

Le stockage ADN pour préserver des données



Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES

par :

Mirko RIVAS

Conseillère au travail de Bachelor :

Chrystel DAYER

Genève, 19 février 2024

Haute École de Gestion de Genève (HEG-GE)

Filière Informatique de Gestion

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre Bachelor.

L'étudiant a envoyé ce document par email à l'adresse remise par son directeur de mémoire afin qu'il l'analyse à l'aide du logiciel de détection de plagiat COMPILATIO.

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 19 février 2024

Mirko RIVAS

Remerciements

Je tiens à exprimer ma sincère gratitude à Mme Dayer pour son suivi attentif et ses conseils précieux tout au long de ce projet. Son expertise et son soutien ont été essentiels.

Un grand merci également à Twistbioscience pour leur aide dans les démarches de commande d'ADN synthétique. Leur accompagnement a été inestimable.

Je remercie aussi l'Institut des Technologies de la Vie de la HES-Valais pour leur assistance dans le séquençage par nanopores, une contribution cruciale pour mon projet.

Enfin, je remercie chaleureusement la société Fasteris pour le temps accordé, leurs nombreux conseils et leur aide précieuse en matière de séquençage Sanger. Leur expertise a grandement contribué à la réussite de mon expérience.

Résumé

L'objectif de ce travail de bachelor est d'examiner l'évolution des technologies de stockage de données, avec une attention particulière portée au stockage de données dans l'ADN. En effet, l'acide désoxyribonucléique (ou ADN) est une molécule présente dans la quasi-totalité des cellules de tous les êtres vivants et contient leur code génétique. Ce projet vise à explorer les caractéristiques distinctives de l'ADN en tant que moyen de stockage, soulignant sa densité importante, sa grande capacité et son impact environnemental réduit, tout en prenant en compte les limites associées à cette technologie. Les aspects théoriques du stockage dans l'ADN, dont notamment la conversion de données numériques en séquences d'ADN et inversement, sont également inclus.

Ensuite, ce travail contient un cas pratique pour lequel j'ai développé un programme dédié à l'encodage et au décodage de texte et fichiers en séquences d'ADN et vice-versa. Grâce à ce programme, j'ai pu obtenir une séquence ADN à partir de texte¹, puis cette même séquence fut ensuite synthétisée par un laboratoire spécialisé afin d'obtenir un échantillon d'ADN. La donnée² était alors stockée ; les molécules d'ADN contenaient le texte que j'avais encodé. L'échantillon d'ADN était alors comme une clé USB, il stockait mes données. Pour récupérer le texte stocké dans l'ADN, l'échantillon fut séquencé par un autre laboratoire, afin d'obtenir la séquence ADN. Celle-ci fut ensuite décodée et m'a ainsi permis de retrouver le texte de base. Au travers ce cette preuve de concept, je souhaite démontrer la viabilité du stockage de données dans l'ADN. En effet, des données sont encodées, synthétisées et stockées sous forme d'ADN, pour qu'enfin l'ADN soit séquencé afin d'accéder à la donnée. Cette démarche vise à évaluer si l'ADN peut se positionner comme une solution de stockage de données future et à identifier les défis et obstacles qui freinent son adoption à grande échelle.

¹ En l'occurrence, une phrase

² La phrase susmentionnée

Table des matières

Déclaration.....	i
Remerciements	ii
Résumé	iii
Liste des figures.....	vii
1. Introduction.....	1
1.1 Contexte et motivations.....	2
1.2 Objectifs de ma recherche.....	3
1.3 Perspectives historiques	4
1.3.1 Cartes perforées	4
1.3.2 Bandes magnétiques	4
1.3.3 Tambours magnétiques	5
1.3.4 Disques durs (HDD).....	5
1.3.5 Disquettes.....	5
1.3.6 CD-ROM et CD-RW.....	6
1.3.7 DVD et Blu-Ray	6
1.3.8 Clés USB	6
1.3.9 Disques à état solide (SSD)	6
1.3.10 Stockage cloud	6
1.3.11 Technologies émergentes : stockage ADN et stockage quantique.....	6
1.3.12 Conclusion.....	7
2. Fondement de l'ADN	8
2.1 ADN en tant que support	8
2.1.1 Pourquoi l'ADN ?	8
2.1.2 Avantages et inconvénients	8
2.1.2.1 Longévité	9
2.1.2.1.1 Papier	9
2.1.2.1.2 Bandes magnétiques	9
2.1.2.1.3 CD/DVD	9
2.1.2.1.4 Disques HDD/SSD.....	9
2.1.2.1.5 ADN	9
2.1.2.2 Densité (et capacité)	10
2.1.2.3 Coûts et accessibilité.....	11
2.1.2.4 Lecture unique.....	12
2.1.2.5 Autres limitations techniques.....	12
2.2 R&D – DNA data storage alliance (SNIA).....	13
3. Le stockage dans l'ADN.....	14
3.1 Encodage de l'information	14
3.1.1 Principe fondamental : le bit.....	14
3.1.1.1 L'origine du bit	14
3.1.1.2 Influence culturelle et sociale du bit	14
3.1.1.3 Rôle des octets dans les systèmes de stockage	15

3.1.1.4	Évolution des unités de mesure de stockage.....	15
3.2	Encodage et synthèse de l'ADN.....	15
3.2.1	Exemple d'encodage : lilasSquare.png	16
3.2.2	Défis et considérations.....	17
3.2.2.1	Sécurité des données.....	17
3.2.2.2	Erreurs de synthèse	17
3.2.2.3	Stabilité des séquences	18
3.2.2.4	Bio-sécurité.....	18
3.3	Stockage et récupération	19
3.4	Séquençage et décodage	19
3.4.1	Séquençage	19
3.4.1.1	Le séquençage Sanger	20
3.4.1.1.1	PCR à terminaison de chaîne.....	20
3.4.1.1.2	Séparation des fragments par électrophorèse	23
3.4.1.1.3	Analyse du gel et détermination de la séquence ADN	23
3.4.1.1.4	Analyse automatisée	24
3.4.1.2	Le séquençage Illumina	24
3.4.1.3	Le séquençage par nanopores.....	25
3.4.2	Décodage des séquences ADN	25
3.4.2.1	Correction d'erreurs.....	25
3.4.2.1.1	Nécessité de correction	25
3.4.2.1.2	Mécanismes et techniques	26
3.4.2.1.3	Optimisation et améliorations	26
4.	Préoccupations environnementales et éthiques	27
4.1	Notre environnement, une priorité.....	27
4.1.1	WORN (Write Once Read Never)	27
4.1.2	Consommation d'énergie	28
4.1.3	Utilisation d'eau	29
4.1.4	Minéraux critiques et e-déchets	30
4.2	Éthique.....	31
4.2.1	Usage malintentionné de l'ADN	31
4.2.2	Inégalités	31
4.2.3	Organismes vivants	31
5.	Conception expérimentale.....	33
5.1	Organismes impliqués.....	33
5.2	Mes hypothèses	33
5.2.1	Obstacles rencontrés	34
5.3	Programme informatique.....	35
5.3.1	Classe Conversion.....	36
5.3.1.1	Encoder et décoder du texte	36
5.3.1.1.1	Gestion de la redondance	37
5.3.1.1.2	Processus d'encodage	37
5.3.1.1.3	Processus de décodage	37
5.3.1.2	Encoder et décoder un fichier	38
5.3.1.2.1	Utilisation d'un marqueur unique	39
5.3.1.2.2	Processus encodage	39

5.3.1.2.3	Processus de décodage	40
5.4	Coûts.....	40
6.	Résultats	41
6.1	Séquençage par nanopores	41
6.2	Séquençage Sanger	41
7.	Conclusion	45
	Bibliographie	47
	Annexe 1 : Java Class Conversion.....	51
	Annexe 2 : Java Class SequencingResults	59
	Annexe 3 : Résultats du séquençage Sanger	61

Liste des figures

Figure 1 : Carte perforée	4
Figure 2 : Tambour magnétique	5
Figure 3 : Distance Terre-Lune.....	8
Figure 4 : Données générées annuellement (en Zo).....	10
Figure 5 : Évolution du coût du séquençage d'un génome	11
Figure 6 : Quelques-uns des membres de l'alliance	13
Figure 7 : Processus global du stockage de données dans l'ADN	14
Figure 9 : Binaire de LilasSquare.png	16
Figure 8 : LilasSquare.png	16
Figure 10 : Séquence ADN de LilasSquare.png	16
Figure 11 : Préparation de la PCR classique	20
Figure 12 : Différence entre dATP et ddATP	20
Figure 13 : Liaisons phosphodiester	21
Figure 14 : Schéma de la PCR à terminaison de chaîne	21
Figure 15 : Nouveaux fragments du tube 1 (ddATP).....	22
Figure 16 : Séparation par électrophorèse	23
Figure 17 : Séquençage Sanger automatisé	24
Figure 18 : Séquençage par nanopores	25
Figure 19 : Répartition de la consommation finale d'énergie en Suisse 2022	28
Figure 20 : Quantité de données stockées dans les centres de données dans le monde de 2015 à 2021 (en exaoctets).	29
Figure 21 : Séquenceur MinION	34
Figure 22 : Encoder et décoder du texte.....	36
Figure 23 : Encoder et décoder un fichier.....	38
Figure 24 : Séquence de Consensus	41
Figure 25 : Séquençage forward	42
Figure 26 : Analyse du séquençage forward avec FinchTV	42
Figure 27 : Séquençage reverse du brin complémentaire.....	43
Figure 28 : Parties fiables des lectures forward et reverse	43
Figure 29 : Décodage de la séquence ADN	44

1. Introduction

Le domaine du stockage de données a grandement évolué au cours du dernier siècle³, notamment grâce à la forte demande d'efficacité, fiabilité et durabilité. Les méthodes de stockage sont devenues de plus en plus petites tout en offrant un espace de rangement grandissant. De nos jours, nous créons et stockons, plus que jamais, un très grand nombre de données. La digitalisation et l'envie de tout conserver créent une demande pour de nouveaux supports de stockage, car ceux actuels ne seront bientôt plus en mesure de couvrir nos besoins. En effet, des études estiment qu'entre 2023 et 2026, la quantité totale de données dans le monde devrait augmenter de 300%. [39] Cependant, étant donné de leur grande consommation en énergie, estimée à 1.5% de la consommation d'électricité annuelle mondiale, la construction de nouveaux centres de données plus grands n'est pas une solution durable. [42] Ces installations, qui stockent d'énormes quantités de données, fonctionnent sans interruption, 24/24h, entraînant ainsi une importante consommation d'énergie. Par exemple, en Suisse, 4% de la consommation totale d'électricité est due aux centres de données. [41] De plus, la longévité limitée, le défi d'extensibilité, le coût, et bien d'autres facteurs, nous poussent aujourd'hui à chercher de nouvelles solutions. Cette réalité souligne l'urgence de développer des alternatives de stockage de données plus éco-responsables et énergétiquement efficaces.

L'émergence de l'ADN comme méthode potentielle de préservation de données suscite donc grand intérêt, et ce, pour de bonnes raisons. L'ADN, qui est l'essence même de la vie, détient des caractéristiques uniques qui s'avèrent être une alternative très attrayante pour préserver des données à très long terme. En effet, l'utilisation de ce support comme moyen de stockage est actuellement au cœur de nombreuses recherches importantes. Des entreprises de renom telles que Microsoft, IBM, Twist Bioscience et bien d'autres, travaillent activement pour développer un processus de stockage ADN fiable, rapide et économique. [16] La Suisse participe également activement à cette innovation. L'Université de Genève et la Haute École Arc Ingénierie collaborent en tant que partenaires académiques au sein du programme DNAMIC, financé par Horizon Europe, le programme de l'Union européenne dédié à la recherche et à l'innovation. Lancé en octobre dernier, ce projet rassemble sept partenaires académiques et industriels issus de cinq pays européens, et vise à élaborer une méthode de stockage de données dans

³ De 1890 à aujourd'hui

l'ADN. Le projet comprend le développement de micro-usines autonomes et économes en énergie pour gérer les différentes phases du stockage de données dans l'ADN. [7]

Dans un contexte où de grandes sommes sont investies dans le domaine, il est légitime de penser que le stockage ADN a du potentiel. Cependant, est-il d'ores et déjà possible de stocker d'user de cette technologie à échelle personnelle ? La synthèse et le séquençage sont des processus complexes et coûteux, nécessitant de l'équipement spécialisé. Ainsi, il est intéressant de comprendre ces processus et de savoir s'il est possible de surmonter l'accès limité à cette technologie. Ensuite, il me faut, évidemment, développer un programme d'encodage et décodage pour user de cette technologie. Cette étape nécessite une compréhension des séquences ADN et de leur structure, afin d'assurer l'intégrité des données lors de leur traitement. Finalement, il est également nécessaire de saisir pourquoi l'ADN soulève autant d'intérêt ? Qu'est-ce qui le différencie des supports de stockage traditionnels ? Puis, surtout, quels sont les obstacles qui empêchent l'usage actif de cette technologie ?

En résumé, la question est de savoir si l'ADN peut être utilisé comme moyen de stockage, bien qu'il faille s'aider de laboratoires et d'équipements spécialisés. Alors que les avantages théoriques du stockage ADN sont impressionnants, il reste des obstacles pratiques à surmonter, ainsi, il est important de les comprendre et d'estimer si ceux-ci pourront bel et bien être vaincus.

1.1 Contexte et motivations

Il se trouve que mon intérêt pour le stockage ADN est arrivé plutôt soudainement. À la suite d'un voyage, pour lequel j'ai pris de nombreuses photos, une mésaventure m'est arrivée ; mon disque dur s'est cassé. Bien que j'aie pu sauvegarder la plupart de mes clichés, il a fallu que je trouve un nouveau disque dur. C'est alors que dans la frustration de mes recherches, pour un disque de remplacement, je suis tombé sur un article portant sur le stockage ADN. Intrigué par cette technologie, je me suis penché davantage sur le sujet. Puis après avoir lu quelques articles, j'ai très vite réalisé que ce n'était pas une idée insolite, mais que ce concept pouvait très certainement représenter l'avenir du stockage de données. Cet épisode a donc été le point de départ de mon intérêt et de ma curiosité pour cette technologie.

De plus, comme cité précédemment, le stockage de données est un sujet d'importance majeure. Notre ère numérique, accélérée par la pandémie de COVID-19, a souligné la nécessité de solutions de stockage de données efficaces et surtout durables. La digitalisation massive à laquelle nous assistons engendre des coûts tantôt financiers, tantôt écologiques. Dans ce contexte, il est devenu urgent et pertinent de trouver de

nouvelles alternatives. Cette technologie pourrait s'avérer être la clé pour répondre à notre forte demande de manière durable, fiable, et écoresponsable.

1.2 Objectifs de ma recherche

Mon projet de recherche se concentre sur l'exploration approfondie d'une technologie qui a capté mon intérêt. Cette technologie, qui représente une symbiose entre la biologie et l'informatique, ouvre des perspectives fascinantes pour le futur du stockage de données. Mon premier objectif est donc d'acquérir une compréhension détaillée de ce domaine, en mettant l'accent sur les aspects tantôt techniques, tantôt théoriques.

Un aspect crucial de ma recherche est la démystification du stockage ADN. Perçu comme complexe, je souhaite clarifier et simplifier ses concepts pour le rendre plus abordable. En démêlant les aspects techniques et scientifiques de cette technologie, mon but est de la rendre plus compréhensible pour un public plus vaste, ouvrant ainsi la voie à une meilleure appréciation de son potentiel. Parallèlement, je m'intéresse aux différents enjeux qui entourent le stockage ADN, notamment les questions éthiques, de sécurité et environnementales. Une part importante de mon travail consiste à dresser un état des lieux des progrès technologiques dans ce domaine. En examinant les avancements récents et les tendances, une autre partie met en lumière les défis et les opportunités présents.

Un volet pratique de cette recherche implique l'évaluation de la faisabilité technique du stockage de données sur l'ADN. Il s'agit de comparer les processus, les coûts et l'efficacité du stockage d'ADN avec l'aspect théorique que j'ai acquis, mettant en balance les avantages et les limites de cette technologie. Cette analyse contribue à une compréhension plus nuancée des implications pratiques.

Enfin, je me propose de démontrer pourquoi le stockage de données sur l'ADN est non seulement viable mais révolutionnaire, particulièrement adapté à notre ère numérique. J'illustre comment cette méthode de stockage peut transformer notre approche de conservation et d'accès à l'information, en soulignant son potentiel en tant que solution durable et à haute capacité pour le stockage de données sur le long terme.

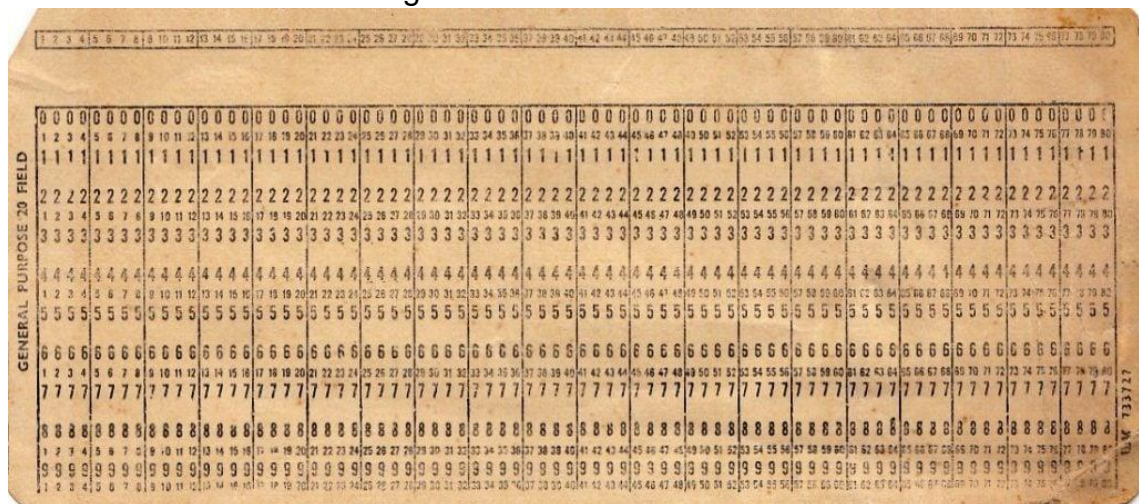
1.3 Perspectives historiques

Tout d'abord, qu'est-ce que représente le stockage de données ? Il consiste notamment à préserver de l'information à l'aide d'une technologie spécifiquement développée pour cela, tout en maintenant ces données accessibles autant que nécessaire. La progression des technologies de stockage a joué un rôle important dans l'évolution numérique.

