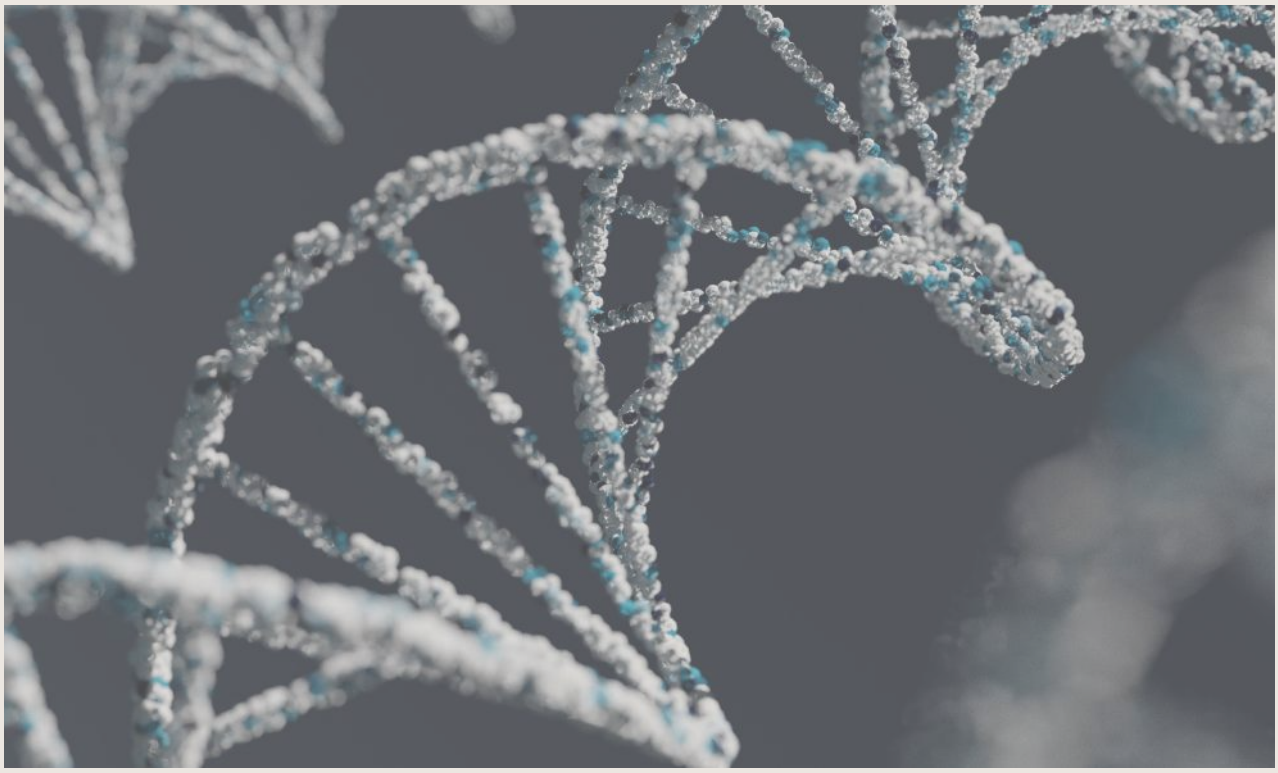
STOCKAGE DES DONNÉES SUR L'ADN

QUEL AVENIR ?



Université de Bretagne Occidentale

Réalisé par :
M. Khaled BELKACEMI
Mme Hassina AOUDIA

Encadré par:
Professeur Gilles BUREL

Dans cette REVUE SCIENTIFIQUE

01 PRESENTATION DU PARCOURS

- Signal et télécommunications Université de Bretagne Occidentale

02 PRÉSENTATION DU PROJET

- Vue globale sur le projet portant sur le stockage des données sur l'ADN (méthode de Goldman)

03 DEROULEMENT

- Problématique
- Aspects biologiques
- Aspect technique : Explication détaillée de la méthode

04 INTÉRÊT

- Pourquoi le stockage des données sur l'ADN ?

PARCOURS

Signal et télécommunications

Université de Bretagne Occidentale

Le parcours Signal et Télécommunications (ST) forme des diplômés spécialisés en télécommunications et en systèmes numériques pour s'insérer dans les métiers du traitement du signal, de l'image et vision que l'on retrouve dans de nombreux secteurs industriels (télécoms, médical, énergie, automobile, aéronautique et spatial, robotique, ...).

