Université de Strasbourg

FACULTÉ DE PHARMACIE

N° d'ordre: 2345

MÉMOIRE DE DIPLÔME D'ÉTAT DE DOCTEUR EN PHARMACIE

–

Stockage de données numériques sur ADN

Présenté par Aymeric Deplace

Soutenu le 3 novembre 2023 devant le jury constitué de

Julien GODET, PU-PH, Directeur de thèse et Président du Jury

Emmanuel BOUTANT, MCU, et Nicolas RISS, Autres membres du jury

Approuvé par le Doyen et
par le Président de l'Université de Strasbourg

Doyen	Esther KELLENBERGER
Directeurs adjoints	Julien GODET
	Béatrice HEURTAULT
	Emilie SICK
Directeur adjoint étudiant	Léo FERREIRA-MOURIAUX

LISTE DU PERSONNEL ENSEIGNANT

Professeurs :

Philippe	BOUCHER	Physiologie
Nathalie	BOULANGER	Parasitologie
Line	BOUREL	Chimie thérapeutique
Pascal	DIDIER	Biophotonique
Saïd	ENNAHAR	Chimie analytique
Valérie	GEOFFROY	Microbiologie
Philippe	GEORGEL	Bactériologie, Virologie
Jean-Pierre	GIES	Pharmacologie moléculaire
Béatrice	HEURTAULT	Pharmacie galénique
Esther	KELLENBERGER	Bio-Informatique
Maxime	LEHMANN	Biologie cellulaire
Eric	MARCHIONI	Chimie analytique
Francis	MEGERLIN	Droit et économie pharm.
Yves	MELY	Physique et Biophysique
Jean-Yves	PABST	Droit Economie pharm.
Françoise	PONS	Toxicologie
Valérie	SCHINI-KERTH	Pharmacologie
Florence	TOTI	Pharmacologie
Thierry	VANDAMME	Biogalénique
Catherine	VONTHRON	Pharmacognosie
Pascal	WEHRLÉ	Pharmacie galénique

Professeurs-praticiens hospitaliers

Julien	GODET	Biostatistiques - science des données
Jean-Marc	LESSINGER	Biochimie
Bruno	MICHEL	Pharm. clinique santé publique
Pauline	SOULAS-SPRAUEL	Immunologie
Geneviève	UBEAUD-SÉQUIER	Pharmacocinétique

Enseignants contractuels

Alexandra	CHAMPERT	Pharmacie d'officine
Matthieu	FOHRER	Pharmacie d'officine
Philippe	GALAIS	Droit et économie pharm.
Philippe	NANDE	Ingéniérie pharmaceutique
Caroline	WILLER - WEHRLÉ	Pharmacie d'officine

Maîtres de Conférences :

Nicolas	ANTON	Pharmacie biogalénique
Fareeha	BATOOL	Biochimie
Martine	BERGAENTZLÉ	Chimie analytique
Elisa	BOMBARDA	Biophysique
Aurélie	BOURDERIOUX	Pharmacochimie
Emmanuel	BOUTANT	Virologie et Microbiologie
Véronique	BRUBAN	Physiologie et physiopath.
Anne	CASSET	Toxicologie
Thierry	CHATAIGNEAU	Pharmacologie
Manuela	CHIPER	Pharmacie biogalénique
Guillaume	CONZATTI	Pharmacie galénique
Marcella	DE GIORGI	Pharmacochimie
Serge	DUMONT	Biologie cellulaire
Gisèle	HAAN-ARCHIPOFF	Plantes médicinales
Célien	JACQUEMARD	Chémoinformatique
Julie	KARPENKO	Pharmacochimie
Sonia	LORDEL	Chimie analytique
Clarisse	MAECHLING	Chimie physique
Rachel	MATZ-WESTPHAL	Pharmacologie
Cherifa	MEHADJI	Chimie
Nathalie	NIEDERHOFFER	Pharmacologie
Sergio	ORTIZ AGUIRRE	Pharmacognosie
Sylvie	PERROTEY	Parasitologie
Romain	PERTSCHI	Chimie en flux
Frédéric	PRZYBILLA	Biostatistiques
Patrice	RASSAM	Microbiologie
Eléonore	REAL	Biochimie
Andreas	REISCH	Biophysique
Ludivine	RIFFAULT-VALOIS	Analyse du médicament
Carole	RONZANI	Toxicologie
Emilie	SICK	Pharmacologie
Yaouba	SOUAIBOU	Pharmacognosie
Maria-Vittoria	SPANEDDA	Chimie thérapeutique
Jérôme	TERRAND	Physiopathologie
Nassera	TOUNSI	Chimie physique
Aurélie	URBAIN	Pharmacognosie
Bruno	VAN OVERLOOP	Physiologie
Maria	ZENIOU	Chimiogénomique

Maîtres de conférences - praticiens hospitaliers

Julie	BRUNET	Parasitologie
Nelly	ÉTIENNE-SELLOUM	Pharmacologie- pharm. clinique

Assistant hospitalier universitaire

Damien	REITA	Biochimie

Faculté de pharmacie
Université de Strasbourg

SERMENT DE GALIEN

Je Jure,

en présence des Maîtres de la Faculté,
des Conseillers de l'Ordre des Pharmaciens
et de mes condisciples :

D'honorer ceux qui m'ont instruit
dans les préceptes de mon art et de
leur témoigner ma reconnaissance en
restant fidèle à leur enseignement ;

D'exercer, dans l'intérêt de la santé publique,
ma profession avec conscience et de respecter non
seulement la législation en vigueur, mais aussi les règles
de l'honneur, de la probité et du désintéressement ;

De ne dévoiler à personne les secrets
qui m'auront été confiés et dont j'aurai eu
connaissance dans la pratique de mon art.

Si j'observe scrupuleusement ce serment,
que je sois moi-même honoré
et estimé de mes confrères
et de mes patients.

Remerciements

Tout d'abord je tiens à remercier la Faculté de Pharmacie de Strasbourg, l'ensemble du corps enseignant et administratif de la Faculté pour les 6 années d'enseignement reçues grâce auxquelles j'ai pu me construire un avenir professionnel prometteur.

Je tiens à remercier le Professeur Julien GODET pour son accompagnement dans le rôle de directeur de thèse. Son regard critique et ses conseils m'ont été utiles pour la rédaction de cet écrit. Aussi, je souhaite remercier le Docteur Emmanuel BOUTANT et le Docteur Nicolas RISS en qualité de membre du jury de cette thèse d'exercice. Je les remercie par avance pour leurs questions pertinentes et leur regard critique.

Je remercie ma famille et mes amis pour leur patience, leur soutien et la confiance qu'ils m'accordent. Je les remercie particulièrement pour leur pression délicate qui m'a permis de garder en tête mon objectif malgré l'exercice de mon activité professionnelle. Aussi, je tiens à remercier tout particulièrement Dilara, grâce à qui j'ai déterminé le sujet de cette thèse au détour d'une conversation entre deux voitures dans un TGV reliant Paris à Strasbourg. Enfin, mes derniers remerciements et non les moindres vont à celle qui partage ma vie. Merci à toi Solène pour ton soutien et tes conseils avisés.

Table des matières

Liste des figures

Liste des abréviations

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

CD : Compact Disc

CD-ROM : Compact Disc Read-Only Memory

cm : Centimètre

CRISPR : Clustered Regularly Interspaced Short Palindromic Repeats

dNTP : Désoxyribonucléotide triphosphate

DVD : Digital Versatile Disc

Eo : Exaoctet

g : Gramme

HDD : Hard Disk Drive

HDS : Hébergeur de Données de Santé

IBM : International Business Machines Corporation

ISO : Organisation Internationale de Normalisation

Ko : Kilo-octet

NGS : Séquençage de nouvelle génération

ONT : Oxford Nanopore Technologies

PCR : Polymerase Chain Reaction

RGPD : Règlement Général sur la Protection des Données

SCRIBE : Synthetic Cellular Recorders Integrating Biological Events

SSD : Solid-State Drive

TdT : Terminal deoxynucleotidyl Transferase

TGS : Séquençage de Troisième Génération

To : Teraoctet

USB : Universal Serial Bus

USD : code de monnaie international établi par l'ISO pour le dollar américain.

ZFN : Zinc Finger Nuclease

ZMW : Zero Mode Waveguide

Zo : Zettaoctet

I. Introduction

A. Stockage de l'information : une nécessité

Depuis des milliers d'années, l'espèce humaine a eu besoin de stocker de l'information afin de la conserver et de la transmettre. En effet, depuis l'apparition du langage et de l'articulation de la pensée il est nécessaire de transmettre et de conserver de la connaissance. La mémoire cérébrale est un formidable outil pour cela mais n'est parfois et souvent pas suffisant. De plus, avec l'apparition de sociétés plus intellectuelles et organisées naquit le besoin de conserver de l'information dans de nombreux secteurs. Nos ancêtres ont d'ailleurs pu utiliser de nombreux supports comme les os, la pierre ou encore le papier (ou ce qui peut s'y apparenter) afin de conserver au mieux les connaissances, l'information. C'est ce qu'on peut appeler des "systèmes artificiels à mémoire" [1]. Par exemple, des écailles du plastron de la carapace de tortues et des omoplates de bovins gravées d'une écriture ossécaille retrouvées en Chine sont les plus anciens témoins écrits de la civilisation chinoise, ils datent du XIIe siècle avant Jésus Christ [2]. L'évolution des technologies et par-dessus tout de l'informatique a révolutionné notre rapport à l'information, son codage, sa lecture, son écriture et son stockage. Aujourd'hui, le stockage de données peut être défini par « toutes les technologies et les méthodes qui permettent de conserver et d'entreposer de l'information numérique, tout support confondu » [3]. Les supports actuels sont multiples : physiques comme les cartes perforées, magnétiques comme les cassettes, les disquettes ou les disques durs, optiques comme les CD ou les Blue-Rays, ou encore à mémoire flash comme les clés USB ou les SSD par exemple. Les utilisateurs particuliers enregistrent des données personnelles qui représentent un volume raisonnable et qui en substance sont généralement des fichiers tels que des documents, du multimédia. Les entreprises, elles, amassent et conçoivent de très grands volumes de données tous secteurs confondus.

B. L'ère informatique

Depuis le début des années 2000, l'essor du « Big Data » (qui signifie « volume massif de données » et fait référence au développement des nouvelles technologies, d'internet et des réseaux sociaux [4]) et des objets connectés avec l'Internet des Objets ont participé à engendrer une augmentation considérable du volume de données mondiales. Le monde s'est très largement digitalisé. L'utilisation de la donnée aujourd'hui a transformé notre manière de vivre, de travailler, de nous amuser. La conséquence directe de ce changement est l'augmentation considérable du volume de la sphère de données mondiales. Le stockage de données a ainsi connu une profonde transformation, les méthodes

de stockage associées ont rapidement évolué, passant par exemple en quelques années du stockage sur bande magnétique dans les années 80 aux nouvelles technologies que nous connaissons aujourd'hui tel que le Cloud. D'après de nombreuses publications analysées depuis plusieurs années [5] la quantité de données mondiales crées en 2020 est estimée à 64,2 zettaoctets (1 Zo = 10^{21} octets) et il est prévu une augmentation de celle-ci à 181 Zo pour 2025 (Figure 1). 181 Zo équivaut à 181 mille milliards de gigaoctets. Pour mieux comprendre ce que représente cette quantité de données : si nous étions capables de stocker tout cela sur des disques Blu-ray (dont la capacité de stockage est d'environ 25 Go), la pile de Blu-ray générée ferait plus de 9300km, ce qui représente pratiquement la distance Paris-Bangkok. Une projection à 2040 estime le volume de la sphère de données mondiales à 5000 Zo [6]. Cette augmentation constante et exponentielle du volume de données à stocker représente une demande de supports de stockage gigantesque. Or, depuis 2010, l'Homme génère plus de données qu'il n'est possible d'en stocker [7]. Un besoin certain se dessine dans ce secteur : il est nécessaire de trouver une alternative aux méthodes de stockage actuelles. Grâce aux progrès considérables de la science dans le secteur des biotechnologies et plus particulièrement en biologie moléculaire, une technologie inspirée du vivant pourrait représenter une alternative ou une méthode complémentaire aux méthodes de stockage déjà existantes : la molécule d'ADN. L'objectif de cet écrit est tout d'abord de présenter les caractéristiques particulières de cette technologie biologique afin de justifier de son intérêt, de présenter les défis techniques à relever dans le cadre de sa mise en application et d'explorer son application dans le cadre du stockage des données de santé.

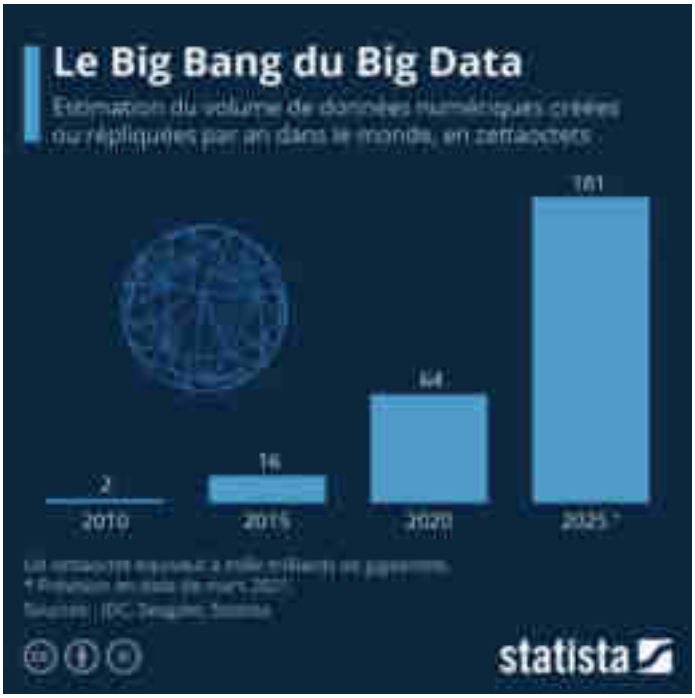

Figure 1 : Évolution de la sphère globale de données de 2010 à nos jours et projection en 2025 [8]

II. Évolution du stockage de données jusqu'à nos jours

A. Évolutions des technologies de stockage de données

À partir de la fin du XIXe siècle, les premières techniques de stockage de données sont apparues, à travers les cartes perforées (Figure 2), offrant une capacité limitée de stockage mais réelle. Elles ont marqué le début des technologies de stockage des données. Les bandes magnétiques sont développées en Allemagne dès 1928 par Fritz Pfleumer [9], initialement leur capacité de stockage n'était que de 960 Ko par support. Si les bandes magnétiques sont originellement construites pour réaliser des enregistrements audios, dans les années 50 leur usage se développe et elles deviennent une technologie populaire pour le stockage et l'enregistrement de données numériques ou analogiques. Elles sont même utilisées dans les premières mémoires secondaires des ordinateurs [10]. Aussi, rapidement, la bande magnétique fut miniaturisée et enroulée avant d'être intégrée dans des boîtiers : les cassettes, audio ou vidéo. La bande magnétique devient alors de plus en plus en vogue car elle représente une méthode de stockage avec une grande capacité et un coût relativement faible pour l'époque. Ensuite, l'apparition des disques durs (aussi appelés Hard Disk Drive, HDD) dans les années 1950, a considérablement augmenté les capacités de stockage et a été l'un des développements majeurs des décennies suivantes. Contrairement aux bandes magnétiques, le disque dur permet un accès instantané aux données qu'il contient [11]. En 1984, Apple introduit le lecteur de disquettes 3.5 pouces, conçu par Sony, dans ses ordinateurs, les Macintosh. Cette innovation gagne rapidement en popularité grâce à sa robustesse, son format de poche – inventé au départ pour tenir dans une poche de chemise – et sa capacité de stockage qui passera aisément de 720 Ko à 1,44 Mo au fil du temps. IBM et d'autres concurrents suivront cette tendance [12]. Le stockage de données sur CD (Compact Discs) a commencé lui aussi dans les années 1980. Les années suivantes verront l'apparition d'une longue suite de supports optiques différents : CD en 1980, CD-ROM en 1984, DVD en 1995, DVD-ROM en 1997, Blu-ray en 2003, etc [13]. Ces supports optiques ont considérablement révolutionné le stockage des données grâce à leur capacité à stocker des informations numériques en grande quantité avec durabilité, une qualité audio élevée et une longue durée de vie. Au milieu des années 1990, les logiciels étant de plus en plus lourds, les disquettes ne font plus le poids – il en faut une dizaine pour installer un programme et lancer un jeu [14]. L'entreprise Iomega invente alors la disquette Zip, contenant l'équivalent de 70 disquettes classiques soit 10 minutes de vidéo, une première pour le grand public. Elle est aussi deux fois plus rapide que la disquette et moins chère que le CD-ROM. Mais Iomega se heurte aux constructeurs qui n'intègreront pas ses lecteurs. Quelques entreprises seulement s'équiperont d'un lecteur externe, avant que le Zip ne soit totalement oublié. Au

début du nouveau millénaire, un nouvel outil de stockage des données portable apparait : Trek 2000 International, une société singapourienne, présente le ThumbDrive à Hanovre en février 2000. C'est la première apparition publique de ce qui sera bientôt surnommé « clé USB ». Avec une capacité initiale de 8 Mo – on monte aujourd'hui à 2 To –, le produit peut stocker deux fois plus de données que les formats de disque étendus de l'époque. La clé USB possède aussi un autre avantage : la mémoire flash permet un transfert de données encore plus rapide [15]. Enfin, au XXIe siècle, les systèmes de stockage basés sur le Cloud ont révolutionné la quantité de données stockées disponibles et l'accessibilité à ces dernières est devenue quasiment instantanée.

L'histoire du Cloud a commencé en 2006, quand Amazon Web Services a eu l'idée d'investir dans un important parc de machines pour louer ces ressources à des tiers et proposer ainsi des services d'infrastructures informatiques aux entreprises. Ce service, maintenant plus connu sous le nom de « Cloud Computing », proposait de nombreux avantages aux entreprises car il leur permettait de réduire les coûts d'infrastructure initiaux par des coûts inférieurs et adaptés spécifiquement à leurs besoins [16]. 2013 est une année marquante car le Cloud en tant que « modèle en tant que service » a largement été adopté par les entreprises [17]. Aujourd'hui, le marché du Cloud est en plein bouleversement. Le choix d'offres pour les entreprises mais aussi pour les particuliers est large et celles-ci sont déclinées en différents forfaits. Elles sont constamment en évolution pour répondre aux besoins de chacun. Depuis les années 2020, diverses options sont disponibles pour les solutions modernes de stockage de données, mais les Cloud Privé ou Public restent les solutions les plus plébiscitées du marché, allant des serveurs physiques sur site conventionnels au stockage Cloud Public ou aux systèmes hybrides. L'interaction complexe de plusieurs technologies de stockage répondant chacune à des objectifs spécifiques caractérise le paysage moderne du stockage de données. Mais, en raison de son évolutivité, de son accessibilité et de son prix abordable, le stockage Cloud s'est très largement démocratisé, permettant un échange de données facile et une collaboration fluide tant qu'un accès internet est disponible. Cette croissance, depuis les années 2010, est attribuée aux demandes d'applications et de charges de travail modernes, qui nécessitent une infrastructure que les centres de données traditionnels ne peuvent pas satisfaire [18]. Il existe deux catégories distinctes : le Cloud Public et le Cloud Privé. Les Clouds Publics utilisent une infrastructure partagée et gérée par un fournisseur tel que Google, Amazon ou Microsoft, tandis que les Clouds Privés fonctionnent sur l'infrastructure d'une entreprise qui en est la principale utilisatrice. En effet, les Clouds Privés, parfois appelés centres de données privés, sont ainsi construits sur l'infrastructure de l'entreprise, gérée par celle-ci, et généralement protégés par un pare-feu et physiquement sécurisé [19]. Le stockage de données avec un Cloud dit « hybride », qui combine un stockage sur Cloud Privé et Public semble répondre plus spécifiquement aux besoins des entreprises actuelles. Ses principaux avantages

reposent sur la flexibilité du contrôle des ressources informatiques plus rapidement accessibles et visibles grâce au Cloud Privé ; mais aussi sur l'évolutivité et la capacité à réduire les coûts d'investissement offert par le Cloud Public [20]. Toutefois, malgré l'essor du Cloud ou des méthodes de stockage hybrides actuelles, la forte progression de la quantité de données – 33 zettaoctets en 2018 (Figure 3) et 64 zettaoctets en 2020 (Figure 4) – questionne notre capacité à pouvoir la gérer avec ces seuls outils. La recherche se poursuit ainsi sur les technologies de pointe et le paysage du stockage de données continue toujours de se remodeler pour permettre répondre aux enjeux liés à l'augmentation considérable de ce volume de données.

