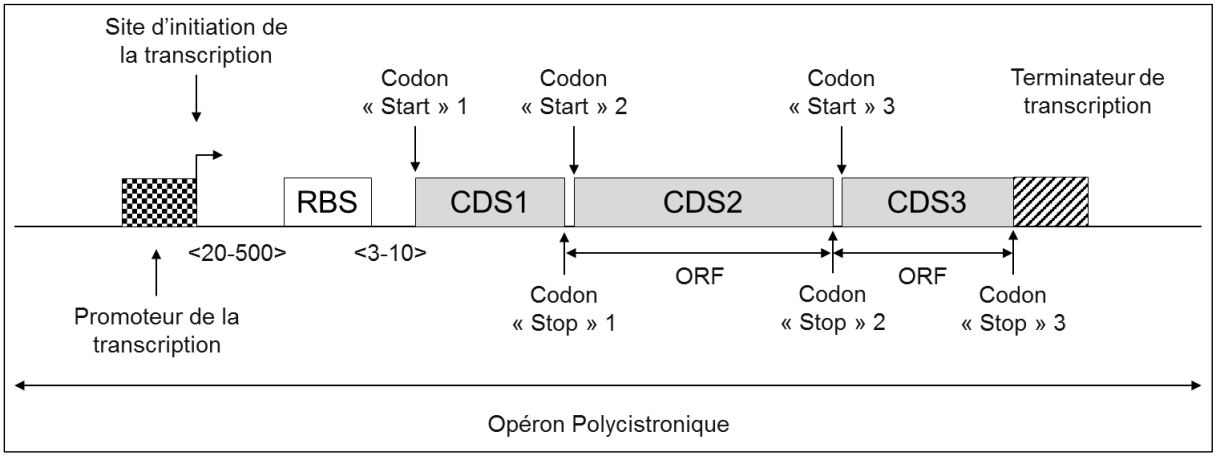


Figure 25. La notion de séquence codante chez les procaryotes [4].

b. Les autres signaux indicateurs de la présence d'une séquence codante chez les procaryotes

La séquence de Shine-Dalgarno ou site de liaison au ribosome (**RBS**) se situe entre **3 à 10** nucléotides en amont du codon « Start ». C'est une région riche en purine de 5-6 nucléotides qui permet au ribosome de se fixer spécifiquement sur les AUG correspondant à un véritable codon « Start » (cf Figure 25). Ce signal permet également à l'annotateur de distinguer un véritable codon « Start » d'un codon ATG codant une méthyonine (figure 26). Chez *Escherichia coli*, la séquence consensus du RBS est : 5' -AGGAGG-3' [4].

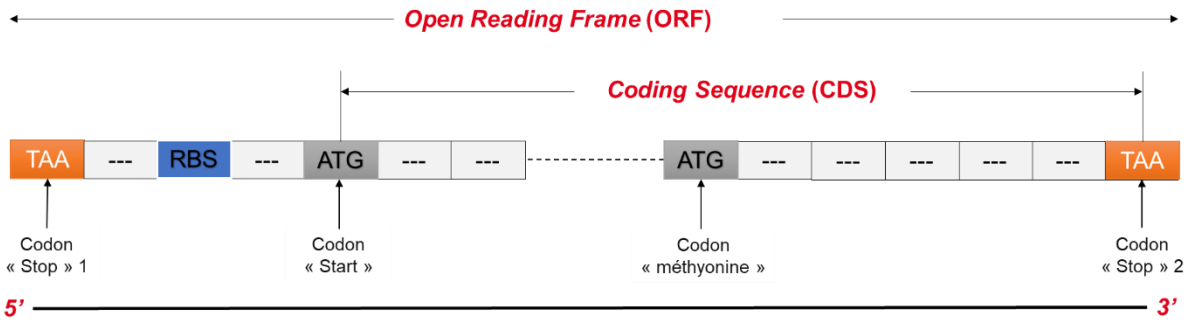


Figure 26. Le site de liaison au ribosome (**RBS**).