A l'issue du master

COMPETENCES

- Traitement du signal et de l'image
- Télécommunications
- Cybersécurité
- Simulations et expérimentations

Unités d'enseignement

- Communications numériques
- Traitement d'images et de vidéos
- Traitement du signal
- Filtrage
- Estimation et modélisation

Biographie du professeur encadrant

Professeur Gilles BUREL

Gilles BUREL

a obtenu

- le diplôme d'ingénieur de Supélec en 1988
- le doctorat de l'université de Brest en 1991
- l'habilitation à diriger des recherches en 1996

a dirigé

- 23 thèses

a travaillé

- à Thomson CSF, puis Thomson MultiMedia, dans le domaine du traitement des images et des communications numériques de 1988 à 1996

a été/ est

- directeur adjoint du Lab-STICC(UMR CNRS 6285), l'un des plus grands laboratoires français dans le domaine des télécommunications et du traitement de l'information de 2008 à 2016
- Professeur à l'université de Bretagne Occidentale (UBO) depuis 1997
- auteur de 180 articles scientifiques en revues et conférences

Principaux domaines d'intérêt de Gilles BUREL

- les communications quantiques
- l'acquisition compressée et la surveillance du spectre

Biographie des étudiants

Khaled BELKACEMI

Khaled BELKACEMI : actuellement étudiant en Master 2 Signal et télécommunications à l'université de Bretagne Occidentale (UBO)

a obtenu

- le diplôme du baccalauréat série mathématiques en Algérie en 2015
- le diplôme licence en télécommunications de l'université de Bejaia, Algérie en 2018

est

- stagiaire au sein d'Airbus SAS

Principaux domaines d'intérêt de Khaled BELKACEMI

- traitement du signal
- les systèmes embarqués

Hassina AOUDIA

Hassina AOUDIA : actuellement étudiante en Master 2 Signal et télécommunications à l'université de Bretagne Occidentale (UBO)

a obtenu

- le diplôme du baccalauréat série Sciences expérimentales en Algérie en 2017
- le diplôme licence en réseaux et télécommunications de l'université de Tizi Ouzou, Algérie en 2021

est

- stagiaire au sein du groupe Orange Innovation

Principaux domaines d'intérêt de Hassina AOUDIA

- les communications numériques et optiques

A PROPOS DU PROJET

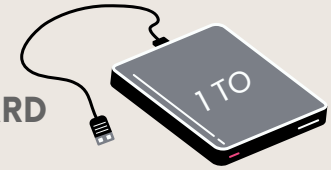
Projets longs (5 mois) proposés aux étudiants par des industriels ou des chercheurs sur des problématiques de recherche actuelles du Lab-STICC. Des semaines complètes sont dédiées aux projets avec un travail en autonomie (par binôme d'étudiants) et un suivi régulier par les tuteurs de projets.

Le projet **"Stockage des données sur l'ADN "** proposé et encadré par le professeur Gilles BUREL consiste fondamentalement à développer des programmes, en suivant la méthode de Nick GOLDMAN, permettant de coder les données et les stocker sur les brins d'ADN pour que le décodage et la reconstitution du contenu des fichiers codés soient possibles par n'importe quel lecteur d'ADN sous la contrainte:

Obtenir des codes génétiques corrects des points de vue chimique et biologique.

Déroulement

1 MILLIARD



7500

- Statista estime à 50 Zettaoctets la quantité de données en 2020. Pour rappel, un Zettaoctet équivaut à mille milliards de milliards d'octets, soit un milliard de disques durs d'un TO ou de 1000 GO.
- IDC (International Data Corporation) : quant à IDC, il prévoit que le volume total de données produites atteindra 175 Zettaoctets en 2025 et 2000 Zettaoctets en 2035.

Pour garder cette gigantesque masse de données , il faut 4800 centres de stockage dans 127 pays.

Mais

✔ **Consommation énorme d'énergie** : une consommation énergétique de 2 ou 3 % de la consommation électrique mondiale.

+ Un data center moyen consomme 100 fois plus d'électricité qu'un grand immeuble commercial

+ Les serveurs et les systèmes de climatisation sont les plus consommateurs d'énergie dans un data center

✔ **Occupation de grands espaces** : les data centers nécessitent beaucoup d'espaces et nécessiteront énormément d'espace.