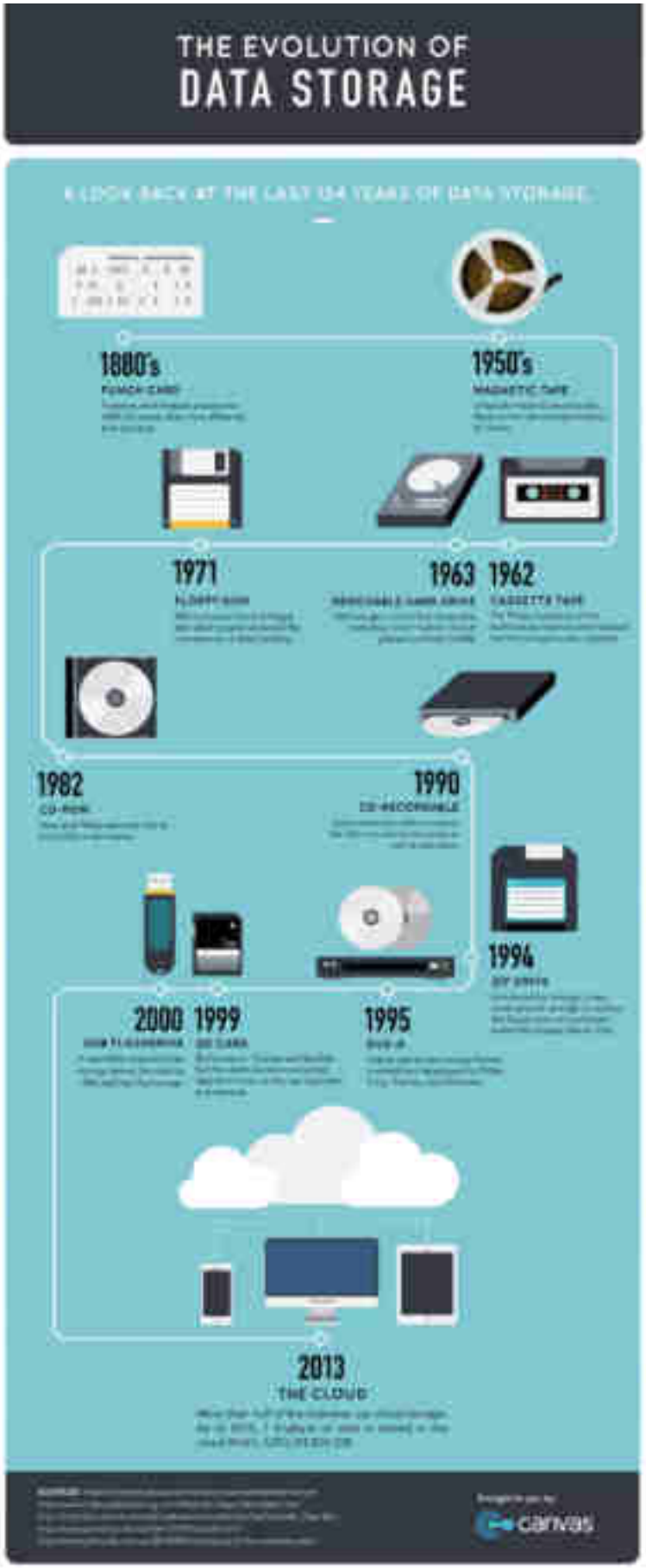

Figure 2 : Évolution des technologies de stockage de données modernes depuis l'apparition des cartes perforées [21]

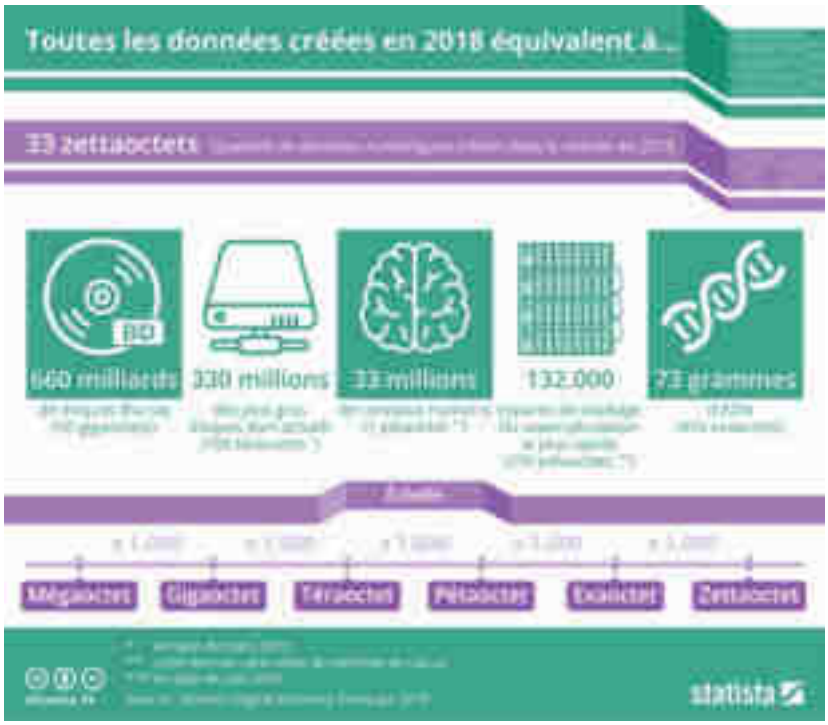

Figure 3 : Comparaison du stockage de la totalité de la totalité des données crées en 2018 stockées dans différents supports [22]

B. Les enjeux actuels associés à l'augmentation du volume de données mondial

La croissance sans précédent des données a commencé dès l'ère du numérique. Ce boom du volume de données, telles que la croissance du contenu multimédia, des appareils liés à l'Internet des Objets, des médias sociaux et des services de streaming a totalement modifié la manière dont nous stockons nos données et présente de nombreux enjeux. Le "Big Data" désigne littéralement les "données massives", nous pouvons aussi parler de "déluge de données". Il est plus précisément défini selon le principe des 3 V : Volume (le volume des données est suffisamment grand), Vélocité (les données arrivent de plus en plus vite et nécessitent d'être traitées rapidement) et Variété (le type de données est varié, elles peuvent être des données structurées ou des applications basées sur des bases de données). Le terme est apparu à la fin des années 1990, à une époque où le volume des données produites commençait à poser de sérieuses inquiétudes. En effet, la production de données a fini par croitre de manière exponentielle et continue encore sa courbe ascendante. En 2018, on estimait qu'on avait produit 33 Zo, soit 33 000 milliards de Go (Figure 3). En 2021, on estimait donc qu'on avait produit 79 Zo soit 40 fois plus qu'en 2010 [23]. D'ici 2025, le volume de données pourrait dépasser 181 Zo (Figure 1 & 4), ce qui posera des problèmes de stockage car ce chiffre ne devrait pas s'arrêter de croitre.

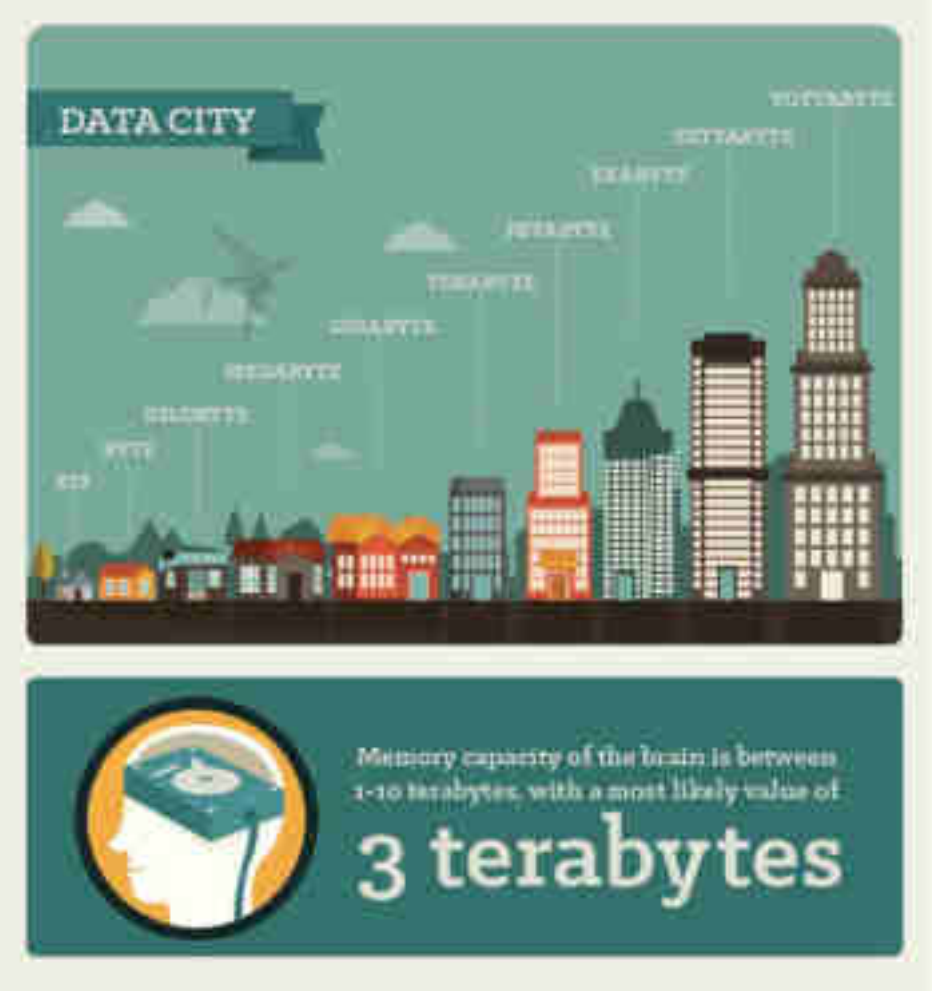

Figure 4 : Représentation de l'échelle des données selon leur taille [24]

Aussi, étant donné que la quantité de données générée augmente quotidiennement, la gestion et le stockage deviennent des défis majeurs. Les entreprises doivent savoir investir dans une infrastructure ou des solutions de stockage adaptées et dans des systèmes de gestion de données efficaces pour stocker, organiser, protéger et accéder à ces données de manière sécurisée et efficace. Il est essentiel de gérer la disponibilité, l'accessibilité et la sécurité des données, qui sont des actifs stratégiques pour les organisations [25]. En effet, à l'ère du Big Data, la sécurité des données constitue aussi l'un des enjeux clés pour les entreprises. En effet, les menaces sur la cybersécurité se multiplient à mesure que la quantité de données disponibles augmente : les entreprises sont de plus en plus exposées aux violations de leurs données. En 2019, l'entreprise spécialisée dans la cybersécurité AV Test GmbH avait recensé plus d'un milliard de cyberattaques par logiciel malveillant dans le monde (Figure 5) et ce chiffre a continué de croitre pour atteindre 1,11 milliard en 2020.

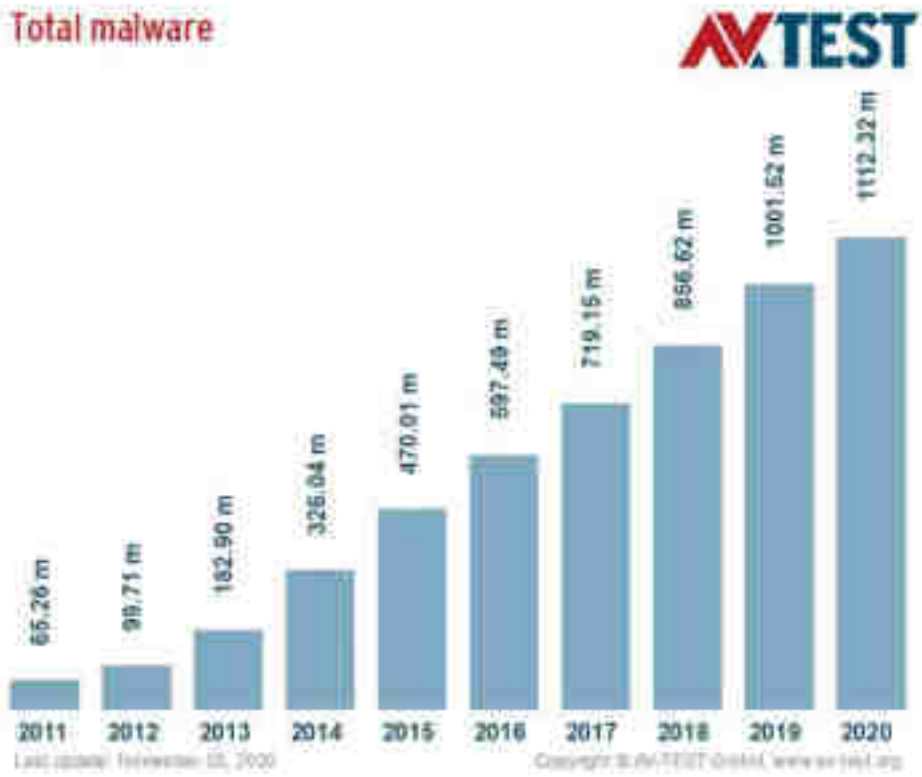

Figure 5 : Nombre de logiciels malveillants enregistrés par année [26]

Les différents types de cyberattaques, du phishing aux failles logiciels sont de plus en plus courantes et difficiles à contrôler (Figure 6). En effet, l'explosion du volume de données accroît la surface d'attaque potentielle car les données massives représentent des cibles intéressantes pour les cybercriminels. Les entreprises doivent ainsi investir et mettre en place des mesures de cybersécurité renforcées mais aussi agir sur la prévention et la détection des violations pour protéger leurs infrastructures de données. Les dépenses en cybersécurité ne cessent alors d'augmenter, L'Observatoire 2022 de l'Alliance pour la Confiance Numérique a notamment fait état d'une croissance de plus de 10 % du marché de la cybersécurité en France [27]. En outre, l'impact environnemental des centres de données est un enjeu majeur résultant de la croissance exponentielle du volume de données. Les préoccupations concernant les effets environnementaux des infrastructures de stockage de données se multiplient à mesure que le volume de données augmente. En effet, la pollution des centres de données provient essentiellement de leur besoin en électricité continu puisqu'ils fonctionnent à toute heure comme leurs systèmes de refroidissement. En 2022, les centres de données consommaient 2% de l'électricité mondiale et étaient responsables de 0,3 % des émissions de gaz à effet de serre à échelle mondiale [28]. De plus, les inquiétudes concernant l'impact environnemental du stockage de données se sont encore exacerbées par l'augmentation des déchets électroniques dû à la dépréciation rapide des équipements électroniques.

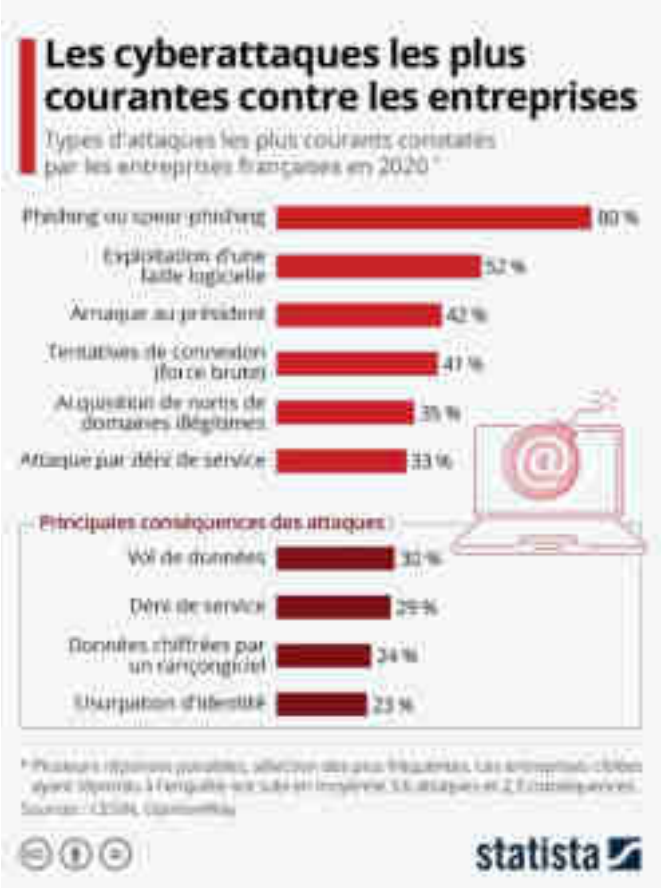

Figure 6 : Cyberattaques les plus fréquentes contre les entreprises [29]

C'est la raison pour laquelle les efforts visant à développer des solutions de stockage de données durables et économes en énergie, y compris dans l'utilisation de sources d'énergie renouvelables, et l'agencement de centres de données éco-énergétiques sont de plus de plus entrepris. Enfin, de nombreux autres défis en lien direct avec le volume croissant de données, tels que la préservation à long terme des données et les coûts associés, la gestion de l'intégrité des données, de leur portabilité ou encore de leur exploitation [30] apparaissent et nécessitent de repenser notre utilisation des méthodes de stockage actuelles. Le stockage de données fait l'objet d'un cycle constant d'innovation stimulé par les progrès technologiques et l'évolution des demandes des consommateurs. Il est certain que les futurs outils de stockage de données tenteront de faire face aux attentes des organisations, des sociétés et des individus tout en construisant un avenir de stockage plus efficace et plus durable.

En conclusion, les enjeux du stockage de données sont nombreux et complexes. L'avenir du stockage de données est néanmoins prometteur grâce à l'innovation : certains concepts émergents représentent de vraies opportunités pour essayer de répondre à ces enjeux et montre la possibilité de développement de nouvelles méthodes de stockage garantissant la gestion d'une nouvelle quantité et

d’une durabilité exponentielle des données. Les recherches approfondies actuelles sur l'informatique de pointe, le stockage quantique (Figure 7), le stockage via l’ADN en sont les démonstrations et pourraient bien révolutionner le panel des outils disponibles pour stocker des données.

Figure 7 : Évolution prévisionnelle de la taille du marché mondial de l’informatique quantique jusqu’en 2030 [31]

III. ADN comme support de l'information

A. L'acide désoxyribonucléique, une molécule essentielle

L'ADN, ou acide désoxyribonucléique, est une macromolécule biologique présente chez tous les êtres vivants. Elle est constituée de sous-unités que l'on appelle les nucléotides. Eux-mêmes sont constitués de trois parties : une groupe phosphate, un groupe désoxyribose et l'une des quatre bases azotées (qu'on appelle également bases nucléiques) que sont l'adénine, la thymine, la cytosine et la guanine. Au niveau structurel les groupements phosphates et désoxyriboses sont reliés les uns aux autres par des liaisons covalentes et s'alternent pour former une longue chaîne à laquelle sont fixées les bases azotées au niveau des sucres : tout cela forme un brin d'ADN. Or, ce qui confère à l'ADN sa stabilité est sa structure générale dite en double brin. Cela signifie que deux brins sont assemblés – on parle alors d'ADN bi-caténaire – par des liaisons hydrogènes au niveau des bases azotées selon une complémentarité qui suit la règle suivante : les thymines forment deux liaisons hydrogènes avec les adénines et les guanines forment trois liaisons hydrogènes avec les cytosines (Figure 8) [32].