Le promoteur ou région promotrice est la région reconnue spécifiquement par le complexe entre l'ARN polymérase (enzyme qui assure la transcription de l'ADN) et le facteur sigma (facteur protéique qui assure la spécificité de l'initiation de la transcription).

Le promoteur est constitué de deux éléments (figure 24) : une séquence très bien conservée qui est localisée environ 10 nucléotides avant le site d'initiation de la transcription, la boîte TATA ou -10 ; une séquence moyennement conservée qui est localisée environ 35 nucléotides avant le site d'initiation de la transcription, la boîte - 35. Dans un opéron, le promoteur ne se localise qu'en amont de la première CDS, puisque l'opéron est une unité de transcription (voir figure 27) [4].

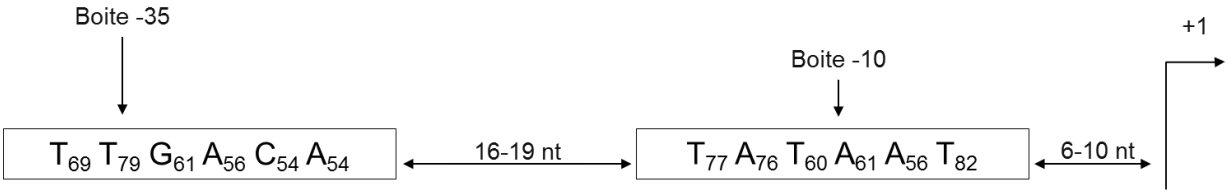


Figure 27. Les séquences consensus des promoteurs reconnus par le facteur sigma 70 chez *E. coli*. Les nombres en indice indiquent le pourcentage d'occurrence de chaque nucléotide dans la définition de la séquence consensus. Le +1 indique le site d'initiation de la transcription [4].

Le terminateur de transcription est une séquence grâce à laquelle le complexe de transcription va se désassembler et ainsi terminer la transcription (voir figure x). Les terminateurs sont des séquences **palindromiques**[2] riches en GC suivies de séquences riches en A (cas des terminateurs Rho-indépendants) ou non (cas des terminateurs Rho-dépendants) [4].

La détection d'un **RBS**, d'un promoteur ou d'un terminateur de transcription peut valider l'existence d'une séquence codante *a posteriori*. Néanmoins, leurs consensus sont trop faiblement conservés pour qu'ils constituent des signaux fiables *a priori* [4].

IV.3.1.2. La recherche de signaux de séquence codante chez les eucaryotes

Chez les génomes eucaryotes, l'annotation syntaxique est nettement plus compliquée. Pour les raisons suivantes [4] :

- Les génomes eucaryotes ont une faible densité de codage. Il y a donc de larges régions génomiques sans séquence codante ;
- Les gènes eucaryotes sont morcelés ; ils subissent des modifications de la séquence nucléotidique (épissage) du pré-ARN messager. L'épissage consiste en l'excision d'une ou plusieurs séquences (introns). Les séquences non excisées (exons) forment après raboutage entre elles la « séquence codante » (Figure 28) ;

[2] Une séquence palindromique est une séquence d'acide nucléique — ADN ou ARN — identique lorsqu'elle est lue dans le sens 5' → 3' sur un brin ou dans le sens 5' → 3' sur le brin complémentaire. Exemple :
5'-GAATTC-3'
3'-CTTAAG-5'

- Enfin, l'épissage peut être alternatif : différents profils d'épissage existent pour un même pré-ARN messager et par conséquent un gène peut produire différentes CDS (voir figure 28).