1.3.1 Cartes perforées

C'est au 18^{ème} siècle que tout commence avec les cartes perforées, qui d'abord utilisées pour les métiers textiles, sont passées à la transformation de données au 19^{ème} siècle. La carte perforée d'Hollerith⁴, qui composée de 80 colonnes contenant chacune dix points de perforation, avait une capacité de 900 bits. [1][5]

Figure 1 : Carte Perforée



(Jean-Marc DELPRATO 2022)

1.3.2 Bandes magnétiques

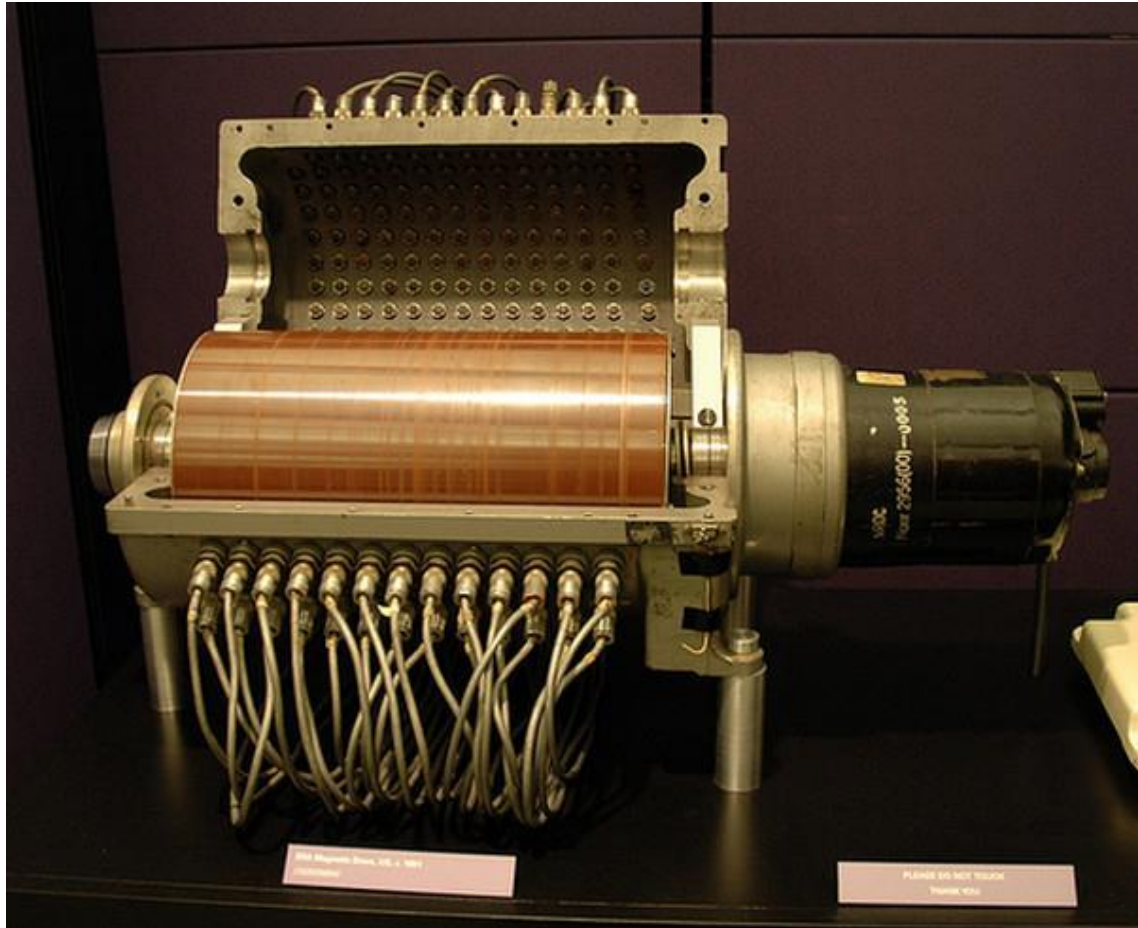
Les bandes magnétiques ont vu le jour dans les années 1930. Elles ont très rapidement révolutionné le stockage de données grâce à leur efficacité. Ces bandes ont joué un rôle vital dans la préservation de l'information pendant plusieurs décennies, démontrant la viabilité du stockage magnétique. [1]

⁴ Herman HOLLERITH, ingénieur américain inventeur de la mécanographie

1.3.3 Tambours magnétiques

C'est dans les années 1950 que les tambours magnétiques font leur apparition en offrant des temps de lecture plus rapides par rapport aux bandes magnétiques. Ils ont même servi de mémoire principale dans les premiers ordinateurs avant d'être dépassés par des technologies plus efficaces. [5]

Figure 2 : Tambour Magnétique



(Lunil Novembre 2020)

1.3.4 Disques durs (HDD)

Le RAMAC 305, introduit dans les années 1950 par IBM, a révolutionné la manière dont les données étaient stockées et accédées, en offrant une capacité et une rapidité sans égal à l'époque. Bien entendu, au début, leur capacité se mesurait en méga-octets plutôt qu'en giga-octets ou téra-octets.⁵

1.3.5 Disquettes

En 1970, c'est l'apparition des disquettes qui désormais offrent une solution portable. Au cours du temps, elles ont évolué du format huit pouces à la version plus compacte

⁵ Un giga-octet est l'équivalent 1'000 méga-octets, un téra-octet équivaut quant à lui à 1000 giga-octets.

de 3.5 pouces, en devenant ainsi un élément incontournable de l'informatique personnelle.

1.3.6 CD-ROM⁶ et CD-RW⁷

Courant 1980, les CD-ROM ont été introduits, offrant ainsi la possibilité de stocker environ 700 Mo de données. Puis, environ dix ans plus tard, le CD-RW donne la possibilité réécrire des données, offrant à son tour une très grande flexibilité aux utilisateurs.

1.3.7 DVD⁸ et Blu-Ray

Le DVD, ainsi que plus tard les Blu-Ray, ont encore augmenté les capacités de stockage, avec des Blu-Ray étant capables de stocker jusqu'à 50 Go de données. Grâce à leur capacité, ils sont très rapidement devenus les principaux supports de stockage de contenu de haute définition.

1.3.8 Clés USB

C'est finalement dans les années 2000 que la clé USB fait son apparition, offrant des capacités de stockage davantage compactes et portables, ainsi qu'une rapidité d'accès aux données sans pareille.

1.3.9 Disques à état solide (SSD)

Apparus au 21^{ème} siècle, ils ont surpassé les HDD en termes de performances, offrant une capacité de lecture et durabilité bien supérieure aux HDD. Cependant, le coût par Go est plus élevé.

1.3.10 Stockage cloud

Le stockage dans le cloud⁹ a une nouvelle fois révolutionné l'accessibilité des données. Il a amené la commodité de l'accès des données à distance, incarnant la transformation numérique du stockage de données.

1.3.11 Technologies émergentes : stockage ADN et stockage quantique

À l'avenir, les technologies émergentes, telles que celles mentionnées ci-dessus, promettent une capacité de stockage et des vitesses sans précédent. Elles promettent une nouvelle fois de révolutionner le secteur en offrant encore et toujours des

⁶ Compact Disc Read-Only Memory

⁷ Compact Disc-Rewritable

⁸ Digital Video Disc

⁹ Il s'agit d'un modèle de stockage de données dans lequel les données sont stockées sur des serveurs distants, mais restent accessibles via le réseau.

améliorations sans pareille, redéfinissant potentiellement l'avenir du stockage de données.

1.3.12 Conclusion

Au cours du dernier siècle et demi, le domaine du stockage de données a connu de nombreuses transformations. La science a constamment repoussé les limites en termes de capacités de stockage. Il y a quelques décennies, il était impensable que nous pourrions avoir des disques durs capables de contenir des Téraoctets. De la même manière, la rapidité avec laquelle nous sommes capables de lire et d'écrire des données a connu une amélioration stupéfiante. Parallèlement, la taille des outils de stockage a considérablement diminué, les immenses bandes magnétiques et premiers disques durs ont laissé place à des outils extrêmement compacts et par conséquent davantage portables. C'est très certainement cette portabilité qui a transformé la manière dont nous interagissons avec nos données, qui doivent être accessibles et disponibles en tout temps et en tout lieu. Désormais, avec les technologies prometteuses, telles que le stockage ADN ou le stockage quantique, il est clair que notre quête d'amélioration du stockage de données est loin d'être terminée et que les prochaines décennies promettent d'être tout aussi révolutionnaires que les précédentes. [2][3][4]

2. Fondement de l'ADN

« Acide du noyau des cellules vivantes, constituant essentiel des chromosomes et porteur de caractères génétiques. »

(LeRobert, 2023)

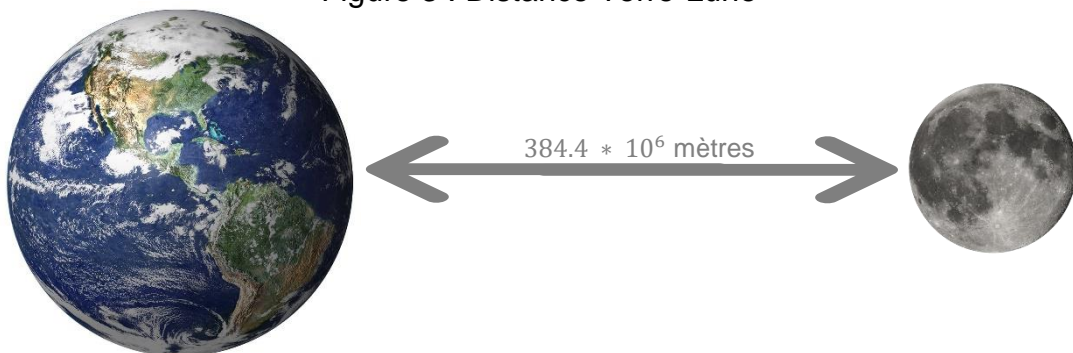
2.1 ADN en tant que support

2.1.1 Pourquoi l'ADN ?

Tout d'abord, l'ADN, ou également acide désoxyribonucléique, est une molécule très longue qui est composée d'un enchaînement complexe de quatre nucléotides : l'adénosine, la cytidine, la guanosine et la thymidine. Chez un être vivant, l'ADN est composé de deux brins dont l'arrangement correspond à l'information génétique de l'être en question. Pour l'être humain, chacune de nos cellules contient tout notre génome, c'est-à-dire l'entièreté de notre information génétique. L'ADN agit donc comme le support de stockage de l'information du vivant. Si on déroulait l'ADN contenu dans une cellule humaine, on obtiendrait, certes, un fil extrêmement fin, mais d'une longueur d'environ deux mètres. [43]

Un corps humain contient en moyenne 37'200 milliards de cellules. [6] Ainsi, si nous déroulions l'ensemble de l'ADN contenu dans un corps et que nous le mettions bout à bout, nous obtiendrions un fil long de 74'400 milliards de mètres, soit environ 192'508 fois la distance Terre-Lune.

Figure 3 : Distance Terre-Lune



(Mirko Rivas, 2024)

2.1.2 Avantages et inconvénients

Dans les points à venir, nous allons traiter en surface les avantages et inconvénients liés au stockage de données dans l'ADN. Ces aspects permettront de comprendre plus en détail l'enjeu de cette technologie tout en mettant en lumière les barrières auxquelles nous faisons actuellement face.

2.1.2.1 Longévité

De nos jours, les outils nous permettant de conserver des données sur le long terme nous offrent une durée de vie d'une centaine d'années, mais ce n'est pas le cas des outils les plus couramment utilisés. La durée de vie des méthodes traditionnelles est très faible comparée à ce que l'on peut obtenir avec l'ADN. Voyons ensemble quelques exemples de différents moyens de stockage pour illustrer le propos.

2.1.2.1.1 Papier

Le papier a une durée de vie qui peut varier de quelques décennies à plusieurs siècles si celui-ci est conservé dans des conditions optimales. Cependant, le papier est très sensible à l'humidité, à la lumière, aux insectes, aux acides et finalement, il se prête pas du tout au stockage de données numériques dont l'enjeu est crucial.

2.1.2.1.2 Bandes magnétiques

Les bandes magnétiques ont une durée de vie d'environ 10 à 30 ans selon la qualité de la bande et des conditions de stockage. Elles ont, notamment, besoin d'une température et humidité stable. Il est également important de noter que plus on utilise une bande, plus elle est susceptible de s'user, elle peut donc devenir obsolète.

2.1.2.1.3 CD/DVD

Ils ont une durée de vie de 10 à 100 ans selon le type de disque et les conditions de stockage. Cependant, ils sont très sensibles à la température, rayures et dégradation du matériau réfléchissant. Tout comme les bandes magnétiques, plus on utilise les CD/DVD, plus on a des risques que ceux-ci se dégradent. Tout média optique est très sensible aux dommages, car la surface lisible n'est que très peu protégée.

2.1.2.1.4 Disques HDD/SSD

Ils ont une durée de vie de cinq à dix ans selon le nombre de cycles de lecture et écriture. Pour les deux types de disques, l'usure est un élément majeur ; tantôt les pièces mécaniques, tantôt les cellules mémoire, peuvent s'user avec le temps. De plus, ils sont extrêmement sensibles aux chocs. Malgré les défauts que comporte cette méthode de stockage, il s'agit du support de stockage dominant.

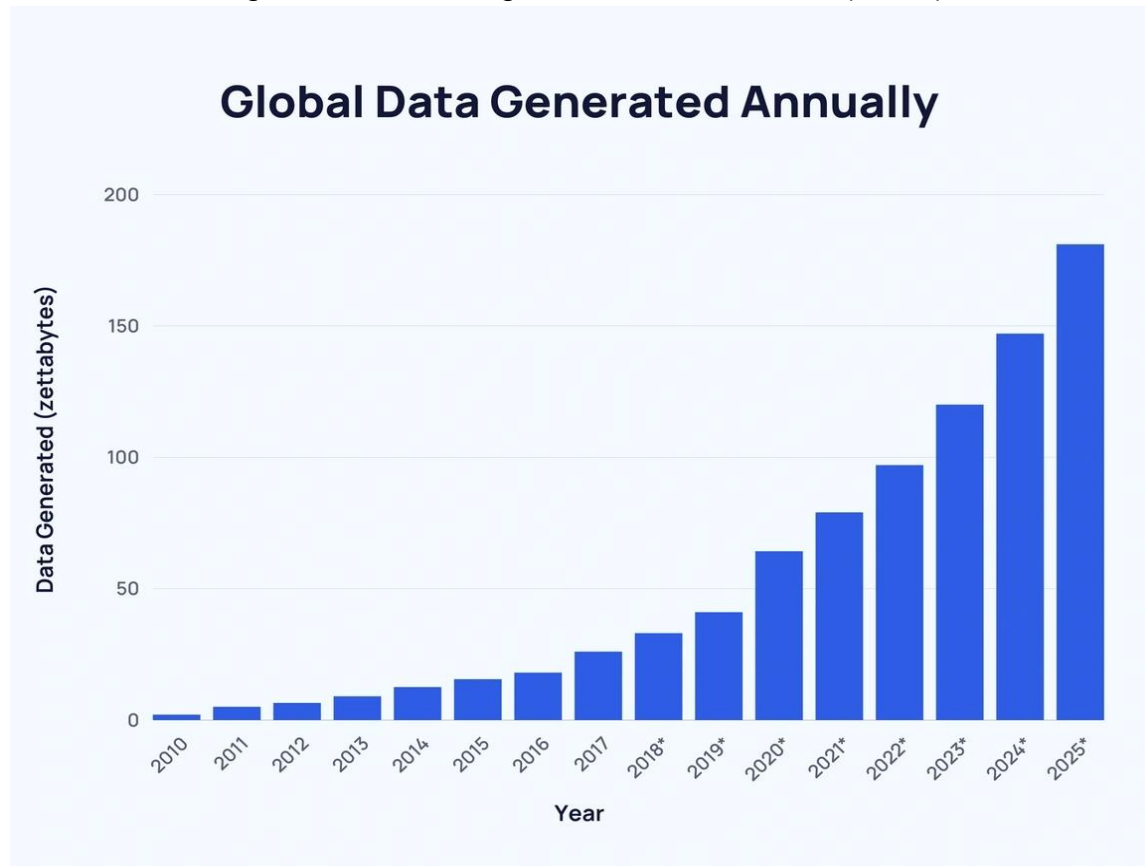
2.1.2.1.5 ADN

L'ADN se distingue remarquablement des autres méthodes traditionnelles de stockage de données grâce à sa longévité. En effet, des études démontrent que l'ADN peut être conservé à très long terme. Le plus vieil ADN découvert et séquencé date d'il y a deux millions d'années. [8] Ainsi, dans de bonnes conditions, l'ADN pourrait être conservé durant des millénaires sans trop de difficultés.

2.1.2.2 Densité (et capacité)

La capacité est une composante fondamentale dans le domaine stockage de données. Comme relevé précédemment, nous générons de plus en plus de données et il devient impératif de trouver de nouvelles solutions de stockage innovantes et efficaces.

Figure 4 : Données générées annuellement (en Zo)



(Statista, 2023) [9]

Les projections actuelles suggèrent qu'en 2025, nous générerons environ 181 zettaoctets¹⁰ de données, soit une augmentation de plus de 60 zettaoctets supplémentaires par rapport à la quantité de données générées en 2023. Face à cette croissance, il nous faut trouver un moyen stockage capable de contenir un tel volume. Dans ce contexte, la capacité et la densité du stockage ADN se révèlent être extrêmement prometteuses. Théoriquement, un seul et unique gramme d'ADN peut contenir près de 215 Po¹¹. [44] Par exemple, si nous souhaitons stocker l'ensemble des données que nous générerons en 2025, il nous faudrait uniquement 841'860¹² grammes d'ADN, soit environ 842 kilogrammes, l'équivalent d'une petite voiture citadine.

¹⁰ 1 Zo correspond à 10¹² Go.

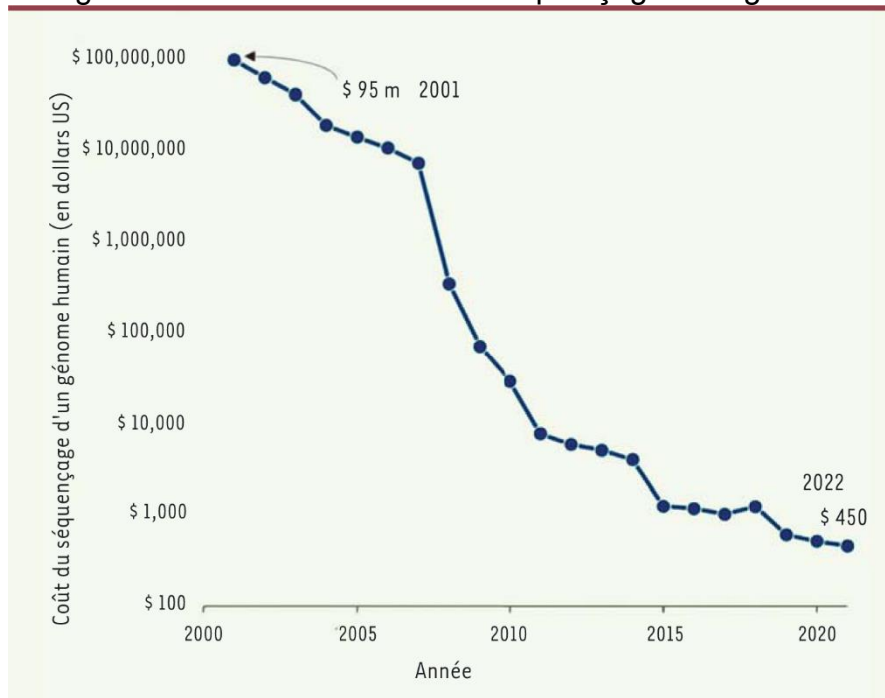
¹¹ 1 Po (pétaoctet) correspond à 10⁶ Go.

¹² $\frac{181 \times 1'000'000}{215} = 841'860 \text{ grammes}$

2.1.2.3 Coûts et accessibilité

Le coût du stockage ADN est un élément très important, car il pourrait déterminer sa viabilité. En effet, s'il s'avère trop coûteux d'user de cette technologie, alors il est difficile d'imaginer l'humanité l'adopter aussi bien que nos supports de stockage actuels (HDD, SSD, ...). Historiquement, les coûts de séquençage et de synthèse ont été bloquants.

Figure 5 : Évolution du coût du séquençage d'un génome



(Bertrand Jordan : biologiste, généticien et immunologiste, 2023) [10]

Le séquençage du premier génome humain, finalisé en 2003 après une quête scientifique de 15 ans et un investissement de 2,7 milliards de dollars, a marqué un tournant décisif dans le domaine de la génomique. [11] Depuis, les coûts associés au séquençage d'un génome humain ont chuté de manière impressionnante. Alors qu'il fallait déboursier environ un million d'euros et allouer trois mois pour un séquençage en 2007, cette somme a été réduite à presque 1'000 euros en 2015, avec une durée de séquençage diminuée à seulement quelques heures, franchissant ainsi un seuil financier et temporel significatif.

L'accessibilité est un autre aspect, non moins important, du stockage de données dans l'ADN. Actuellement, les technologies nécessaires pour ce faire, telles que les équipements de synthèse et séquençage, sont principalement confinées aux laboratoires de recherche et développement spécialisés. Si l'on souhaite que cette technologie se démocratise et soit accessible au grand public, il est impératif, d'une part, d'élargir les accès à ce genre d'outils, puis d'autre part, de simplifier leur utilisation. Certes, il n'est pas simple de faire des projections, car ces technologies n'en sont

qu'au stade de recherche et développement. Cependant, nous pouvons déjà constater qu'au fur et à mesure que la technologie se développe, les progrès continuent, puis les coûts et le temps de travail diminuent. Ainsi, il est tout à fait raisonnable d'imaginer et prévoir une amélioration de l'accessibilité dans les années à venir. Ces avancées pourraient même ouvrir la voie à des applications nouvelles, allant de la préservation de données à long terme, à des usages plus quotidiens dans d'autres secteurs.

Globalement, il est encore tôt pour avoir des perspectives claires en termes de coûts et d'accessibilité puisque les technologies sont encore difficilement utilisables par le grand public. Cependant, nous n'en sommes pas loin et les projets de recherche et développement sont nombreux.

2.1.2.4 Lecture unique

Un des défis du stockage de données dans l'ADN concerne le nombre de lectures possibles. Une fois l'ADN synthétisé, sa forme la plus stable et durable est celle de l'ADN lyophilisé¹³. Toutefois, le processus de lecture implique une étape de réhydratation de l'ADN. Bien que cette réhydratation soit nécessaire pour lire la séquence, elle a pour conséquence de rendre l'ADN inutilisable pour des lectures ultérieures. Ainsi, chaque séquence d'ADN ne peut être lue qu'une seule fois, ce qui représente une limitation significative pour l'utilisation d'ADN comme support de stockage à long terme.

2.1.2.5 Autres limitations techniques

Comme cité précédemment, les méthodes utilisées pour lire et écrire des données sur l'ADN, bien que révolutionnaires, sont encore complexes et chronophages, surtout lorsqu'on les compare à la rapidité et facilité d'utilisation des disques durs électroniques.