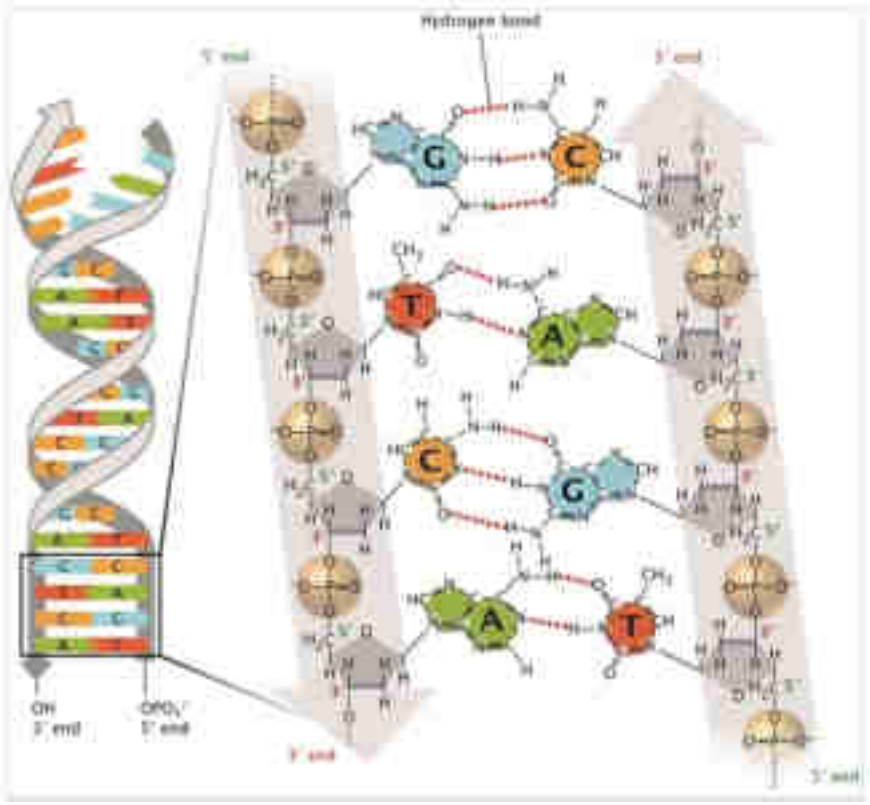

Figure 8 : Appariement des bases entre les deux brins de la molécule d'ADN [32]

La double hélice asymétrique est constituée de deux brin d'ADN complémentaires. Le squelette de chaque brin est composé de successions de phosphates et de désoxyriboses auxquels sont fixés l'une des 4 bases azotées. L'adénine et la thymine sont reliées par 2 liaisons hydrogènes et la cytosine et la guanine sont reliées par 3 liaisons hydrogènes. On peut également observer les deux extrémités des brins qui portent une dénomination indiquant le sens de chaque brin (5' et 3').

Ainsi, si l'on connaît la séquence de l'un des brins, on connaît également directement la séquence de son brin complémentaire. Cette structure unique permet à l'ADN d'être répliqué lors de la division cellulaire : les deux brins se séparent pour laisser place aux enzymes chargées de répliquer le code génétique et permettre de transmettre aux nouvelles cellules l'exacte copie de l'intégralité du code génétique de l'individu [33]. En 1953, Watson, Crick et Rosalind Franklin ont mis en évidence que la structure globale de l'ADN était en fait une double hélice ce qui signifie que les deux brins s'enroulent l'un avec l'autre de sorte à former une longue hélice, dont l'orientation des deux brins est anti-parallèle, ce qui signifie qu'ils sont orientés dans le sens opposé [34]. L'orientation de l'hélice est dite droitière quand elle présente une hélice qui tourne vers la droite (Figure 8). Cette structure dite de type B est la forme dans laquelle on retrouve majoritairement l'ADN en conditions physiologiques. *In vivo*, on peut également retrouver la forme dite Z (double hélice gauche dont l'espacement entre les boucles est plus court). Il existe enfin la forme A qu'on ne retrouve pas *in vivo* mais dans des échantillons peu hydratés, qui est une double hélice droite, plus large et courte [32]. Chez les organismes eucaryotes, l'ADN se retrouve dans un compartiment cellulaire : le noyau. En son sein, les longues molécules d'ADN sont compactées et prennent la forme de chromosomes (Figure 9). Chaque noyau de chaque cellule contient toute l'information génétique de chaque individu, répartie dans 46 chromosomes, ce qui représente environ 1,6 Go de données dans chaque noyau cellulaire. Si on multiplie ce chiffre par $3,0*10^{13}$ ce qui correspond au nombre total de cellules humaines dans un corps humain de 70kg en moyenne [35], on obtient un total d'environ 100 Zo de données stockées dans chaque corps humain [7]. Ce qui représente un volume considérable de données. Chez les organismes procaryotes, être vivants unicellulaires dépourvus de noyau, l'ADN est alors dans le cytoplasme et est sous forme circulaire, il forme une structure qu'on appelle nucléoïde [36]. Par exemple, la densité de données de la bactérie *Escherichia coli* est de 10^{19} bits/cm [37]. On remarque ici que la capacité de compaction de l'ADN est réellement stupéfiante et que la densité d'information qu'il contient dans un si petit espace est très intéressante.

A la lumière de toutes ces informations l'ADN présente des caractéristiques très intéressantes en ce qui concerne sa capacité à contenir de l'information. Très compacte, très dense, présent partout. La question qui peut alors venir à l'esprit est ainsi : Pourquoi ne pas utiliser la technologie que représente l'ADN afin de subvenir aux besoins grandissant d'une alternative aux méthodes de stockage des données numériques actuelles ?

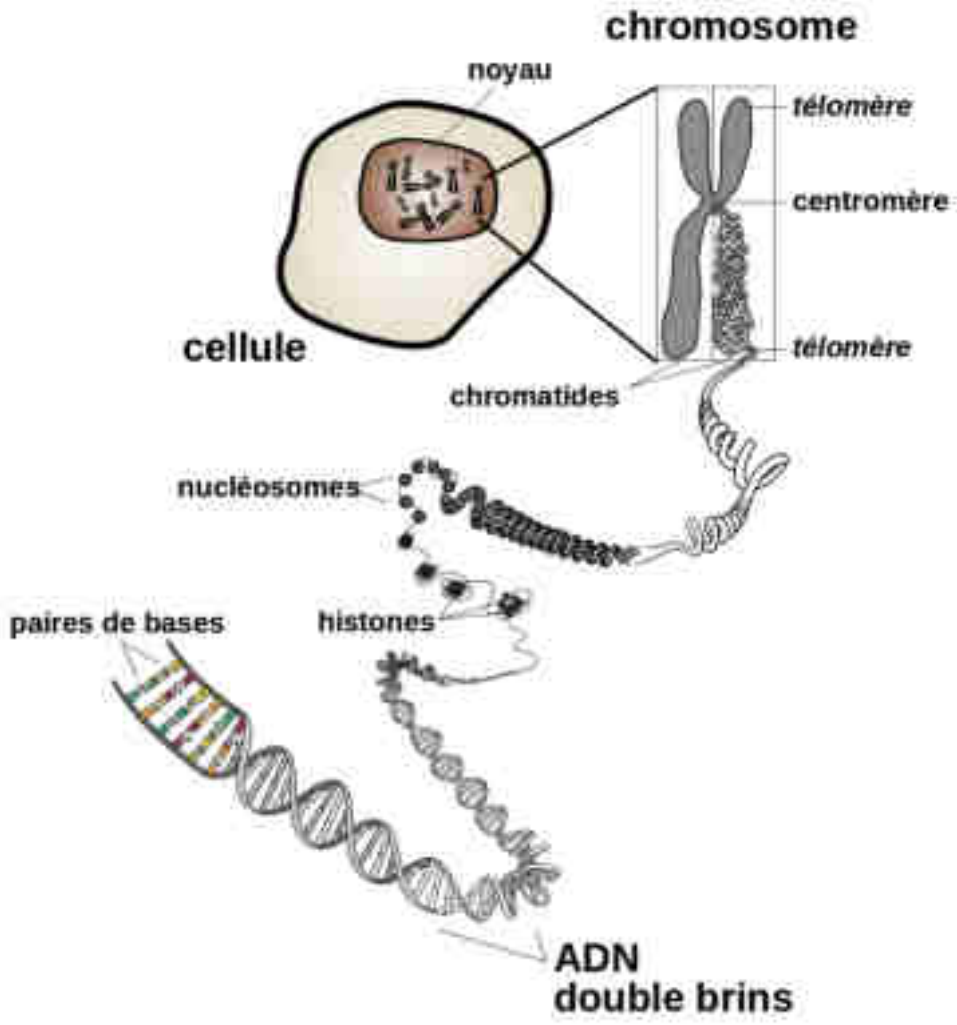

Figure 9 : Représentation des différents états d'organisation de l'ADN [38]

Dans les cellules eucaryotes et dans certaines bactéries (les archées), l'ADN est présent sous plusieurs états de condensation. Les chromosomes sont formés par la longue molécule d'ADN double brin compactée grâce à une protéine nucléaire spécifiques des cellules eucaryotes et des archées : les histones. Ensemble, ils forment une structure appelée nucléosome qui est la structure de base de la chromatine qui compose les chromosomes.

B. Intérêt

Tout d'abord, il est nécessaire de rappeler que dans le contexte biologique, le rôle de l'ADN est de conserver l'information génétique. Il doit donc disposer d'au moins deux caractéristiques essentielles à la bonne transmission et conservation de cette information : la stabilité physico-chimique afin de minimiser le risque d'altération de sa séquence, et la possibilité d'être répliqué [39]. Ces deux caractéristiques permettent déjà de remarquer que l'ADN respecte des critères essentiels au stockage de données.

De plus, la densité de la molécule a déjà été abordée mais elle mérite d'être approfondie davantage. La masse d'un bit sur ADN est de $5{,}1*10^{-22}$ g. Cette valeur est plus de 10 fois inférieure à celle de la mémoire flash. Si la même comparaison est faite concernant le volume on observe une différence de l'ordre de 10^3 bit/cm^3 (densité volumétrique ADN : 10^{19} bit/cm^3 ; mémoire flash : 10^{16} bit/cm^3). Selon Zhirnov et al., s'il était réalisable d'atteindre une pareille densité par voie de synthèse, alors la totalité des données prévues pour 2040 pourraient être stockées dans une boîte aux dimensions suivantes : 100 cm x 100 cm x 10 cm [37]. Toutefois, cette projection n'est pas réaliste. En effet, dans la réalité tout n'est pas stocké sur une longue molécule unique d'ADN mais sur de très nombreux fragments répliqués et qui contiennent des séquences de contrôle de la qualité et de classification. Une prévision plus réaliste du volume total d'ADN qui serait nécessaire pour stocker ce yottabit (YB) de données correspondrait plutôt au coffre d'une petite camionnette [6].

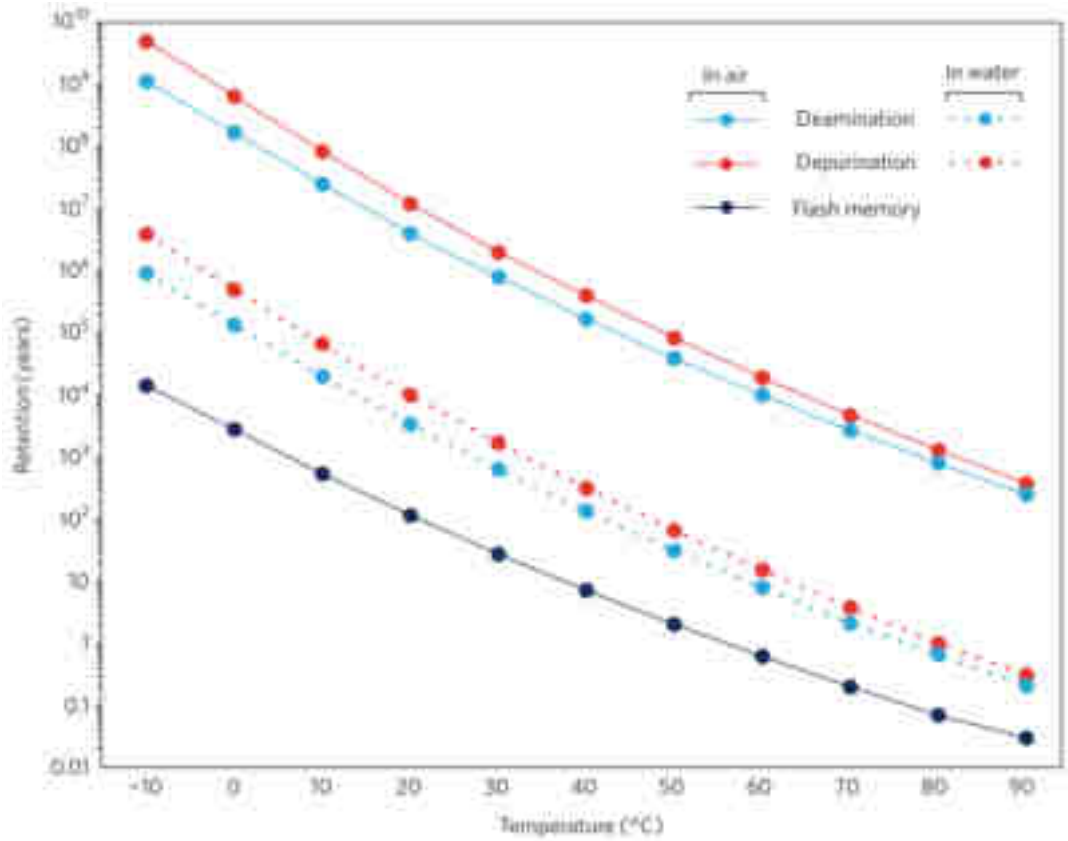

Figure 10 : Représentation des temps de conservation calculés de l'ADN et de la mémoire flash dans l'air et l'eau selon la température [37]

Temps de rétention théorique calculé de l'ADN dans deux milieux de stockage différents (air et eau) et soumis à deux mécanismes de dégradation tels que sont la déamination et la dépurination comparé au temps de rétention théorique calculé de la mémoire flash, support de stockage informatique le plus performant existant à ce jour.

Enfin, la consommation d'énergie du stockage de l'ADN est extrêmement faible : sa demi-vie, c'est-à-dire le temps au bout duquel la moitié de sa quantité est détruite, est de 521 années dans des conditions de conservation non adaptées [40]. Si on le compare au stockage flash, le stockage sur ADN démontre une réelle supériorité dans la conservation de données à long terme (Figure 10) [37]. Ces données impliquent que pour sa conservation dans le cadre de son utilisation en tant que support de l'information, très peu d'énergie serait nécessaire. Ainsi, la seule consommation importante proviendrait des technologies utilisées pour l'écrire et la lire. A ce sujet, une réduction de cette consommation de l'ordre d'un facteur mille a été évoquée dans le cadre d'un projet en cours, faisant passer la consommation d'un centre de données de plusieurs mégawatts à quelques kilowatts pour stocker la même quantité d'information sur l'ADN et y accéder régulièrement [41].

C. Étapes historiques

Le concept de l'ADN support de la donnée informatique est né dans les années 60 lors d'une interview du Dr. Norbert Wiener dans laquelle il aborde le sujet d'une possibilité qu'un jour l'Homme utilise une technologie liée à l'ADN pour stocker de l'information [42]. A ce moment-là les technologies d'écriture (synthèse) et de lecture (séquençage) de l'ADN n'en étaient qu'à leur commencement. La première preuve de concept fut réalisée en 1988 par l'artiste Joe Davis et des collaborateurs scientifiques de Harvard : ils insérèrent dans le génome d'une bactérie (*Escherichia coli*) une séquence de 35 bits codant pour une image d'une ancienne rune germanique représentant la vie et la terre féminine [43,44]. Ce n'est qu'en 1999 que pour la première fois de l'information fut stockée *in vitro* sur ADN : Clelland, Risca et Bancroft ont créé un micropoint (une photo très réduite d'une page dactylographiée collée sur un point dans une lettre anodine) refermant un message codé par l'ADN et lui-même camouflé dans la complexité de l'ADN humain [45]. Entre 2001 et 2008 plusieurs équipes publièrent à ce sujet sans avancée majeure et toujours *in vivo* [46–51]. C'est en 2012 et 2013 que Church et al. [52] ainsi que Goldman et al. [53] publient chacun de leur côté deux publications qui marquent le début d'une nouvelle ère pour la recherche dans ce secteur. En effet, pour la première fois depuis 1999 des scientifiques sont parvenus à stocker *in vitro* des megabits de données numériques grâce à l'ADN et à lire ces données par séquençage en récupérant la totalité des fichiers stockés. Ces travaux marquent une étape majeure dans l'avancement de cette technologie : ils sont en quelque sorte la première preuve de concept du stockage de données numériques dans l'ADN. Les premiers travaux concluants sur l'accès localisé à l'information dans le cadre de l'utilisation de l'ADN comme support sont publiés en 2015 par Yazdi et al. [54]. Ils sont basés sur la synthèse de nombreux blocs d'ADN flanqués de séquences connues qui permettent une

reconnaissance localisée de l'information. En 2017, Yazdi, Gabsy et Milenkovic [55] mettent en place le premier système portable, il est également doté d'une technique de correction des erreurs lors de la phase de lecture afin de retrouver l'intégralité de l'information recherchée. Jusqu'alors, la méthode pour synthétiser l'ADN dans ces travaux était chimique. En 2018, Lee et al. [56] publièrent leur travail sur la synthèse enzymatique de l'ADN dans le cadre de son utilisation pour stocker de la donnée. Leur système permet une économie de coût, de vitesse de synthèse et de séquençage mais n'est qu'un essai à petite échelle (moins de 100 bits). La synthèse enzymatique reste toutefois un sujet de recherche majeur et prometteur. La même année, un groupe de chercheurs [57] publia la première preuve de concept à plus grande échelle en réalisant expérimentalement le stockage de 35 fichiers (200 MB), tout en ayant une gestion de l'accès localisé et de la correction des erreurs. A ce jour, le record de la plus grande quantité de données stockée est détenu par l'entreprise Catalog, basée à Boston aux Etats-Unis. Ils sont parvenus à enregistrer l'intégralité des textes de la version anglaise de la plateforme Wikipédia (ce qui correspond à environ 16 Go) sur ADN. La machine était alors capable d'écrire la donnée à une vitesse de 4 MB par seconde [58]. Ils travaillent toujours sur ce sujet et ont réussi à lever des fonds à hauteur de 35 millions de dollars américains [59]. En France, l'entreprise DNA Script qui a développé une machine à synthétiser l'ADN par voie enzymatique pour le laboratoire a aussi levé 200M dollars [60]. Ils travaillent également sur le sujet du stockage de données sur ADN. Enfin, l'Intelligence Advanced Research Project Activity est une agence gouvernementale américaine qui finance des projets de recherche universitaires ou industriels dans une large panel de domaine dont la science. Lancé en 2019, le programme MIST (Molecular Information Storage) et financé par l'« Intelligence Advanced Research Projects Activity » (une organisation du Bureau du directeur du renseignement national américain) vise à réussir à écrire 1 To de données, en lire 10 (To) par jour avec accès localisé à la donnée, pour moins de 1000$ et consommant moins de 1 kilowatt [41].

D. Étapes majeures du processus de stockage de données dans l'ADN

Le processus général de la technologie a déjà été déterminé [52,53]. Il est nécessaire de bien le comprendre afin de saisir quelles en sont les étapes limitantes aujourd'hui et les défis à relever.

