DNAdrive

Imaginez le nombre de fois que votre dispositif vous a informé que vous êtes resté sans espace de stockage. Il n'est pas surprenant que dans quelques années cela sera le cas pour des milliers d'entreprises spécialisées dans le stockage d'informations telles que Google, Amazon ou Microsoft. On estime qu'à cause de l'accumulation exponentielle de données produites par l'Homme, la quantité d'information produite annuellement augmentera de 16.3 ZB (zettabytes = 10^{21} bytes) à 163 ZB d'ici 2025. Malheureusement, la rupture imminente entre la demande de stockage et la production de capacité de stockage n'est pas un problème qui peut être facilement résolu. En effet, il est beaucoup plus difficile de construire de la capacité que de générer des données. Construire de la capacité capable de surmonter cette demande stratosphérique impliquerait des investissements en milliards d'euros, ce qui n'est pas une option rentable ou même réaliste. De plus, une seule éruption solaire pourrait endommager tous les circuits électroniques en un clin d'œil. Cependant, la nature nous a offert une alternative très efficace au stockage de données : l'ADN. Disque dur du vivant, cette molécule incroyablement stable et résistante pourrait permettre de stocker l'entièreté d'Internet dans un tube à essai.

Problématique : L'encapsulation dans un organisme d'un ADN modifié présente-t-elle une alternative efficace et durable au stockage courant des données ?

L'objectif de notre projet est de montrer comment la manipulation d'une bactérie capable d'intégrer de l'ADN étranger dans son génome peut être utilisée à des fins de stockage.

PLAN:

1. Le système immunitaire CRISPR-Cas **2**

1.1. Histoire 2
1.2. Structure du locus 2
1.3. Les gènes *cas* 2
1.4. La séquence chef 2
1.5. La suite répétition-espaceur 2
1.6. Le mécanisme du système immunitaire adaptatif bactérien 3

2. Notre projet **6**

2.1. L'idée 6
2.2. Comparaison aux autres technologies 7
2.3. L'algorithme de conversion 9
2.4. L'expérience 11
2.5. Conclusion 20

Remerciements et liens utiles **21**

Glossaire **21**

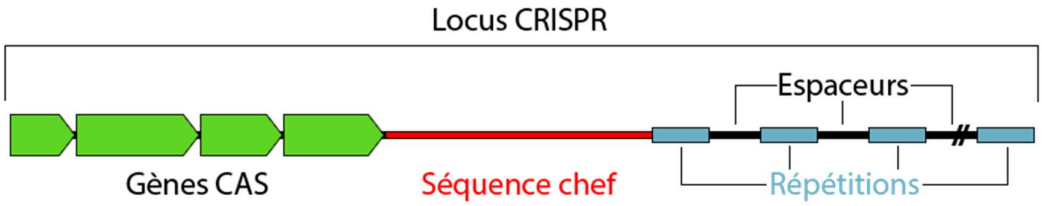
Références et bibliographie **21**

1. Le système CRISPR-Cas

1.1. Histoire

En 1987, les généticiens ont observé un motif étrange dans certains génomes bactériens. Une séquence d'ADN serait répétée plusieurs fois, avec des séquences uniques entre les répétitions (Ishino et al. 1987). Ils ont appelé cette configuration étrange « courtes répétitions palindromiques groupées et régulièrement espacées », ou CRISPR. En observant que les espaceurs présents dans les locus CRISPR correspondent en grande partie à des fragments d'éléments extra-chromosomiques (tels que du matériel génétique de bactériophages ou de plasmides), CRISPR fut enfin reconnu comme étant un système immunitaire adaptatif procaryote 19 ans après sa découverte. (Bolotin et al. 2005). Autrement dit, CRISPR correspond à une région d'ADN spécialisée dans le génome d'un organisme procaryote (bactérie ou archée) qui possède deux caractéristiques distinctes : la présence de répétitions de nucléotides de 23 à 55 paires de bases et entre celles-ci, d'espaceurs généralement uniques de 21 à 72 paires de bases (Grissa et al. 2007). Les bactéries acquièrent la mémoire des envahisseurs viraux en incorporant des extraits d'ADN invasif dans le locus CRISPR de leur génome, générant ainsi un nouvel espaceur entre deux séquences d'ADN répétitives. L'ARN hébergeant la séquence d'espacement aide les protéines Cas à reconnaître et à couper l'ADN de virus similaires lors d'une invasion ultérieure (Bolotin et al. 2005; Barrangou et al. 2007; Brouns et al. 2008; Barrangou and Marraffini 2014).