+ Un data center occupe en moyenne une surface de 5000 mètres carrés

+ La surface occupée par les 7500 data centers qui existent dans le monde est supérieure à 37 500 000 mètres carrés

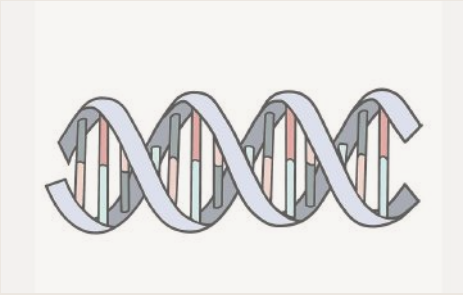
✔ **Augmentation de la pollution** : aujourd'hui, les data centers sont responsables de 2 % des émissions de carbone dans le monde.

+ Avec la même croissance qu'aujourd'hui et sans amélioration, les data centers produiront 359 mégatonnes de CO_2 ce qui est équivalent à 125 millions voitures

Déroulement

“Aspects biologiques ”

ADN , "acide désoxyribonucléique" une macromolécule chimique qui se trouve au cœur de toutes les cellules , cette molécule est le support de l'information génétique.
Chaque molécule d'ADN est contenue dans un chromosome sous forme d'un long fil enroulé sur lui même.



Gènes

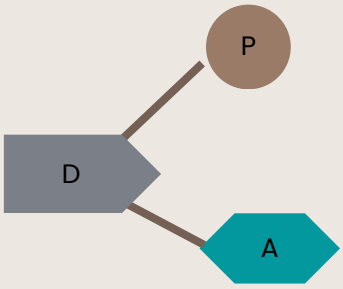
La molécule d'ADN est constituée d'un ensemble de fragments , nommés également gènes , ces derniers sont les porteurs d'information génétique.
Le fragment d'ADN, gène, se compose de deux brins enroulés l'un autour de l'autre ce qui donne la forme double hélice à l'ADN.

Nucléotides

Les brins d'ADN se composent d'une suite de nucléotides (Adénine, Cytosine, Guanine, Thymine) A,C,G,T respectivement. Dans un brin d'ADN chaque nucléotide a son complément c'est à dire A avec T, C avec G.
Cet enchainement de nucléotides dans un brin d'ADN définit le message codé et donc l'information génétique.

Nucléotides

Un nucléotide est un assemblage d'une molécule de Glucide (le Désoxyribose), un phosphate et une base azotée " A C G T"



Structure chimique d'un nucléotide

Le séquençage d'ADN

Le séquençage d'ADN consiste à lire et déterminer l'enchainement des nucléotides dans une molécule d'ADN, ce qui nous permet de récupérer l'information génétique stockée dans cette molécule .

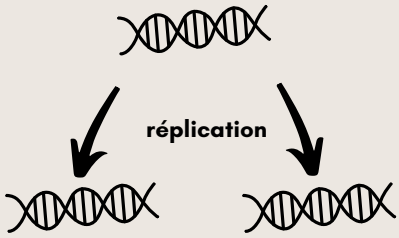
L'information génétique est portée sur une molécule d'ADN. La succession et l'enchainement des nucléotides définissent l'information.

Comment ces molécules sont-elles produites ?

Réplication et la Synthèse d'ADN

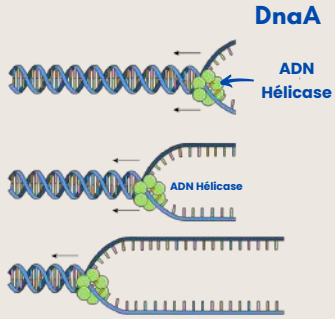
La réplication d'ADN est une opération qui permet de produire deux copies de la molécule d'ADN à partir d'une seule molécule appelée l'ADN père en gardant la même information génétique . C'est l'écriture !