Il est constitué de cinq étapes majeures (Figure 11) :

- L'encodage, qui consiste à passer du code binaire à un code déterminé compatible avec les caractéristiques de l'ADN ;

- La synthèse, il s'agit de synthétiser la molécule d'ADN encodée à l'étape précédente selon une méthode de synthèse qui peut être chimique ou enzymatique.
- Le stockage, qui est la finalité de la technologie.
- La lecture, nécessaire à l'accessibilité de l'information. Là aussi plusieurs technologies existent.
- Et enfin, le décodage, selon la méthode d'encodage qui a été utilisée.

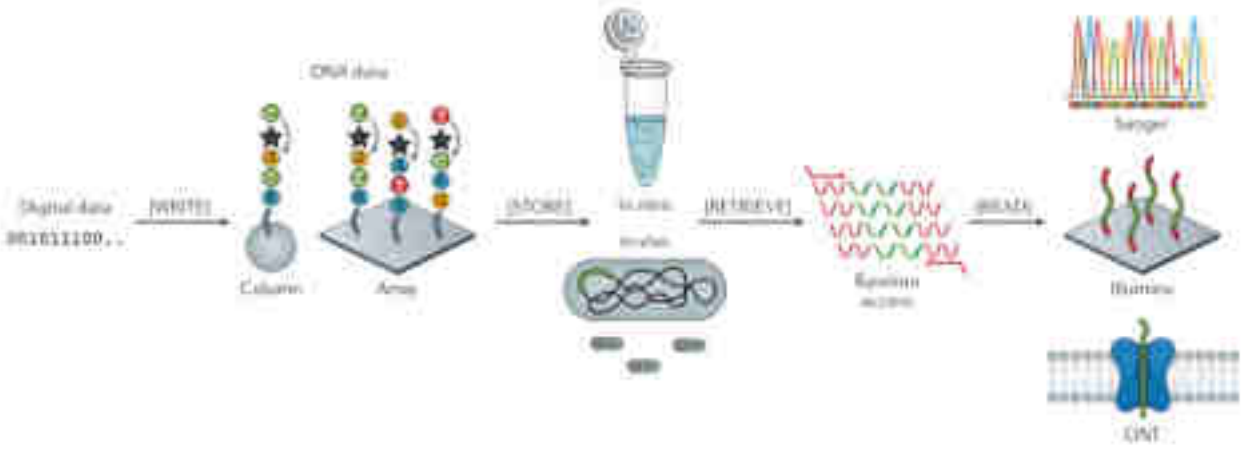

Figure 11 : Principales étapes du stockage de données de l'ADN [61]

Pour être stockée sur ADN, la donnée doit passer par des étapes essentielles : codage de la donnée binaire en langage compatible avec les 4 bases qui constituent la molécule d'ADN, synthèse des séquences d'ADN correspondantes, stockage, accès localisé, lecture par séquençage puis décodage selon la méthode d'encodage utilisée.

E. Encodage de la donnée et écriture par synthèse : où en sommes-nous, quelles pistes pour l'avenir ?

La question élémentaire à se poser avant d'entreprendre la synthèse est : "Quoi écrire ?". En effet, il est nécessaire de transposer l'information numérique en un code qui peut être utilisé dans la séquence d'ADN pour adapter le langage du matériel informatique à celui de la technologie à base d'ADN.

Encodage

Le langage utilisé par le matériel informatique est le système binaire. Il utilise un système de numérotation en base 2, qu'on appelle bit, du chiffre binaire anglais ("binary digit"), le nombre de notation binaire. Un bit peut prendre deux valeurs, qui par convention sont "0" et "1". De son côté, l'ADN est constitué de 4 acides nucléiques (A, T, C et G) : une possibilité d'encodage logique serait alors d'utiliser chaque nucléotide un à un pour coder une paire de bit ; soit la paire 00,

soit 11, soit 10 soit 01. Ainsi, à la place d'avoir des séquences de 0 et de 1 on aurait des séquences d'ADN qui codent pour les séquences de 0 et de 1 désirées avec une densité de deux bits par nucléotide. Cela correspond à un gain de taille de deux fois en termes de densité de stockage. De plus, les séquences d'ADN sont relativement courtes : autour de 200 paires de bases, car il n'est pas encore possible de synthétiser des fragments très longs avec un faible taux d'erreur [6]. Pour identifier chaque séquence et pouvoir les classer dans le bon ordre, des séquences étiquettes sont ajoutées à chaque séquence codée. La nécessité d'accéder à l'information de manière localisée implique également l'ajout d'une séquence d'identification dans chaque fragment d'ADN synthétisé. Enfin, la présence de séquences identiques d'un fragment à l'autre permet l'amplification des fragments [62,63] grâce à une technologie de biologie moléculaire appelée réaction de polymérisation en chaine (PCR) et qui permet à l'aide de deux courts brins d'ADN monocaténaires d'amplifier de manière exponentielle un ou plusieurs séquences d'ADN. Ainsi, chaque fragment est constitué de la séquence codant pour l'information numérique (qu'on appelle charge utile), de séquences étiquettes, de séquences d'identification et de séquences d'amplification. Il est essentiel d'optimiser au maximum l'information encodée, une approche possible qui pourrait permettre de réduire l'espace utilisé par la donnée est la compression des données. C'est une technique informatique régulièrement utilisée de nos jours et qui a pour but de réduire la taille des données en éliminant la redondance et en encodant les informations de manière plus compacte. Dans le contexte du stockage de données sur ADN, cela impliquerait de préalablement transformer les données avant l'encodage et la synthèse des molécules d'ADN pour occuper moins d'espace. Cela pourrait avoir comme effet de réduire les coûts de synthèse et de séquençage ou simplement de stocker davantage de données. Des algorithmes de compression sans perte, tels que celui de Lempel-Ziv-Welch ou encore de Huffman [64], ainsi que des méthodes de compression avec perte qui peuvent être adaptés pour être utilisés avec les données stockées dans l'ADN. L'algorithme de Huffman a par exemple déjà été utilisé dans le cadre de la réalisation d'une preuve de concept afin de comprimer les données stockées [53]. Néanmoins, il est évident que le choix de la méthode de compression devra soigneusement être évalué afin de tenir compte des compromis entre la complexité de la décompression, les ressources nécessaires pour la lecture des données et la taille finale des données stockées.

Synthèse

Les deux principales méthodes de synthèse de l'ADN existant à ce jour sont la méthode de synthèse chimique et la méthode enzymatique.

La **première**, la **méthode chimique** a été proposée en 1981 par Beaucage et Carhuters [65]. A ce jour, elle est la méthode la plus communément utilisée pour la synthèse d'oligonucléotides. Elle est basée sur la chimie des phosphoramidites et comporte plusieurs étapes chimiques spécifiques (Figure 12). Lors de sa synthèse, l'ADN est accroché sur un support de résine duquel il sera décroché une fois la synthèse terminée. Ce n'est d'ailleurs pas un brin unique mais de multiples copies qui sont synthétisées en même temps. Chaque nucléotide est ajouté un à un, le nucléotide précédent est protégé par un groupe labile (4,4'-diméthoxytrityle) en 5' : lors de son exposition à la lumière ou à l'acide, le groupe protecteur est éliminé (étape n°1 dite de déprotection) et rend accessible le groupe hydroxyle en 5'. Cela permet au nucléotide libéré de former un phosphite triester avec la phosphoramidite introduite dans le milieu réactionnel (étape n°2 de couplage). Comme tous les nucléotides libérés dans l'étape n°2 n'ont pas été couplés car la réaction n'est jamais totale, une étape de blocage (étape n°3) est nécessaire : elle permet d'empêcher que les hydroxyles encore libres ne soient plus réactifs en étant acétylés par ajout d'acide acétique. Enfin, la dernière étape (n°4) est une étape d'oxydation afin d'obtenir une liaison phosphodiester à la place de la liaison phosphite, ce qui est plus stable. Une fois la synthèse terminée, une étape de purification a lieu pour éliminer les groupements de protection et les solvants du milieu, retirer les oligonucléotides synthétisés du support et retirer ceux ayant été synthétisés tronqués [66]. Bien qu'étant la plus répandue à ce jour, la synthèse chimique présente un taux d'erreur d'environ 0,5-0,7 %. De plus, la technologie ne permet pas de synthétiser avec une fiabilité suffisante des molécules d'ADN de plus de 200 nucléotides. Pour augmenter la taille des fragments, des extrémités chevauchantes sont alors ajoutées en début et fin de séquence afin d'assembler bout à bout les fragments synthétisés une fois la synthèse terminée. Cette méthode basée sur la chimie des phosphoramidites est la méthode utilisée par tous les façonniers de fragments d'ADN tels que Twist Bioscience, Integrated DNA Technologies ou encore GenScript. Bien que le monde de la biologie moléculaire n'ait eu besoin d'utiliser que de courts fragments d'ADN, depuis plusieurs années sont nécessaires des fragments d'ADN bien plus longs. Il a alors fallu trouver une méthode afin de synthétiser des fragments plus longs [67].

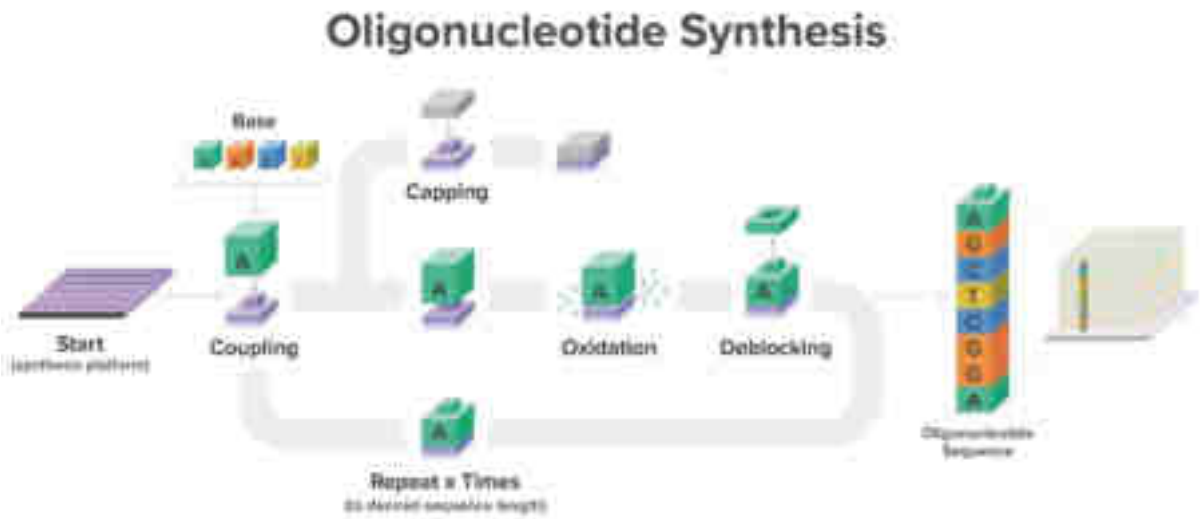

Figure 12 : Représentation d'un cycle de synthèse chimique d'ADN [68]

La **seconde méthode** est basée sur l'utilisation d'**enzymes** et est donc une méthode dite biologique. Son processus est intrinsèquement plus simple et efficace car il est basé sur une enzyme dont le rôle est la synthèse d'ADN dans le vivant : l'ADN polymérase (Figure 13). Grâce à une version de l'ADN polymérase particulière provenant de cellules immunitaires, la synthèse enzymatique est possible depuis les années 2010 [6,69]. Cette enzyme s'appelle la *Terminal deoxynucleotidyl Transferase* (TdT), elle agit sans matrice : si on lui fournit un à un les nucléotides voulus, il est alors possible de lui faire synthétiser un brin d'ADN de séquence voulue. C'est cette enzyme qui est utilisée dans la plupart des méthodes de synthèses enzymatiques en raison de ces caractéristiques particulières. Plusieurs techniques basées sur cette enzyme existent, par exemple, l'une d'elles consiste à fixer à chaque unité de l'enzyme TdT une molécule désoxyribonucléoside triphosphate (dNTP) qui sera ensuite incorporée à une amorce. Une fois la molécule de dNTP incorporée, l'extrémité 3' de l'amorce reste liée de manière covalente à la TdT, ce qui empêche l'ajout d'autres dNTP et permet ainsi un ajout sélectif. Il suffit alors de couper la liaison covalente pour libérer l'extrémité de l'amorce et continuer la synthèse avec les dNTP voulus [70]. Une autre technique existe, elle est toujours basée sur l'action de la terminal désoxynucléotidyl transférase (TdT) et d'une autre enzyme : une apyrase. Le principe reste relativement similaire aux autres techniques car il s'agit d'un ajout successif de nucléotides au brin d'ADN synthétisé, les nucléotides libres dans le milieu sont inactivés par l'apyrase puis éliminés avant de passer au nucléotide suivant. Cette technique reste néanmoins beaucoup moins précise car elle ne permet pour le moment pas de maîtriser la quantité de nucléotides ajoutés à chaque étape [71]. Elle semble être une méthode adaptée au stockage de données dans l'ADN selon Henry Lee, fondateur de Kern System, start-up dont l'objectif est le développement de cette méthode dans le cadre du stockage de données sur ADN [67].

Plusieurs entreprises tel que DNA Script en France, Ansa Biotechnologies ou encore Molecular Assemblies aux Etats-Unis travaillent activement à l'amélioration et l'optimisation de cette méthode. Par exemple, DNA Script a annoncé avoir réussi à atteindre 99,7% de fidélité avec sa méthode de synthèse enzymatique, ce qui est supérieur aux 99,3% obtenus généralement avec la méthode chimique [67]. De plus, la méthode enzymatique pourrait permettre de synthétiser des brins allant jusqu'à 1000 nucléotides soit 5 fois plus que la méthode chimique. Elle permet également la synthèse d'ADN sans la production de déchets nocifs, c'est donc une méthode plus verte et respectueuse de l'environnement. Enfin, la synthèse enzymatique d'ADN double brin est moins chère que celle par voie chimique, ce qui a un intérêt particulier dans le cadre du stockage de données sur ADN.

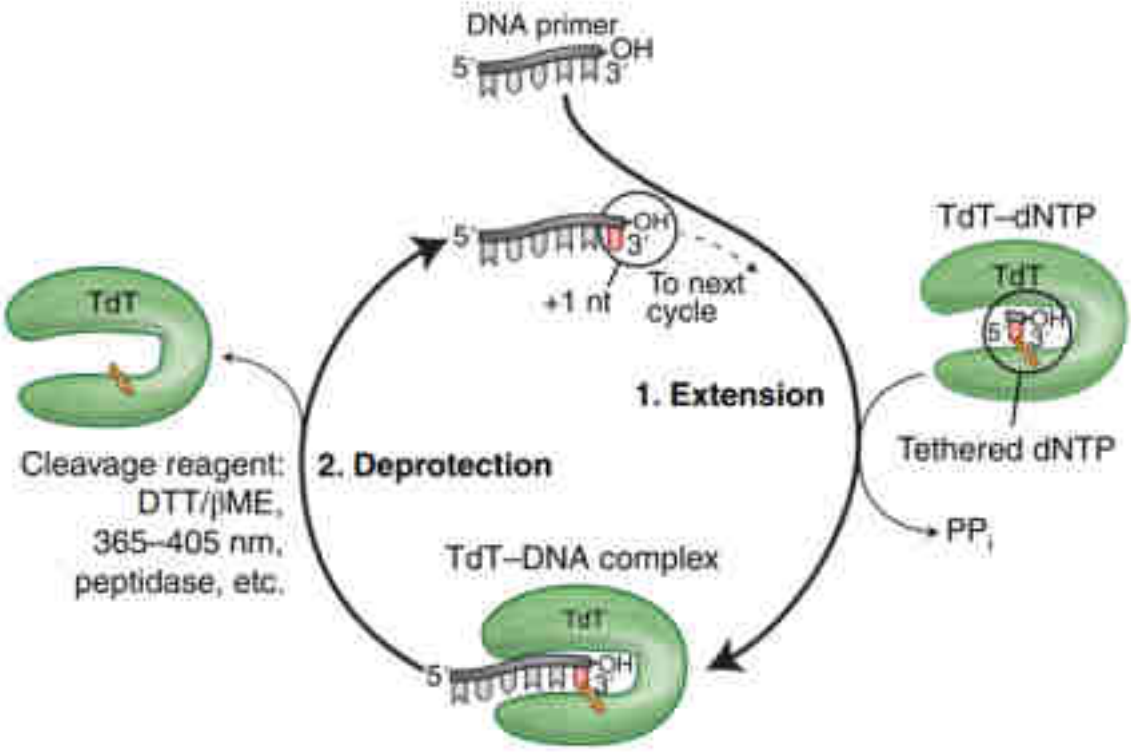

Figure 13 : Représentation d'un cycle de synthèse enzymatique d'ADN grâce à la méthode de la TdT couplée à un dNTP [70]

F. Stockage

Il a déjà été évoqué que la molécule d'ADN est extrêmement stable et qu'elle possède une demi-vie de 521 ans dans de mauvaises conditions de stockage (cf. III. B.). Il est alors aisé de comprendre qu'à l'échelle d'une vie humaine, même si stocké dans des conditions inadaptées, la

donnée stockée pourrait être conservée très longtemps. Plusieurs méthodes de stockage ont été mises au point et peuvent être utilisées dans ce cadre.

Il faut tout d'abord identifier la donnée. Sera-t-elle une donnée d'archive à laquelle on souhaite accéder très rarement ? Sera-t-elle une donnée qu'on souhaite consulter régulièrement ? Très souvent ? Doit-elle être accessible à tout moment ? Les données peuvent être classées en plusieurs catégories : les données chaudes (que l'on souhaite solliciter régulièrement) et les données froides (que l'on souhaite conserver mais qui ne sont consultées que rarement). En effet, la technologie de stockage utilisée sera fonction de l'objectif du stockage et de ce que souhaite en faire l'utilisateur, car chaque méthode possède ses propres caractéristiques énergétiques, logistiques, et surtout ses propres propriétés de conservation. Une étude de 2021 a permis de déterminer quelles seraient les meilleures techniques de stockage de l'ADN dans le cadre du stockage de données [72]. Ces méthodes sont déterminées selon une balance bénéfice/risque qui prend en compte la fréquence d'accès à la donnée en regard de la possible dégradation de l'ADN selon la méthode de stockage utilisée. Dans le cas de données froides (données d'archives) auquel la fréquence d'accès ne serait que d'une fois par an, la méthode de stockage serait l'encapsulation. Les facteurs de dégradation dans ce cas sont l'oxydation, l'humidité, la température et les radiations. Pour les données auxquelles on ne souhaite accéder qu'une dizaine de fois par an, la méthode de stockage pourrait être sous forme lyophilisée congelée. Les facteurs de dégradations sont alors la congélation/décongélation, la réhydratation et la lyophilisation. Enfin, s'il s'agit de données dont l'accès est très fréquent, la proposition est le stockage sous forme solubilisée en milieux aqueux tamponné. Les facteurs de dégradation sont la température, le pH, le tampon et les cisaillements mécaniques dûs aux pipetages (Figure 14).