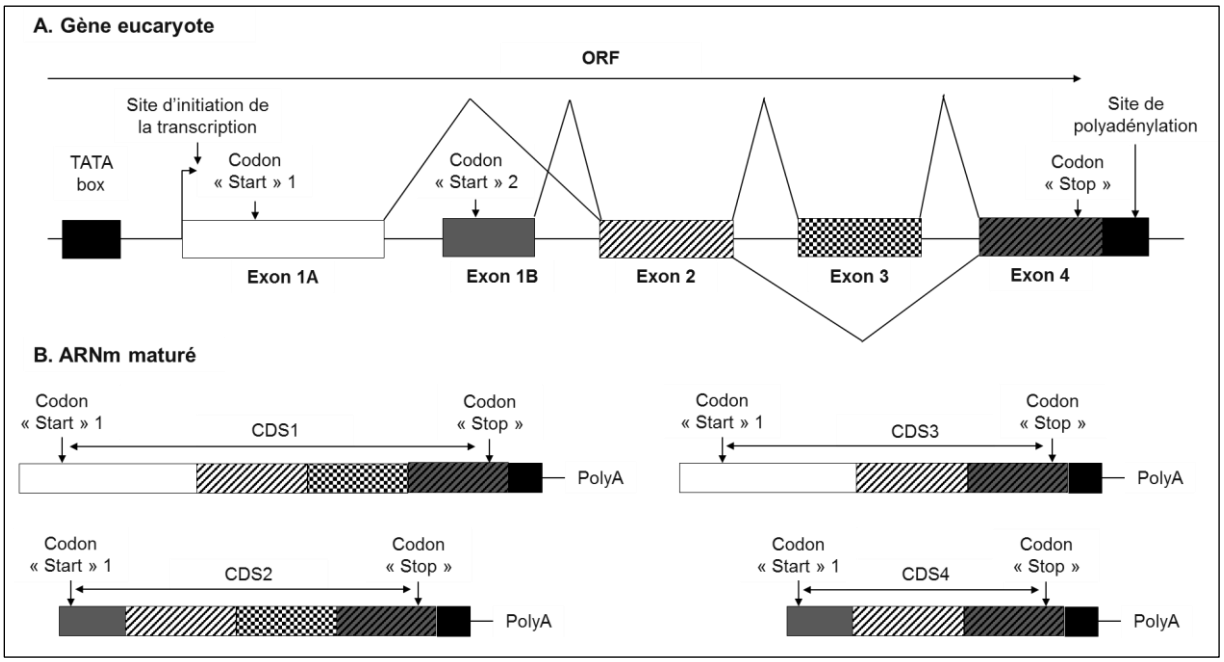


Figure 28. La notion de séquence codante chez les eucaryotes [4].

a. Promoteurs et signaux 5'

Outre les codons « Start » et les codons « Stop », différents types de signaux marquant la région 5' de la séquence codante peuvent être recherchés. Les signaux de transcription sont les suivants [4] :

- Les séquences promotrices reconnues par les ARN polymérases. Chez les eucaryotes, il y a trois types d'ARN polymérase (ARNpolI, ARNpolII et ARN polIII). Chaque ARN polymérase reconnaît un type de promoteur. Les promoteurs PolI se trouvent en amont des gènes d'ARN ribosomaux 18S et 28S. Les promoteurs PolII se trouvent en amont des gènes d'ARN messagers. Les promoteurs PolIII se trouvent en amont des gènes d'ARN ribosomaux 5S et d'ARN de transfert. En amont des séquences codant des protéines, on recherche donc une boîte TATA, une séquence conservée, riche en AT et de 8 nucléotides, qui se trouve dans les promoteurs PolII 25 à 30 nucléotides en amont du site d'initiation de la transcription (Figure 29) ;
- Les sites de liaison aux facteurs de transcription ;
- L'initiateur (INR), une séquence, faiblement conservée, qui se trouve près du site de début de transcription entre les positions - 3 et + 5 ;

- Les îlots CpG. Ce sont des régions de 1 à 2 kb, riches en dinucléotides CG, qui sont fréquemment associées aux régions 5' des gènes de vertébrés et qui s'étendent sur le promoteur et le premier exon.
- Pour les signaux de traduction, on recherche en particulier le site de liaison au ribosome ou séquence de **Kozak** localisée en amont du codon « Start » [4].