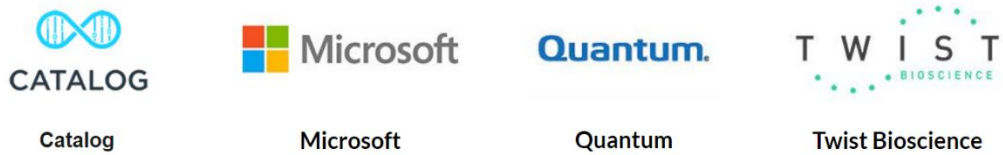
De plus, l'ADN est extrêmement sensible aux mutations même lorsqu'il s'agit d'ADN synthétique. Les mutations peuvent survenir en raison de divers facteurs. Il se peut que des erreurs de synthèse se produisent, à cause de l'incorporation de mauvaises bases nucléiques¹⁴. Il se peut également qu'à cause de facteurs environnementaux, tels que les radiations, de hautes températures ou l'exposition à certains produits, la structure de l'ADN soit altérée entraînant ainsi des mutations. Puis, la manipulation physique de l'ADN peut également parfois entraîner les mêmes conséquences, dont notamment lors des processus de lyophilisation et de réhydratation.

¹³ « La lyophilisation, ou anciennement cryodessiccation, est la dessiccation d'un produit préalablement congelé, par sublimation. » (Wikipédia, 2023)

¹⁴ Blocs de construction qui composent la molécule d'ADN. Il en existe quatre : l'adénine (A), la thymine (T), la cytosine (C) et la guanine (G).

2.2 R&D – DNA data storage alliance (SNIA)

Figure 6 : Quelques-uns des membres de l'alliance



(SNIA, 2023) [16]

Comme nous l'avons vu précédemment, l'ADN est porteur de toute l'information nécessaire au développement de chaque être vivant, cette information est encodée dans chaque cellule de tout être. Le décodage de celle-ci présente un enjeu crucial pour la recherche scientifique. L'étude de la génétique s'avère fondamentale pour analyser et saisir une vaste gamme de phénomènes, incluant les pathologies, nos racines ancestrales, l'évolution des espèces, le mécanisme héréditaire des traits génétiques et bien d'autres. Les progrès technologiques de ces 20 dernières années ont considérablement enrichi et transformé ce champ d'étude. Nos avancées technologiques en termes de séquençage et de synthèse d'ADN suscitent désormais un questionnement quant à la viabilité du stockage de données sur l'ADN.

De nombreuses compagnies telles que Microsoft, Quantum, Twist Bioscience, Western Digital, Dell Technologies, IBM, Lenovo, Illumina, et j'en passe, ont rejoint une alliance créée par la SNIA¹⁵ dans le but de standardiser au maximum les technologies émergentes et avancements à propos du stockage de données dans l'ADN. [16] Leurs objectifs principaux étant de :

- Créer une feuille de route technologique de l'industrie, afin d'identifier les défis pour l'ensemble des processus.
- Élaborer un DARS¹⁶, afin de créer une méthode universelle permettant de décoder de l'ADN.
- Spécifier l'interopérabilité entre différents fournisseurs et éviter que certains verrouillent des formats.
- Établir des méthodologies de préservation des données ainsi que des méthodes de comparaison, afin d'évaluer et de comparer les solutions de stockage de données dans l'ADN. [17][18]

Ces objectifs reflètent l'engagement de l'alliance SNIA à développer le stockage de données dans l'ADN en tant que solution viable et durable pour l'avenir du stockage de données à grande échelle.

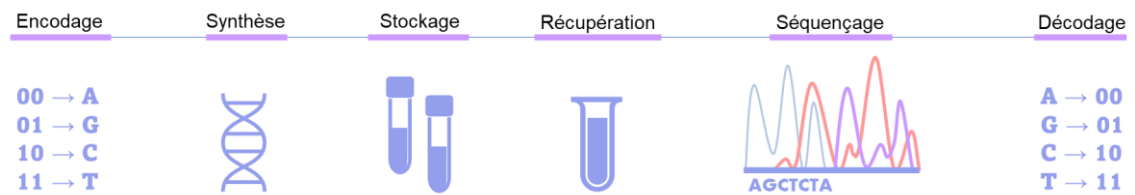
¹⁵ Storage Networking Industry Association

¹⁶ DNA Archive Rosetta Stone

3. Le stockage dans l'ADN

Afin de stocker les données dans de l'ADN, il est nécessaire de suivre différentes étapes. Il faut tout d'abord encoder les données, ensuite les synthétiser pour obtenir l'échantillon de la séquence ADN souhaitée. Puis, lorsqu'on souhaite avoir accès aux données, il faut séquencer l'ADN, afin d'en obtenir la séquence, puis décoder cette même séquence pour en obtenir son contenu.

Figure 7 : Processus global du stockage de données dans l'ADN



(Mirko Rivas, 2023)

3.1 Encodage de l'information

3.1.1 Principe fondamental : le bit

Les 'binary digits', ou 'bits', constituent la pierre angulaire du stockage des données numériques. Au-delà de la représentation binaire classique, popularisée par des œuvres telles que le film 'Matrix', où s'affichent des suites de 0 et de 1, les bits représentent l'élément central du langage informatique.

3.1.1.1 L'origine du bit

Claude Shannon a été le premier à définir formellement le concept de bit dans son article révolutionnaire de 1948, posant ainsi les fondations de la théorie de l'information. [40] En définissant le bit comme unité fondamentale de l'information, Shannon a exercé une influence majeure sur des domaines tels que l'informatique, la télécommunication et la cryptographie, en les transformant en sciences rigoureuses.

3.1.1.2 Influence culturelle et sociale du bit

Le bit a largement dépassé les frontières du monde technologique, devenant un emblème de l'ère numérique. Il symbolise l'idée que toute information, textuelle, visuelle ou sonore, peut être convertie en un langage universel de zéros et d'uns. Ce principe, véhiculé par la culture populaire et les médias, a démocratisé la notion de numérisation et de codification de l'information auprès du grand public.

3.1.1.3 Rôle des octets¹⁷ dans les systèmes de stockage

Dans les systèmes de stockage de données, les octets, composés de huit bits, sont l'unité standard pour mesurer la capacité et la taille des fichiers. Grâce à leur capacité à représenter 256 états différents, les octets permettent de coder une vaste gamme de caractères et de valeurs, essentiels pour la création de documents, d'images, de logiciels, etc.

3.1.1.4 Évolution des unités de mesure de stockage

Avec le temps, les unités de stockage se sont diversifiées et agrandies, évoluant des kilooctets et mégaoctets des premiers ordinateurs vers des tailles bien plus conséquentes comme les giga-octets et téraoctets. Cette évolution reflète la croissance fulgurante des besoins en stockage numérique et les avancées technologiques pour y répondre.

3.2 Encodage et synthèse de l'ADN

La transcription de données numériques en ADN consiste à transformer les données sous forme binaire - généralement des suites de bits - en séquences de nucléotides d'ADN, constituées des bases : adénine (A), thymine (T), cytosine (C) et guanine (G). Cette transformation s'effectue grâce à un algorithme d'encodage, qui associe chaque paire de bits à un nucléotide. Le code binaire peut être converti comme ceci : A pour 00, G pour 01, C pour 10 et T pour 11, ainsi une séquence binaire 00011011 devient AGCT. Il est essentiel de concevoir cet algorithme d'encodage avec précision pour éviter les séquences répétitives ou problématiques, qui pourraient compliquer la lecture ou la synthèse de l'ADN. Une fois cet algorithme défini, la séquence d'ADN obtenue est synthétisée, aboutissant à la création d'un brin d'ADN réel renfermant les données numériques encodées.

¹⁷ Également connu sous le nom de « Bytes », en anglais.

3.2.1 Exemple d'encodage : lilasSquare.png

Pour illustrer la conversion d'une image en une séquence binaire, puis en séquence d'ADN. Je commence avec le fichier lilasSquare.png, créé à l'aide de l'application Paint, qui représente un carré lilas de 20x20 pixels. Tout d'abord, il faut convertir cette image en une séquence binaire à l'aide d'un programme. Pour convertir cette séquence binaire en une séquence d'ADN, nous devons suivre une norme de conversion établie, où chaque paire de bits est remplacée par une base nucléique spécifique, dans notre cas A pour 00, G pour 01, C pour 10 et T pour 11. Une fois la séquence ADN obtenue, celle-ci représente fidèlement le fichier original lilasSquare.png sous forme de séquence ADN.

Figure 8 :
LilasSquare.png

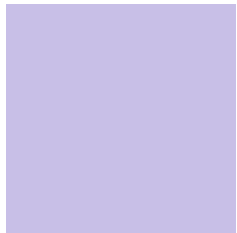


Figure 9 : Binaire de LilasSquare.png

[illegible]

Figure 10 : Séquence ADN de LilasSquare.png

CACGGGAAGATCGAGTAATGAACGACCAACCAAAAAAAAAAAAAAAAA
TGGACGGACAGAGAGGACAAAAAAAAAAAAAGGAAAAAAAAAAAAAA
AGGAAACAAAGCAAAAAAAAAAAAAACATGCACGAGTGAATGAAAAAA
AAAAAAAAAAGGTATGGACGAGTGAACAAACCTCTATCAGTATCCG
AAAAAAAAAAAAAGAGCGTGAAGGATGGAAGAAAAAACTAGCA
TTAACTTTTAGCAGAAGGAAAAAAAAAAAAAACGTAAGACAGGCG
GTATAAAAAAAGACGTGAAAAAAAAAAGACGTGAAAGTGTGCG
GCAGTTGTCAAAAAAAAAAAAAACCCGACGGAGAGAAGGGGAATCA
GATTGCATATTACTAGTTTTTTCGTTTAAGCACCCAAACACGCAC
CAGATGATGGATAAGCCACAATCGGGCAATGGAGAAATACCTAAG
CCCACAAGCGGACAATGGAGAAATACCGAAGTCAATTGTGGAAAA
GCAAGCAAAACCCTTCTAAATCGGGGATCCTCATGTGTGTGAAAA
AAAAAAAAAAAAAGACGGAGGGATCGAGACCTCGAACGCAACAAC

3.2.2 Défis et considérations

Lors de l'encodage et de la synthèse des séquences, plusieurs défis et considérations doivent être pris en compte pour assurer l'intégrité et la sécurité des données.

3.2.2.1 Sécurité des données

L'encodage de données dans une séquence ADN, similaire au binaire, n'est pas en soi un défi technique majeur, car il peut être réalisé aisément avec une clé d'encodage appropriée. Cependant, un aspect crucial à considérer est la possibilité de stocker du malware dans ces séquences ADN, tout comme dans les supports de données actuels. Un exemple est l'expérience menée en 2017 par une équipe de l'Université de Washington. Ils ont réussi à prendre le contrôle d'un ordinateur en utilisant un malware inséré dans une séquence ADN. Bien que l'exécution du malware ait été facilitée par une vulnérabilité introduite dans le logiciel de traitement de l'ADN, cette démonstration illustre que le changement de support de stockage ne diminue pas les risques de sécurité informatique. Le potentiel de malveillance doit être pris en considération, peu importe le médium de stockage. [14]

3.2.2.2 Erreurs de synthèse

En effet, l'exactitude de la synthèse est cruciale. Des erreurs lors de la synthèse peuvent mener à des inexactitudes dans les données stockées. Une des solutions pouvant quelque peu pallier ce défi est la redondance des fragments. Habituellement, on ne travaille pas qu'avec un seul fragment d'ADN ; la duplication de l'ADN est un processus relativement simple et économique. En séquençant plusieurs fragments de la même séquence, il est possible de comparer les résultats et d'identifier aisément les brins présentant des erreurs de synthèse.

Par ailleurs, la non-redondance dans certaines parties de la séquence d'ADN peut également contribuer à une synthèse plus précise. Les séquences où les nucléotides se répètent fréquemment peuvent rendre la synthèse plus complexe, voire impraticable. Pour pallier ce problème, l'introduction contrôlée et aléatoire de nucléotides dans la séquence peut être bénéfique. Ainsi, une séquence, avec des répétitions, comme `AAAAAATTTTTTCCCCCGGGGGG` est plus difficile à synthétiser qu'une séquence plus variée comme `ATGACTAGGCTACGCTTACGATCA`.

Pour l'expliquer de manière plus concrète, voici la façon dont mon programme, conçu pour encoder du texte en séquence ADN, aborde la question des redondances. Par exemple, imaginons que je souhaite encoder cette phrase :

« Ceci est un message qui sert d'exemple de formatage afin d'éviter la redondance de caractères. »¹⁸.

Le processus de formatage initial transforme la phrase en :

« Ceci Est_un\$meSsagE\$qUI^SerT\$d'ExeMplE_De+foRmAtaGE#AFiN
+d'evlTER^La+ReDOndANCE^De_carACtERes. »

Cette version modifiée reste parfaitement intelligible – bien qu'elle paraisse très désordonnée à première vue – tout en minimisant les répétitions. Les espaces sont substitués par une alternance de caractères spéciaux, et la casse des lettres est inversée à chaque occurrence pour une plus grande variabilité.

3.2.2.3 Stabilité des séquences

Comme cité précédemment, l'ADN n'est pas toujours stable. Divers facteurs externes, ainsi que sa composition, peuvent en compromettre sa structure et par le même cas compromettre les données.

3.2.2.4 Bio-sécurité

Dans le cas de ce travail de Bachelor, les fragments d'ADN sont synthétisés et fournis par Twist Bioscience, entreprise américaine spécialisée, entre autres, dans la synthèse d'ADN. En vue de la protection et atténuation de risques liés à l'ADN, des directives de l'HHS (Health and Human Services, instance américaine) ont été émises pour encourager les fournisseurs de dsDNA¹⁹ à filtrer toutes les commandes. [19][20]

En effet, il est actuellement difficile d'évaluer la capacité à détecter la pathogénicité et de prédire les fonctions des séquences ADN synthétiques. Les méthodes actuelles de filtrage ne sont pas en mesure de détecter des menaces bio-sécuritaires qui ne sont pas déjà répertoriées ou qui ne présentent pas une corrélation élevée avec des agents pathogènes déjà connus. Ainsi, la bio-sécurité dans ce contexte exige une surveillance rigoureuse pour prévenir l'utilisation malveillante de matériel génétique.

Historiquement, les tentatives d'utilisation de la biologie pour le développement d'armes biologiques par les États ont rencontré des limites, principalement dues à la vulnérabilité des agents biologiques aux facteurs environnementaux et aux mutations. Toutefois, les progrès récents en biotechnologie suggèrent que la conception d'agents pathogènes améliorés ou entièrement nouveaux pourrait devenir une réalité. Par conséquent, le développement de bases de données exhaustives recensant un maximum d'agents pathogènes est nécessaire, ainsi qu'une amélioration continue des algorithmes de

¹⁸ J'ai délibérément omis les accents pour simplifier le traitement.

¹⁹ Double stranded DNA – (ADN à double brin)

reconnaissance de ceux-ci. Une collaboration internationale et un échange d'informations sont également des éléments cruciaux, cela pourrait assurer une norme globale de bio-sécurité dans le domaine de la biologie synthétique et bien entendu, du stockage ADN.

3.3 Stockage et récupération

Une fois synthétisé, l'ADN requiert peu pour son stockage et sa conservation à long terme. Protégé de facteurs environnementaux tels que l'eau, la lumière et l'air, il a le potentiel de rester intact pendant des millénaires, sans nécessité d'énergie pour sa préservation. Cette durabilité, combinée à sa compacité exceptionnelle, rend l'ADN particulièrement attrayant pour le stockage de données. Utiliser l'ADN comme moyen de stockage pourrait non seulement révolutionner ce secteur en réduisant significativement les émissions de CO₂ et également libérer une quantité considérable d'espace physique, dû à sa grande densité, offrant ainsi une solution à la fois écologique et efficace.

3.4 Séquençage et décodage

Le séquençage, tout comme la synthèse, joue un rôle crucial dans le processus de stockage de données dans l'ADN. Il est impératif que les données lues soient fidèles à leur original. La présence de la moindre erreur peut compromettre l'intégrité des données, pouvant même les rendre illisibles. Par conséquent, l'efficacité et la précision des équipements de séquençage ainsi que des processus de décodage sont des aspects primordiaux, car ils garantissent la fiabilité et la validité des données récupérées.

3.4.1 Séquençage

Comme nous l'avons vu précédemment, l'ADN est porteur de toute l'information nécessaire au développement de chaque être vivant, cette information est encodée dans chaque cellule de tout être. Le séquençage est le processus utilisé pour découvrir les séquences de bases nucléiques composant l'ADN. L'étude de la génétique s'avère fondamentale pour analyser et saisir une vaste gamme de phénomènes, incluant les pathologies, nos racines ancestrales, l'évolution des espèces, ainsi que le mécanisme héréditaire des traits génétiques, et bien d'autres. Les progrès technologiques de ces vingt dernières années ont considérablement enrichi et transformé ce champ d'étude. De nombreuses méthodes existent et se développent, nous en sommes aujourd'hui à la troisième génération.

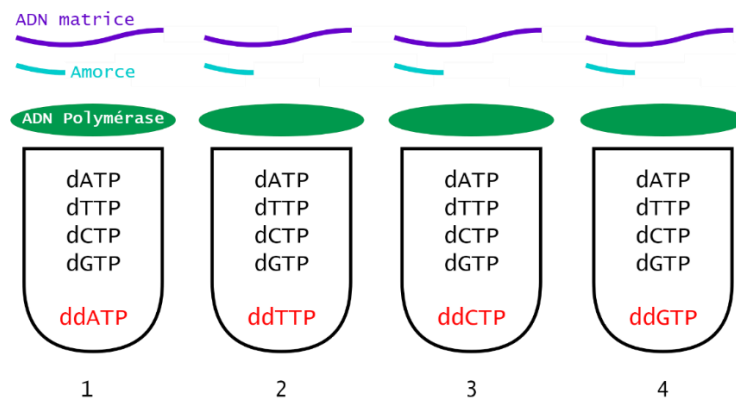
3.4.1.1 Le séquençage Sanger

Le séquençage Sanger, du nom de son inventeur Frederick Sanger²⁰, est une méthode de première génération pour lire l'ordre des bases dans un brin d'ADN. Cette méthode se décline en deux variantes : le séquençage Sanger traditionnel et sa version automatisée. De nos jours, c'est la forme automatisée qui est principalement utilisée. [38]

3.4.1.1.1 PCR à terminaison de chaîne

La PCR à terminaison de chaîne est la première étape dans le séquençage Sanger. Elle commence par la préparation de quatre solutions contenant le fragment d'ADN à séquencer, l'ensemble des désoxyribonucléotides (dNTPs²¹), un seul type de didésoxynucléotide (ddNTPs²²) en faible quantité, de l'ADN polymérase et une amorce.

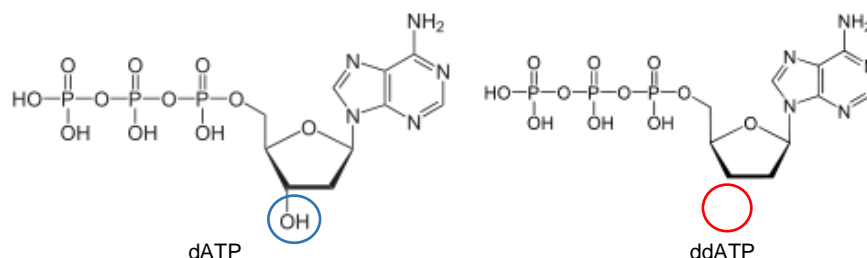
Figure 11 : Préparation de la PCR classique



(Mirko Rivas, 2024)

Durant la réaction, l'ADN polymérase lit le brin matrice et commence à synthétiser un nouveau brin d'ADN en incorporant des dNTPs. Cependant l'ajout aléatoire de ddNTPs, qui manquent d'un groupe hydroxyde²³ en position 3', interrompent la réaction.

Figure 12 : Différence entre dATP et ddATP



(Wikipédia, 2024) [33][34]

²⁰ Biochimiste anglais, ayant reçu deux prix Nobel (en 1958 et 1980)

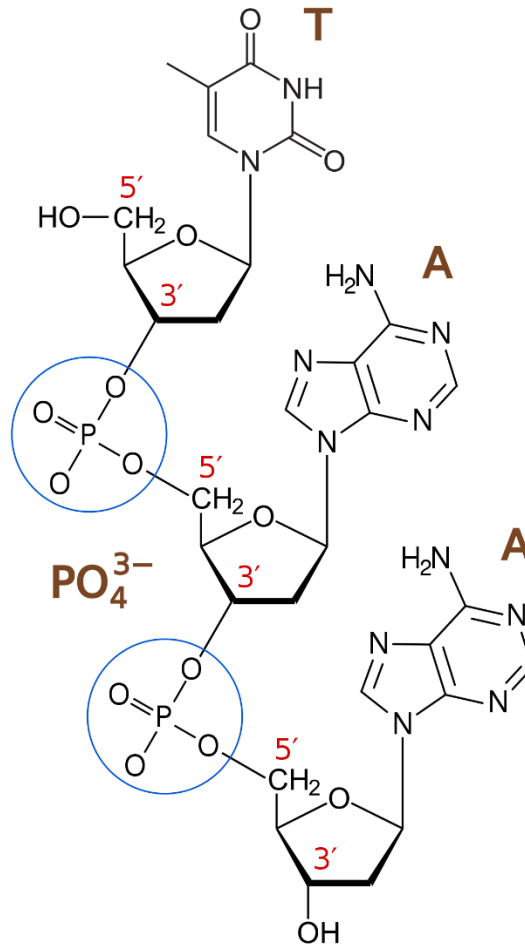
²¹ dATP, dTTP, dCTP et dGTP

²² ddATP, ddTTP, ddCTP ou ddGTP

²³ OH⁻

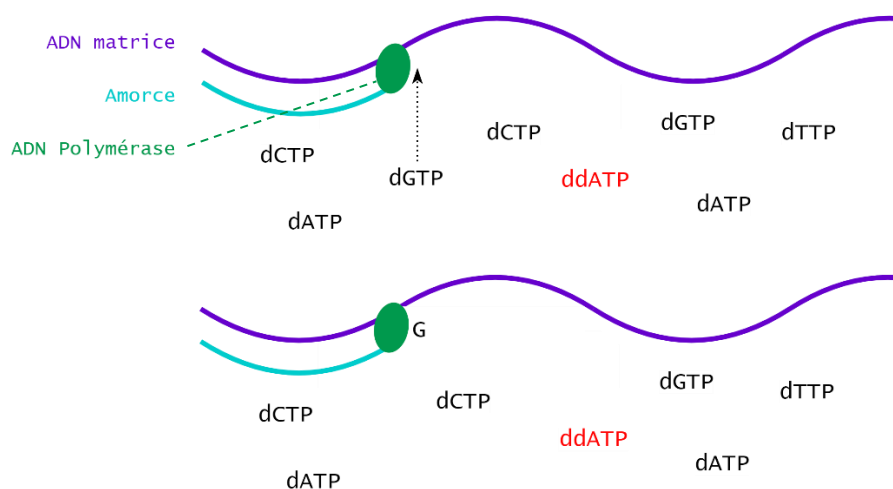
L'absence du groupement OH^- dans un ddNTP empêche une nouvelle liaison phosphodiester de se réaliser, arrêtant ainsi la synthèse du brin d'ADN.