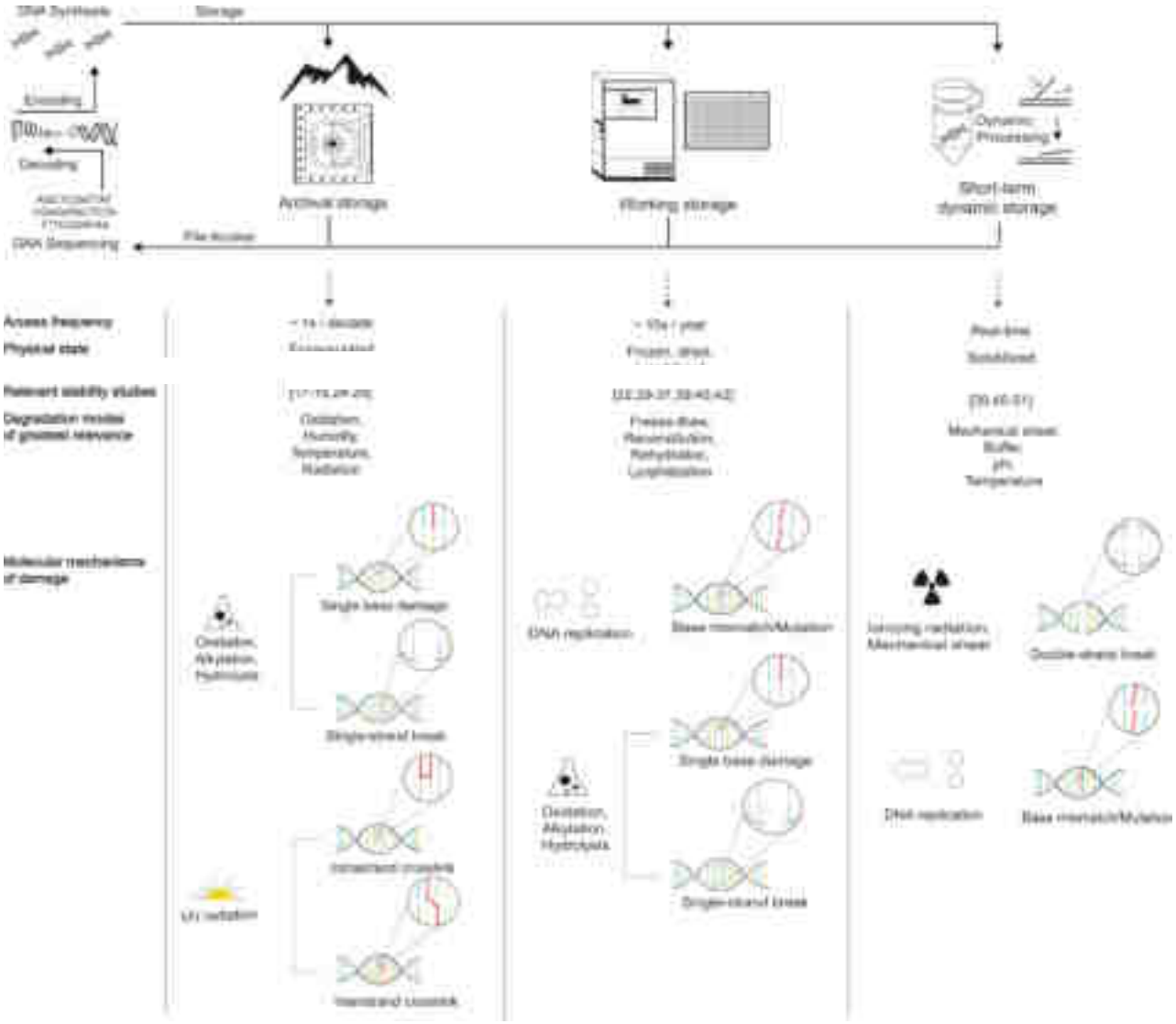

Figure 14 : Types de stockages d'ADN et leurs dommages potentiels associés [72]

L'ADN pourra être stocké différemment selon la fréquence d'accès prévue aux données qui le constituent. Chaque méthode de stockage dispose de ses propres caractéristiques et expose à différents mécanismes de dégradation. Ces mécanismes peuvent induire plusieurs types de dommages au niveau moléculaire : cassure d'un brin, altération d'une base unique, réticulation intra- ou interbrins, mauvais appariement de bases, mutations, coupure simple brin ou encore coupure double brin.

L'application première du stockage de données dans l'ADN sera très certainement le stockage des données d'archives car le coût de synthèse et de lecture de la donnée lorsque stockée sur ADN ne peut être amorti qu'après plusieurs années de stockage [73]. Dans leur revue de 2021, Lim et al. [74] ont référencé plusieurs formes de stockage de l'ADN en analysant en parallèle leur profil de durabilité, de facilité d'utilisation, la densité de l'ADN dans son contenu et le coût (Figure 15).

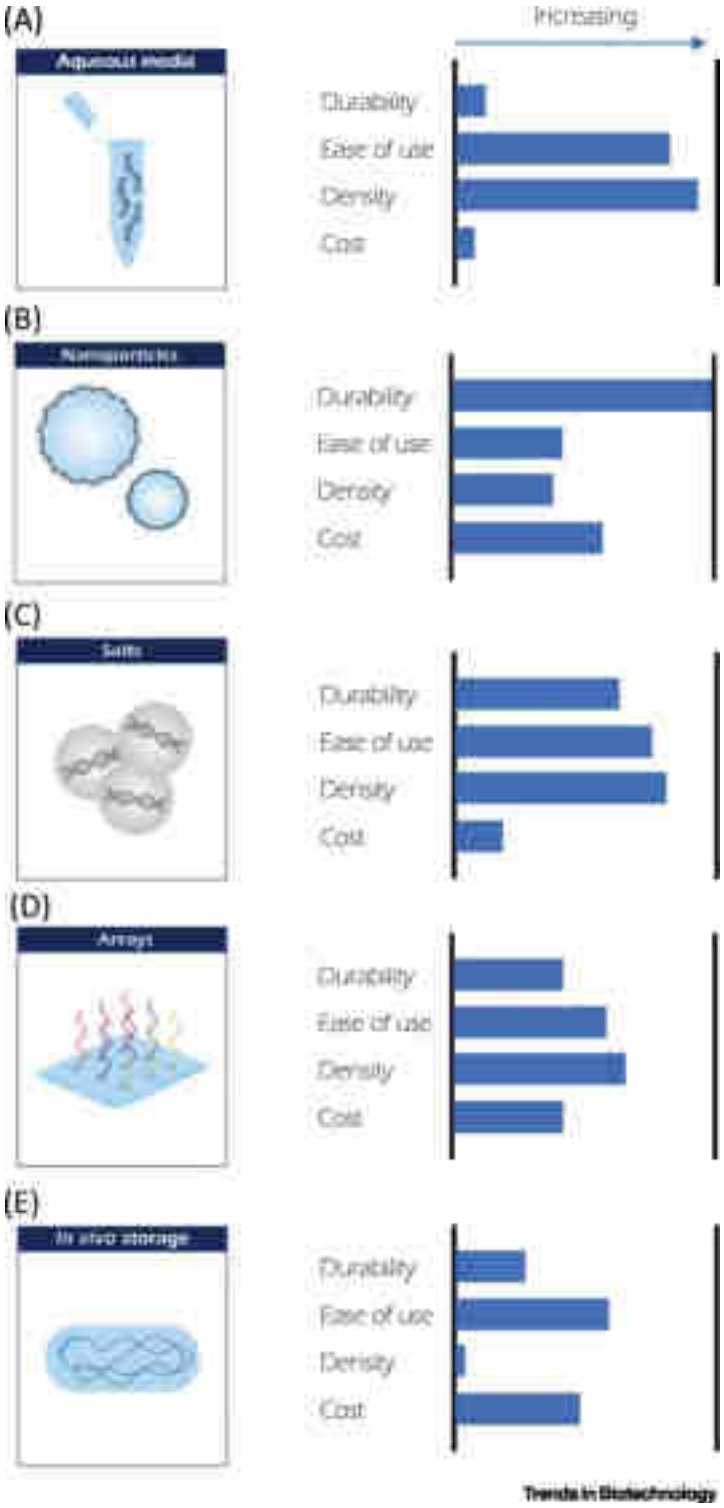

Figure 15 : Méthodes de stockage de l'ADN et leur caractéristiques techniques principales [74]

Chaque méthode de conservation de l'ADN possède ses propres caractéristiques et ses propres avantages et inconvénients. Le stockage en milieu aqueux possède une facilité d'utilisation ainsi qu'une densité élevée pour un faible coût mais une faible durabilité. Les nanoparticules présentent une durabilité la plus haute mais ont un coût élevé et une densité et facilité d'utilisation moyennes. Le stockage sous forme de sels présente une bonne durabilité, densité et facilité d'utilisation pour un coût maitrisé. Les puces à ADN présentent une durabilité moyenne ainsi qu'une facilité d'utilisation et une densité moyenne mais un coût relativement élevé. Enfin, le stockage in vivo présente une faible durée de vie, une facilité d'utilisation moyenne, une très faible densité et un coût élevé.

Il existe une multitude de choix pour le stockage de l'ADN, chaque technique possédant ses propres caractéristiques, avantages et inconvénients. En outre, le choix de la méthode de stockage

devra dépendre de l'objectif du stockage (comme évoqué plus haut : donnée chaude ou donnée froide). Quelle qu'elle soit, la méthode utilisée doit pouvoir préserver l'ADN de l'eau, de l'oxygène, de la lumière et des températures élevées qui sont les principaux facteurs qui peuvent induire des dommages aux molécules. Classiquement, la méthode utilisée pour préserver l'ADN de ces facteurs de dégradation est le stockage à basse température qui est cher, qui nécessite de l'espace et qui peut induire une dégradation du matériel en cas de dysfonctionnement [75]. Ainsi, une solution a été développée par une entreprise française. Elle permet d'encapsuler de manière automatique l'ADN préalablement séché sous vide dans une capsule à atmosphère inerte. Une stabilité de plusieurs dizaines de milliers d'années de conservation à température ambiante est estimée. De plus, chaque capsule peut contenir 0,8g d'ADN, les capsules peuvent être stockée sous formes de plaques contenant 96 unités. A usage uniques, elles sont constituées d'acier inoxydable à l'extérieur et de verre à l'intérieur. La manufacturier américain Twist Bioscience utilise déjà depuis 2019 ces capsules pour son travail sur le stockage de données sur ADN.

G. Lecture et décodage

Avant de pouvoir décoder les informations stockées il est nécessaire de lire la séquence d'ADN. La méthode de lecture des séquences des brins d'ADN est appelée le séquençage. Trois techniques principales existent.

La **première**, dite de **Sanger** du nom de son inventeur, est basée sur la polymérisation de brins d'ADN avec pour matrice le brin d'ADN à séquencer. Elle repose sur la synthèse d'un brin complémentaire à un brin de la molécule à séquencer. Les désoxyribonucléotides sont incorporés au brin synthétisé un à un. Une faible quantité de didésoxyribonucléotides marqués avec un fluorophore sont présent dans le mix réactionnel qui une fois ajoutés stoppent la polymérisation. De multiples séquences sont ainsi synthétisées en même temps permettant alors d'obtenir une multitude de brins qui auront des longueurs différentes dont le dernier nucléotide est marqué d'un fluorophore. Une étape de lecture permettra de déterminer quel nucléotide est présent à la fin de chaque brin. Grâce à leurs longueurs différentes, l'ordre des nucléotides (la séquence) sera alors déterminé lors d'une étape de séparation des brins selon leur longueur grâce à ce que l'on appelle une électrophorèse capillaire. Cette technique de première génération est principalement utilisée pour des séquençages de molécules uniques car elle n'est pas très adaptée au haut débit.

La **seconde** est appelée NGS pour "**New Generation Sequencing**" en anglais, qui signifie séquençage de nouvelle génération. Elle fait référence à des techniques plus récente qui ont émergé

en 2004 [76] et qui sont en mesure de réaliser le séquençage de centaines de milliers à plusieurs millions de courtes molécules d'ADN en parallèle. On peut les qualifier de méthodes à haut débit.

Il existe plusieurs techniques de séquençage de nouvelle génération mais la plus utilisée, et la plus décrite dans le cadre du stockage de données sur ADN est une technique basée sur les travaux des docteurs Shankar Balasubramanian et David Klenerman qui fondèrent l'entreprise Solexa en 1998, rachetée par la société Illumina en 2007. Cette technique se déroule selon les étapes suivantes : au sein de ce que l'on appelle une cellule d'écoulement, la première étape consiste à fragmenter les molécules d'ADN en de nombreux courts fragments qui sont complétés de part et d'autre par des séquences qu'on appelle adaptateurs et qui servent de point de départ à la polymérisation, au séquençage et à l'analyse. Les fragments sont ensuite fixés sur un support grâce aux adaptateurs, ce qui permet ainsi de réaliser l'amplification et le séquençage : les deux étapes ont lieu en parallèle. En effet, les nucléotides libres dans la cellule d'écoulement possèdent un bloqueur fluorescent réversible, à chaque ajout et après lavage des nucléotides libres un lecteur de fluorescence et un ordinateur analysent le nucléotide ajouté sur chaque fragment grâce à la longueur d'onde de chaque tag fluorescent. Le bloqueur est ensuite retiré par voie chimique afin de libérer l'extrémité 3' pour la suite de la synthèse et du séquençage. Ces étapes s'enchainent jusqu'à la fin de la polymérisation qui concorde donc avec la fin du séquençage (Figure 16).

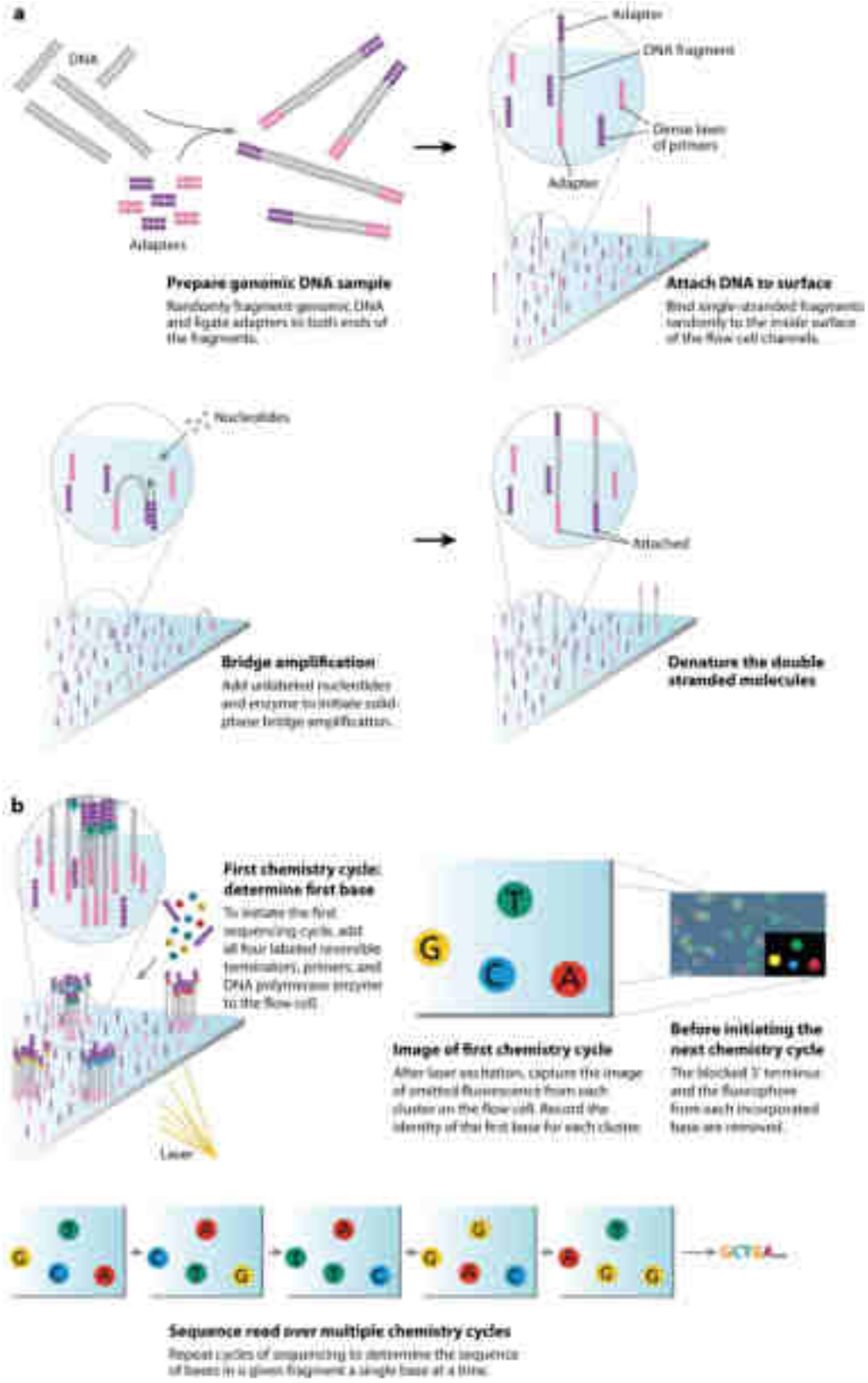

Figure 16 : Représentation de la technique de séquençage NGS développée par Illumina [77]

A. *Une fois préparés par fragmentation aléatoire et ajout d'adaptateurs de part et d'autre de chaque fragment, les échantillons d'ADN mono-brins se fixent sur la surface de la cellule d'écoulement. Les brins sont synthétisés par étape : chaque nucléotide supplémentaire est ajouté un à un. Après chaque ajout les sondes fluorescentes sont excitées par laser et l'émission de fluorescence enregistrée pour chaque emplacement (**B**). Une fois la lecture effectuée, le nucléotide ajouté est libéré par ajout de solvant et le cycle se répète jusqu'à la fin de la séquence.*

Également à haut débit, les techniques Roche 454 et Ion Torrent sont fondées sur des techniques de lecture différentes mais conservent des principes similaires concernant la fragmentation de l'ADN, l'ajout d'adaptateurs et la fixation sur un support. La première est basée sur le pyroséquençage qui consiste à détecter la libération de pyrophosphate par bioluminescence lors de l'ajout de nouveau nucléotide. La seconde ne fait pas intervenir de mesure de lumière, la lecture à lieu grâce à la détection de l'émission de proton à chaque ajout de nucléotide par changement de pH du milieu [78].

La **troisième** méthode est appelée TGS (« Third Generation Sequencing », séquençage de troisième génération en anglais) et a émergé en 2011, proposé par les sociétés Oxford Nanopore Technologies (ONT) et Pacific Bioscience [76]. Elle repose sur le passage de la molécule d'ADN à séquencer dans un nanopore. Ce dernier est une protéine insérée dans une membrane à haute résistance électrique immergée dans une solution d'électrolytes (Figure 17). Une perturbation du courant électrique caractéristique est mesurée lors du passage de chaque nucléotide dans le nanopore et permet d'identifier la molécule. La technique développée par Pacific Bioscience est ce qu'on appelle le « single-molecule real-time sequencing » en anglais, qui signifie séquençage en temps réel d'une molécule unique. Elle est fondée sur une technologie de guidage de l'énergie lumineuse vers un espace de faibles dimensions en regard de la longueur d'onde, qu'on appelle guide d'onde en mode zero (ZMW pour « zero mode waveguide » en anglais) et qui a été développé par l'entreprise pour le séquençage. Une ADN polymérase est fixée au bas d'un ZMW, ce qui crée un volume d'observation assez petit pour observer un nucléotide unique. Chaque nucléotide qui se fixe libère une sonde fluorescente lorsqu'elle sort de la zone d'observation du ZMW ce qui, grâce à un détecteur, permet de déterminer quel nucléotide a été ajouté (Figure 18) [79]. Enfin, le TGS présente l'avantage de ne pas nécessiter de phase d'amplification et permet un séquençage en temps réel de longues molécules d'ADN ou d'ARN pour un coût assez faible et un temps très raisonnable.