1.2. Structure du locus



Les trois principaux composants du locus CRISPR sont les gènes *cas*, la séquence chef et la suite répétition-espaceur.

1.2.1. Les gènes *cas*

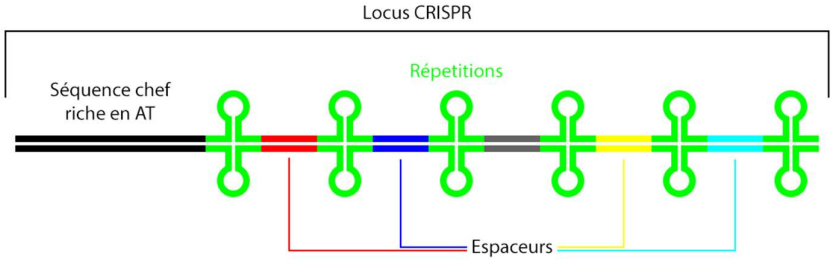
Le système immunitaire CRISPR-Cas nécessite la présence d'un ensemble de gènes associés à CRISPR (*cas*), rencontrés uniquement dans les génomes porteurs de CRISPR et situés à proximité de la suite répétitions-espaceurs du locus. Ceux-ci codent pour les protéines Cas, essentielles à la réponse immunitaire de la bactérie. Collectivement, 93 gènes *cas* sont regroupés en 35 familles en fonction de la similarité de séquence des protéines codées. 11 des 35 familles forment le noyau *cas*, qui comprend les familles de protéines Cas1 à Cas9. Un locus CRISPR-Cas complet possède au moins un gène appartenant au noyau cas (Grissa et al. 2007; Makarova et al. 2015). Cas1 et Cas2 sont les deux seules protéines Cas universellement conservées dans tous les systèmes CRISPR-Cas, nécessaires à l'étape d'adaptation du système immunitaire (Makarova et al., 2011).

1.2.2. La séquence chef

Une séquence chef riche en AT (adénine et thymine) d'environ 60 pb située en amont de la première répétition est essentielle pour l'acquisition d'espaceurs (Yosef et al. 2012) et favorise la transcription de la suite répétitions-espaceurs du locus (Pougach et al. 2010). De plus, celle-ci contient le site de liaison de la protéine IHF (5' - WATCAANNNNTTR - 3', où W est A ou T, R est A ou G et N est un nucléotide quelconque) qui est essentiel dans l'étape d'adaptation – en l'absence de celui-ci, il n'y aurait pas d'intégration d'espaceurs dans le locus (Nuñez et al. 2016).

1.2.3. La suite répétition-espaceur

La séquence chef est immédiatement suivie d'une suite de séquences répétitives séparées par des espaceurs uniques. Chaque espaceur correspond à un fragment unique d'ADN exogène (par exemple celui introduit par un virus) et sert à la bactérie comme marqueur stratégique afin de l'aider à détruire des séquences similaires lors d'un contact futur. Certaines répétitions possèdent des propriétés palindromiques (lorsqu'un brin d'ADN possède aussi l'inverse de sa séquence complémentaire), ce qui implique la formation d'une structure secondaire telle qu'une tige-boucle (« épingle à cheveux ») dans l'ARN (Brouns et al. 2008), alors que d'autres répétitions n'ont pas de structure spécifique.