Figure 13 : Liaisons phosphodiester



(Wikipédia, 2024) [45]

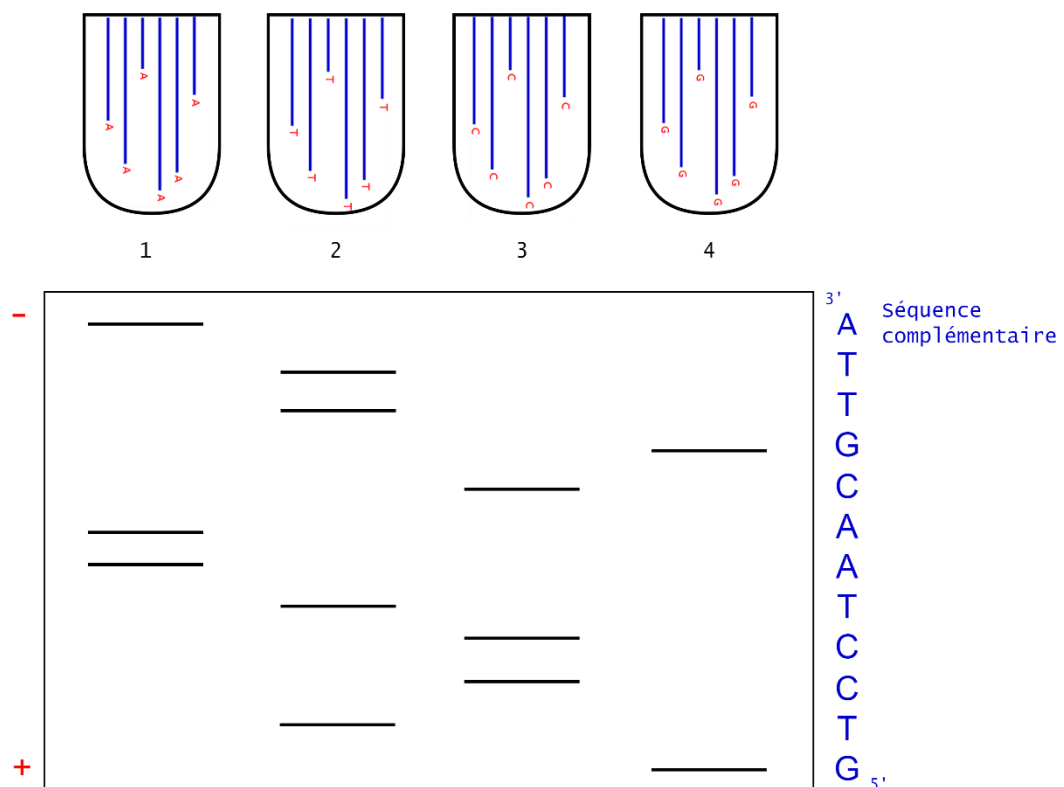
Figure 14 : Schéma de la PCR à terminaison de chaîne



3.4.1.1.2 Séparation des fragments par électrophorèse

La deuxième étape du séquençage consiste à déposer le contenu des tubes sur un gel de polyacrylamide. On procède ensuite à une électrophorèse. Étant donné que l'ADN est chargé négativement dû au groupement phosphate, celui-ci va donc migrer vers le pôle positif lors de l'électrophorèse. La séparation des fragments dans le gel se base uniquement sur leur taille : plus le fragment est petit, plus rapidement et plus loin, il se déplace à travers le gel. Cette propriété permet donc de trier les fragments selon leur longueur facilitant ainsi leur identification ultérieure. [38]

Figure 16 : Séparation par électrophorèse



(Mirko Rivas, 2024)

3.4.1.1.3 Analyse du gel et détermination de la séquence ADN

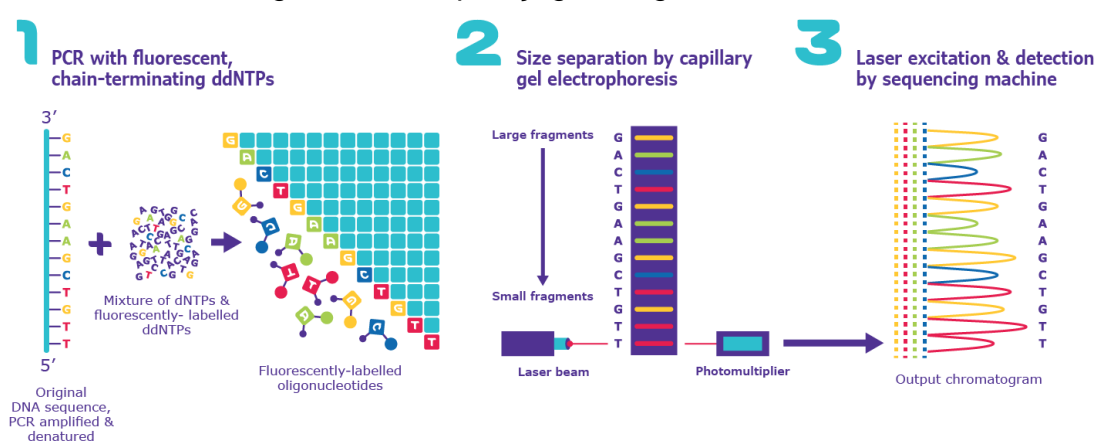
Pour déterminer la séquence ADN, on utilise la radiographie ou une lumière UV, afin d'identifier le déplacement des fragments. Ensuite, il est alors facile de déterminer la séquence, il suffit d'effectuer une lecture des fragments. Celle-ci se fait du plus petit fragment au plus grand²⁴. Chaque fragment se termine par un ddNTP spécifique, qui indique la fin de la synthèse. En retraçant l'ordre des fragments, ordonnés par taille, finissant par le même ddNTP, on peut reconstituer la séquence de base.

²⁴ G, T, C, ..., puis A sur la figure 16.

3.4.1.1.4 Analyse automatisée

Cette méthode, plus courante actuellement, utilise des ddNTPs marqués chacun par un fluorophore différent, permettant leur mélange dans une seule réaction. La séparation des fragments d'ADN se fait par électrophorèse capillaire, une méthode qui permet également de séparer les fragments d'ADN en fonction de leur taille. Excepté que cette fois-ci, un laser excite les fluorophores attachés à chaque ddNTP, chacun émettant une longueur d'onde spécifique captée par un détecteur CCD. Les résultats sont ensuite représentés sur un chromatogramme sous forme de pics de fluorescence, facilitant la détermination de la séquence nucléique de l'ADN d'intérêt.

Figure 17 : Séquençage Sanger automatisé



(Merck KGaA, 2024) [37]

Bien que le séquençage Sanger soit une méthode répandue et abordable, elle demande du temps et n'est pas entièrement fiable pour les premières bases lues.

3.4.1.2 Le séquençage Illumina²⁵

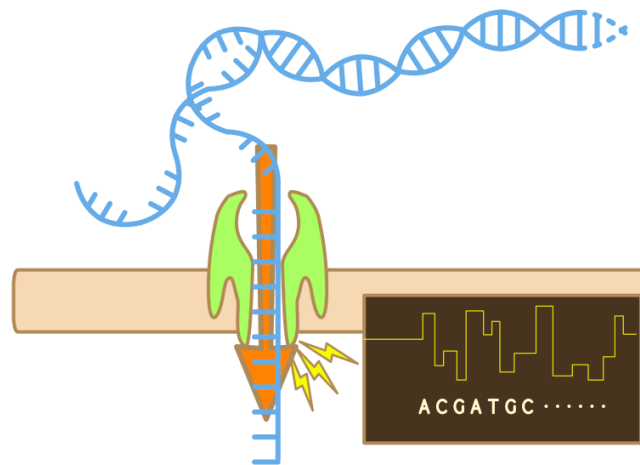
Le séquençage Illumina, appartenant à la deuxième génération des techniques de séquençage, repose sur le concept de séquençage par synthèse. Le procédé débute par la fragmentation de l'ADN, suivi de l'ancrage des fragments sur une surface solide. Pendant le séquençage, des nucléotides, chacun marqué d'un fluorophore distinct, sont ajoutés. À chaque étape, un seul nucléotide est intégré au brin d'ADN en formation. Les émissions fluorescentes capturées par une caméra indiquent le type de nucléotide ajouté. Ces signaux sont ensuite traités pour établir la séquence d'ADN. Ce processus, effectué sur plusieurs cycles, permet de séquencer simultanément des millions de fragments. [35]

²⁵ Société américaine spécialisée dans la biotechnologie.

3.4.1.3 Le séquençage par nanopores

Le séquençage par nanopores, une méthode de séquençage de la troisième génération, se distingue par son haut débit, sa capacité à fournir des résultats en temps réel et son coût avantageux. Cette technologie repose sur le passage d'un brin d'ADN à travers un nanopore extrêmement petit situé dans une membrane. Ce passage modifie spécifiquement le courant électrique en fonction de chaque base nucléique (A, T, C, G). Cette spécificité permet la lecture continue de longues séquences d'ADN. En mesurant ces changements de courant, il est possible de déterminer la séquence de l'ADN, ce qui rend le séquençage par nanopores idéal pour les projets de séquençage à grande échelle et les applications nécessitant une réponse rapide. [36]

Figure 18 : Séquençage par nanopores



(Wikipédia, 2024) [47]

3.4.2 Décodage des séquences ADN

Une fois la séquence ADN obtenue, il est primordial de transformer les séquences ADN en langage binaire pour que ces données puissent être interprétées par un ordinateur. Pour ce faire, il faut utiliser un programme – semblable au programme d'encodage vu précédemment – qui décode les données en suivant les bien évidemment les mêmes règles que celles d'encodage. Ainsi, une séquence AGCT devient 00011011, puis cette même séquence binaire peut enfin être lue et interprétée par un ordinateur.

3.4.2.1 Correction d'erreurs

3.4.2.1.1 Nécessité de correction

Comme cité précédemment, il y a une réelle nécessité de gérer les erreurs. Les processus de synthèse et de séquençage ne sont pas à l'abri d'erreurs, dues à des facteurs comme des inexactitudes chimiques, des erreurs de lecture, ou des mutations naturelles. Ainsi, corriger ces erreurs est crucial pour maintenir l'intégrité et la fiabilité

des données stockées, en vue d'une conservation à long terme sans risques de corruption.

3.4.2.1.2 Mécanismes et techniques

Les mécanismes de correction d'erreurs dans le stockage de données dans l'ADN s'inspirent souvent des techniques utilisées dans les méthodes de stockage de données traditionnelles.

Tout d'abord, la duplication de séquences ADN est une technique fondamentale pour préserver plusieurs copies d'une même séquence. Ainsi, même si certains fragments se détériorent, il reste possible de se référer à d'autres fragments encore intacts.

Ensuite, comme mentionné précédemment, il existe des programmes, ou codes, de correction d'erreurs. Par exemple, les « codes d'effacement » s'avèrent extrêmement utiles dans les situations où il est probable que certaines parties des données soient manquantes ou altérées durant la transmission ou le stockage. [21] Ces codes opèrent en intégrant des données supplémentaires et redondantes lors de l'encodage, permettant ainsi la reconstitution des informations égarées au moment de la récupération des données. Par exemple, imaginons un cas simple ; nous souhaitons stocker trois chiffres : 3, 10, et 25. Si nous voulons être en mesure de les retrouver quand bien même un d'entre eux venait à manquer, nous pourrions alors stocker la somme de ces chiffres, c'est-à-dire 38.

$$\begin{array}{rcl} 3 & + & 10 & + & ? & = & 38 \\ 38 & - & 10 & - & 3 & = & 25 \end{array}$$

Ainsi, mathématiquement, on peut récupérer une valeur perdue en inversant l'équation. Bien que cela représente un cas simplifié, ce concept peut être étendu et appliqué à des situations plus complexes. En d'autres termes, cette méthode de récupération de données via l'inversion d'équations s'avère un outil puissant, capable de s'adapter à des scénarios de stockage de données plus élaborés et volumineux.

3.4.2.1.3 Optimisation et améliorations

Dans le contexte de l'optimisation, ce sont les avancées dans les processus de synthèse et de séquençage de l'ADN, ainsi que les progrès réalisés dans les algorithmes de correction d'erreurs, qui vont jouer un rôle crucial dans la révolution du stockage de données dans l'ADN. En effet, à mesure que ces technologies se rapprochent de la perfection en termes de rapidité, précision, intégrité des données et coût, l'utilisation pratique et courante du stockage de données sur l'ADN devient de plus en plus une réalité tangible.

4. Préoccupations environnementales et éthiques

4.1 Notre environnement, une priorité

Les conséquences du dérèglement climatique s'intensifient et s'accroissent : incendies de forêt, températures record, inondations, sécheresses, ouragans, et bien d'autres phénomènes deviennent de plus en plus fréquents. Aujourd'hui plus que jamais, nous sommes à un tournant crucial où l'action est impérative. L'ONU tire la sonnette d'alarme, révélant que la fréquence de ces événements a été multipliée par cinq en seulement 50 ans. Sans mesures concrètes pour inverser cette tendance, ces conséquences pourraient devenir irréversibles, faisant de ces phénomènes extrêmes notre nouvelle réalité quotidienne. [46]

Parallèlement, la perte de biodiversité est également alarmante, de nombreuses espèces se sont éteintes, soulignant l'urgence d'agir pour protéger notre patrimoine naturel. Face à ces défis environnementaux, la transition écologique devient une nécessité impérieuse. Dans ce contexte, le stockage de données sur l'ADN se présente comme une partie intégrante de la solution. Cette technologie a le potentiel d'être bien plus écologique que les supports de stockage traditionnels, offrant ainsi une alternative prometteuse dans notre quête de solutions durables face au changement climatique.

4.1.1 WORN (Write Once Read Never)

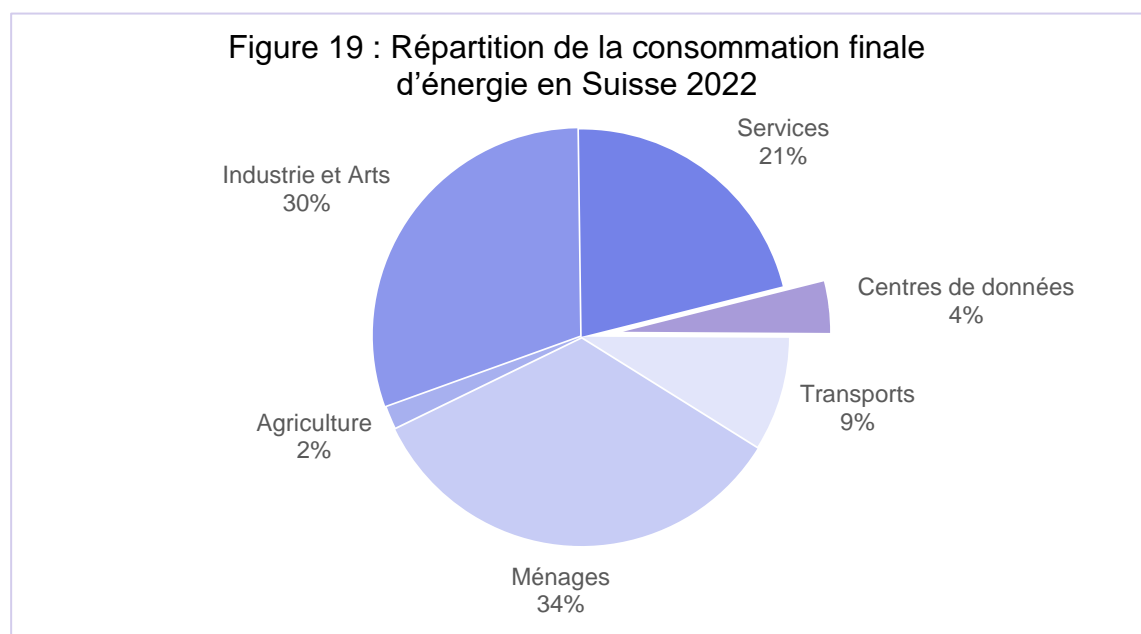
L'étude de Seagate²⁶ en 2020 révèle une utilisation sous-optimale des données dans les entreprises. D'après le rapport, « seulement 32% des données disponibles sont effectivement utilisées, laissant 68% inexploitées » (Seagate 2020, p. 5). [31] Ce constat met en lumière un potentiel considérable dans les données d'entreprise encore inexploré. Il souligne l'importance pour les entreprises d'élaborer des stratégies plus efficaces pour gérer et exploiter leurs données. Cela implique l'identification des données pertinentes, l'amélioration des techniques d'analyse, et l'adoption de technologies avancées pour maximiser l'usage de ces ressources informationnelles. Le rapport suggère que les entreprises doivent améliorer la gestion de leurs données pour en tirer un meilleur profit. Cependant, il faut reconnaître que de nombreuses données stockées sont très certainement inutiles puisqu'elles ne sont jamais consultées. En conséquence, une gestion plus judicieuse des données pourrait également conduire à une réduction significative de l'espace de stockage nécessaire.

²⁶ Seagate Technology Holdings plc est une société américaine spécialisée dans le stockage de données de masse.

4.1.2 Consommation d'énergie

Même si certains centres de données s'efforcent d'adopter des énergies renouvelables, leur consommation globale d'énergie reste importante. Ces infrastructures requièrent une énergie substantielle, non seulement pour opérer les serveurs qui stockent nos données, mais aussi pour en assurer le refroidissement. Par exemple, en 2019, les centres de données en Suisse ont consommé 2,1 térawattheures (TWh), ce qui correspond à 3,6% de la consommation électrique totale du pays, évaluée à 57,2 TWh cette année-là. La Suisse se distingue en Europe par un nombre élevé de centres de données par habitant, avec un total de 86 en 2023. [13][22][23]

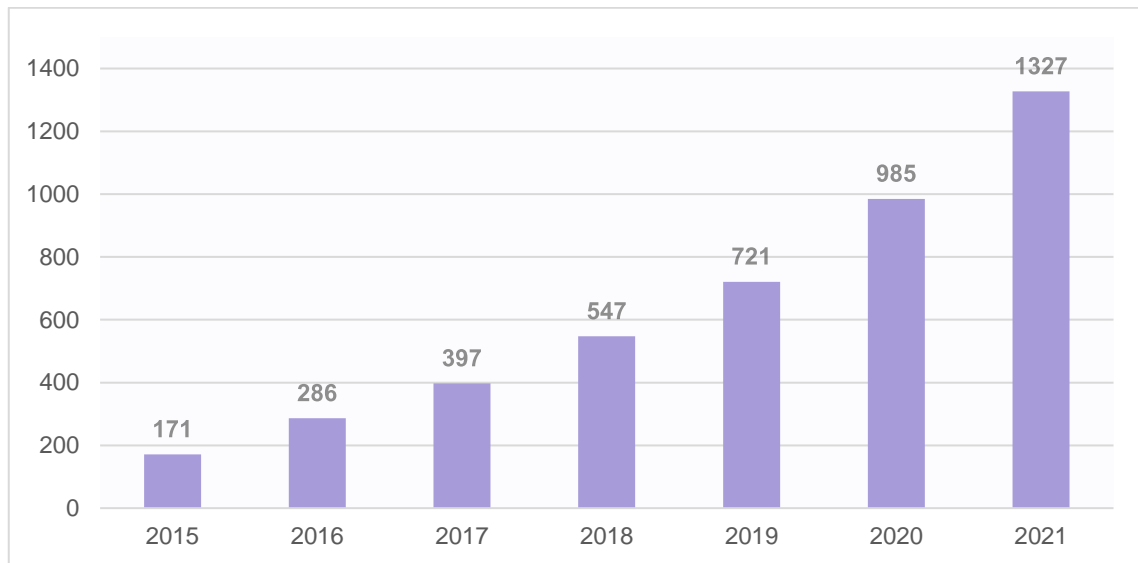
Le graphique ci-dessous offre un aperçu de la répartition de la consommation énergétique dans divers secteurs. Cette visualisation aide à contextualiser la part de consommation attribuable aux centres de données, en comparaison avec d'autres secteurs d'activité.