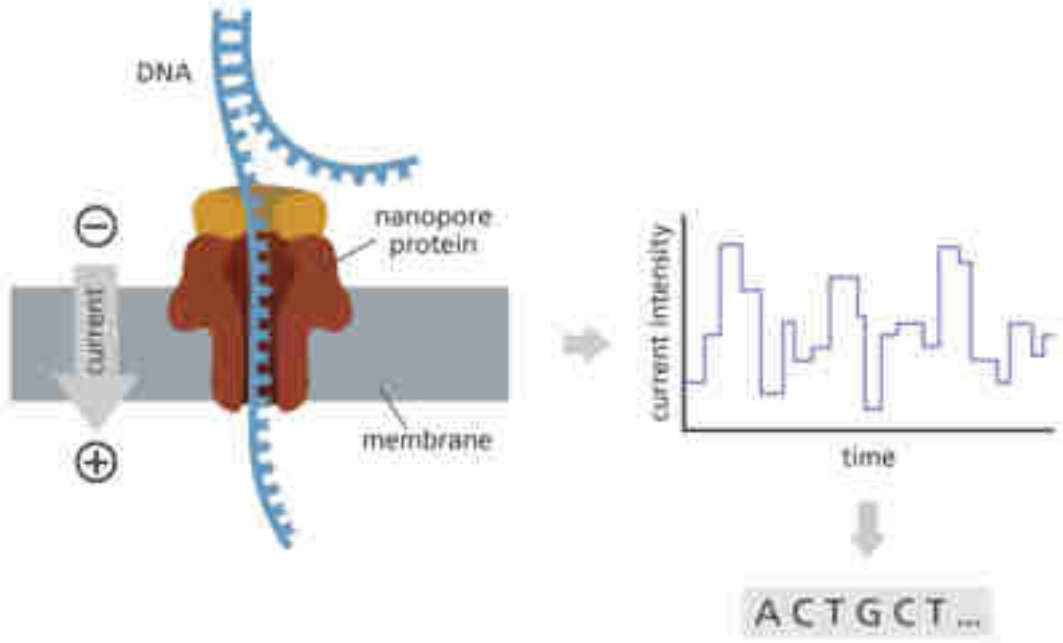

Figure 17 : Représentation de la technique TGS développée par Oxford Nanopore Sequencing [80]

Lors du séquençage, le brin d'ADN chargé négativement passe à travers la protéine transmembranaire, poussée par le courant appliqué à la membrane dans laquelle est inséré le nanopore. Ce passage induit une différence de potentiel qui est collectée par voie informatique et décodée grâce à un algorithme qui traduit la variation de potentiel en séquence ADN.

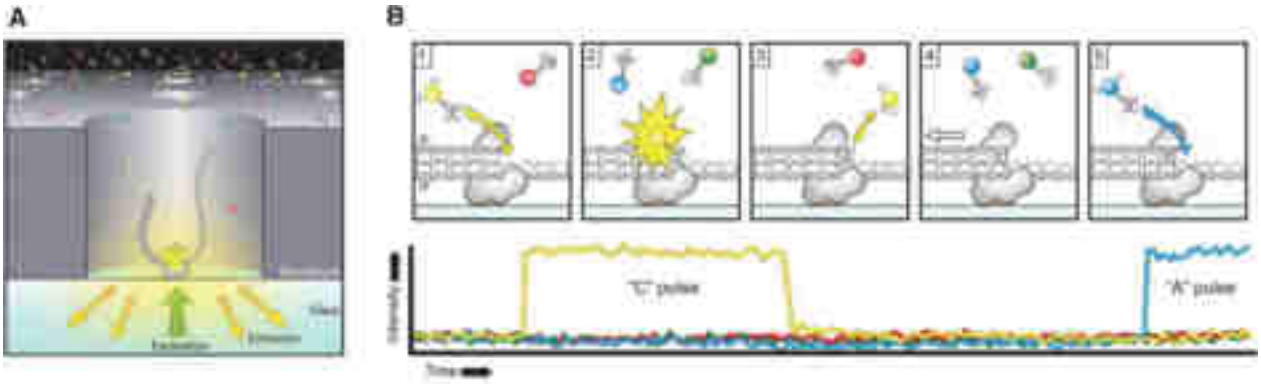

Figure 18 : Représentation de la technique TGS développée par Pacific Bioscience [79]

A. *Représentation d'un ZMW dans lequel est fixé une ADN polymérase qui synthétisera le brin complémentaire de la molécule séquencée.* ***B.*** *Chacun des 4 nucléotides est marqué avec une molécule fluorescente (G en rouge, C en jaune, T en vert et A en bleu). Lorsqu'un nucléotide est ajouté, la sonde fluorescente qui y est fixée émet à une certaine longueur d'onde.*

IV. Mise en place de la technique : défis et enjeux

A. Méthode d'accès localisée aux données

Les techniques de stockages utilisées de nos jours permettent un accès très localisé aux données : si l'on souhaite ouvrir un dossier en question, accéder à un fichier particulier, il suffit de le demander via une interface telle qu'un ordinateur et la machine ira chercher l'information demandée, en un instant. L'un des défis de la mise en place de la méthode réside en cet accès localisé à la donnée une fois codée et synthétisée sous forme de molécule d'ADN. Il n'est pas envisageable de devoir séquencer l'intégralité d'un pool d'ADN afin d'accéder à une donnée précise, cela impliquerait des coûts élevés et une nécessité de performance déraisonnable. Il existe heureusement un domaine scientifique appelé biologie moléculaire et au sein duquel l'extraction sélective de fragments d'ADN est une pratique courante réalisée en routine. Deux principales techniques existent : l'extraction par billes magnétiques et la réaction de polymérisation en chaine ou plus communément appelée PCR.

Billes magnétiques

L'extraction d'ADN par billes magnétiques est une méthode créée dans les années 1990 et qui est utilisée couramment afin d'isoler l'ADN de différentes sources biologiques, telles que le sang, les cellules, les tissus ou les échantillons environnementaux. Le principe de base de cette méthode est de lier l'ADN à des billes magnétiques et d'y appliquer un champ magnétique afin d'isoler les billes des autres composants de l'échantillon. Ensuite, l'ADN est libéré en utilisant une solution de désorption, qui permet de récupérer l'ADN purifié. Les principales étapes sont la préparation de l'échantillon afin de rendre accessible l'ADN aux billes magnétiques (lyse cellulaire par exemple). Ensuite l'ADN est lié aux billes magnétiques qui sont elles-mêmes préalablement fonctionnalisées avec des molécules qui peuvent se lier à l'ADN tel que des oligonucléotides. Un champ magnétique est ainsi appliqué afin de retenir exclusivement les billes et permettre de les séparer du reste de l'échantillon par lavage, ce qui permet d'éliminer les contaminants et les impuretés. La dernière étape consiste en l'élution des molécules d'ADN d'intérêt en utilisant une solution de désorption qui va séparer l'ADN des billes.

PCR

La PCR est une technique de biologie moléculaire inventée dans les années 1980 par Kary Mullis. Son principe (Figure 19) est de répliquer de manière exponentielle une région spécifique de l'ADN grâce à une enzyme, l'ADN polymérase et à des amorces (de courtes séquences d'ADN qui

se lient de manière spécifique de part et d'autre de la région d'ADN à amplifier). La PCR se déroule en cycles répétitifs qui sont composés de trois étapes principales : la dénaturation, l'hybridation et l'extension. Lors de la dénaturation, l'ADN double brin est chauffé à une température élevée pour séparer les deux brins d'ADN. Ensuite, lors de l'hybridation, les amorces se lient aux sites d'amorçage sur le de monobrin complémentaire. Enfin, lors de l'extension, l'ADN polymérase synthétise une nouvelle chaîne d'ADN complémentaire à chaque brin simple, en utilisant les amorces comme point de départ. A chaque cycle de PCR, le nombre total de copies de la région d'intérêt est doublé à chaque cycle de synthèse ce qui permet d'amplifier rapidement et efficacement l'ADN cible générant ainsi des millions de copies.

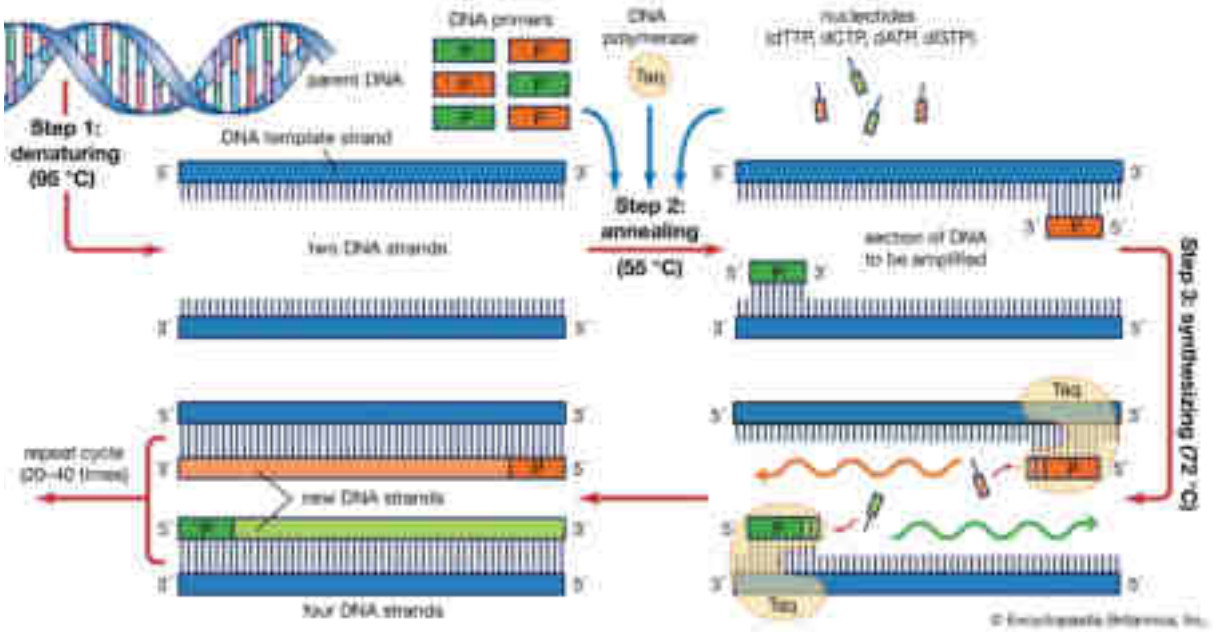

Figure 19 – Illustration des étapes de la PCR [81]

La première étape de dénaturation à 95°C sert à séparer les deux brins d'ADN afin de les rendre accessibles aux amorces spécifiques de la zone concernée. L'étape 2 d'hybridation permet aux amorces de se fixer sur chaque brin pour ensuite laisser la polymérase assembler le nouveau brin en prenant comme matrice le brin sur lequel elle est fixée et en commençant à partir de l'endroit où est fixé l'amorce. Ici l'exemple donné est celui de la Taq polymérase qui est une enzyme classique utilisée en biologie moléculaire. Une fois la synthèse effectuée, une molécule unique d'ADN a permis de générer deux molécules identiques. La même chose se répète ainsi entre 20 et 40 fois de sorte à amplifier en grande quantité la molécule désirée.

Son utilisation dans le cadre du stockage des données sur ADN pourrait fonctionner comme suit : lors du codage, le système attribue des paires d'amorces uniques aux différentes données et inclut ces sites d'amorce lors de la synthèse des séquences d'ADN correspondantes pour les données. Quand un utilisateur souhaite accéder à une donnée particulière, le système trouve les amorces correspondantes, amplifie les séquences d'ADN qui contiennent les données désirées et séquence un échantillon du résultat de PCR [61]. Le travail d'une équipe a mené au développement d'une technique d'accès localisé basé sur ce concept d'amorces spécifique à chaque brin d'ADN synthétisé

[54]. L'un des principaux défis de cette approche consiste à concevoir des amorces différentes des zones codant pour de la donnée et qui peuvent permettre une PCR multiplexe (c'est-à-dire une amplification de plusieurs fragments aux séquences différentes dans la même réaction de polymérisation en chaine) dans le cas où plusieurs éléments de données sont demandés simultanément. Cela qui implique des amorces toutes différentes les unes des autres. Une autre étude a développé une méthode utilisant les deux techniques à la fois en fixant l'ADN stocké à des billes magnétiques, y accède en réalisant une PCR directement sur ce matériel et en séquençant le résultat de cette PCR. Cela permet des lectures répétées des données tout en préservant les molécules d'ADN originelles fixées aux billes et en maintenant la qualité de lecture des données [82]. Toutefois, malgré l'utilisation d'une méthode d'accès aléatoire au sein d'un pool, il n'est pas pratique d'avoir à collecter toutes les données dans un seul pool d'ADN : des mélanges trop complexes d'ADN auront des temps de diffusion longs et conduiront à des extractions moins spécifiques. Récemment, une équipe a mis au point une technique basée sur la PCR qui est en mesure d'établir des pools physiquement isolés codant pour plusieurs téraoctets de données. Ces pools pourraient être organisés sous forme déshydratée en un réseau dense : il serait alors possible d'obtenir un système dont la densité serait de l'ordre de la bande magnétique [57]. Il sera certainement nécessaire de créer une bibliothèque de pools d'ADN physiquement isolés qui seraient récupérés à la demande afin de dépasser cette limite (avec un automate par exemple). Idéalement, il faudrait être en mesure de faire quelque chose de compact afin de continuer à bénéficier de la densité de l'ADN. C'est actuellement un domaine de recherche actif et l'un des principaux défis à relever.

B. Méthodes de correction des erreurs de synthèse

Le taux d'erreur relevé dans la littérature est de 1% par base et par position, ce qui signifie que lorsqu'un brin d'ADN est synthétisé et séquencé à nouveau, environ 1 % des lectures comporteront une erreur à une position donnée, dans le cas de la synthèse par voie chimique et de séquençage par la technique Illumina [61]. En fait, la plupart des erreurs surviendraient à l'étape de séquençage. Par exemple, dans le cas de l'utilisation de la technologie de séquençage par nanopores de la société Oxford Nanopore Technologies, le taux d'erreur peut grimper jusqu'à environ 12% [57]. Enfin, il a été démontré que les erreurs proviennent principalement de la synthèse et du séquençage et que la manipulation de l'ADN, son amplification par PCR et le stockage peuvent provoquer des effacements, c'est-à-dire une sous-représentation disproportionnée de certaines séquences dans un mélange [61].

Bien que les supports de stockage utilisés de nos jours présentent de nombreux avantages, ils n'ont pas une fiabilité de 100% : le taux d'erreur d'un support magnétique est d'environ 1% [83]. Il existe un domaine entier de l'informatique appelé théorie de l'information - ou théorie du codage - qui se concentre sur le développement de schémas de codage permettant de fournir des données numériques de manière fiable sur des supports et des canaux de communication chargés [61]. Ainsi, tout comme c'est déjà le cas pour les supports de stockages existants, il est essentiel d'ajouter des codes correcteurs d'erreurs aux supports afin de ne pas exposer l'utilisateur à un taux d'erreur qui peut impacter son utilisation. La particularité du stockage sur molécules d'ADN est qu'en plus des erreurs de substitutions qu'on retrouve dans les supports classiques, on peut faire face à des insertions ou des délétions de bases nucléiques ce qui complique le codage.

Les stratégies de correction d'erreurs sont des techniques utilisées afin de trouver et corriger les erreurs qui peuvent survenir lors de la transmission ou du stockage de données numériques. Dans les supports de stockages majoritairement utilisés de nos jours, elles peuvent provenir de plusieurs facteurs tels que les bruits dans les canaux de communication, les interférences électromagnétiques ou les défaillances matérielles. Les stratégies de corrections permettent de garantir l'intégrité de la donnée malgré les perturbations. Dans le cas du stockage de données dans l'ADN, plusieurs stratégies existent. La principale est la redondance physique qui consiste à incorporer plusieurs copies identiques de la même séquence d'ADN dans le support. Cela permet d'augmenter la probabilité de retrouver les informations originellement codées malgré des données manquantes dues aux erreurs de synthèse et de séquençage. Il existe deux types fondamentaux de redondance : la redondance physique qui consiste en la présence de nombreuses copies physiques d'une séquence d'ADN donnée ; et la redondance logique dont il existe plusieurs types et qui provient de l'intégration d'informations supplémentaires lors de l'encodage de la donnée en séquence d'ADN [61]. Les deux types sont complémentaires : des travaux ont démontré qu'en absence de redondance logique et malgré un taux de redondance physique élevé il n'était pas possible d'obtenir un taux d'erreur de l'ordre de zero bit [52]. Toutefois, nombreux sont les projets ayant développé un code permettant de réduire les erreurs de lecture et d'écriture et faire face à la dégradation de l'ADN dans le temps [54,57,84–86]. Pour être plus exact, c'est au moment où les données numériques sont codées sous forme de séquences d'ADN que les bits subissent une série de transformations et de contrôles. Une fois les séquences encodées, des séquences étiquettes sont ajoutées afin de permettre de retrouver l'emplacement de la donnée codée dans le fichier d'origine. Dans le but de générer une variation entre les différents brins d'ADN encodés, les séquences peuvent être soumises à une opération logique appelée « disjonction exclusive » ou encore « OU exclusif », la redondance est ensuite ajoutée via l'utilisation de codes particuliers développés pour cela [61]. La redondance est ajoutée par un algorithme de codage qui ajoutera des bits de redondance aux données d'origine. Cet encodage sera

réalisé de manière différente selon l'algorithme utilisé. Ainsi, si des erreurs sont détectées au moment du décodage, l'algorithme sera en mesure d'en corriger un maximum afin d'obtenir les données dans leur état d'origine.

Les codes correcteurs d'erreurs ont été développés à partir des années 1940, il en existe plusieurs aujourd'hui mais le premier du nom vit le jour en 1950 : le code de Hamming [87]. Il est basé sur l'ajout de bits de redondance. Il est efficace pour corriger des erreurs simples mais ne peut corriger qu'une seule erreur à la fois, si le nombre d'erreurs augmente son efficacité diminue. L'un des algorithmes le plus utilisé pour le stockage traditionnel comme avec les techniques de stockage optiques (DVD, Blu-ray) est le codage de Reed-Solomon. C'est une technique dont l'origine provient des années 1960 [88] et qui est également régulièrement utilisée dans le cadre du stockage de données dans l'ADN [57,84,89]. Le concept du code de Reed-Solomon repose sur la conversion des données originales en un ensemble de symboles. Ces symboles sont ensuite associés aux coefficients d'un système d'équations linéaires, dont les solutions sont liées aux données d'origine. Ce système a la capacité de résoudre les problèmes liés à la perte de symboles (lorsqu'un symbole manque) et aux erreurs (lorsqu'un symbole original est altéré). C'est là l'avantage principal de ce code : il est capable de corriger à la fois les erreurs de perte de symboles et les erreurs de symboles corrompus, le tout en utilisant un seul et même code [61].

D'autres méthodes existent pour la correction des erreurs comme celle proposée par Goldman et al. [53] dont le principe repose sur la réplication du code d'une donnée précise et son placement sur plusieurs fragments différents et à des positions différentes. Cette stratégie permet d'éviter les erreurs systématiques liées à la position sur le brin lors des étapes de synthèse et de séquençage. Une autre méthode basée sur ce que l'on appelle les codes fontaines a été utilisée par Erlich et Zielinski [86]. Les codes ADN fontaines sont des codes correcteurs d'erreurs qui ont été conçus spécialement pour le stockage de données sur ADN. Leur nom provient de l'idée de "fontaines" desquelles les gouttes d'eau sont émises de manière aléatoire et indépendante. De la même manière, les codes ADN fontaines divisent les données en petites unités surnommées "gouttes" et en créent un grand nombre contenant des morceaux de l'information et de la redondance. Ces gouttes sont réparties aléatoirement dans des brins d'ADN synthétisés. Lorsque les données sont récupérées, les brins d'ADN sont séquencés en parallèle et les informations des différentes gouttes sont combinées pour restaurer les données originales. Cette approche permet une détection et une correction efficaces des erreurs tout en utilisant efficacement la capacité de stockage de l'ADN. Enfin, en 2018, une étude a proposé une méthode reposant sur l'utilisation, en plus des quatre bases nucléiques, de onze bases dégénérées pour l'encodage des données. Cette méthode a permis d'augmenter quasiment d'un facteur deux la densité

de stockage en comparaison à la méthode utilisant des codes fontaines expliquée précédemment. Les auteurs estiment qu'il serait possible d'obtenir une réduction du coût du stockage d'environ 50% [90].