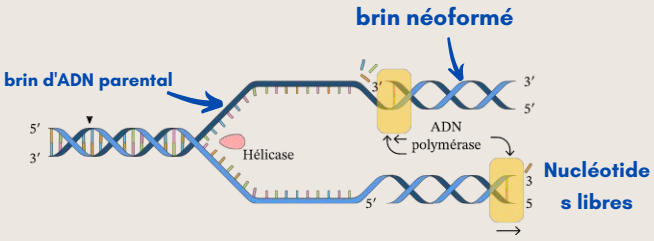
La réplication d'ADN se fait en plusieurs étapes:



1 INITIATION

La protéine ***DnaA*** vient se fixer sur un endroit bien précis de l'ADN, c'est la définition de l'**origine de réplication**.
Une fois l'origine de la réplication est définie, un autre type d'enzyme intervient, c'est la protéine ***Hélicase***, elle déroule toute la molécule d'ADN et prolonge l'œil de réplication tout au long en détruisant toutes les liaisons en formant ***l'œil de réplication***, le bout de chaque œil de réplication est appelé ***fourche*** de réplication.





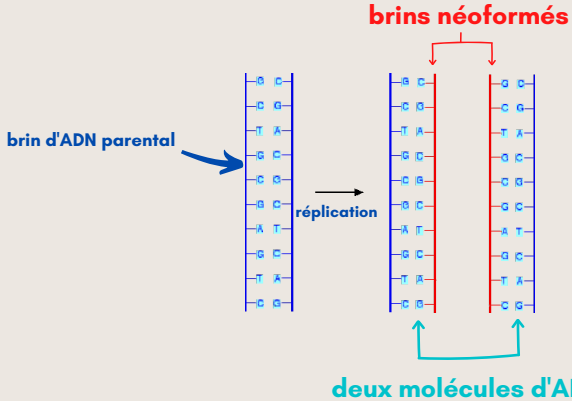
2 ELONGATION OU SYNTHESE

Une fois l'ouverture de l'œil est prolongée, chaque brin d'ADN sera traduit en un autre brin complémentaire, cette traduction ou écriture est assurée par un autre type d'enzyme par exemple ***"ADN Polymérase"***.
à l'issue de cette étape, deux brins d'ADN sont formés, ce sont les brins ***néoformés***.

3 TERMINAISON

Lorsque une fourche de réplication rencontre une autre fourche ou rencontre un signal de terminaison, la réplication prend fin.
Deux molécules d'ADN sont formées à partir d'une seule,, chaque ***"brin parent"*** s'associe avec son complémentaire ***"brin néoformé"*** pour former une molécule .

C'est pourquoi on appelle la réplication d'ADN un processus ***semi-conservatif***



“Processus d'écriture et lecture des données ADN”