(Office fédéral de l'énergie, 2023) [24]

En comparaison avec les centres de données traditionnels, le stockage de données sur ADN se révèle nettement plus avantageux en termes de consommation d'énergie. En effet, une fois l'ADN synthétisé, il peut être stocké et préservé pendant des millénaires avec un besoin énergétique quasi nul, et ce, à température ambiante. Des estimations suggèrent même que, conservées à -18°C, les données stockées dans l'ADN pourraient atteindre une durée de conservation avoisinant le million d'années. De plus, la haute densité de stockage de l'ADN permet une réduction supplémentaire de la consommation énergétique, bien au-delà de ce qu'exiger le stockage classique.

Figure 20 : Quantité de données stockées dans les centres de données dans le monde de 2015 à 2021 (en exaoctets).



(Statistica, 2023) [25]

Ce graphique illustre qu'en 2021, les centres de données à travers le monde stockaient déjà 1'327 exaoctets²⁷ de données. Étant donné la croissance constante de la création et du stockage de données, il est évident qu'en 2023, cette quantité a augmenté. Comme mentionné précédemment, théoriquement, un seul gramme d'ADN peut stocker jusqu'à 215 pétaoctets (Po), soit 0,215 exaoctets (Eo). [44] Par conséquent, l'ensemble des données stockées dans les centres de données en 2021, soit 1'327 Eo, pourrait théoriquement être contenu dans environ

$$\frac{1327}{0,215} \cong 6'172 \text{ grammes d'ADN.}$$

Bien entendu, il est bien illusoire de penser que nous pourrions rapidement atteindre ces taux de rendement théoriques. Néanmoins, même si nous réduisons nos attentes par un facteur de 1'000 ou même de 1'000'000, le stockage d'ADN resterait extrêmement compact comparé aux millions de mètres carrés actuellement nécessaires²⁸ pour héberger les centres de données requis pour stocker nos données.

4.1.3 Utilisation d'eau

L'intensification des phénomènes climatiques extrêmes exerce une pression croissante sur nos ressources en eau. Alors que la nécessité de modérer la consommation globale d'eau devient de plus en plus pressante, les centres de données, en revanche,

²⁷ 1 Eo correspond à 10⁹ Go.

²⁸ Il est difficile d'obtenir des données précises, mais selon certaines estimations, Amazon Web Services (AWS) posséderait à lui seul environ 20 millions de mètres carrés dédiés à ses centres de données.

continuent de consommer des quantités astronomiques de cette ressource, souvent sans opposition notable. Pour refroidir les serveurs et pallier les températures élevées, de nombreux centres de données se tournent vers l'utilisation de circuits d'eau. Même si certains de ces centres s'efforcent d'atteindre une neutralité carbone, leur consommation d'eau demeure conséquente. Par exemple, en 2021, Google a révélé sa consommation en eau aux États-Unis : environ 4,338 millions de gallons, soient près de 16 milliards de litres, sur un an. En 2022, Microsoft a enregistré une consommation de plus de 6,4 milliards de litres d'eau, non sans controverse. En septembre de cette même année, il a été révélé aux Pays-Bas que Microsoft avait utilisé 84 millions de litres d'eau, bien au-delà des 12 à 20 millions estimés initialement. Cette consommation élevée d'eau représente donc un problème significatif pour de nombreux centres de données. [26][27]

4.1.4 Minéraux critiques et e-déchets

L'utilisation de matériaux rares est effectivement au centre de nombreuses controverses environnementales et éthiques. Les centres de données détiennent une grande quantité de minéraux, dits critiques, dans leur équipement. Par exemple, les serveurs intègrent souvent du titane, de l'étain ou du platine, utilisés respectivement comme pigment blanc, revêtements protecteurs et agents catalytiques. Le cas du cobalt en République Démocratique du Congo²⁹ illustre parfaitement les problèmes liés à l'extraction de ces ressources, notamment l'exploitation minière inhumaine impliquant le travail forcé d'enfants, posant ainsi un grave problème de droits humains. En outre, le recyclage des déchets électroniques représente un défi environnemental et sanitaire majeur, notamment lorsque ces déchets sont expédiés en Afrique de l'Ouest. Les travailleurs locaux, souvent dépourvus de protection adéquate, y sont exposés à des substances toxiques, augmentant ainsi le risque de maladies graves. [28]

Dans le cadre du stockage de données dans l'ADN, les perspectives semblent plus prometteuses. Bien que les équipements de synthèse et de séquençage d'ADN nécessitent certains composants rares, ils en requièrent probablement moins que les technologies traditionnelles. Grâce à la densité de stockage élevée de l'ADN, la quantité de composants électroniques nécessaires pour stocker un volume important de données est considérablement réduite. Le stockage dans l'ADN pourrait conduire à des avancées significatives en matière de recyclage et de gestion des déchets électroniques. Si les composants requis pour la lecture et l'écriture dans l'ADN sont produits de manière plus durable, cela pourrait révolutionner la manière dont les données sont stockées et traitées, tout en atténuant les impacts sur l'environnement et la santé humaine.

²⁹ Près de 60% de la production mondiale provient de la région.

4.2 Éthique

4.2.1 Usage malintentionné de l'ADN

Actuellement, la synthèse d'ADN est relativement contrôlée dans le sens où très peu d'instances possèdent les technologies nécessaires pour produire des séquences avec un minimum d'erreurs. Un des risques majeurs concerne la capacité des centres de synthèse à identifier rapidement les séquences d'ADN potentiellement pathogènes. En effet, une personne ayant accès à des séquences virales pourrait théoriquement déclencher des épidémies si des contrôles stricts ne sont pas appliqués. Il est donc impératif que le développement et l'exploitation des centres de synthèse soient rigoureusement réglementés, avec des normes de vérification et de validation strictes.

4.2.2 Inégalités

À l'heure actuelle, le coût de la synthèse d'ADN reste élevé. Par exemple, chez Twistbioscience, le coût de synthèse pour une séquence de 500 paires de bases est d'environ USD 50, hors frais de livraison. Si le développement de la technologie de synthèse et de séquençage continue sans avancées majeures, l'utilisation de l'ADN comme support de stockage risque d'être initialement plus accessible aux entités aisées. Dans un contexte où la transition écologique est essentielle, ce sont probablement les sociétés et pays les plus riches qui pourront se permettre d'assumer ces coûts.

4.2.3 Organismes vivants

L'idée d'utiliser des organismes vivants comme dupicateurs de séquences d'ADN est activement étudiée. Cette méthode implique d'insérer de l'ADN à double brin, préalablement synthétisé, dans des organismes tels que la bactérie *Escherichia Coli*. La bactérie utilise alors ses processus biologiques naturels pour reproduire la séquence d'ADN. Ce procédé exploite la capacité naturelle de réplication de l'ADN de l'organisme hôte, offrant un moyen potentiellement plus efficace et moins onéreux de dupliquer l'ADN que les méthodes de laboratoire. L'avantage majeur de cette approche est sa capacité à générer un grand nombre de copies. Une fois l'archive d'ADN établie au sein de l'organisme, il devient possible de produire des milliers, voire des millions de copies. En résumé, utiliser les organismes vivants pour la réplication de l'ADN ouvre des perspectives captivantes pour le stockage et la duplication de grandes quantités de données, en exploitant les processus biologiques naturels, pour une duplication efficace et économique. [29] D'un point de vue technique, cette approche exploite la capacité naturelle de réplication de l'ADN des organismes hôtes pour produire de multiples copies d'une séquence. Du point de vue éthique, il est difficile de considérer que l'utilisation des bactéries pour le stockage de données soulève un problème moral significatif. Malgré

leur statut d'êtres vivants, les bactéries sont déjà largement utilisées dans divers domaines sans susciter de préoccupations éthiques majeures. De plus, dans le contexte du stockage de données, l'objectif n'est pas du tout lié à la modification ou création de nouvelles formes de vie.

Cependant, il est important de reconnaître que les avancées dans le domaine de la biologie synthétique, telles que la capacité à modifier ou à créer de nouvelles formes de vie, soulèvent des questions éthiques complexes. Ces progrès technologiques amènent certains à questionner les limites morales de nos actions : avons-nous le droit de créer de nouvelles formes de vie pour répondre à nos besoins ? Cette interrogation met en lumière un débat plus large sur les implications éthiques de la manipulation de la vie à des fins technologiques.

5. Conception expérimentale

Pour enrichir ma recherche sur le stockage de données dans l'ADN, j'ai intégré une expérience pratique. Celle-ci m'a permis d'obtenir une compréhension plus complète des processus et technologies impliqués. Cette expérience a été une occasion précieuse de mesurer les coûts réels et d'évaluer la durabilité du stockage ADN. Elle m'a également permis de valider et remettre en question certaines de mes hypothèses de base, tout en mettant en lumière des aspects que j'avais omis.

5.1 Organismes impliqués

Pour mener à bien cette expérimentation, j'ai dû collaborer avec plusieurs institutions et entreprises spécialisées dans différents aspects du stockage dans l'ADN.

- Twist Bioscience : une société américaine spécialisée dans la synthèse d'ADN.
- Haute Ecole de Gestion HES-SO Genève : chargée de la réception d'un fragment d'ADN.
- Institute of Life Technologies HES-SO Valais : chargé du séquençage par nanopores.
- Fasteris : société suisse spécialisée dans la biotechnologie, chargée du séquençage Sanger.

5.2 Mes hypothèses

Initialement, j'avais sous-estimé la complexité du stockage de données dans l'ADN, je m'attendais à une plus grande simplicité et accessibilité. En effet, je pensais qu'il serait beaucoup plus simple de trouver des entreprises compétentes dans ce domaine, cependant, je me suis rapidement rendu compte que même parmi les sociétés spécialisées en synthèse et en séquençage d'ADN, la notion de stockage de données ADN n'était pas une pratique courante.

De plus, avant de commencer mon projet, j'avais une confiance exagérée dans la fiabilité des technologies de synthèse et séquençage, pensant qu'elles étaient extrêmement précises. Or, bien que certaines aient des taux de précision plus élevés que d'autres, aucune n'atteint une perfection absolue. Cette prise de conscience m'a amené à saisir l'importance cruciale des mécanismes de correction d'erreurs. Dans mon projet, l'absence de développement approfondi de cet aspect a soulevé un très grand risque d'échec de l'expérimentation pratique en raison des potentielles erreurs de synthèse ou lecture de l'ADN³⁰. Bien que les séquences que j'utilise soient relativement courtes,

³⁰ L'absence de correction d'erreurs dans mon approche expérimentale rend chacune d'entre elles extrêmement significative pour la récupération de la séquence binaire correspondante.

minimisant ainsi le risque d'erreurs, la gestion de ces erreurs n'en reste pas moins importante.

5.2.1 Obstacles rencontrés

- Synthèse d'ADN : la synthèse d'ADN a présenté des défis logistiques notables. La société Twist Bioscience, en charge de la livraison des séquences ADN, n'assure pas la distribution aux particuliers. J'ai dû recourir à des méthodes alternatives, comme la livraison d'un fragment à la HEG-GE, ainsi qu'à l'Institute of Life Technologies. Un autre défi de synthèse concernait la redondance des bases nucléiques dans les séquences ADN. Une répétition excessive rend la synthèse difficile et peut compromettre la stabilité de la séquence.
- Sociétés de séquençage d'ADN : trouver des entreprises capables de séquencer un seul échantillon pour mon projet a été un obstacle majeur. La majorité des centres de séquençage traitent des commandes en grand volume. De plus, grand nombre d'entre eux opèrent uniquement en B2B, rendant l'accès difficile pour les privés. Heureusement, mon adresse email universitaire a facilité l'obtention de leurs services.
- Séquençage par nanopores : mon intention initiale était d'effectuer moi-même le séquençage des fragments d'ADN en utilisant un séquenceur d'Oxford Nanopore Technologies. Bien que financièrement abordable, avec des séquenceurs disponibles à partir d'USD 1'000, je n'avais pas prévu les préparatifs complexes nécessaires. [12]

Figure 10 : Séquenceur MinION



(Oxford Nanopore Technologies, 2023)

Après consultation avec des experts en génétique forensique et génomique du CHUV, nous avons conclu qu'il n'était pas essentiel que je réalise le séquençage moi-même. En effet, l'industrie actuelle dispose de différents acteurs spécialisés pour chaque étape du processus de stockage de données dans l'ADN. L'avenir repose plutôt sur des avancées technologiques qui permettront de séquencer l'ADN sans nécessiter de connaissances spécialisées, ainsi, il n'était alors pas primordial que je procède au séquençage de mon échantillon.

En outre, le séquençage par nanopores a révélé un autre problème critique : le traitement des résultats. Ce type de séquençage génère des centaines de fichiers³¹, chacun contenant des centaines de lectures. Initialement, je m'attendais à recevoir uniquement la séquence synthétisée en guise de résultat concluant, mais j'ai dû comprendre comment ces vastes données étaient traitées. J'ai ensuite découvert qu'il fallait obtenir une *séquence de consensus*³² ; ça consiste à comparer les bases nucléiques lues et conserver les plus fréquentes, éliminant ainsi les potentielles erreurs de lecture. [30]

- Séquençage Sanger : le séquençage Sanger n'est pas totalement fiable pour la lecture des 20-25 premières bases nucléiques, entraînant des erreurs fréquentes dans cette section, voire une non-identification de nucléotides.

5.3 Programme informatique

Dans la conception de mon programme informatique, j'ai opté pour la simplicité, en visant la création d'un outil permettant l'encodage de données en séquences ADN. J'ai choisi Java comme langage de programmation, étant donné ma familiarité avec celui-ci. Initialement, mon focus était sur l'encodage de texte. Néanmoins, poussé par la curiosité, j'ai étendu mon exploration à l'encodage de fichiers. Cette démarche m'a semblé plus alignée sur les applications réelles du stockage de données dans l'ADN.

Cependant, lors de l'expérimentation, j'ai été confronté à un obstacle financier majeur : la synthèse d'un fragment d'ADN stockant un fichier s'est révélée inabordable. En effet, les fichiers encodés produisent des séquences ADN beaucoup plus longues que du simple texte, entraînant ainsi une augmentation significative des coûts et de la complexité de synthèse. Ainsi, cette contrainte financière a limité mon expérience à l'encodage de texte dans l'ADN.

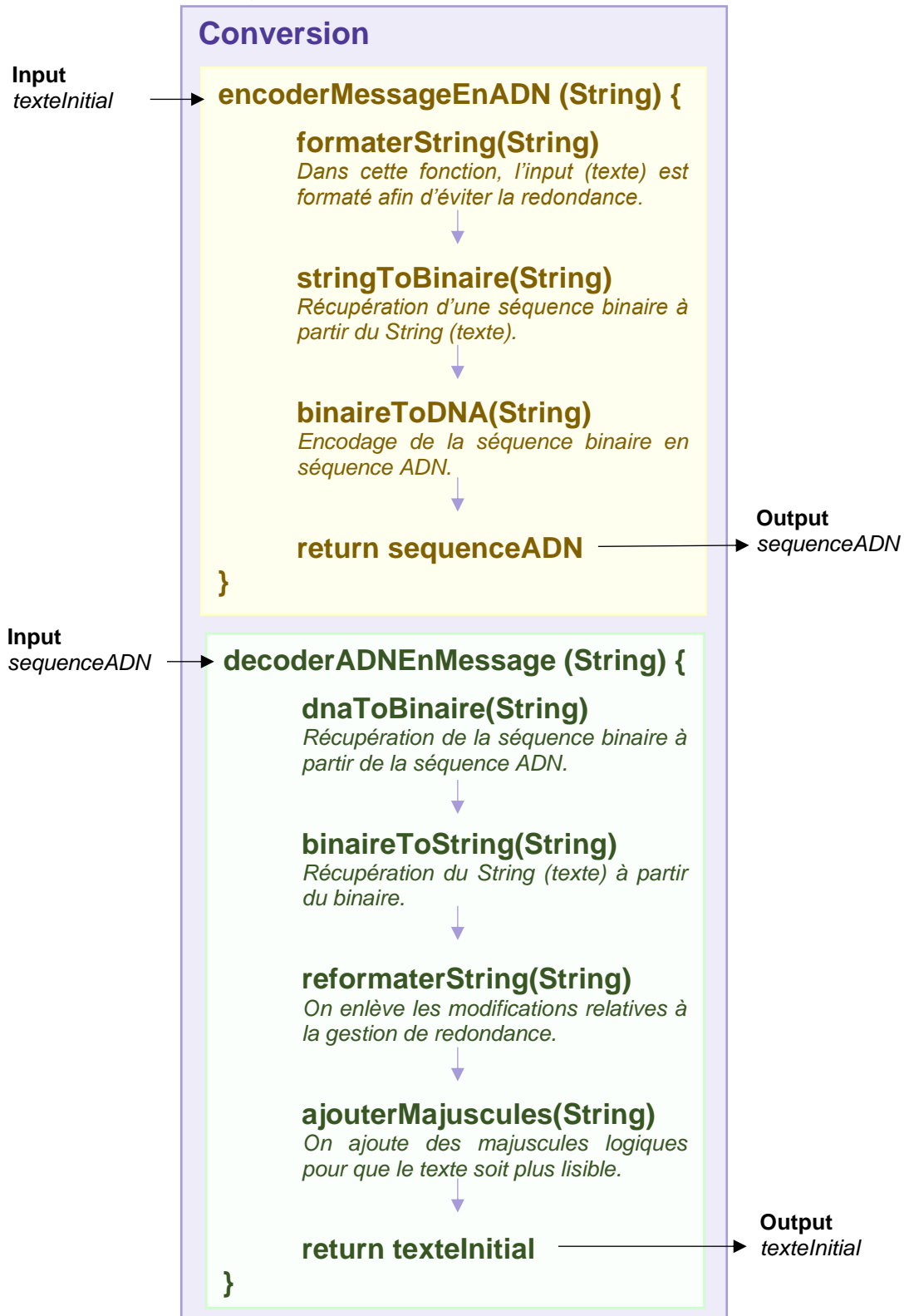
³¹ Près de 800 fichiers distincts dans mon cas.

³² Voir figure 24 à la page 41

5.3.1 Classe Conversion³³

5.3.1.1 Encoder et décoder du texte

Figure 11 : Encoder et décoder du texte



(Mirko Rivas, 2023)

³³ Vous trouverez le code à l'annexe 1.

Dans le schéma présenté, on observe deux fonctions clés de la classe Conversion : l'encodage et le décodage de *texte*, accompagnées de leurs sous-fonctions associées. Ce schéma illustre de façon simplifiée les processus d'encodage et de décodage de texte. Voici quelques précisions :

5.3.1.1.1 Gestion de la redondance

Comme expliqué précédemment, la fonction *formaterString()* joue un rôle crucial en évitant la redondance dans le texte. Une redondance non gérée peut rendre les séquences ADN générées difficiles, voire impossibles à synthétiser. Par exemple, l'omniprésence de l'espace " " dans les langues peut entraîner des complications dans la séquence binaire et, par conséquent, dans la séquence ADN. Un défi rencontré avec cette méthode est la nécessité de remplacer les espaces par d'autres caractères, ce qui les rend inutilisables pour leurs fonctions d'origine³⁴.

5.3.1.1.2 Processus d'encodage

Lors de l'encodage d'un texte, plusieurs étapes sont impliquées. Prenons l'exemple de la phrase « Salut comment tu vas ? ». Après le formatage pour minimiser les redondances, et l'ajout des marqueurs DEBUT et FIN, la phrase est transformée en « DEBUTSalut comMenT tU\$vAs_?FIN ». Chaque caractère est ensuite converti en sa séquence binaire correspondante, complétée à huit bits si nécessaire. Par exemple pour la lettre D, la séquence binaire correspondante est 1000100, or cette séquence ne fait que sept bits. Il est impératif de rajouter des zéros afin que les séquences fassent toutes huit bits : 1000100 → 01000100. Cette séquence binaire est ensuite convertie en séquence ADN à l'aide d'une table de conversion établie³⁵, où D devient 'GAGA'. Ce processus est appliqué à l'ensemble du texte. Ainsi, la phrase formatée devient, 01000100010001010101000010010101010101000101001101100001011011000111010101110100001000000110001101101111011011010100110101100101011011100101010000100000011101000101010100101011011101100100000101110011001001100011111010001100100100101001110.