1.3. Le mécanisme du système immunitaire adaptatif bactérien

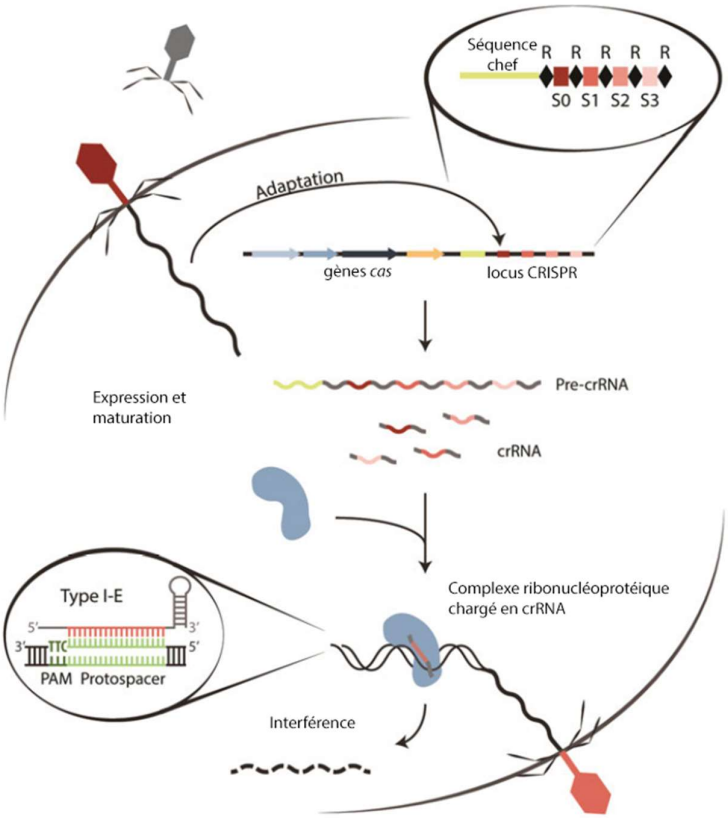


Image modifiée à partir d'une illustration par Amlinger, Lina.

Schéma général des étapes du système immunitaire CRISPR-Cas : Le système immunitaire CRISPR-Cas est divisé en trois étapes (van der Oost et al. 2014). Dans la première étape, l'adaptation, les endonucléases Cas1 et Cas2 stockent la mémoire des envahisseurs dans le locus CRISPR sous la forme de courtes séquences intermédiaires, appelées espaceurs (Barrangou et al. 2007; Fineran and Charpentier 2012). Au cours de la deuxième étape, l'expression, la suite répétitions-espaceurs est transcrite et ensuite traitée en petites séquences d'ARN CRISPR (crRNA), chacune constituée d'un espaceur et d'une ou deux portions d'une répétition. Les crRNA sont liés par le complexe multiprotéique Cascade (Brouns et al. 2008). Pendant la dernière étape, l'interférence, Cascade cherche des chaines d'ADN complémentaires au crRNA. Lorsqu'une correspondance est trouvée, l'ADN cible est clivé plusieurs fois par la nucléase Cas3 recrutée et l'attaque est arrêtée.

L'étude présentée dans ce projet a été réalisée dans le système CRISPR-Cas de type I-E de Escherichia coli, un des systèmes CRISPR-Cas les mieux étudiés. Les informations suivantes décrivent donc le mécanisme spécifique de ce système, sauf si indiqué autrement.

Le système CRISPR-Cas de type I-E est constitué de la nucléase Cas3 (qui a pour rôle la dégradation de la séquence cible), du complexe effecteur multiprotéique Cascade, des protéines Cas1 et Cas2 et de la suite répétitions-espaceurs du locus CRISPR associé. La suite contient des répétitions palindromiques de 29 paires de bases (pb) et des espaceurs intermédiaires de 32 pb.

Le locus doit contenir au moins une répétition, puisque la répétition précédée de la séquence chef est utilisée comme modèle pour la synthèse d'une nouvelle répétition (Yosef et al. 2012). Les mutations dans la frontière séquence chef - répétition empêchent définitivement l'intégration (Nuñez et al. 2016) et les structures secondaires d'ADN en épingle à cheveux potentiellement formées dans les répétitions peuvent également jouer un rôle dans la reconnaissance par Cas1-Cas2 (J.K. Nuñez et al. 2015).

La formation du complexe ainsi que l'activité nucléase de Cas1 sont nécessaires à l'adaptation. L'activité de la nucléase Cas2 est dispensable, indiquant un rôle plutôt structurel pour celle-ci (Nuñez et al. 2014).

1.3.1. L'adaptation - la construction d'un répertoire immunitaire

L'adaptation du système CRISPR-Cas est un processus complexe à plusieurs étapes dans lequel un proto-espaceur doit être extrait d'un ADN étranger envahissant et ensuite stocké dans le locus CRISPR comme espaceur. Premièrement, l'ADN étranger doit être reconnu comme cible pour l'acquisition d'espaceurs. Deuxièmement, une séquence d'une taille spécifique (généralement 30-40 pb, selon le sous-type du système CRISPR-Cas) doit être acquise à partir de l'ADN étranger. Enfin, la séquence acquise doit être intégrée en tant que nouvel espaceur dans le locus CRISPR et la séquence répétée adjacente doit être dupliquée. Bien que les composants et les conditions préalables de la machinerie d'acquisition des espaceurs varient entre les organismes et les sous-types du système CRISPR-Cas, plusieurs composants semblent être universellement conservés et sont essentiels parmi tous les sous-types du système CRISPR-Cas. Ces composants sont les protéines Cas Cas1 et Cas2, la séquence chef et la première répétition du locus CRISPR.