Analogie bits/nucléotides

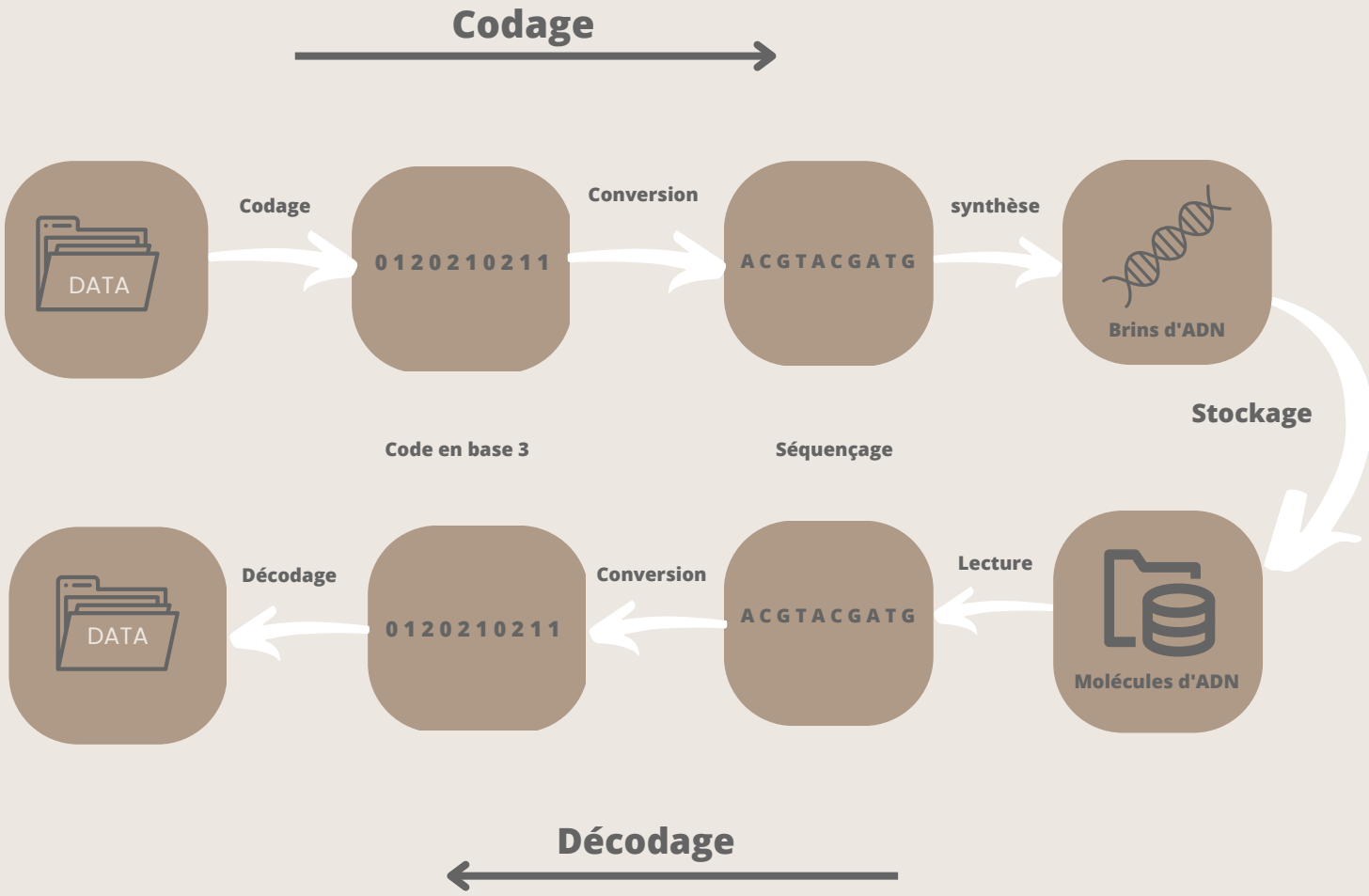
De manière générale

Actuellement, les données sont numériques et elles sont stockées dans des supports électroniques sous la forme binaire, c'est à dire l'information est représentée par une suite de bits (des 0 et des 1).
Le stockage ADN consiste à encoder ces données binaires afin de les stocker sur l'ADN. Pour ce faire, chaque bit, (0 ou 1), qui représente l'élément le plus petit de l'information est équivalent à un nucléotide, c'est à dire, A,C,G ou T.
Ces lettres représentent les quatre principaux composants d'ADN, Adénine, Cytosine ,Guanine et Thymine.
L'information est donc stockée comme dans le vivant, sur de longs fragments d'ADN en double hélice. Ces fragments sont ensuite stockés dans des capsules métalliques. La lecture des données qui sont agencées comme dans un disque dur, se fait par un séquenceur ADN.

De manière précise

Dans la méthode de Nick Goldman, les données numériques qui sont représentées par une suite de bits
(0 et 1) seront plutôt représentées par une série de trits (0,1 et 2) en utilisant le dictionnaire faisant la correspondance entre chaque code ASCII et son code équivalent en base 3, sachant que la longueur des codes du dictionnaire est entre 5 et 6 trits .Ce passage de la base 2 vers la base 3 permet d'une part d'éviter la contrainte de la répétition des nucléotides et ce en utilisant le raisonnement de Table de conversion illustrée au dessous. D'une autre part, ce passage permet de réduire la taille des données codées car chaque code ASCII est codé sur
5 ou 6 éléments ternaires (trits) au lieu de 8 éléments binaires (bits).

Schéma simplifié du processus



Codage : Avoir les codes ASCII des données et utilisation du codage de Huffman pour avoir la suite de trits correspondante

Conversion : Convertir la suite de trits en nucléotides ACGT sans qu'il y ait de répétition