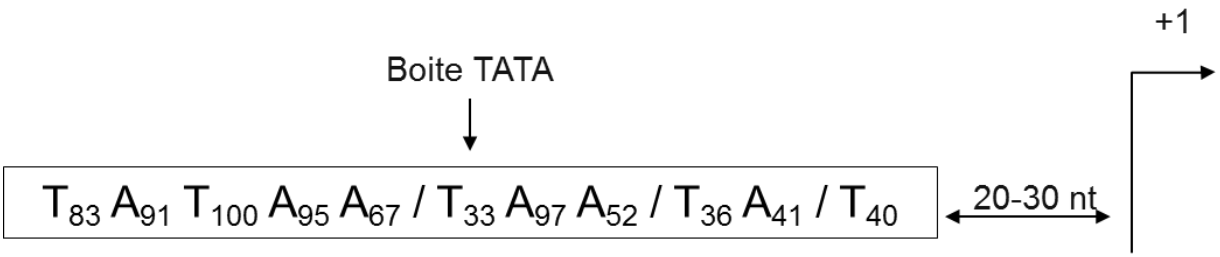


Figure 29. La séquence consensus de la boîte TATA chez les eucaryotes. Les nombres en indice indiquent le pourcentage d'occurrence de chaque nucléotide dans la définition de la séquence consensus. Le +1 indique le site d'initiation de la transcription [4].

b. Jonctions exons-introns

Les introns possèdent quatre signatures importantes dont les deux premiers indiquent les jonctions exons-introns (Figure 30) [4] :

- Le site donneur /GTRAGT à l'extrémité 5' de l'intron ;
- Le dinucléotide GT est systématiquement exclu des ARNm matures ;
- Le site accepteur NYAG/G à l'extrémité 3' de l'intron ;
- Le dinucléotide AG est systématiquement exclu des ARNm matures ;
- Le point de branchement CTRAY avec un A qui joue un rôle central dans le processus d'épissage ;
- Une région riche en pyrimidine entre le point de branchement et le site récepteur.

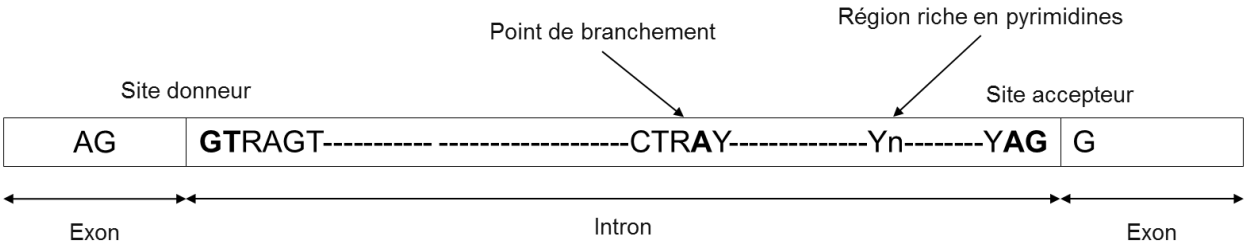


Figure 30. Signaux de jonction exon-intron chez les vertébrés. Y : pyrimidines ; R : purines [4].

c. Signaux 3'

L'exon terminal contient un ensemble de signaux indicateurs de la terminaison de la transcription. Pour les ARN messagers transcrits par l'ARN polymérase II, ces signaux sont également nécessaires à la polyadénylation car ce dernier événement est couplé à la terminaison de la transcription [4] :

- Le signal de polyadénylation : 5'-AAUAAA-3' ou 5'-AUUAAA-3' ;
- Un signal de clivage. C'est un dinucléotide CA assez mal conservé, localisé 10 à 30 bases en aval du signal de polyadénylation.
- À ce niveau, l'ADN est clivé avant l'ajout de la queue polyA ;
- Une région riche en GU de séquence variable, 20 à 40 bases après le site de clivage.