En conclusion, il existe de nombreuses méthodes de correction des erreurs de synthèse de l'ADN dans le cadre de son utilisation pour stocker des données. Les algorithmes de correction d'erreurs basés sur la redondance ou les répétitions de séquences sont tous deux des moyens efficaces de minimiser les erreurs de synthèse de l'ADN, mais le choix de la méthode dépend des besoins et des contraintes du système en question. Il est important de continuer à explorer et à développer de nouvelles méthodes pour améliorer la qualité et l'intégrité des données stockées dans l'ADN.

C. Stockage *in vivo*

Bien que la plupart des groupes de recherche aient travaillé sur le développement d'une méthode de stockage *in vitro*, certains ont élaboré une méthode *in vivo* du stockage de donnée dans l'ADN [91,92]. Cela implique de stocker les données dans l'ADN génomique de cellules vivantes, qui est devenu un support idéal en raison de sa durabilité et sa compatibilité bio-fonctionnelle. A la différence du stockage *in vitro*, la méthode *in vivo* tire parti des mécanismes cellulaires de réplication, de relecture et de maintenance de l'ADN. Le développement de la biologie synthétique et d'édition des gènes nous a permis de modifier l'information génétique de manière très précise. Il est possible d'utiliser les enzymes artificielles ainsi que les enzymes naturelles de ciblage et de modification de l'ADN en tant que moyen d'écriture dans les systèmes de stockage d'ADN. De plus, les possibilités sont en constante évolution et amélioration en ce qui concerne la manipulation de l'ADN *in vivo*, les outils de manipulation de l'ADN évoluent rapidement [93,94]. A titre d'exemple, une équipe de chercheurs américain a réussi à insérer le code d'images en noir et blanc (2,6 Ko d'informations) dans le génome d'une population bactérienne vivante grâce au système CRISPR-Cas [91]. Pour la récupération des données, l'équipe a collecté et séquencé l'ADN des différentes cellules puis procédé à un alignement de séquences.

Plusieurs types d'outils d'écriture de l'ADN existent et ils peuvent être classés sur la base de mutations qui résultent de leur utilisation : il y a les outils précis et ceux qu'on peut qualifier de pseudo-aléatoires [94]. Les outils d'écritures d'ADN dits précis tels que les transcriptases inverses ou les recombinases à site spécifique génèrent des mutations déterminées. En ce qui concerne les outils pseudo-aléatoires comme les nucléases à site spécifique ou bien le système Cas1-Cas2, ils induisent des mutations ciblées mais stochastiques.

Outils précis

Les recombinases site-spécifique sont des enzymes efficaces et précises qui sont capables d'insérer, exciser ou encore inverser un morceau d'ADN entre les sites de reconnaissances apparentés à l'enzyme. Pour citer un exemple, le système de recombinaison Cre-LoxP est un système connu et régulièrement utilisé en biologie synthétique. L'utilisation de ces enzymes a pour effet de stocker les informations de manière héréditaire dans un emplacement du génome spécifique [95]. Néanmoins, une écriture réversible de l'information est possible grâce aux propriétés d'une autre enzyme : l'excisionase. Elle est en mesure d'effacer l'information préalablement intégrée et permet une sorte de réinitialisation de l'état de l'ADN [96]. La seconde classe des outils précis est les transcriptases inverses avec par exemple le système SCRIBE (Synthetic Cellular Recorders Integrating Biological Events) dont l'activation répond à un stimulus tel qu'un produit chimique et qui a pour effet l'introduction de mutations précises et programmables [97].

Outils pseudo-aléatoires

Les outils pseudo-aléatoires reposent sur des cassures d'ADN double-brin ciblées qui sont réalisées par des nucléases site-spécifiques [94] tel que Cas9, les nucléases à doigt de zinc (ZFN) ou encore les TALEN (nucléases effectrices de type activateur de transcription). Une seconde classe utilise la fonction immunitaire cellulaire du système Cas1-Cas2 qui permet l'intégration de courts fragments d'ADN simple brin qui codent pour de l'information et longs d'environ 20-30pb dans le réseau CRISPR (qui est lui-même localisé dans le locus CRISPR) de manière orientée [98].

Il est important de considérer, pour le stockage de donnée dans l'ADN génomique de cellules vivantes (*in vivo*), la quantité maximale d'informations qu'une seule cellule peut transporter. A ce jour c'est *Escherichia coli* qui est le procaryote le plus étudié à ce sujet en raison de sa facilité d'utilisation et de son utilisation massive en biologie moléculaire. Toutefois, d'autres micro-organismes pourraient être utilisés tels que *Bacillus subtilis* - dont le génome fait 4,2 Mb - dans lequel une équipe a réussi à encoder le génome d'une autre souche (*Synechocystis*) dont le génome fait 3,5 Mb [99]. Ceci démontre la capacité de certains micro-organismes à supporter beaucoup plus d'information que leur propre génome. Toutefois, une incompatibilité certaine existe entre l'ADN génomique de l'hôte et l'ADN synthétique et représente l'un des principaux défis pour le stockage de l'ADN *in vivo*. En outre, l'ADN synthétique n'est théoriquement pas en mesure de former des cadres de lectures ouverts (donc n'est pas en mesure de conduire à l'expression de protéines) mais une mauvaise expression peut apparaitre lorsque le volume de stockage augmente à cause des erreurs, les conséquences biologiques doivent être examinées. Pour les cellules eucaryotes le défi est encore plus vaste car ce type de cellules à un fonctionnement beaucoup moins rudimentaire que les cellules procaryotes.

D. Exploration des enjeux et potentiels du stockage de données de santé dans l'ADN

Générées par les centres de recherche et par l'industrie pharmaceutique, les données de santé sont recueillies pour un usage particulier et son cruciales pour la prise de décisions médicales, la recherche, le développement de médicaments, les soins aux patients, les remboursements de ces derniers, etc. Selon le règlement général sur la protection des données (RGPD) qui est une réglementation de l'union européenne, les données de santé sont définies comme « les données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de soins de santé, qui révèlent des informations sur l'état de santé de cette personne » [100]. On y retrouve des informations relatives à une personne physique, à un examen médical, à une maladie ou encore des données génomiques ou d'imagerie qui sont stockées sous la forme de dossiers médicaux. Elles peuvent être utilisées par des systèmes aux puissances de calculs importantes tels que des systèmes dotés d'intelligence artificielle ou d'apprentissage automatique (« machine learning » en anglais) afin de préciser un diagnostic ou pour aider les médecins dans leur prise de décision [101]. Leur caractère confidentiel en fait des données particulières qui doivent être protégées car elles contiennent des données personnelles : chaque dossier médical contient les antécédents de santé et de traitements d'un patient, des informations essentielles à sa santé et à la réussite de son parcours de soin [102]. Le stockage de ces données à long terme est aussi un élément important comme leur évolutivité face à l'augmentation constante de leur volume qui dans le monde a été multiplié par 15 en 7 ans passant de 153 Exaoctets (Eo) en 2013 à 2314 Eo en 2020 [103]. A titre d'exemple, l'article L. 1142-28 du Code de la Santé publique porte à 10 ans le stockage d'un dossier médical à partir de la consolidation d'une maladie ou d'une blessure. De plus, l'archivage des documents essentiels d'une étude clinique par le promoteur et l'investigateur doit être de 10 à 40 années lorsque l'étude concerne un produit dérivé du sang. Enfin, comme ces données de santé sont souvent partagées entre différents acteurs tels que les fournisseurs de soins de santé, les chercheurs, les institutions de recherche et l'industrie pharmaceutique, il est essentiel que la technologie de stockage utilisée facilite la collaboration tout en garantissant des niveaux de contrôle et de confidentialité appropriés. Toutes ces caractéristiques particulières nécessitent une méthode de stockage qui puisse s'adapter et respecter ces critères. L'utilisation d'une technologie telle que le stockage de données numériques de santé dans des molécules d'ADN présente de nombreux enjeux complexes et cruciaux afin de garantir le respect de la vie privée des individus, la protection et la conformité de ces données aux normes éthiques et juridiques.

Comme le prévoit la réglementation, le traitement des données de santé est soumis au consentement libre, spécifique, éclairé et univoque du patient. Étant donné le caractère confidentiel

de ces données, il est essentiel qu'elles bénéficient d'un stockage disposant d'un niveau de sécurité élevé : aujourd'hui en France les hébergeurs de données de santé doivent être certifiés selon la norme ISO27001. Cette norme fournit un cadre afin de mettre en œuvre, gérer et améliorer en continu un système de gestion de la sécurité des données au sein d'une organisation pour en protéger le contenu sensible contre les risques éventuels et les menaces. Ces entités doivent également obtenir l'accréditation HDS (Hébergeur de Données de Santé) par un organisme agréé par le Comité Français d'Accréditation [104,105] qui s'assure que les données de santé sont stockées et gérées de manière sécurisée et conforme à la réglementation. Aujourd'hui la réglementation ne permet pas de certifier de cette manière une technologie telle que le stockage sur ADN car elle n'existe pas mais à mesure qu'elle évolue et devient plus répandue, et en considérant les besoins croissants il est possible que la réglementation soit adaptée ou qu'une réglementation spécifique soit élaborée pour encadrer l'utilisation de l'ADN comme support total ou partiel des données de santé. De plus, l'accès à la donnée stockée sur ADN impliquera une interaction entre le support biologique et un opérateur (humain ou robotisé) : cela ajoute une strate de sécurité supplémentaire face au stockage sur un support informatique classique la protégeant davantage face à des cyberattaques ou piratages. Ainsi, du point de vue de la sécurité et de la confidentialité, l'ADN représente une solution compatible et prometteuse pour le stockage de données de santé, comme le propose une équipe de chercheurs indiens [106].

Néanmoins, le défi ne s'arrête pas là : le volume croissant de ces données, généré par des systèmes médicaux de plus en plus sophistiqués ou la recherche génomique rendent impératif le besoin d'une méthode qui serait en mesure de faire face à cette croissance exponentielle et à long terme [107]. Les dossiers médicaux mais également les séquences génomiques de plus en plus étudiées ou encore l'imagerie médicale de haute résolution occupent de massifs espaces de stockage et il est nécessaire de les conserver sur de très longues périodes. La technologie semble être ici aussi une solution adaptée en raison de sa grande densité et de sa durabilité. Aussi, le secteur de la santé nécessite une interopérabilité transparente entre les multiples acteurs que sont les hôpitaux, les cliniques, les laboratoires de recherche et l'industrie pharmaceutique afin d'assurer une prise en charge optimale des patients, favoriser la collaboration en recherche et accélérer le développement de médicaments. Toutefois, cette interopérabilité est souvent entravée par des problèmes techniques, des protocoles différents et des systèmes qui ne sont pas compatibles [108]. Bien que le stockage de données sur ADN pourrait éventuellement atténuer certains de ces obstacles en fournissant un support de stockage universel dans le cas de son adoption généralisée et pourrait faciliter la coordination entre les différents acteurs, il est clair que la technologie est pour le moment loin d'être adaptée à cette utilisation.

En conclusion, le stockage des données de santé dans l'ADN présente un fort potentiel afin de répondre aux besoins de sécurité, de confidentialité et de stockage à long terme mais la technologie est loin d'être au point pour permettre d'être adoptée de manière généralisée et ainsi de s'adapter à la problématique d'interopérabilité des données de santé.

V. Discussion et conclusion

Selon l'état actuel de la technologie, l'utilisation de l'ADN pour stocker la donnée ne peut être dans un premier temps adapté qu'au stockage de données froides. En effet, le stockage d'archives peut tolérer des latences plus élevées et pourrait alors bénéficier d'une empreinte énergétique et environnementale bien inférieure à celle d'aujourd'hui. De plus, les centres de données d'aujourd'hui stockent les données de manière redondantes, c'est-à-dire qu'une même information est stockée plusieurs fois dans plusieurs centres différents afin de palier à d'éventuels problèmes techniques ou d'origine environnementale. Ainsi, une application crédible du stockage sur ADN serait par exemple d'encoder une ou plusieurs copies de données encodées en centre de données sur des supports classiques, ce qui aurait pour effet une diminution du volume de données à stocker sur les supports déjà existants. Il est cependant plus que nécessaire que la technologie évolue et s'améliore afin de concurrencer les supports actuels dont l'ordre de grandeur du débit d'écriture et de lecture est du gigaoctet par seconde. De plus, un autre aspect et qui est un frein à son utilisation est son coût. En 2016, le stockage sur disque dur coûtait environ 39 USD par téraoctet (contre 272 USD en 2008) et a diminué d'environ 20% par an entre 2013 et 2016 [109]. Il est difficile d'estimer précisément les coûts de synthèse de l'ADN car c'est une information généralement confidentielle. Toutefois, Robert Carlson, un analyste de l'industrie, estime que le coût de la synthèse de matrice ADN est d'environ 0,0001 USD par base [110], ce qui correspond à 800 millions d'USD par téraoctet soit plus de 10 millions de fois plus cher que pour un disque dur. Malgré de grands écarts, plusieurs axes peuvent être optimisés pour les réduire : la précision peut être sacrifiée au profit de la vitesse ; la redondance physique peut également être considérablement réduite (grâce à l'utilisation de codes d'erreurs). Aussi, l'amélioration de la mise à l'échelle et des performances des méthodes de synthèse et de séquençage seront accompagnées de réductions des coûts proportionnelles. Il faut enfin prendre en considération un aspect important de la technologie : le stockage physique. Bien qu'il ait été découvert de l'ADN vieux de plusieurs millions d'années [111], il peut se dégrader beaucoup plus rapidement que cela s'il est stocké dans de mauvaises conditions. Les solutions sont multiples (encapsulation, milieu aqueux, sels, nanoparticules, …). Le choix de la méthode de stockage aura également un impact important sur l'automatisation nécessaire. En effet, aujourd'hui, en dehors des étapes de séquençage et de synthèse, la majorité des manipulations sont encore réalisées par un opérateur en laboratoire.

Pour conclure, notre relation à la donnée s'est totalement transformée avec la révolution numérique, nous faisant ainsi entrer dans l'ère de l'information. Les quantités et les types de données que nous créons et que nous stockons sont en constante évolution et expansion. Ce que nous générons

dépassera rapidement nos capacités de stockages technologiques actuelles. De nouvelles formes sont donc nécessaires afin de suivre le rythme et l'utilisation de l'ADN dans ce but semble représenter une alternative prometteuse. En outre, les technologies relatives à l'ADN (séquençage, synthèse, récupération), initialement développées pour des applications en science de la vie, peuvent et seront réutilisées dans les systèmes de stockage de données. Elles sont notamment en constante amélioration pour les besoins scientifiques. Dans le même temps, les recherches sur le stockage de données sur ADN continuent de progresser, promettant de réduire progressivement les obstacles à son adoption à plus grande échelle. Enfin, le stockage de données sur ADN présente un fort potentiel afin de répondre aux besoins de sécurité, de confidentialité et de stockage à long terme des données de santé mais la technologie est loin d'être au point pour permettre d'être adoptée de manière généralisée et ainsi de s'adapter aux problématiques des données de santé.

Bibliographie

1. D'Errico F. L'origine du stockage de l'information [Internet]. Pourlascience.fr. 2001 [cité 7 juill 2022]. Disponible sur: https://www.pourlascience.fr/sd/informatique/https:https://www.pourlascience.fr/sd/informatique/l-origine-du-stockage-de-l-information-4486.php

2. China's Warrior Queen - Fu Hao [Internet]. China's Warrior Queen - Fu Hao. 2022 [cité 27 août 2022]. Disponible sur: https://www.arte.tv/fr/videos/102277-000-A/la-chine-a-l-age-du-bronze-fu-hao-reine-et-guerriere/

3. Bastien L. Stockage de données : mais en fait, qu'est-ce que c'est ? [Internet]. LeBigData.fr. 2019 [cité 1 avr 2022]. Disponible sur: https://www.lebigdata.fr/stockage-donnees

4. Big data | CNIL [Internet]. [cité 28 août 2022]. Disponible sur: https://www.cnil.fr/fr/definition/big-data#:~:text=On%20parle%20depuis%20quelques%20ann%C3%A9es,%2C%20photos%2C%20vid%C3%A9os%2C%20etc.

5. Petroc T. Data growth worldwide 2010-2025 [Internet]. Statista. 2023 [cité 5 oct 2023]. Disponible sur: https://www.statista.com/statistics/871513/worldwide-data-created/

6. Académie des Technologies. Archiver les mégadonnées au-delà de 2040 : la piste de l'ADN [Internet]. Académie des technologies. 2020 [cité 28 mai 2022]. Disponible sur: https://www.academie-technologies.fr/publications/archiver-les-megadonnees-au-dela-de-2040-la-piste-de-ladn/

7. Twist Bioscience. DNA-Based Digital Storage [Internet]. 2022. Disponible sur: https://www.twistbioscience.com/sites/default/files/resources/2019-03/WhitePaper_DataStorage_29Oct18_Rev1.pdf

8. Infographie: Le Big Bang du Big Data [Internet]. Statista Daily Data. 2021 [cité 17 oct 2023]. Disponible sur: https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde

9. Futura. Définition | Bande magnétique | Futura Maison [Internet]. Futura. [cité 13 sept 2023]. Disponible sur: https://www.futura-sciences.com/maison/definitions/maison-bande-

magnetique-11072/

10. Polytech. L'automatisme dans l'ordinateur [Internet]. [cité 13 sept 2023]. Disponible sur: http://users.polytech.unice.fr/~strombon/Formation/TL.2000/Groupe2/Memoires_Secondaires/index.html

11. Royas C. L'évolution du stockage des données [Internet]. Nectar du Net. 2023 [cité 13 sept 2023]. Disponible sur: https://www.nectardunet.com/7943/evolution-du-stockage-des-donnees/

12. Chaillot M. La folle évolution du stockage informatique [Internet]. Capital.fr. 2014 [cité 13 sept 2023]. Disponible sur: https://www.capital.fr/economie-politique/la-folle-evolution-du-stockage-informatique-953110

13. Delprato JM. De la carte perforée à la mémoire flash : la grande histoire du stockage [Internet]. Tom's Hardware. 2022 [cité 13 sept 2023]. Disponible sur: https://www.tomshardware.fr/de-la-carte-perforee-a-la-memoire-flash-la-grande-histoire-du-stockage-des-donnees/

14. Morizur M. L'EVOLUTION DES SUPPORTS ET DU STOCKAGE DES DONNEES [Internet]. [cité 13 sept 2023]. Disponible sur: http://icn13.alwaysdata.net/evolution%20support%20stockage.html

15. Bohic C. { SILICON - 20 ANS } - Clé USB : quand le stockage devient vraiment mobile [Internet]. Silicon. 2020 [cité 13 sept 2023]. Disponible sur: https://www.silicon.fr/silicon-20-ans-cle-usb-quand-le-stockage-devient-vraiment-mobile-352676.html

16. Amazon. Présentation d'Amazon Web Services [Internet]. Amazon Web Services; 2023. Disponible sur: https://docs.aws.amazon.com/fr_fr/whitepapers/latest/aws-overview/aws-overview.pdf

17. Grandmontagne Y. Silicon [Internet]. Silicon. 2013 [cité 13 sept 2023]. Disponible sur: https://www.silicon.fr/

18. Fontaine S. Le cloud public, un marché à 227,8 milliards en 2019 [Internet]. Techniques de l'Ingénieur. 2019 [cité 13 sept 2023]. Disponible sur: https://www.techniques-ingenieur.fr/actualite/articles/le-cloud-public-un-marche-a-2278-milliards-en-2019-73365/

19. Cloud privé et Cloud public : quelle est la différence ? | Glossaire VMware | FR [Internet].

[cité 13 sept 2023]. Disponible sur: https://www.vmware.com/fr/topics/glossary/content/private-cloud-vs-public-cloud.html#:~:text=Pour%20r%C3%A9sumer%2C%20les%20Clouds%20publics,pare%2Dfeu%20et%20physiquement%20s%C3%A9curis%C3%A9e