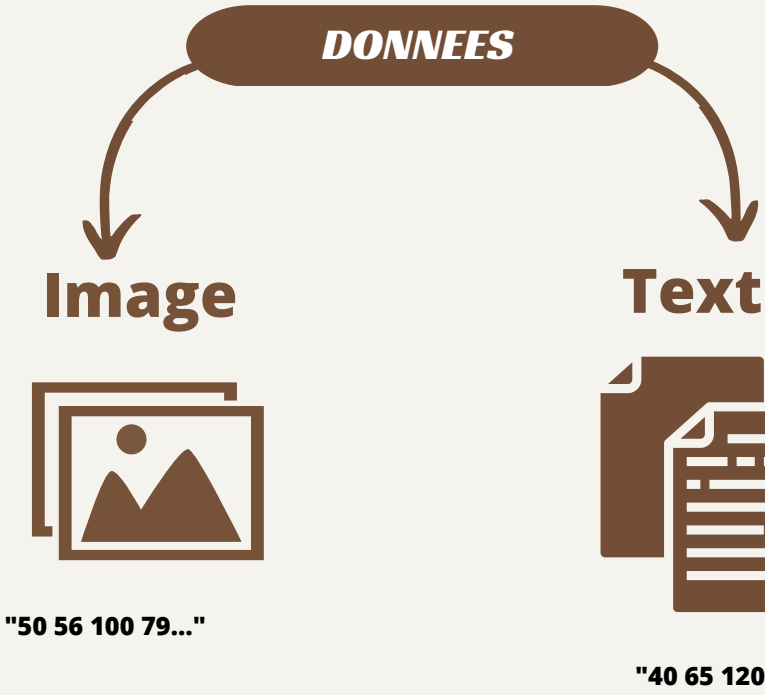
Synthèse : Synthétiser des molécules d'ADN à partir du code ADN obtenu

Stockage : Stocker les molécules d'ADN portant les données dans des capsules

En effectuant les opérations inverses du codage, les données seront reconstruites correctement

CODAGE EN ASCII

Explication détaillée de la méthode développée au cours de ce projet



ASCII : " Americain Standard Code for Information Interchange"

Un standard pour l'affichage de caractères à travers des dispositifs électroniques

- Décimal : entre 0 et 255
- Binaire : entre 0000 0000 et 1111 1111

Description étendue de l'usage de quelques caractères de contrôle

Dec	Hex	Binary	HTML	Char	Description
0	00	00000000	&#0	NUL	Null
1	01	00000001	&#1	SOH	Start of Header
2	02	00000010	&#2	STX	Start of Text

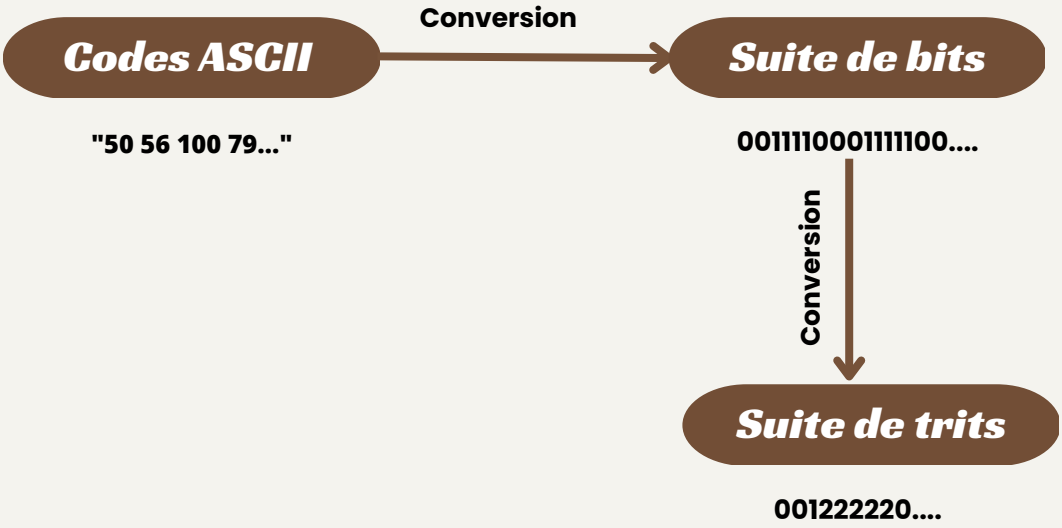
Un fichier contient une représentation d'informations(données), il peut contenir un texte, une image, un son...etc. Les machines permettant la lecture des fichiers interprètent ces derniers comme étant une suite de 0 et 1, des bits.

Dans un fichier texte à titre d'exemple, chaque lettre ou de manière générale chaque caractère est codé.

La table ASCII fait la correspondance entre les caractères et leurs équivalents en binaire.