IV.3.1.3. Analyse du contenu en base des séquences codantes

La recherche de signaux indiquant la présence de séquences codantes est insuffisante. Ceci est vrai chez tous les génomes, mais encore plus marqué chez les génomes eucaryotes. Chez ces derniers, les signaux peuvent être très dégénérés (consensus peu conservé) et la structure mosaïque des gènes peut être une source d'erreur. Une seconde approche est utilisée pour rechercher les séquences codantes dans les génomes : l'analyse du contenu des séquences et des biais de ce contenu dans les régions codantes par rapport aux régions non codantes [4].

a. La composition en base

Il existe des biais entre les séquences codantes et non codantes dans la composition en base de séquences di-nucléotidique, hexa-nucléotidique, etc. Ce biais est utilisé pour la recherche de séquences codantes et en particulier pour distinguer les introns des exons chez les eucaryotes [4].

b. Le biais d'usage des codons

L'abondance et l'utilisation des acides aminés est variable d'un organisme à l'autre. Cela entraîne des fréquences différentes pour chacun des codons au sein d'un génome. Néanmoins, on pourrait s'attendre à ce que les codons synonymes soient employés avec la même fréquence. Or ce n'est pas le cas. On parle alors de biais d'usage des codons [4].

Il existe un biais d'usage des codons propre à chaque espèce. Au sein d'un génome, il existe également des biais propres à certaines régions génomiques et même à certains gènes. Il a été observé que les gènes les plus exprimés sont les plus biaisés, les codons les plus

utilisés étant ceux pour lesquels les ARN de transfert sont les plus nombreux. Le biais d'usage des codons est utilisé pour repérer [4] :

- Des signatures de CDS chez les procaryotes ;
- Des signatures d'exons codants chez les eucaryotes ;
- Des signatures de transferts horizontaux de matériel génétique puisque le biais d'usage du code est souvent différent entre espèces différentes.

IV.3.1.4. Programmes bioinformatiques d'annotation syntaxique

Les meilleurs programmes informatiques d'annotation syntaxique sont celles qui combinent la détection de signaux de gènes et l'analyse du contenu des gènes (Tableau 6). Elles utilisent des algorithmes qui, après une phase d'apprentissage sur un ensemble de données, permettent de différencier les régions géniques des régions intergéniques. Dans la plupart des cas, ces méthodes font appel aux modèles cachés de Markov **(HMM)**. Les HMM permettent de modéliser et de prédire des caractéristiques locales de séquences moléculaires (par exemple, composition en base) tout en intégrant des connaissances provenant de recherches antérieures dans les analyses (par exemple la séquence du RBS ou du promoteur) [4].

Tableau 6. Exemples de programmes de recherche de séquences codantes et leurs applications [4].

Programmes	Procaryotes	Eucaryotes
GenMark-P	+	
Glimmer	+	
Genemark-E		+
Grail		+
Genescan		+
Genie		+

Ces méthodes ont évidemment des défauts :

- Caractérisation incomplète ou imprécise de la structure du gène ;
- Caractérisation de faux positifs (séquences prédites comme des CDS ou des exons et qui n'en sont pas) ;
- Incapacité à identifier certains vrais gènes lorsque ceux-ci ne sont pas canoniques.

En particulier chez les eucaryotes, seule l'annotation fonctionnelle par comparaison avec des banques de données d'expression (voir plus bas) permettra de valider les gènes prédits lors de l'annotation syntaxique et de trouver des gènes que les programmes informatiques de prédiction ont échoué à identifier [4].

IV.3.2. Annotation fonctionnelle : la recherche de fonctions potentielles

L'annotation fonctionnelle permet d'attribuer à des objets génomiques prédits par l'annotation syntaxique des fonctions potentielles. En aucun cas, l'annotation ne permet d'avoir accès à la fonction réelle. Seule l'expérimentation l'autorise [4].