L'adaptation dans le système CRISPR-Cas de type I-E est le processus d'adaptation CRISPR-Cas le mieux compris. Malgré une étude approfondie, l'adaptation reste l'étape la moins comprise du système immunitaire

CRISPR-Cas par rapport aux étapes d'expression et d'interférence. Par exemple, il reste à élucider comment les produits de dégradation d'ADN viral (ou proto-espaceurs) sont produits, comment le complexe Cas1-Cas2 capture le proto-espaceur, si celui-ci est à simple ou à double brin, s'il possède des surplombs (un brin débordant sur l'autre à une extrémité) et comment il est coupé par le complexe Cas1-Cas2 avant son intégration en tant qu'espaceur dans le locus CRISPR (Wang et al. 2015; Künne et al. 2016).

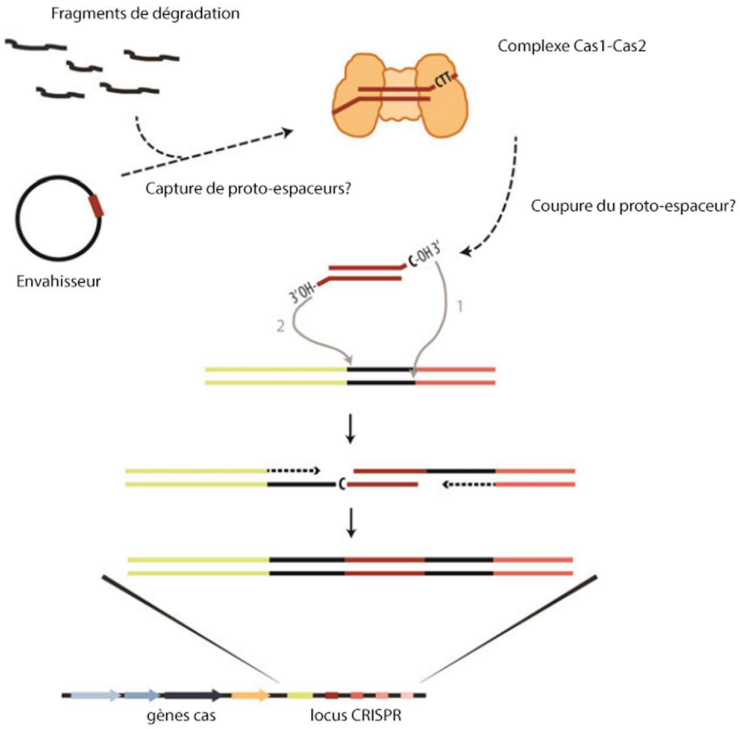


Schéma général du processus d'adaptation : Les fragments d'espaceurs, générés par Cas1 et Cas2 ou d'autres systèmes immunitaires, sont capturés par le complexe Cas1-Cas2. L'espaceur est inséré dans l'extrémité de la séquence chef suivie de la suite répétitions-espaceurs du locus à travers deux attaques nucléophiles (1 et 2). Le PAM (ici CTT) aide à orienter le nouvel espaceur dans le locus et la dernière base de la répétition provient de l'ADN de l'envahisseur intégré. C'est cette séquence d'espacement intégrée dans le locus CRISPR qui forme la mémoire immunitaire. La nouvelle répétition est synthétisée et l'intégration est complète.

Image modifiée à partir d'une illustration par Amlinger, Lina.

L'espaceur acquis est préférentiellement intégré à l'extrémité de la séquence chef suivie de la suite répétitions-espaceurs. Le locus est alors étendu par une unité de répétition-espaceur (28 + 33 = 61 pb) (Yosef et al. 2012). Le facteur d'intégration de l'hôte (la protéine IHF) et son site de liaison dans la séquence chef sont aussi nécessaires à l'adaptation (Nuñez et al. 2016).