5.3.1.1.3 Processus de décodage

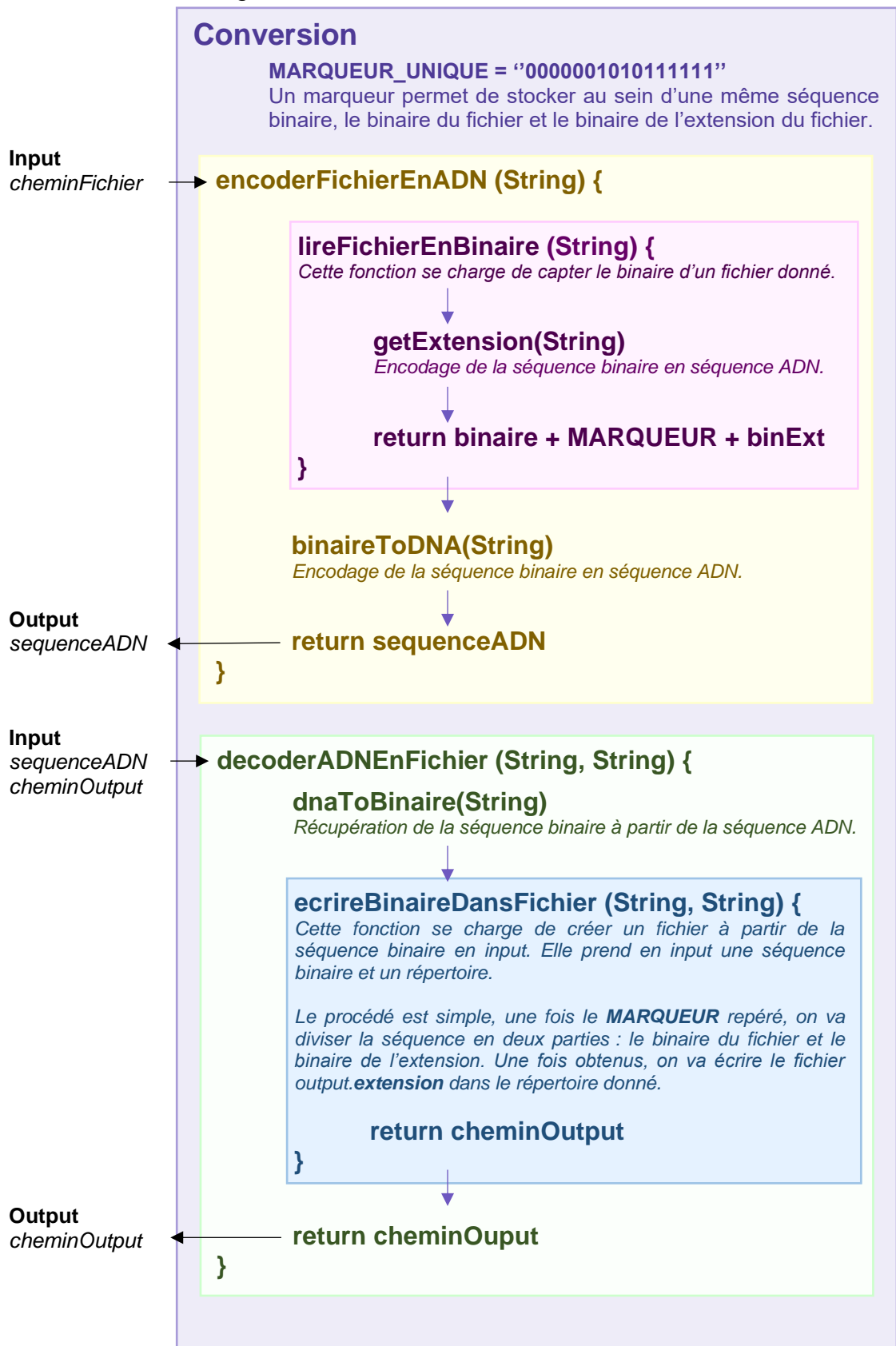
Le décodage suit le même processus que l'encodage, mais dans le sens inverse. La séquence ADN est d'abord reconvertie en séquence binaire, puis en texte, avec restitution des caractères originaux et élimination des marqueurs DEBUT et FIN.

³⁴ List(" ", "#", "&", "_", "+", "^", "\$", "="), je remplace des " " par des caractères de cette liste. Si mon texte d'entrée contient un des éléments ci-dessus, ils seront remplacés par des " " lors du reformatage.

³⁵ A pour 00, G pour 01, C pour 10 et T pour 11.

5.3.1.2 Encoder et décoder un fichier

Figure 12 : Encoder et décoder un fichier



(Mirko Rivas, 2023)

Dans le deuxième schéma, nous examinons les processus d'encodage et de décodage de *fichiers*. Bien que similaire à l'encodage de texte, l'encodage de fichier dans la classe Conversion présente des spécificités liées au traitement binaire du fichier.

5.3.1.2.1 Utilisation d'un marqueur unique

J'ai intégré un MARQUEUR_UNIQUE³⁶ dans la séquence binaire pour séparer le binaire du fichier de l'extension du fichier (également encodée en binaire), afin de pouvoir récupérer et lire le fichier³⁷ question sans problème. Un défi majeur a été de garantir l'unicité du marqueur pour son identification. Initialement, je pensais qu'un marqueur de type 1111111111111111 serait adéquat, car il est extrêmement rare d'avoir autant d'occurrences de 1, du moins du cadre de mes essais. Cependant, ce marqueur s'est avéré vulnérable à des erreurs d'identification, notamment si un « 1 » apparaissait juste avant le marqueur. Par exemple dans la séquence suivante « 0101001 1111111111111111 01001001101011000 »³⁸, si on essaie d'identifier le marqueur alors on identifiera le marqueur, mais décalé d'une place, comme ceci « 0101001 1111111111111111 1 01001001101011000 ». Ce type d'erreur pouvait donc compromettre la récupération du fichier et de l'extension.

Un autre élément compromettant est le fait que dans un cas d'usage réel, le marqueur doit impérativement être suffisamment aléatoire pour qu'il puisse être synthétisé sans problème. Or, la redondance de caractères dans mon marqueur, le rend potentiellement inexploitable pour un emploi concret.

5.3.1.2.2 Processus encodage

Cette fonction contient la fonction *lireFichierEnBinaire()* qui est chargée de récupérer l'extension d'un fichier donné et de la convertir en binaire. Par exemple pour le fichier « test.txt », on obtiendra 00101110011101000111100001110100 qui en binaire signifie « .txt ». Ensuite, elle se charge de lire un fichier octet par octet pour obtenir la séquence binaire relative au fichier. Finalement, cette fonction va retourner une séquence binaire qui contient le **binaire du fichier**, le **marqueur** et le **binaire de l'extension**, comme ceci : 0101100101000111000000101011111101001011. Finalement, comme pour l'encodage de texte, le binaire sera transformé en séquence ADN à l'aide de la fonction *binaireToDNA()*.

³⁶ 00000000000000010101111111111111

³⁷ Lorsqu'on récupère les données binaires d'un fichier, son extension n'est pas encodée. Ainsi il est impératif de la stocker quelque part, si on veut pouvoir récupérer lire le fichier.

³⁸ Marqueur en gras

5.3.1.2.3 Processus de décodage

Pour le décodage, le processus est tout simplement inversé. On convertit la séquence ADN en binaire, puis on utilise le marqueur pour séparer et identifier correctement les différentes parties de la séquence. Enfin, le fichier est reconstruit octet par octet dans le répertoire spécifié, avec la bonne extension de fichier.

5.4 Coûts

Comme mentionné précédemment, les technologies actuelles nécessaires au stockage de texte dans l'ADN restent relativement coûteuses. Pour illustrer cela, voici un aperçu détaillé des coûts engendrés par mon expérience :

- **Synthèse de l'ADN** : le coût pour chacun des deux fragments d'ADN commandés s'est élevé à environ USD³⁹ 100. Ce montant inclut le prix de la synthèse d'ADN, environ USD 50 par fragment, c'est-à-dire environ USD 0.09 par paire de bases, et inclus également les frais de livraison estimés à USD 50.
- **Séquençage par nanopores** : le coût du séquençage, réalisé par l'ITV HES-Valais, a été de CHF 215.
- **Séquençage Sanger** : ce séquençage, effectué par Fasteris, a coûté CHF 75.

Ces chiffres soulignent le budget significatif nécessaire pour entreprendre une telle expérience dans le domaine actuel du stockage de données ADN.

³⁹ CHF 1 = USD 1.15 au moment de l'expérience.

6. Résultats

Dans les sections suivantes, je compare les résultats obtenus avec les deux méthodes de séquençage utilisées : par nanopores et Sanger.

6.1 Séquençage par nanopores

Pour ce qui est du séquençage par nanopores, j'ai malheureusement dû écarter les résultats d'analyse, faute de compétences nécessaires pour traiter l'ensemble des données. Malgré plusieurs tentatives de collaboration avec des sociétés spécialisées en séquençage par nanopores, une seule a accepté de traiter les résultats pour établir une séquence de consensus. Cette démarche visait à identifier la séquence la plus fréquente à chaque position nucléique.

Figure 24 : Séquence de Consensus

n ° 1	:	A T C G C A T C A A T
n ° 2	:	A T T G C A T C G A A
n ° 3	:	A A C G C A T G G A A
n ° 4	:	A T C C C A C C G A A
n ° 5	:	T T C G G T T C G A A
n ° 6	:	A C C C C A C C G T A
n ° 7	:	T T A G G T T C G A A
n ° 8	:	A T C C C A C C G A A
n ° 9	:	A T C G C A T A C A G
Consensus		: A T C G C A T C G A A

(Mirko Rivas, 2024)

Toutefois, le coût de ce service, estimé à CHF 600, dépassait mes moyens financiers. J'ai donc choisi de retirer cette partie de mon expérience.

6.2 Séquençage Sanger

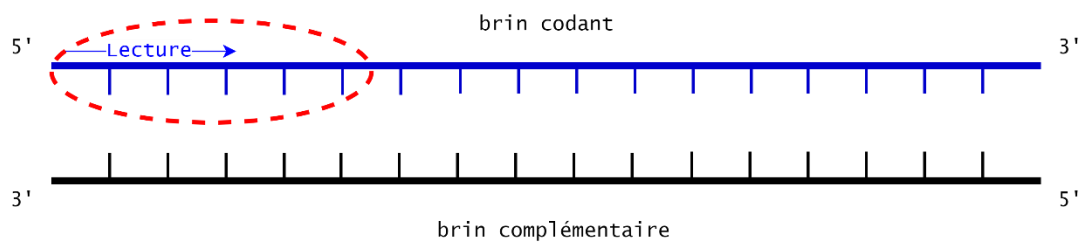
La méthode Sanger, particulièrement adaptée pour les fragments de 500 à 1000 paires de bases, a donné des résultats très satisfaisants et rapides⁴⁰. Cependant, un malentendu avec la société Fasteris a conduit à l'inclusion d'un adaptateur inutile sur un seul côté 5' de ma séquence⁴¹ lors de ma commande. Normalement, cet adaptateur

⁴⁰ Résultats reçus en 2 jours ouvrables après avoir déposé l'échantillon d'ADN.

⁴¹ Cet adaptateur était inséré avant ma séquence codante afin de permettre l'accroche d'une amorce et ainsi faciliter le processus de séquençage.

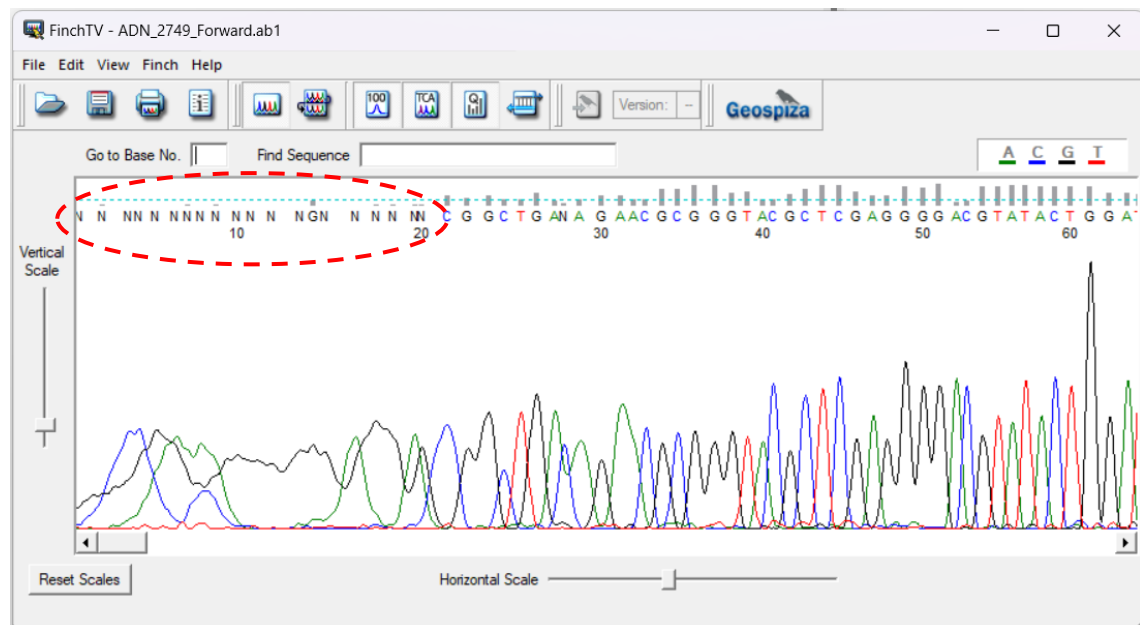
servait à faciliter la lecture du brin dans un sens spécifique, cependant, si la lecture n'avait été faite que dans un sens, il aurait alors été impossible de récupérer les premières bases nucléiques du brin⁴² (en rouge). Sur la figure 26, les premières bases non déterminées par le séquençage sont facilement identifiables : elles sont représentées par la lettre 'N' au lieu des nucléotides spécifiques (A, T, G, ou C).

Figure 25 : Séquençage forward



(Mirko Rivas, 2024)

Figure 26 : Analyse du séquençage forward avec FinchTV



(Mirko Rivas, 2024)

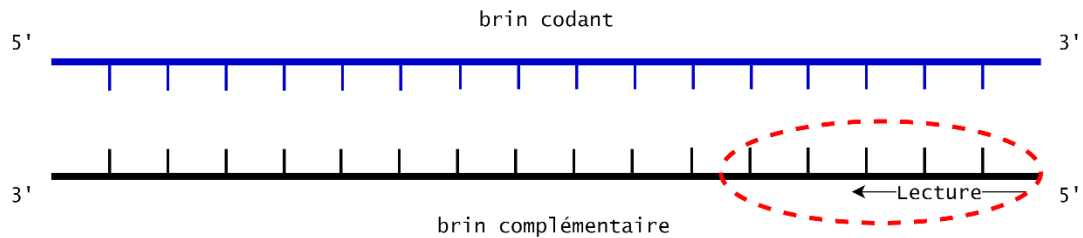
Ceci aurait grandement posé un problème puisque mon programme, ne comportant pas de correction d'erreurs, aurait été incapable de retrouver les bases nucléiques manquantes.

Pour résoudre ce problème, nous avons décidé avec Fasteris de lire le fragment dans le sens inverse à partir de l'autre extrémité 5'. En exploitant la complémentarité de l'ADN,

⁴² Les 20-25 premières bases lues lors d'un séquençage Sanger ne sont pas fiables.

il devient alors possible d'obtenir la séquence du brin codant par une lecture en reverse du brin complémentaire⁴³.

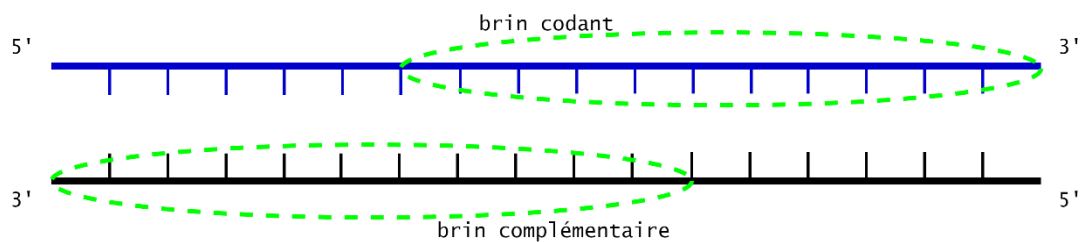
Figure 27 : Séquençage reverse du brin complémentaire



(Mirko Rivas, 2024)

Cette méthode permet de compenser les erreurs potentielles en combinant les deux lectures et en ne retenant que les parties fiables.

Figure 28 : Parties fiables des lectures forward et reverse



(Mirko Rivas, 2024)

La séquence obtenue était presque identique à celle commandée, à l'exception d'une base nucléique supplémentaire non-définie en fin de séquence. En éliminant l'adaptateur⁴⁴ et la base supplémentaire non-définie⁴⁵, il était ensuite possible de définir la séquence ADN codante.

GGCATTTTGCTGCC**GGTCACG**GAGAGAGGGAACGGGGGGGAGGGAGCCGGCTG
ACAAGAACGCGGGTACGCTCGAGGGGACGTATACTGGATAGCGGGAGGACAAG

[...] ⁴⁶

TGGGATCGAGGACGCGCACGCTTGCTCGATCGCGGACATGACGGCGAGAGGGC
GGA**CT**CGAGCGACGGATC**N**

⁴³ Les paires de bases sont toujours les mêmes : A-T et C-G.

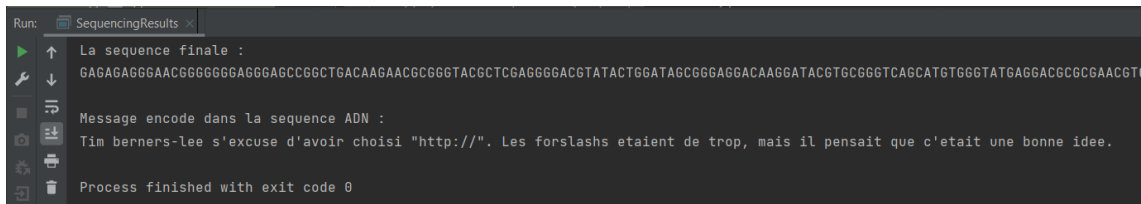
⁴⁴ En bleu

⁴⁵ En rouge

⁴⁶ La séquence de référence complète se trouve à l'annexe 3.

Après avoir obtenu cette séquence, il suffit de la traiter avec mon programme Java pour décoder le message texte qu'elle contient.

Figure 29 : Décodage de la séquence ADN



```
Run: SequencingResults x
La sequence finale :
GAGAGAGGGAACGGGGGGGAGGGAGCCGGCTGACAAGAACGCGGGTACGCTCGAGGGGACGTATACTGGATAGCGGGAGGACAAGGATACGTGCGGGTCAGCATGTGGGTATGAGGACGCGCGAACGT
Message encode dans la sequence ADN :
Tim berners-lee s'excuse d'avoir choisi "http://". Les forslashes etaient de trop, mais il pensait que c'etait une bonne idee.
Process finished with exit code 0
```

(Mirko Rivas, 2024)

Le message encodé était le suivant :

« Tim berners-lee s'excuse d'avoir choisi "http://". Les forslashes etaient de trop, mais il pensait que c'etait une bonne idee. »

L'inventeur du web, Tim Berners-Lee, a reconnu, près de 20 ans après sa création, que l'inclusion de « // » dans l'URL n'était pas nécessaire. En effet, leur présence a engendré de nombreuses erreurs et confusions, notamment avec les « \ » chez certains utilisateurs. [32]

7. Conclusion

Dans le contexte actuel, marqué par une augmentation exponentielle des données numériques, il est crucial de trouver des méthodes de stockage plus viables et durables. Les limites des supports de stockage traditionnels, tant en termes de capacité que d'impact environnemental, et bien d'autres aspects, rendent l'exploration de nouvelles technologies telles que le stockage sur ADN non seulement innovante, mais aussi nécessaire. [15] Cette recherche de solutions plus efficaces et écologiques est primordiale pour faire face aux besoins croissants en matière de stockage et de conservation des données dans notre monde de plus en plus numérisé.

Au cours de mon travail de recherche sur le stockage de données dans l'ADN, ainsi que grâce à mon essai pratique, j'ai pu constater et mettre en lumière certains points clés de cette technologie. Premièrement, le coût élevé de la synthèse d'ADN constitue un obstacle majeur à l'évolution de ce moyen de stockage. Actuellement, le coût de synthèse s'élève à environ USD 0.07 par paire de bases nucléiques, tandis que le séquençage coûte environ USD 500 pour un génome complet, soit environ trois milliards de paires de bases (génome humain), équivalant à $\text{USD } 1.67 \times 10^{-7}$ par base.

Ensuite, comme cité précédemment, le temps requis pour la synthèse ou le séquençage des données reste encore un défi majeur, car il est relativement long comparé aux méthodes de stockage traditionnelles. Par exemple, pour mon expérience pratique, une fois l'échantillon synthétisé, il a fallu le faire parvenir aux différents centres de séquençage pour que celui-ci puisse être séquencé. Ensuite, une fois les résultats de séquençage obtenus, il faut analyser l'information et déterminer la séquence ADN afin de la décoder. Ce processus est relativement long⁴⁷ et ne répond pas aux normes habituelles d'accessibilité et de disponibilité de nos données. La difficulté d'accès rapide aux données empêche l'ADN de devenir un support de stockage dynamique à l'instar des disques durs conventionnels. De plus, l'ADN, une fois synthétisé et stocké, doit être détruit pour être lu. Même lorsqu'une petite information est nécessaire, il est tout de même indispensable de séquencer l'intégralité de l'échantillon. Ainsi, il est donc essentiel de développer des méthodes de stockage d'ADN permettant un accès facilité aux données.