PROJET DEVELOPPE SOUS MATLAB

CODAGE DE HUFFMAN BASE-3



Explication détaillée de la méthode développée au cours de ce projet

Exemple

Caractère	Base-2	Base-3
a (minuscule)	01100001	01112
M (majuscule)	01001101	01211
4 (quatre)	00000100	22112

Le codage de Huffman est un codage de compression sans pertes.

La correspondance d'un caractère dans un premier message n'est pas toujours la même dans un second message.

Dans notre cas, il s'agit d'un codage permettant d'avoir des suites de 5 à 6 trits encodant chaque caractère au lieu de 8 bits, ça permet alors de réduire la taille du message.

PROJET DEVELOPPE SOUS MATLAB

03

CODAGE ADN (ACGT)

Explication détaillée de la méthode développée au cours de ce projet

Conversion de la suite des trits en nucléotides sans qu'il y ait des nucléotides répétés

Suite de trits → **Conversion** → **Suite de nucléotides**

001222220.... → **AGTGTCTCG....**

Table permettant de faire la conversation en évitant la répétition

Nucléotide précédent	Prochain trit à coder		
	0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

Table de conversion

PROJET DEVELOPPE SOUS MATLAB

MATLAB

04

CONSTRUCTION DES FRAGMENTS (ACGT)

Explication détaillée de la méthode développée au cours de ce projet

La suite de nucléotides précédente sera découpée en fragments

Longueur des fragments et chevauchement

Chaque fragment est d'une longueur de 100 nucléotides

100

AGTGTCTCTC...TACGCACGCT **TGAGAGTCGC...GTCTCTCTCT** ...

CGCTCTGTCG...ATAGCACGCT **ACTGTCTCTC...GACGATCGCT**

Chaque deux fragments qui se suivent subissent un chevauchement de 75 %

AGTGTCTCTC...TACGCACGCT

TACGCACGCT...ATAGCACGCT

ATAGCACGC...GTCTCTCTCT

GTCTCTCTCT...GACGATCGCT

100

PROJET DEVELOPPE SOUS MATLAB

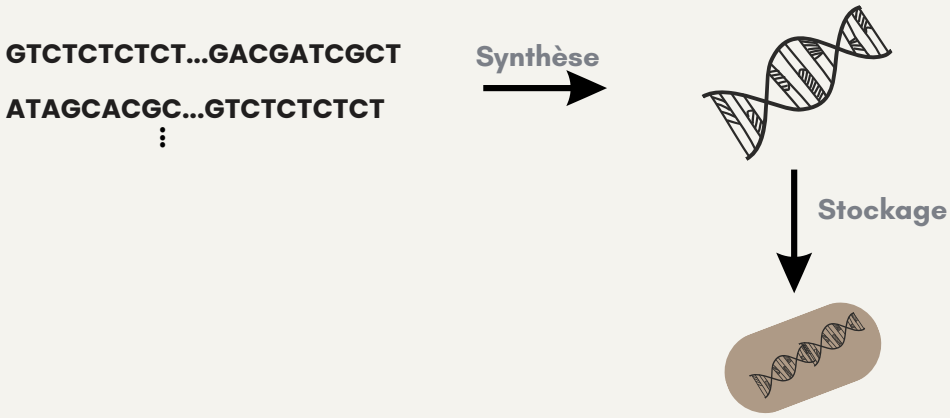
05

SYNTHÈSE, STOCKAGE ET LECTURE

Explication détaillée de la méthode développée au cours de ce projet

PROJET DEVELOPPE SOUS MATLAB

Synthèse et stockage



La suite de nucléotides obtenue à l'issue de l'étape de codage sera synthétisée,
des liaisons entre chaque nucléotides de la suite seront créées,
c'est la ***synthèse de l'ADN***

L'ADN synthétisé sera ensuite conservé dans des capsules métalliques,
C'est le ***stockage de l'ADN***

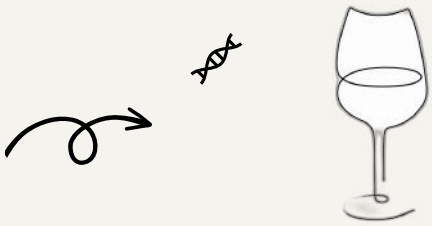
Lecture et récupération des données

L'information stockée dans les capsules peut être lue et ceci à l'aide d'un séquenceur d'ADN .