L'annotation fonctionnelle est fondée sur la recherche de similarité avec des séquences nucléotidiques, des séquences d'acides aminés ou éventuellement des structures déjà décrites dans les bases de données. Plus le nombre de données augmente dans les bases de données internationales, plus la chance de trouver des éléments déjà décrits dans la nouvelle séquence d'un génome augmente [4].

En général, l'étape d'annotation s'effectue en deux étapes : une phase automatique qui s'effectue grâce à des programmes informatiques de comparaison et une phase manuelle au cours de laquelle l'annotateur peut corriger le cas échéant la première phase [4].

IV.3.2.1. Les outils bioinformatique de comparaison de séquence

Les séquences peuvent être comparées avec des programmes comme **FASTA** (**FAST-ALL)** ou **BLAST**. Ces instruments de recherche de similarité reposent sur la notion d'alignement local. Les algorithmes d'alignement local recherchent dans des paires de séquences des régions isolées qui ont un haut degré de similitude.

Nous décrirons ici l'usage du programme le plus couramment utilisé (cité dans google scholar 68461 fois), **BLAST** (Altschul et al. 1990). L'utilisateur fournit une séquence-requête qui est alors comparée à toutes les séquences d'une base de données choisie. Différents sous-programmes existent selon la nature de la séquence-requête et des séquences de la base de données (Tableau 8) [4].

Tableau 7. Les différents programmes BLAST [4].

Nom du programme	Nature de la séquence-requête	Nature des séquences des bases de données
Blast ou Blastn	Nucléotides	Nucléotides
Blastp	Acides aminés	Acides aminés
Blastx	Nucléotides traduits dans les 6 phases de lectures	Acides aminés
Blastn	Acides aminés	Nucléotides traduits dans les 6 phases de lectures
Blastx	Nucléotides traduits dans les 6 phases de lectures	Nucléotides traduits dans les 6 phases de lectures

Le résultat de BLAST classe les résultats de similitude en fonction d'un indice de significativité appelé E-value (valeur attendue), le résultat le plus significatif étant le premier de la liste. La valeur E est le nombre d'alignements différents ayant le même degré de

similitude, et que l'on peut espérer trouver au hasard, s’il n'existait pas de vraie séquence similaire dans la base de données. Donc, pratiquement, plus la valeur E est proche de 0, moins la similitude est due au hasard [4].

Références bibliographiques

[1] - Rechenmann, F., & Quinkal, I. (2004). Entre biologie, informatique et mathématiques : la bioinformatique. Interstices. https://hal.inria.fr/inria-00000932v1

[2] - Pevsner, J. (2015). Bioinformatics and functional genomics. Third Edition, John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc. Companion Website: www.wiley.com/go/pevsnerbioinformatics

[3] - Jamet P. (2006). Analyse bioinformatique des séquences – Cours de l'Université de Tours_ Génétique. France. http://genet.univ-tours.fr/fichiers_de_base/gen001400.HTM

[4] - Gaudriault S., & Vincent R. (2009). Génomique. Editions De Boeck Université, Bruxelles, Belgique.

[5] - Oezratty O. (2013). Les technologies de séquençage du génome humain. https://www.oezratty.net/wordpress/2012/technologies-sequencage-gnome-humain-3/?output=pdf

[6] - Fondrat C., Granger G., Brunet M., Lheureux C., Fermanian C., Mortazavi R., Latour C., Kalfon J. (2017). Cours d'autoformation en bioinformatique. Site Web de l'Université René Descartes Paris 5. France. http://www.dsi.univ-paris5.fr/bio2/autof2/index.htm

[7] - Rechenmann F. & Parmentelat T. (2017). Cours de Bioinformatique : algorithmes et génomes. Inria. Université de Technologie Ouverte Pluri-partenaires. https://www.fun-mooc.fr/courses/course-v1:inria+41003+session03/about

[8] - Maftah, A., & Petit, J. M. (2018). Mini Manuel de Biologie moléculaire-4e éd. Cours+ QCM+ QROC. Dunod.