Cependant, pour le stockage de données à long terme ne nécessitant pas un accès fréquent, cette technologie présente un potentiel très prometteur. Supposons que dans

⁴⁷ 10 à 15 jours ouvrables pour la synthèse et réception de l'ADN. Ensuite, le séquençage Sanger a été le plus rapide, les résultats ont été obtenus en 2 jours ouvrables.

le cadre de mon expérience, l'ADN synthétisé représentait des données de sauvegarde, autrement dit un backup. Dans de grands nombres de cas, ceux-ci ne sont que très peu consultés et ne nécessitent que rarement une accessibilité constante. Ainsi, il aurait tout à fait été concevable de conserver une sauvegarde dans l'ADN, car cette méthode n'est pas du tout énergivore en comparaison avec des centres de stockage. Toutefois, encore en termes de coûts, cette technologie est bien loin de rivaliser avec les méthodes conventionnelles. Dans le cadre de mon expérience, en excluant la synthèse et le séquençage par nanopores qui n'ont malheureusement pas abouti, les coûts ont atteint environ CHF 160 pour stocker seulement 125 octets, soit 1.25×10^{-7} Go. Cette observation souligne la nécessité de progrès significatifs afin de rendre cette technologie viable et plus largement utilisable.

Finalement, un dernier point qui mérite d'être relevé est le fait que l'ADN constitue le pilier de l'information génétique, et ce depuis des millions d'années. Développer des moyens de stockage dans l'ADN est une idée révolutionnaire et prometteuse. À l'instar des supports de stockage classiques dont la pertinence peut rapidement devenir obsolète – par exemple, les lecteurs de cassettes sont presque introuvables de nos jours – la structure de l'ADN ne risque pas de changer au fil du temps. Ensuite, notre besoin de décrypter l'ADN constitue également un moyen de rendre cette technologie intemporelle, car il est difficile d'imaginer un monde où les technologies de séquençage deviendraient obsolètes ou inutiles. En effet, l'étude de l'ADN est d'une importance cruciale dans grand nombre de domaines. Ainsi, le développement de techniques de stockage basées sur celui-ci ouvre la voie à une révolution du stockage de données, avec une promesse inédite de fiabilité et durabilité.

Bibliographie

- [1] De la carte perforée à la mémoire flash : la grande histoire du stockage, [online]. Retrieved from : <https://www.tomshardware.fr/de-la-carte-perforee-a-la-memoire-flash-la-grande-histoire-du-stockage-des-donnees/> [accessed 31 October 2023].
- [2] La folle évolution du stockage informatique, 2014 *Capital.fr* [online]. Retrieved from : <https://www.capital.fr/economie-politique/la-folle-evolution-du-stockage-informatique-953110> [accessed 31 October 2023].
- [3] L-ÉPROUVETTE, 2017. *History of Digital Storage - Part 1* [online]. 1 December 2017. Retrieved from : https://www.youtube.com/watch?v=5Oz_iSRw5XM [accessed 31 October 2023].
- [4] L-ÉPROUVETTE, 2018. *L'Histoire du Stockage Numérique - 2ème Partie* [online]. 1 February 2018. Retrieved from : <https://www.youtube.com/watch?v=oi4vM-WLtqY> [accessed 31 October 2023].
- [5] LUNIL, 2020. Mémoire à tambour magnétique. *LUNIL - L'innovation dans le monde* [online]. 29 November 2020. Retrieved from : <https://www.lunil.com/invention-memoire-tambour-magnetique-atanasoff/> [accessed 31 October 2023].
- [6] @NATGEOFRANCE, 2023. Combien y a-t-il de cellules dans le corps humain ? National Geographic [online]. 31 October 2023. Retrieved from : <https://www.nationalgeographic.fr/sciences/2023/10/combien-y-a-t-il-de-cellules-dans-le-corps-humain> [accessed 1 November 2023].
- [7] Des universités suisses vont faire sortir le stockage ADN des laboratoires, 2023 [online]. Retrieved from : <https://www.ictjournal.ch/news/2023-10-20/des-universites-suisses-vont-faire-sortir-le-stockage-adn-des-laboratoires> [accessed 1 November 2023].
- [8] Le plus vieil ADN, découvert au Groenland, a 2 millions d'années, 2022 rts.ch [online]. Retrieved from : <https://www.rts.ch/info/sciences-tech/13612361-le-plus-vieil-adn-decouvert-au-groenland-a-2-millions-dannees.html> [accessed 7 November 2023].
- [9] Data growth worldwide 2010-2025, Statista [online]. Retrieved from : <https://www.statista.com/statistics/871513/worldwide-data-created/> [accessed 7 November 2023].
- [10] JORDAN, Bertrand, 2023. Séquençage d'ADN, la fin d'un quasi-monopole ? - Chroniques génomiques. médecine/sciences. Vol. 39, no. 5, pp. 469–473. DOI 10.1051/medsci/2023061.
- [11] Séquençage de l'ADN, 2023 Wikipédia [online]. Retrieved from : https://fr.wikipedia.org/w/index.php?title=S%C3%A9quen%C3%A7age_de_l'ADN [accessed 7 November 2023]. Page Version ID: 208799477
- [12] Nanopore store: MinION, [online]. Retrieved from : <https://store.nanoporetech.com/minion.html> [accessed 16 December 2023].
- [13] La Suisse est le deuxième pays d'Europe le plus dense en datacenters, 2021 [online]. Retrieved from : <https://www.ictjournal.ch/etudes/2021-02-11/la-suisse-est-le-deuxieme-pays-deurope-le-plus-dense-en-datacenters> [accessed 18 December 2023].
- [14] Allen School researchers expose cybersecurity risks of DNA sequencing software, Allen School News [online]. Retrieved from : <https://news.cs.washington.edu/2017/08/10/allen-school-researchers-expose-cybersecurity-risks-of-dna-sequencing-software/> [accessed 19 December 2023].
- [15] GOLDMAN, Dr Nick. Enregistrer des données sur l'ADN. Pictet Asset Management [online]. Retrieved from : <https://am.pictet.fr/france/mega/informations-adn> [accessed 10 November 2023].

- [16] Current Members , [online]. Retrieved from : <https://dnastoragealliance.org/members/current-members/> [accessed 27 November 2023]
- [17] DNA Data Storage | SNIA, [online]. Retrieved from : <https://www.snia.org/groups/snia-dna-technology-affiliate> [accessed 27 November 2023].
- [18] THE WHY, 2022 [online]. Retrieved from : <https://dnastoragealliance.org/why/> [accessed 27 November 2023].
- [19] ELWORTH, R. A. Leo et al., 2020. Synthetic DNA and biosecurity: Nuances of predicting pathogenicity and the impetus for novel computational approaches for screening oligonucleotides. PLoS Pathogens. Vol. 16, no. 8, p. e1008649. DOI 10.1371/journal.ppat.1008649.
- [20] U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES, Administration for Strategic Preparedness and Response, 2023. Screening Framework Guidance for Providers and Users of Synthetic Nucleic Acids. p. 11.
- [21] Code d'effacement, 2023Wikipédia [online]. Retrieved from : https://fr.wikipedia.org/w/index.php?title=Code_d%27effacement&oldid=207350051 [accessed 3 December 2023]. Page Version ID: 207350051
- [22] La consommation d'électricité des centres de calcul en Suisse continue d'augmenter, [online]. Retrieved from : <https://www.bfe.admin.ch/bfe/fr/home/actualites-et-medias/communiqués-de-presse/mm-test.msg-id-83072.html> [accessed 4 December 2023].
- [23] L'enjeu énergétique autour des datacenters, 2022rts.ch [online]. Retrieved from : <https://www.rts.ch/info/suisse/13549788-lenjeu-energetique-autour-des-datacenters.html> [accessed 4 December 2023]. Last Modified: 2022-11-21T09:23:07Z
- [24] BFE Publikationen, [online]. Retrieved from : <https://pubdb.bfe.admin.ch/de/suche?keywords=400#collapseForm> [accessed 4 December 2023].
- [25] Global data center storage capacity used 2015-2021, Statista [online]. Retrieved from : <https://www.statista.com/statistics/638613/worldwide-data-center-storage-used-capacity/> [accessed 5 December 2023].
- [26] MICROSOFT, 2023. 2022 Environmental Sustainability Report. Global sustainability. p. 81. Retrieved from : <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW15mgm> [accessed 5 December 2023]
- [27] GOOGLE, 2022. Google Data Centers 2021 Annual Water Metrics. . p. 1. Retrieved from : <https://www.gstatic.com/gumdrop/sustainability/2022-us-data-center-water.pdf> [accessed 5 December 2023]
- [28] E-Déchets : Data Centers, extraction des terres rares et traitement des déchets numériques tuent notre planète., 2020Buzzles [online]. Retrieved from : <https://buzzles.org/2020/04/18/e-dechets-data-centers-extraction-des-terres-rares-et-traitement-des-dechets-numerique-tuent-notre-planete/> [accessed 6 December 2023].
- [29] Biomemory relance le stockage sur ADN via des cellules vivantes, LeMagIT [online]. Retrieved from : <https://www.lemagit.fr/actualites/252524795/Biomemory-relance-le-stockage-sur-ADN-via-des-cellules-vivantes> [accessed 6 December 2023].
- [30] Séquence consensus, 2016Wikipédia [online]. Retrieved from : https://fr.wikipedia.org/w/index.php?title=S%C3%A9quence_consensus&oldid=131921182 [accessed 16 December 2023]. Page Version ID: 131921182

- [31] SEAGATE, 2020. RETHINK DATA, Exploitez davantage vos données d'entreprise de la périphérie jusqu'au cloud. p.56. Retrieved from : https://www.seagate.com/files/www-content/our-story/rethink-data/files/Rethink_Data_Report_2020_fr_FR.pdf [accessed 18 December 2023]
- [32] World wide web inventor admits forward slashes "a mistake," 2009 The Economic Times [online]. Retrieved from : <https://economictimes.indiatimes.com/tech/internet/world-wide-web-inventor-admits-forward-slashes-a-mistake/articleshow/5123863.cms?from=> [accessed 13 January 2024].
- [33] Deoxyadenosine triphosphate, 2023Wikipedia [online]. Retrieved from : https://en.wikipedia.org/w/index.php?title=Deoxyadenosine_triphosphate&oldid=1173820018 [accessed 21 January 2024]. Page Version ID: 1173820018
- [34] Dideoxynucleotide, 2024Wikipedia [online]. Retrieved from : <https://en.wikipedia.org/w/index.php?title=Dideoxynucleotide&oldid=1195932881> [accessed 21 January 2024]. Page Version ID: 1195932881
- [35] ILLUMINA, 2016. Overview of Illumina Sequencing by Synthesis Workflow [online]. 5 October 2016. Retrieved from : <https://www.youtube.com/watch?v=fCd6B5HRaZ8> [accessed 21 January 2024].
- [36] Nanopore sequencing, 2023Wikipedia [online]. Retrieved from : https://en.wikipedia.org/w/index.php?title=Nanopore_sequencing&oldid=1189334284 [accessed 21 January 2024]. Page Version ID: 1189334284
- [37] Sanger-Sequenzierung – Schritte & Methode, [online]. Retrieved from : <https://www.sigmaaldrich.com/CH/de/technical-documents/protocol/genomics/sequencing/sanger-sequencing> [accessed 21 January 2024].
- [38] Séquençage de Sanger | Biochimie Facile - YouTube, [online]. Retrieved from : https://www.youtube.com/watch?v=0Am85rajth4&ab_channel=BiochimieFacile [accessed 21 January 2024].
- [39] Too much information - Aston University researchers to tackle global data storage crisis, [online]. Retrieved from : <https://www.aston.ac.uk/latest-news/too-much-information-aston-university-researchers-tackle-global-data-storage-crisis> [accessed 11 February 2024].
- [40] Claude Shannon, 2024Wikipédia [online]. Retrieved from : https://fr.wikipedia.org/w/index.php?title=Claude_Shannon&oldid=211925341 [accessed 24 February 2024]. Page Version ID: 211925341
- [41] SWISSINFO.CH, S. W. I., 2023. Data centres account for 4% of Swiss electricity usage. SWI swissinfo.ch [online]. 8 January 2023. Retrieved from : <https://www.swissinfo.ch/eng/sci-&-tech/data-centres-account-for-4-of-swiss-electricity-usage/48188702> [accessed 24 February 2024].
- [42] GARCIA, Clarissa, 2023. Data Center Energy Use - AKCP Monitoring. AKCP Remote Sensor Monitoring | Data Center Monitoring [online]. 17 July 2023. Retrieved from : <https://www.akcp.com/blog/the-real-amount-of-energy-a-data-center-use/> [accessed 24 February 2024].
- [43] Génétique humaine, 2023Wikipédia [online]. Retrieved from : https://fr.wikipedia.org/w/index.php?title=G%C3%A9n%C3%A9tique_humaine&oldid=209887554 [accessed 24 February 2024]. Page Version ID: 209887554

- [44] Stockage de données numériques sur ADN, 2023Wikipédia [online]. Retrieved from : https://fr.wikipedia.org/w/index.php?title=Stockage_de_donn%C3%A9es_num%C3%A9riques_sur_ADN&oldid=209326129 [accessed 24 February 2024]. Page Version ID: 209326129
- [45] Phosphodiester bond, 2024Wikipedia [online]. Retrieved from : https://en.wikipedia.org/w/index.php?title=Phosphodiester_bond&oldid=1195356613 [accessed 24 February 2024]. Page Version ID: 1195356613
- [46] Climate and weather related disasters surge five-fold over 50 years, but early warnings save lives - WMO report | UN News, 2021 [online]. Retrieved from : <https://news.un.org/en/story/2021/09/1098662> [accessed 24 February 2024].
- [47] Séquençage par nanopores, 2023Wikipédia [online]. Retrieved from : https://fr.wikipedia.org/w/index.php?title=S%C3%A9quen%C3%A7age_par_nanopores&oldid=205041989 [accessed 25 February 2024]. Page Version ID: 205041989

Annexe 1 : Java Class Conversion⁴⁸

```
package utils;
import org.apache.log4j.Logger;

import java.io.FileInputStream;
import java.io.FileOutputStream;
import java.io.IOException;
import java.util.*;

/**
 * Cette classe permet de convertir des données String en données
 * ADN et vice-versa en passant par le format binaire.
 * String --> Binaire --> ADN
 * ADN --> Binaire --> String
 * Cette méthode ne gère que très peu les redondances (espaces,
 * répétitions de caractères) mais elle sert tout de même de
 * proof of concept.
 *
 * Elle permet également de convertir des fichiers en données
 * binaires puis de les convertir à leur tour en données ADN.
 * Fichier --> Binaire --> ADN
 * ADN --> Binaire --> Fichier
 * Pour ce traitement, il n'y pas de gestion de redondance de
 * séquences dans le binaire ou la séquence ADN.
 * Ce qui peut rendre un synthèse beaucoup plus complexe,
 * mais il était relativement complexe à implémenter.
 */

public class Conversion {
    static Logger logger = Logger.getLogger(Conversion.class);
    private Map<String, String> chiffrement;
    private ArrayList<String> remplacement;
    private String DEBUT = "DEBUT";
    private String FIN = "FIN";

    public Conversion() {
        /**
         Initialisation la clé de chiffrement
         "A" <--> "00"
         "G" <--> "01"
         "C" <--> "10"
         "T" <--> "11"
         */
        chiffrement = new HashMap<String, String>();
        chiffrement.put("A", "00");
        chiffrement.put("G", "01");
        chiffrement.put("C", "10");
        chiffrement.put("T", "11");
        chiffrement.put("00", "A");
        chiffrement.put("01", "G");
        chiffrement.put("10", "C");
        chiffrement.put("11", "T");
        remplacement = new ArrayList<String>(Arrays.asList(" ", "#", "&", "_",
            "+", "^", "$", "="));
    }
}
```

⁴⁸ L'ensemble du programme se trouve à https://github.com/Komirstone/conversion_DNA.
La classe Conversion contient l'ensemble des méthodes utilisées pour convertir du texte ou des fichiers en séquence ADN.

```

public String formaterString(String message){
    /**
     * Il est important de formater le message pour qu'il soit :
     * - encapsulé par DEBUT et FIN
     * - éviter les répétitions de caractères :
     *   (en mettant en majuscule le caractère répété ou en
     *   remplaçant les espaces répétitifs par un caractère spécial)
     */
    logger.info("FormaterString - Debut du formatage du message.");
    StringBuilder sb = new StringBuilder();
    Map<Character, Integer> caracteresCompte = new HashMap<Character,
    Integer>();

    for (int i = 0; i < message.length(); i++) {
        char c = message.charAt(i);
        if (!caracteresCompte.containsKey(c)) {
            caracteresCompte.put(c, 1);
            sb.append(c);
        } else if (c == ' ') { /* Dès le deuxième ' ' */
            int random = (int) (Math.random()*(remplacement.size()-1));
            String nouveau_c = remplacement.get(random);
            caracteresCompte.put(c, caracteresCompte.get(c) + 1);
            sb.append(nouveau_c);
        } else if (caracteresCompte.get(c) % 2 != 0) {
            /* éviter la répétition de caractères */
            sb.append(Character.toUpperCase(c));
            caracteresCompte.put(c, caracteresCompte.get(c) + 1);
        } else {
            sb.append(c);
            caracteresCompte.put(c, caracteresCompte.get(c) + 1);
        }
    }
    /* ENCAPSULATION afin de définir le debut et la fin du message*/
    String phrase_encapsulee = DEBUT + sb.toString() + FIN;
    logger.info("FormaterString - Message formate : " +
    phrase_encapsulee);
    return phrase_encapsulee;
}

public static String stringToBinaire(String messageFormate){
    /**
     * Renvoie le message formaté en binaire
     */
    logger.info("StringToBinaire - Conversion... formate en "
    + "binaire.");
    StringBuilder binaire = new StringBuilder();
    for (char ch : messageFormate.toCharArray()) {
        String binStr = Integer.toBinaryString(ch);
        //S'assurer que le message fait bien 8 bits de long, en
        rajoutant des 0 au début
        while (binStr.length() < 8) {
            binStr = "0" + binStr;
        }
        binaire.append(binStr);
    }
    logger.info("StringToBinaire - Conversion reussie, sortie binaire"
    + " : " + binaire.toString());
    return binaire.toString();
}

```

```

public String binaireToDNA(String messageBinaire) {
    /**
     * Renvoie le message binaire en ADN, en reprenant la clé de
     * chiffrement définie dans le constructeur
     * "00" --> "A"
     * "01" --> "G"
     * "10" --> "C"
     * "11" --> "T"
     */
    logger.info("BinaireToDNA - Conversion du binaire en sequence "
+ "ADN.");
    StringBuilder adn = new StringBuilder();
    //IMPORTANT : nous devons lire tous les 2 bits !
    for (int i = 0; i < messageBinaire.length(); i += 2) {
        String binStr = messageBinaire.substring(i, i + 2);
        adn.append(chiffrement.get(binStr));
    }
    logger.info("BinaireToDNA - Conversion reussie, sequence ADN : " +
adn.toString());
    return adn.toString();
}

public String dnaToBinaire(String messageADN) {
    /**
     * Renvoie le message ADN en binaire, en reprenant la clé de
     * chiffrement définie dans le constructeur
     * "A" --> "00"
     * "G" --> "01"
     * "C" --> "10"
     * "T" --> "11"
     */
    logger.info("DnaToBinaire - Conversion de la sequence ADN en "
+ "binaire.");
    StringBuilder binaire = new StringBuilder();
    for (int i = 0; i < messageADN.toUpperCase().length(); i++) {
        String binStr = chiffrement.get(messageADN.substring(i,i+1));
        binaire.append(binStr);
    }
    logger.info("DnaToBinaire - Conversion reussie, sortie binaire : "
+ binaire.toString());
    return binaire.toString();
}

public static String binaireToString(String messageBinaire) {
    /**
     * Renvoie le message binaire en String
     */
    logger.info("BinaireToString - Conversion du binaire en String.");
    StringBuilder sb = new StringBuilder();
    // IMPORTANT : nous devons lire tous les 8 bits !!!
    for (int i = 0; i < messageBinaire.length(); i += 8) {
        String binStr = messageBinaire.substring(i, i + 8);
        char c = (char) Integer.parseInt(binStr, 2);
        sb.append(c);
    }
    logger.info("BinaireToString - Conversion reussie : " +
sb.toString());
    return sb.toString();
}