20. Intel. Qu'est-ce que le Cloud hybride ? [Internet]. Qu'est-ce que le Cloud hybride ? Disponible sur: https://www.intel.fr/content/www/fr/fr/cloud-computing/what-is-hybrid-cloud.html#:~:text=Les%20avantages%20du%20Cloud%20hybride,d'investissement%20du%20Cloud%20public

21. The Evolution of Data Storage (Infographic) [Internet]. [cité 17 oct 2023]. Disponible sur: https://www.slideshare.net/GoCanvas/historyofdatastor

22. Infographie: La totalité des données créées dans le monde équivaut à... [Internet]. Statista Daily Data. 2019 [cité 17 oct 2023]. Disponible sur: https://fr.statista.com/infographie/17793/quantite-de-donnees-numeriques-creees-dans-le-monde

23. Bressange G. Le Big Data [Internet]. Acadys. 2022 [cité 16 sept 2023]. Disponible sur: https://www.acadys.com/le_big_data/

24. L'évolution du stockage en 50 ans – infographie [Internet]. Access Group - conseil et solutions informatiques, ingénierie réseaux, télécoms, cloud et câblage informatique. 2013 [cité 17 oct 2023]. Disponible sur: https://www.access-group.fr/news/news-evolution-du-stockage-en-infographie/

25. Accompagner l'explosion des volumes de données : les nouveaux enjeux du stockage - Livre blanc - ZDNet [Internet]. [cité 13 sept 2023]. Disponible sur: https://www.zdnet.fr/livre-blanc/accompagner-l-explosion-des-volumes-de-donnees-les-nouveaux-enjeux-du-stockage-63748637.htm

26. Blue Soft Group. Cybersécurité : 3 types de cyberattaques auxquelles votre entreprise est exposée [Internet]. Blue Soft Group. 2021 [cité 17 oct 2023]. Disponible sur: https://www.bluesoft-group.com/cybersecurite-3-types-de-cyberattaques-auxquelles-votre-entreprise-est-exposee/

27. Alliance pour la confiance numérique. Observatoire de la Filière de la Confiance Numérique. 2022.

28. Achite-Henni M. Pollution des data centers : comment la réduire ? [Internet]. Carbo. 2022

[cité 14 sept 2023]. Disponible sur: https://www.hellocarbo.com/blog/communaute/pollution-des-data-centers-comment-la-reduire/

29. Infographie: Les cyberattaques les plus courantes contre les entreprises françaises [Internet]. Statista Daily Data. 2021 [cité 17 oct 2023]. Disponible sur: https://fr.statista.com/infographie/15871/types-de-cyberattaques-les-plus-courantes-entreprises-francaises

30. Avenier S. Le big data ne peut pas être un défi dans le monde d'aujourd'hui [Internet]. 2022 [cité 14 sept 2023]. Disponible sur: https://www.journaldunet.com/ebusiness/crm-marketing/1515349-le-big-data-ne-peut-pas-etre-un-defi-dans-le-monde-d-aujourd-hui/

31. Infographie: L'informatique entre dans l'ère quantique [Internet]. Statista Daily Data. 2021 [cité 17 oct 2023]. Disponible sur: https://fr.statista.com/infographie/26275/prevision-evolution-taille-chiffre-affaires-marche-informatique-quantique

32. Pray LA. Discovery of DNA Double Helix: Watson and Crick | Learn Science at Scitable [Internet]. 2008 [cité 21 juin 2022]. Disponible sur: http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397

33. NIH. Deoxyribonucleic Acid (DNA) Fact Sheet [Internet]. Genome.gov. 2020 [cité 21 juin 2022]. Disponible sur: https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet

34. Watson JD, Crick FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. avr 1953;171(4356):737-8.

35. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. août 2016;14(8):e1002533.

36. Thanbichler M, Wang SC, Shapiro L. The bacterial nucleoid: A highly organized and dynamic structure. J Cell Biochem. 15 oct 2005;96(3):506-21.

37. Zhirnov V, Zadegan RM, Sandhu GS, Church GM, Hughes WL. Nucleic acid memory. Nat Mater. avr 2016;15(4):366-70.

38. Épigénétique et cancer [Internet]. Planet-Vie. [cité 1 août 2022]. Disponible sur: https://planet-vie.ens.fr/thematiques/sante/pathologies/epigenetique-et-cancer

39. Nguyen HH, Park J, Park SJ, Lee CS, Hwang S, Shin YB, et al. Long-Term Stability and Integrity of Plasmid-Based DNA Data Storage. Polymers. janv 2018;10(1):28.

40. Kaplan M. DNA has a 521-year half-life. Nature [Internet]. 10 oct 2012 [cité 23 juin 2022]; Disponible sur: https://www.nature.com/articles/nature.2012.11555

41. Office of the Director of National Intelligence. Molecular Information Storage [Internet]. Intelligence Advanced Research Projects Activity. [cité 13 oct 2023]. Disponible sur: https://www.iarpa.gov/research-programs/mist

42. Wiener N. Machines Smarter Than Men? Interview with Dr. Norbert Wiener, Noted Scientist [Internet]. 1964 [cité 21 juin 2022]. Disponible sur: https://profiles.nlm.nih.gov/spotlight/bb/catalog/nlm:nlmuid-101584906X7699-doc

43. Extance A. How DNA could store all the world's data. Nature. 1 sept 2016;537(7618):22-4.

44. Davis J. Microvenus. Art J. 1 mars 1996;55(1):70-4.

45. Clelland CT, Risca V, Bancroft C. Hiding messages in DNA microdots. Nature. juin 1999;399(6736):533-4.

46. Bancroft C, Bowler T, Bloom B, Clelland CT. Long-Term Storage of Information in DNA. Science. 7 sept 2001;293(5536):1763-5.

47. Wong PC, Wong KK, Foote H. Organic data memory using the DNA approach. Commun ACM. janv 2003;46(1):95-8.

48. Arita M, Ohashi Y. Secret signatures inside genomic DNA. Biotechnol Prog. oct 2004;20(5):1605-7.

49. Yachie N, Sekiyama K, Sugahara J, Ohashi Y, Tomita M. Alignment-based approach for durable data storage into living organisms. Biotechnol Prog. avr 2007;23(2):501-5.

50. Ailenberg M, Rotstein O. An improved Huffman coding method for archiving text, images, and music characters in DNA. BioTechniques. sept 2009;47(3):747-54.

51. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science. 2 juill 2010;329(5987):52-6.

52. Church GM, Gao Y, Kosuri S. Next-Generation Digital Information Storage in DNA. Science. 28 sept 2012;337(6102):1628-1628.

53. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature. 7 févr 2013;494(7435):77-80.

54. Yazdi SMH, Yuan Y, Ma J, Zhao H, Milenkovic O. A Rewritable, Random-Access DNA-Based Storage System. Sci Rep. 18 sept 2015;5(1):14138.

55. Yazdi SMHT, Gabrys R, Milenkovic O. Portable and Error-Free DNA-Based Data Storage. Sci Rep. 10 juill 2017;7(1):5011.

56. Lee HH, Kalhor R, Goela N, Bolot J, Church GM. Enzymatic DNA synthesis for digital information storage [Internet]. bioRxiv; 2018 [cité 22 juin 2022]. p. 348987. Disponible sur: https://www.biorxiv.org/content/10.1101/348987v1

57. Organick L, Ang SD, Chen YJ, Lopez R, Yekhanin S, Makarychev K, et al. Random access in large-scale DNA data storage. Nat Biotechnol. mars 2018;36(3):242-8.

58. Shankland S. Startup Catalog has jammed all 16GB of Wikipedia's text onto DNA strands [Internet]. CNET. 2019 [cité 23 juin 2022]. Disponible sur: https://www.cnet.com/tech/computing/startup-packs-all-16gb-wikipedia-onto-dna-strands-demonstrate-new-storage-tech/

59. Press - CATALOG Website [Internet]. CATALOG. [cité 23 juin 2022]. Disponible sur: https://www.catalogdna.com/press

60. Bouleau C. Pourquoi la biotech DNA Script a levé 200 millions de dollars [Internet]. Challenges. 2022 [cité 23 juin 2022]. Disponible sur: https://www.challenges.fr/club-entrepreneurs/pourquoi-la-biotech-dna-script-a-leve-200-millions-de-dollars_812810

61. Ceze L, Nivala J, Strauss K. Molecular digital data storage using DNA. Nat Rev Genet. août 2019;20(8):456-66.

62. Hao Y, Li Q, Fan C, Wang F. Data Storage Based on DNA. Small Struct. 2021;2(2):2000046.

63. Xu C, Zhao C, Ma B, Liu H. Uncertainties in synthetic DNA-based data storage. Nucleic

Acids Res. 4 juin 2021;49(10):5451-69.

64. Bony S. La compression de données - La compression de Lempel Ziv Welch [Internet]. [cité 13 août 2023]. Disponible sur: http://igm.univ-mlv.fr/~dr/XPOSE2013/La_compression_de_donnees/lzw.html

65. Beaucage SL, Caruthers MH. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. Tetrahedron Lett. 1 janv 1981;22(20):1859-62.

66. Hughes RA, Ellington AD. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. Cold Spring Harb Perspect Biol. 3 janv 2017;9(1):a023812.

67. Eisenstein M. Enzymatic DNA synthesis enters new phase. Nat Biotechnol. 1 oct 2020;38(10):1113-5.

68. Phosphoramidite Chemistry for DNA Synthesis | Twist Bioscience [Internet]. [cité 17 oct 2023]. Disponible sur: https://www.twistbioscience.com/fr/blog/science/simple-guide-phosphoramidite-chemistry-and-how-it-fits-twist-biosciences-commercial

69. Perkel JM. The race for enzymatic DNA synthesis heats up. Nature. 25 févr 2019;566(7745):565-565.

70. Palluk S, Arlow DH, de Rond T, Barthel S, Kang JS, Bector R, et al. De novo DNA synthesis using polymerase-nucleotide conjugates. Nat Biotechnol. août 2018;36(7):645-50.

71. Lee HH, Kalhor R, Goela N, Bolot J, Church GM. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. Nat Commun. 3 juin 2019;10(1):2383.

72. Matange K, Tuck JM, Keung AJ. DNA stability: a central design consideration for DNA data storage systems. Nat Commun. 1 mars 2021;12(1):1358.

73. Byron J, Long D, Miller E. Measuring the cost of reliability in archival systems. In: Proceeding of the Conference on Mass Storage Systems and Technologies (MSST'20). 2020.

74. Lim CK, Nirantar S, Yew WS, Poh CL. Novel Modalities in DNA Data Storage. Trends Biotechnol. oct 2021;39(10):990-1003.

75. Haye L. DNA data storage. Assemblée Nationale - Sénat; 2021.

76. Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr. avr 2020;9(2):163-73.

77. Mardis ER. Next-Generation DNA Sequencing Methods. Annu Rev Genomics Hum Genet. 1 sept 2008;9(1):387-402.

78. EMBL-EBI. What is Next Generation DNA Sequencing? | Functional genomics II [Internet]. [cité 10 août 2022]. Disponible sur: https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/

79. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 1 oct 2015;13(5):278-89.

80. What is Oxford Nanopore Technology (ONT) sequencing? [Internet]. @yourgenome · Science website. [cité 14 août 2022]. Disponible sur: https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/

81. Polymerase chain reaction | Definition & Steps | Britannica [Internet]. [cité 12 mars 2023]. Disponible sur: https://www.britannica.com/science/polymerase-chain-reaction

82. Lau B, Chandak S, Roy S, Tatwawadi K, Wootters M, Weissman T, et al. Magnetic DNA random access memory with nanopore readouts and exponentially-scaled combinatorial addressing [Internet]. bioRxiv; 2021 [cité 11 mars 2023]. Disponible sur: https://www.biorxiv.org/content/10.1101/2021.09.15.460571v1

83. Albrecht TR, Arora H, Ayanoor-Vitikkate V, Beaujour JM, Bedau D, Berman D, et al. Bit-Patterned Magnetic Recording: Theory, Media Fabrication, and Recording Performance. IEEE Trans Magn. mai 2015;51(5):1-42.

84. Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed Engl. 16 févr 2015;54(8):2552-5.

85. Bornholt J, Lopez R, Carmean DM, Ceze L, Seelig G, Strauss K. (Null) Toward a DNA-Based Archival Storage System. IEEE Micro. 2017;37(3):98-104.

86. Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture.

Science. 3 mars 2017;355(6328):950-4.

87. Hamming RW. Error detecting and error correcting codes. Bell Syst Tech J. avr 1950;29(2):147-60.

88. Reed IS, Solomon G. Polynomial Codes Over Certain Finite Fields. J Soc Ind Appl Math. juin 1960;8(2):300-4.

89. Antkowiak PL, Lietard J, Darestani MZ, Somoza MM, Stark WJ, Heckel R, et al. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. Nat Commun. 22 oct 2020;11(1):5345.

90. Choi Y, Ryu T, Lee AC, Choi H, Lee H, Park J, et al. Addition of Degenerate Bases to DNA-based Data Storage for Increased Information Capacity [Internet]. bioRxiv; 2018 [cité 22 août 2023]. Disponible sur: https://www.biorxiv.org/content/10.1101/367052v1

91. Shipman SL, Nivala J, Macklis JD, Church GM. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature. juill 2017;547(7663):345-9.

92. Yim SS, McBee RM, Song AM, Huang Y, Sheth RU, Wang HH. Robust direct digital-to-biological data storage in living cells. Nat Chem Biol. mars 2021;17(3):246-53.

93. Kim H, Kim JS. A guide to genome engineering with programmable nucleases. Nat Rev Genet. mai 2014;15(5):321-34.

94. Farzadfard F, Lu TK. Emerging applications for DNA writers and molecular recorders. Science. 31 août 2018;361(6405):870-5.

95. Wirth D, Gama-Norton L, Riemer P, Sandhu U, Schucht R, Hauser H. Road to precision: recombinase-based targeting technologies for genome engineering. Curr Opin Biotechnol. oct 2007;18(5):411-9.

96. Bonnet J, Subsoontorn P, Endy D. Rewritable digital data storage in live cells via engineered control of recombination directionality. Proc Natl Acad Sci U S A. 5 juin 2012;109(23):8884-9.

97. Farzadfard F, Lu TK. Synthetic biology. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. Science. 14 nov 2014;346(6211):1256272.

98. Nuñez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. Cas1-Cas2

complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat Struct Mol Biol. juin 2014;21(6):528-34.

99. Itaya M, Tsuge K, Koizumi M, Fujita K. Combining two genomes in one cell: stable cloning of the Synechocystis PCC6803 genome in the Bacillus subtilis 168 genome. Proc Natl Acad Sci U S A. 1 nov 2005;102(44):15971-6.

100. CNIL. RGPD - Chapitre I [Internet]. CNIL. [cité 14 août 2023]. Disponible sur: https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1#Article4

101. LEEM. Les données de santé [Internet]. LEEM. 2019 [cité 13 août 2023]. Disponible sur: https://www.leem.org/les-donnees-de-sante

102. Conseil National de l'Ordre des Médecins. Le dossier du patient [Internet]. Conseil National de l'Ordre des Médecins. 2019 [cité 2 sept 2023]. Disponible sur: https://www.conseil-national.medecin.fr/medecin/exercice/dossier-patient

103. Vie Publique RF. Données numériques de santé : entre enjeux médicaux, technologiques et juridiques [Internet]. 2023 [cité 14 août 2023]. Disponible sur: https://www.vie-publique.fr/eclairage/289281-donnees-numeriques-de-sante-quels-enjeux-pour-quel-progres-medical#:~:text=En%202020%2C%202%20314%20exaoctets,aurait%20%C3%A9t%C3%A9%20multipli%C3%A9%20par%20dix.

104. AFNOR. Certification Hébergeurs des données de santé – AFNOR Certification [Internet]. [cité 14 août 2023]. Disponible sur: https://certification.afnor.org/numerique/hebergement-des-donnees-de-sante-hds#:~:text=Qu%27est%2Dce%20que%20la,de%20sant%C3%A9%20%C3%A0%20caract%C3%A8re%20personnel.

105. ANS. Certification des hébergeurs de données de santé [Internet]. Agence du Numérique en Santé. [cité 14 août 2023]. Disponible sur: https://esante.gouv.fr/labels-certifications/hds/certification-des-hebergeurs-de-donnees-de-sante

106. Sharma D, Kumar R, Gupta M, Saxena T. Encoding scheme for data storage and retrieval on DNA computers. IET Nanobiotechnol. sept 2020;14(7):635-41.

107. Sanofi/Challenges. La révolution du Big Data en santé [Internet]. Le Lab/Santé par Sanofi – Challenges.fr. 2018 [cité 14 août 2023]. Disponible sur: https://sanofi.challenges.fr/esante/la-

revolution-du-big-data-en-sante/

108. IBM. L'interopérabilité dans le secteur de la santé | IBM [Internet]. [cité 14 août 2023]. Disponible sur: https://www.ibm.com/fr-fr/topics/interoperability-in-healthcare

109. Fontana RE, Decad GM. Moore's law realities for recording systems and memory storage components: HDD, tape, NAND, and optical. AIP Adv. mai 2018;8(5):056506.

110. Carlson R. Guesstimating the Size of the Global Array Synthesis Market [Internet]. Guesstimating the Size of the Global Array Synthesis Market. 2017 [cité 12 mars 2023]. Disponible sur: http://www.synthesis.cc/synthesis/2017/8/guesstimating-the-size-of-the-global-array-synthesis-market

111. Kjær KH, Winther Pedersen M, De Sanctis B, De Cahsan B, Korneliussen TS, Michelsen CS, et al. A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. Nature. déc 2022;612(7939):283-91.

Stockage de données numériques sur ADN

Des écailles du plastron de la carapace de tortues aux centres de données d'aujourd'hui, l'humanité a toujours nécessité un support pour transmettre, partager et conserver de l'information. Les rapides progrès dans le secteur de l'informatique au cours du dernier siècle ont conduit à la création de quantités de données considérables dont les volumes ne cessent d'augmenter. Il est crucial de trouver une alternative aux technologies de stockages utilisées actuellement qui ne pourront pas être en mesure de supporter les quantités massives de données que nous aurons à stocker dans un futur proche. Ces progrès informatiques ont été accompagnés de progrès dans de nombreux autres secteurs dont en science, et plus particulièrement dans le secteur des biotechnologies. Notre connaissance de la molécule d'ADN a ainsi considérablement évolué, nous connaissons dorénavant les caractéristiques physico-chimique de cette dernière qui en font un candidat technologique potentiel et sérieux pour le stockage de données. Le développement de la technologie que représente de stockage de données numériques sur molécules d'ADN étant en plein essor, il parait important d'en expliquer les raisons. Aussi, de nombreux défis restent à relever et sont également décrits dans ce document. Enfin, le stockage de données sur ADN présente un fort potentiel afin de répondre aux besoins de sécurité, de confidentialité et de stockage à long terme des données de santé mais la technologie est loin d'être au point pour permettre d'être adoptée de manière généralisée et ainsi de s'adapter aux problématiques des données de santé.

From the scales on the plastron of turtle shells to today's data centers, mankind has always needed a medium to transmit, share and store information. Rapid advances in computing over the last century have led to the creation of vast quantities of data, the volumes of which continue to grow. It is crucial to find an alternative to the storage technologies currently in use, which will not be able to cope with the massive quantities of data we will have to store in a near future. These advances in computing have been accompanied by advances in many other sectors, including science, and particularly biotechnology. Thus, our knowledge of the DNA molecule has evolved considerably, and we now know its physical and chemical characteristics, which make it a serious potential technological candidate for data storage. The development of digital data storage on DNA molecules is booming, and it's important to explain why. There are also several challenges ahead, which are also described in this document. Finally, DNA data storage has great potential to meet the needs of healthcare data security, confidentiality and long-term storage, but the technology is far from ready for widespread adoption and thus for adaptation to healthcare data issues.