La suite de lettres obtenue est convertie en une suite binaire, ce qui permet de reconstruire l'information codée.

INTERETS DU STOCKAGE SUR ADN

Totalité des données mondiales



100 grammes d'ADN

✔ **Durée de vie des données :**

d'après les recherches :

+ Les données stockées sur l'ADN pourraient se conserver entre 700 000 et 1 million d'années

✔ **Espace occupé :**

+ De mini capsules peuvent stocker des millions de brins (donc de données)
+ Théoriquement, une seule capsule pourrait remplacer un data center

✔ **Capacité de stockage :**

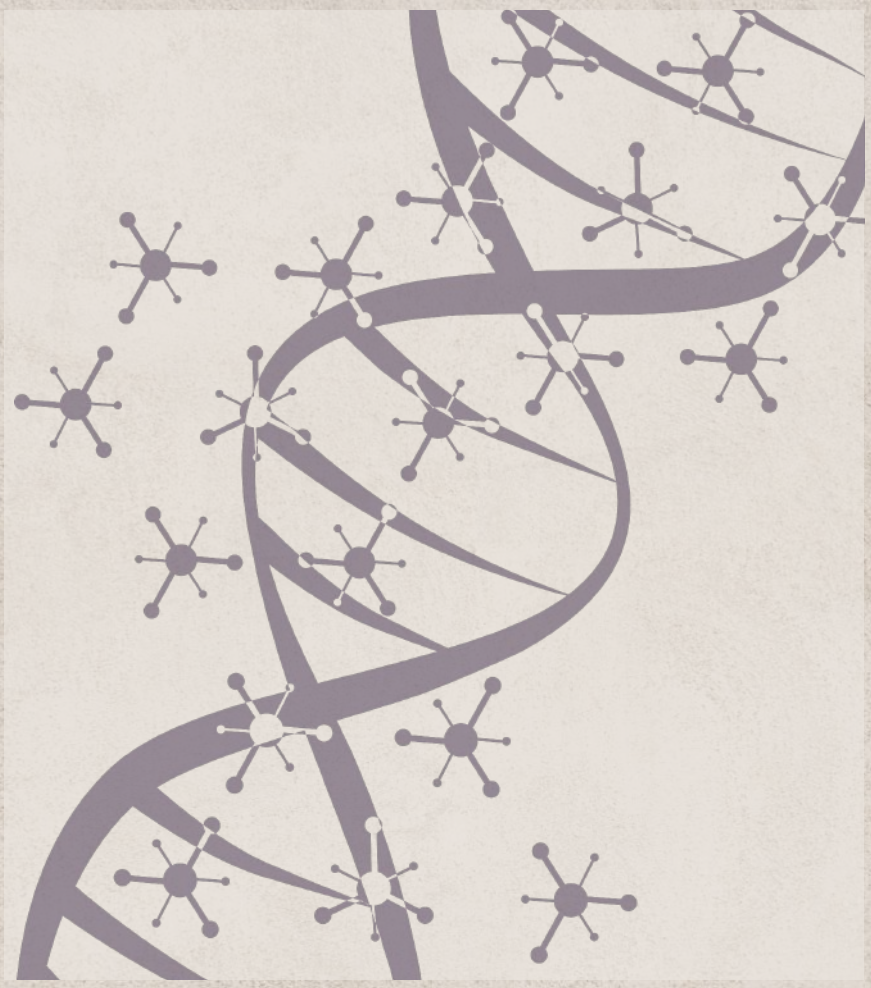
+ 1 gramme d'ADN peut contenir 450 millions TO de données
+ La totalité des données mondiales pourrait être conservée dans 100 grammes d'ADN

✔ **Consommation énergétique :**

+ L'ADN permet de stocker de gigantesques quantités de données sur des petits volumes sans consommer de l'énergie

Références

- Nick GOLDMAN EMBL-European Institute, Hinxton, United Kingdom
- Stéphane Lemaire et Pierre Crozet, SORBONNE Université, stockage des données sur l'ADN
- Nagwa startup
- EU research and innovation magazine
- Ciena
- ABB
- RTFLASH recherche & technologie
- Vecteezy
- Illustration et designe : CANVA



ADN
SUPPORT DE L'INFORMATION

Université de bretagne occidentale

Faculté des sciences et techniques

2022/2023