```

```

public String reformaterString(String messageFormate) {
    /**
     * Cette méthode permet de reformater le message en enlevant les
     * caractères spéciaux et les majuscules ajoutées
     */
    logger.info("ReformaterString - Debut de reformatage du "
+ "message.");
    StringBuilder sb = new StringBuilder();
    for (int i = DEBUT.length(); i < messageFormate.length() -
FIN.length(); i++) {
        char c = messageFormate.charAt(i);
        if (remplacement.contains(Character.toString(c))) {
            sb.append(" ");
        } else {
            sb.append(c);
        }
    }
    logger.info("ReformaterString - Message reformate : " +
sb.toString().toLowerCase(Locale.ROOT));
    return ajouterMajuscules(sb.toString().toLowerCase(Locale.ROOT));
}

private static String ajouterMajuscules(String phrase) {
    logger.info("AjouterMajuscules - Debut de la fonction avec la "
+ "phrase : " + phrase);
    if (phrase == null || phrase.isEmpty()) {
        logger.warn("AjouterMajuscules - La phrase est nulle ou "
+ "vide.");
        return phrase;
    }

    StringBuilder phraseModifiee = new StringBuilder(phrase.length());
    boolean majusculeNecessaire = true; //Indique s'il faut mettre
une majuscule
    for (int i = 0; i < phrase.length(); i++) {
        char c = phrase.charAt(i);
        if (majusculeNecessaire && Character.isLetter(c)) {
            phraseModifiee.append(Character.toUpperCase(c));
            majusculeNecessaire = false; //Désactive la majuscule pour
les caractères suivants
            logger.debug("AjouterMajuscules - Majuscule ajoutée au "
+ "caractere : " + c);
        } else {
            phraseModifiee.append(c);
        }

        // Réactive la majuscule après un point suivi d'un espace
        if ((c == '.' || c == '?' || c == '!') && i < phrase.length()
- 1 && phrase.charAt(i + 1) == ' ') {
            majusculeNecessaire = true;
            logger.debug("AjouterMajuscules - \". \", \"? \" \"! \", \"
+ \"majuscule necessaire pour le prochain caractere.");
        }
    }
    logger.info("AjouterMajuscules - Fin de la fonction, phrase "
+ "modifiee : " + phraseModifiee);
    return phraseModifiee.toString();
}

```

```

public String encoderMessageEnADN(String message) {
    /**
     * Cette méthode permet d'encoder un message
     */
    logger.info("EncoderMessageEnADN - Demarrage du processus "
    + "d'encodage du message en ADN.");
    String messageFormate = formaterString(message);
    String messageBinaire = stringToBinaire(messageFormate);
    String sequenceADN = binaireToDNA(messageBinaire);
    logger.info("EncoderMessageEnADN - Fin du processus d'encodage, "
    + "sequence ADN obtenue : " + sequenceADN);
    return sequenceADN;
}

public String decoderADNEnMessage(String sequenceADN) {
    /**
     * Cette méthode permet de décoder un message
     */
    logger.info("DecoderADNEnMessage - Demarrage du processus de "
    + "decodage de la sequence ADN en message.");
    String messageBinaire = dnaToBinaire(sequenceADN);
    String messageFormate = binaireToString(messageBinaire);
    String message = reformaterString(messageFormate);
    logger.info("DecoderADNEnMessage - Fin du processus de decodage, "
    + "message obtenu : " + message);
    return message;
}

/**
 * CETTE CLASSE PERMET EGALEMENT DE TRANSFORMER DES FICHIERS EN
 * BINAIRE ET VICE-VERSA
 */
//Ce marqueur sert à délimiter le binaire relatif au fichier et celui
relatif à l'extension du fichier (.txt, .jpg, etc.)
public static final String MARQUEUR_UNIQUE =
"00000000000000010101111111111111";

public static String getExtensionFichier(String cheminFichier) {
    /** Cette méthode permet de récupérer l'extension d'un fichier */
    logger.info("GetExtensionFichier - Recuperation de l'extension du "
    + " fichier a partir du chemin : " + cheminFichier);
    int indexPoint = cheminFichier.lastIndexOf('.'); //Le dernier
point du filepath correspond à l'extension du fichier
    if (indexPoint >= 0 && indexPoint < cheminFichier.length() - 1) {
        logger.info("GetExtensionFichier - Extension du fichier "
        + "recuperee : " + cheminFichier.substring(indexPoint));
        return cheminFichier.substring(indexPoint); //Retourne ".txt",
        ".jpg", etc.
    }
    return "";
}

```

```

public static String lireFichierEnBinaire(String cheminFichier) {
    /**
     * Cette méthode permet de lire un fichier en binaire
     * Elle retourne un String contenant le binaire du fichier
     * Le binaire est composé de 3 parties :
     * 1) Le binaire de l'extension du fichier
     * 2) Un marqueur unique
     * 3) Le binaire du fichier
     */
    logger.info("LireFichierEnBinaire - Lecture du fichier en "
+ "binaire a partir du chemin : " + cheminFichier);
    String extensionFichier = getExtensionFichier(cheminFichier);
    String binaireExtension = stringToBinaire(extensionFichier);

    StringBuilder binaire = new StringBuilder();
    /**Ouverture du fichier en mode read*/
    try (FileInputStream fis = new FileInputStream(cheminFichier)) {
        int byteData;
        //Lecture du fichier byte par byte
        //System.out.println("lecture du binaire byte par byte...");
        logger.info("lireFichierEnBinaire - lecture du binaire byte "
+ "par byte...");
        while ((byteData = fis.read()) != -1) {
            String binaireString = Integer.toBinaryString(byteData);
            //Ajouter des 0 devant la chaine binaire pour être certain
            qu'elle fasse toujours 8 bits
            while (binaireString.length() < 8) {
                binaireString = "0" + binaireString;
            }
            //Ajouter le byte lu au "binaire"
            binaire.append(binaireString);
        }
    } catch (IOException e) {
        logger.error("erreur lors de la lecture du fichier : " +
cheminFichier);
        e.printStackTrace();
    }
    /**Recomposition de la séquence binaire : binaire + MARQUEUR +
extension */
    logger.info("LireFichierEnBinaire - Lecture en binaire terminée, "
+ "donnees binaires : " + binaire.toString());
    return binaire.toString() + MARQUEUR_UNIQUE + binaireExtension;
}

```



```

public static String ecrireBinaireDansFichier(String binaireData,
String cheminOutput) {
    /**
     * Transforme le binaire en fichier en conservant l'extension.
     * Identifie le marqueur pour séparer les données de l'extension.
     * Les données binaires sont ensuite écrites dans un fichier de
     * sortie, byte par byte.
     */
    //Déterminer la position de l'extension dans le binaire
    logger.info("EcrireBinaireDansFichier - Ecriture des donnees"
+ " binaires dans le fichier au chemin : " + cheminOutput);
    int marqueurIndex = binaireData.indexOf(MARQUEUR_UNIQUE);
    if (marqueurIndex == -1) {
        logger.error("EcrireBinaireDansFichier - Le MARQUEUR n'a pas"
+ " ete trouve dans le binaire. Impossible de continuer.");
        return null;
    }
    //Capturer l'extension (en binaire)
    String binaireExtension = binaireData.substring(marqueurIndex +
MARQUEUR_UNIQUE.length());
    //Convertir l'extension binaire en String
    String extensionFichier = binaireToString(binaireExtension);

    //Capturer le binaire du fichier (sans marqueur, sans extension)
    String binaireFichier = binaireData.substring(0, marqueurIndex);
    //On peut décider du nom du fichier de sortie, ici "output"
    String cheminOutputFichier = cheminOutput + "output" +
extensionFichier;

    //Création d'un nouvel objet FileOutputStream pour écrire le
    fichier de sortie cheminOutputFichier
    try (FileOutputStream fos = new
FileOutputStream(cheminOutputFichier)) {
        /**Il faut lire tous les 8 bits !**/
        logger.info("EcrireBinaireDansFichier - Recomposition du "
+"fichier a partir du binaire, en vue de l'écriture dans : " +
cheminOutputFichier);

        for (int i = 0; i < binaireFichier.length(); i += 8) {
            String byteString = binaireFichier.substring(i, i + 8);
            //Convertir le groupe de 8 bits en int
            int valeur = Integer.parseInt(byteString, 2);
            //Écrire la valeur entière dans le fichier
            fos.write(valeur);
        }
    } catch (IOException e) {
        logger.error("EcrireBinaireDansFichier - Erreur lors de"
+ " l'écriture du fichier dans : " + cheminOutputFichier);
        e.printStackTrace();
    }
    logger.info("EcrireBinaireDansFichier - Ecriture terminee, chemin"
+ " du fichier ecrit : " + cheminOutputFichier);
    return cheminOutputFichier;
}

```

```

public String encoderFichierEnADN(String cheminFichier){
    /**
     * Cette méthode permet d'encoder un fichier
     */
    logger.info("EncoderFichierEnADN - Encodage du fichier en"
    + " sequence ADN a partir du chemin : " + cheminFichier);
    String binaireFichier = lireFichierEnBinaire(cheminFichier);
    String sequenceADN = binaireToDNA(binaireFichier);
    logger.info("EncoderFichierEnADN - Encodage termine, sequence ADN"
    + " : " + sequenceADN);
    return sequenceADN;
}

public String decoderADNEnFichier(String sequenceADN, String
cheminOutput) {
    /**
     * Cette méthode permet de décrypter un fichier
     */
    logger.info("DecoderADNEnFichier - Decodage de la sequence ADN en"
    + " fichier, ecriture au chemin : " + cheminOutput);
    String binaireFichier = dnaToBinaire(sequenceADN);
    String cheminNouveauFichier = ecrireBinaireDansFichier(
    binaireFichier, cheminOutput);
    logger.info("DecoderADNEnFichier - Decodage termine, chemin du"
    + " fichier decode : " + cheminNouveauFichier);
    return cheminNouveauFichier;
}
}

```

Annexe 2 : Java Class SequencingResults⁴⁹

```
package utils;
import java.util.ArrayList;

public class SequencingResults {

    public static String sequenceComplementaire(String seqADN){
        //On parcourt la séquence ADN de la fin vers le début vu que la
        //lecture s'est faite dans le sens inverse
        String seqADNComplementaire = "";
        for (int i = seqADN.length() - 1; i >= 0; i--) {
            if (seqADN.charAt(i) == 'A') {
                seqADNComplementaire += 'T';
            } else if (seqADN.charAt(i) == 'T') {
                seqADNComplementaire += 'A';
            } else if (seqADN.charAt(i) == 'C') {
                seqADNComplementaire += 'G';
            } else if (seqADN.charAt(i) == 'G') {
                seqADNComplementaire += 'C';
            } else {
                //On laisse l'élément non reconnu tel que nucléotide
                seqADNComplementaire += seqADN.charAt(i);
            }
        }
        return seqADNComplementaire;
    }

    public static String validerSequence(String seq){
        /**vérifier que notre séquence ADN ne contient que des nucléotides
        A T C G*/
        ArrayList nucleotidesFaux = new ArrayList();
        boolean estValide = true;
        for (int i = 0; i < seq.length(); i++) {
            if (seq.charAt(i) != 'A' && seq.charAt(i) != 'T' &&
                seq.charAt(i) != 'C' && seq.charAt(i) != 'G') {
                estValide = false;
                nucleotidesFaux.add(i+1);
            }
        }
        if (!estValide) {
            return "La sequence ADN n'est pas valide, elle contient des "
                + "nucleotides non reconnus aux positions suivantes : " +
                nucleotidesFaux;
        }
        return "La sequence ADN est valide";
    }

    public static void main(String[] args) {
        String SEQUENCEREERENCE =
            "GAGAGAGGGAACGGGGGGGAGGGAGCCGGCTGACAAGAACGCGGGTACGCTCGAGGGGACGTATAC
            TGGATAGCGGGAGGACAAGGATACGTGCGGGTCAGCATGTGGGTATGAGGACGCGGAACGTGCAGGTGC
            GCTTGACGGTACACAAGAATGCCAGATTGCCGGGATGACGACGCACACGACAGTGAGGGAGTAAATCCAC
            TTACTTTACACACTCGGTTGATAGCGGGTATACATGCGCGCTTGGACGGATGCTAGAAGGTATGCCAGGAT
            ACGAGAGGGTGAGCAGGCCGCGGGATCGGGAACGCGAGAGAGGACGCGTGAGTACGATTGGAAACTAGG
            TTGATGGAAGGACGGTATACCTGCCGATAACGAGTAAGCGGGCTCGGATGCAGGACGGGGAGGTGTAG
            GGGGGAGGGGTTGCATACGTGCGGGTGAGAAGGCCGGGGAACCTGTGGGATCGAGGACGCGCACGCTTGC
            TCGATCGCGGACATGACGGCGAGAGGGCGGACTCGAGCGACGGATC";
    }
}
```

⁴⁹ La classe SequencingResults est celle qui m'a permis d'analyser mes résultats.

```

String adapter5 = "GGCATTTTGCTGCCGGTCACG";

String seqForward =
"NNNNNNNNNNNNNNNNNNNNCGGCTGANAGAACGCGGGTACGCTCGAGGGGACGTATACTGGATAG
CGGGAGGNCAAGGATACGTGCGGGTCAGCATGTGGGTATGAGGACGCGCAACGTGCAGGTGCGCTTGAC
GGTACACAAGAATGCCAGATTGCCGGGATGACGACGCACACGACAGTGAGGGAGTAAATCCACTTACTTA
CACACTCGGTTGATAGCGGGTATACATGCGCGCTTGGACGGATGCTAGAAGGTATGCCAGGATACGAGAG
GGTGAGCAGGCCGCGGGATCGGGAACGCGAGAGAGGACGCGTGAGTACGATTGGAAGTATAGGTTGATGG
AAGGACGGTATACCTGCCGGATAACGAGTAAGCGGGCTCGGATGCAGGACGGGGAGGTTGTAGGGGGGAG
GGGTTGCATACGTGCGGGTGAGAAGGCCGGGGAACCTGTGGGATCGAGGACGCGCACGCTTGCTCGATCG
CGGACATGACGCGAGAGGGCGGACTCGAGCGACGGATCN";

//Dans seqReverse la qualité de la dernière base est mauvaise, on
//la supprime
String seqReverse =
"NNNNNNNNNNNNNNNNNNNANNGNNNAAGCGTGCGCGTCCTCGATCCCACAGGTTCCCCGGCCTTCT
CACCCGCACGTATGCAACCCCTCCCCCTACAACCTCCCCGTCCTGCATCCGAGCCCGCTTACTCGTTAT
CCGGCAGGTATACCGTCCTTCCATCAACCTAGTTTCCAATCGTACTCACGCGTCCTCTCTCGCGTTCCCG
ATCCCGCCGCGCTGCTCACCTCTCGTATCCTGGCATACTTCTAGCATCCGTCCAAGCGCGCATGTATA
CCCGCTATCAACCGAGTGTGTAAGTAAGTGGATTTACTCCCTCACTGTGCGTGTGCGTCGTCATCCCGGCA
ATCTGGCATTCTTGTGTACCGTCAAGCGCACCTGCACGTTTCGCGCGTCCTCATACCCACATGCTGACCCG
CACGTATCCTTGTCTCCCGCTATCCAGTATACGTCCCTCGAGCGTACCCGCGTTCTTGTGACCCGGCT
CCCTCCCCCCCCGTTCCCTCTCTCCGTGACCGGCAGCAAATGCCA";

/**Trouver la séquence complémentaire à la lecture reverse*/
seqReverse = seqReverse.substring(0, seqReverse.length() - 1);
String seqComplementaire = sequenceComplementaire(seqReverse);

/** Ne garder que d'environ la [moitié, fin] pour les 2 lectures
 * Le début de lecture contient souvent quelques erreurs
 * Cela permet de réduire les erreurs*/
String seqPart2 = seqForward.substring(seqForward.length()/2,
seqForward.length());
String debutSeqPart2 = seqPart2.substring(0, seqPart2.length()/10);
//Ne conserver que la partie avant le début de la partie 2
String seqPart1 = seqComplementaire.substring(0,
seqComplementaire.indexOf(debutSeqPart2));

/**Assembler les deux parties et enlever l'adaptateur et ce qui
peut précéder*/
String seqFinale = seqPart1 + seqPart2;
seqFinale = seqFinale.substring(seqFinale.indexOf(adapter5) +
adapter5.length());

/**Afficher la séquence finale*/
System.out.println("\nValidation de sequence : ");
System.out.println(validerSequence(seqFinale));
//On supprime le dernier nucléotide car il est en fin de séquence
est il est non-défini (N)
seqFinale = seqFinale.substring(0, seqFinale.length() - 1);
//On affiche la séquence finale
System.out.println("\nLa sequence finale : ");
System.out.println(seqFinale + "\n");

/**Vérifier si la séquence ADN contient bien notre message une
fois décodée*/
Conversion conversion = new Conversion();
String message = conversion.decoderADNEnMessage(seqFinale);
System.out.println("\nMessage encode dans la sequence ADN : ");
System.out.println(message);
}
}

```

Annexe 3 : Résultats du séquençage Sanger⁵⁰

	10 20 30 40 50
Seq_refere	~GGCATTTTG CTGCCGGTCA CGGAGAGAGG GAACGGGGGG GAGGGAGCCG
Seq_2749	~~~~~ ~~~~~ ~~~~~~N NNNNNNNNNN NNGNNNNNNC
Seq_2750rc	TGGCATTTTG CTGCCGGTCA CGGAGAGAGG GAACGGGGGG GAGGGAGCCG

	60 70 80 90 100
Seq_refere	GCTGACAAGA ACGCGGGTAC GCTCGAGGGG ACGTATACTG GATAGCGGGA
Seq_2749	GGCTGANAGA ACGCGGGTAC GCTCGAGGGG ACGTATACTG GATAGCGGGA
Seq_2750rc	GCTGACAAGA ACGCGGGTAC GCTCGAGGGG ACGTATACTG GATAGCGGGA

	110 120 130 140 150
Seq_refere	GGACAAGGAT ACGTGCGGGT CAGCATGTGG GTATGAGGAC GCGCGAACGT
Seq_2749	GGNCAAGGAT ACGTGCGGGT CAGCATGTGG GTATGAGGAC GCGCGAACGT
Seq_2750rc	GGACAAGGAT ACGTGCGGGT CAGCATGTGG GTATGAGGAC GCGCGAACGT

	160 170 180 190 200
Seq_refere	GCAGGTGCGC TTGACGGTAC ACAAGAATGC CAGATTGCCG GGATGACGAC
Seq_2749	GCAGGTGCGC TTGACGGTAC ACAAGAATGC CAGATTGCCG GGATGACGAC
Seq_2750rc	GCAGGTGCGC TTGACGGTAC ACAAGAATGC CAGATTGCCG GGATGACGAC

	210 220 230 240 250
Seq_refere	GCACACGACA GTGAGGGAGT AAATCCACTT ACTTACACAC TCGGTTGATA
Seq_2749	GCACACGACA GTGAGGGAGT AAATCCACTT ACTTACACAC TCGGTTGATA
Seq_2750rc	GCACACGACA GTGAGGGAGT AAATCCACTT ACTTACACAC TCGGTTGATA

	260 270 280 290 300
Seq_refere	GCGGGTATAC ATGCGCGCTT GGACGGATGC TAGAAGGTAT GCCAGGATAC
Seq_2749	GCGGGTATAC ATGCGCGCTT GGACGGATGC TAGAAGGTAT GCCAGGATAC
Seq_2750rc	GCGGGTATAC ATGCGCGCTT GGACGGATGC TAGAAGGTAT GCCAGGATAC

	310 320 330 340 350
Seq_refere	GAGAGGGTGA GCAGGCCGGC GGGATCGGGA ACGCGAGAGA GGACGCGTGA
Seq_2749	GAGAGGGTGA GCAGGCCGGC GGGATCGGGA ACGCGAGAGA GGACGCGTGA
Seq_2750rc	GAGAGGGTGA GCAGGCCGGC GGGATCGGGA ACGCGAGAGA GGACGCGTGA

	360 370 380 390 400
Seq_refere	GTACGATTGG AAAC TAGGTT GATGGAAGGA CCGTATACCT GCCGGATAAC
Seq_2749	GTACGATTGG AAAC TAGGTT GATGGAAGGA CCGTATACCT GCCGGATAAC
Seq_2750rc	GTACGATTGG AAAC TAGGTT GATGGAAGGA CCGTATACCT GCCGGATAAC

⁵⁰ Seq_refere est ma séquence de référence

Seq_2749 est la lecture de mon échantillon en forward (brin codant)

Seq_2750 est le brin complémentaire la lecture en reverse

	410 420 430 440 450
Seq_refere	GAGTAAGCGG GCTCGGATGC AGGACGGGGA GGTGTAGGG GGGAGGGGTT
Seq_2749	GAGTAAGCGG GCTCGGATGC AGGACGGGGA GGTGTAGGG GGGAGGGGTT
Seq_2750rc	GAGTAAGCGG GCTCGGATGC AGGACGGGGA GGTGTAGGG GGGAGGGGTT

	460 470 480 490 500
Seq_refere	GCATACGTGC GGGTGAGAAG GCCGGGGAAC CTGTGGGATC GAGGACGCGC
Seq_2749	GCATACGTGC GGGTGAGAAG GCCGGGGAAC CTGTGGGATC GAGGACGCGC
Seq_2750rc	GCATACGTGC GGGTGAGAAG GCCGGGGAAC CTGTGGGATC GAGGACGCGC

	510 520 530 540 550
Seq_refere	ACGCTTGCTC GATCGCGGAC ATGACGGCGA GAGGGCGGAC TCGAGCGACG
Seq_2749	ACGCTTGCTC GATCGCGGAC ATGACGGCGA GAGGGCGGAC TCGAGCGACG
Seq_2750rc	ACGCTTNNNC NNTNNNGNN NNNNNNNNNN N.....

	560
Seq_refere	GATC.....
Seq_2749	GATCN.....
Seq_2750rc