1.3.1.1. Le PAM

Dans les milieux naturels, l'acquisition accidentelle d'espaceurs à partir de l'ADN « propre » - du génome de la cellule - plutôt que de l'ADN envahissant est généralement nuisible car elle entraîne la dégradation de l'ADN du génome par les mécanismes d'interférence de CRISPR-Cas. Un tel auto-ciblage conduit à l'auto-immunité CRISPR-Cas (Stern et al. 2010) et il a été montré que l'échappement à cette auto-immunité implique généralement l'inactivation mutationnelle des gènes cas, des mutations dans les répétitions à côté de l'espaceur auto-dérivé ou des mutations échappées dans le PAM (Stern et al. 2010). Par conséquent, il est nécessaire que les systèmes CRISPR-Cas évitent d'intégrer des espaceurs provenant du génome afin de minimiser ces effets nocifs. En effet, les premières observations du système CRISPR-Cas de E. coli de type I-E ont montré une forte préférence pour l'acquisition d'espaceurs à partir d'ADN étranger et un évitement de l'ADN « propre » (Yosef et al. 2012; Díez-Villaseñor et al. 2013; Levy et al. 2015).

Le PAM, séquence de 2 à 4 nucléotides, est d'une importance critique pour la reconnaissance et la sélection du proto-espaceur pendant l'adaptation et l'interférence chez Escherichia coli. En effet, celui-ci est un élément de ciblage essentiel (absent dans le locus CRISPR bactérien) qui distingue le génome bactérien de l'ADN étranger, empêchant ainsi le locus CRISPR d'être ciblé et détruit par le mécanisme d'interférence (Deveau et al. 2008; Westra et al. 2013). Il a été constaté qu'un proto-espaceur pourrait être incorporé dans le locus CRISPR en tant qu'espaceur si et seulement si il était flanqué du PAM approprié (Mojica et al. 2009). 4 séquences PAM garantissent l'immunité CRISPR : CAT, CTT, CCT et CTC (Westra et al. 2012). Un PAM plus général a été proposé, 5' – CHH - 3' (H = A, C ou T) (Hayes et al. 2016), qui inclut tous les PAM précédents. Chez Escherichia coli, le PAM le plus commun est 5' – CTT – 3'.

Remarque intéressante : chez Escherichia coli, le dernier nucléotide de la nouvelle répétition est dérivé du premier nucléotide de l'espaceur acquis. Ce nucléotide est en effet le dernier nucléotide de la séquence PAM (Datsenko et al. 2012). Comme il est presque invariablement un C, il est généralement considéré comme la dernière base de la répétition une fois intégrée (explication du modèle de la répétition de 29 pb (28+1) et de l'espaceur de 32 pb (33 - 1)) (Datsenko et al. 2012; Swarts et al. 2012).

1.3.1.2. Les exigences du proto-espaceur

Les exigences structurelles du proto-espaceur pour une intégration complète et certaine sont mal-caractérisées. Cependant, les données structurelles et biochimiques du complexe Cas1-Cas2 indiquent que, lorsque le proto-espaceur d'ADN est capturé par le complexe, celui-ci adopte une forme à double fourche avec une tige centrale à double brin (duplex) flanquée de surplombs monocaténaires issus de la séparation des deux brins par des résidus de tyrosine (Tyr22) conservés dans chaque monomère Cas1. Cas2 reconnaît la région à double brin tandis que Cas1 se lie aux surplombs monocaténaires des extrémités 3'. La séquence du PAM dans le surplomb de l'extrémité 3' est reconnue par les sous-unités catalytiques de Cas1a d'une manière spécifique aux bases et un clivage ultérieur à 5 nt de la frontière du duplex génère un ADN intermédiaire de 33 nt (comprenant le duplex de dsDNA de 23 pb et deux surplombs de ssDNA de 5 nt avec des groupes 3'-OH) qui est incorporé dans le locus CRISPR via un mécanisme de couper-coller (J. Nuñez et al. 2015; Wang et al. 2015; Amitai and Sorek 2016). Il est suggéré que cet intermédiaire forme le substrat pour l'intégration de l'espaceur dans le réseau CRISPR. Les propriétés du complexe Cas1-Cas2 montrent que le substrat préféré *in vitro* du proto-espaceur est l'ADN à double brin flanqué de surplombs monocaténaires de 7 nt aux extrémités 3' dont l'un contient un PAM à 5 nt de la frontière du duplex (J. Nuñez et al. 2015; Wang et al. 2015).

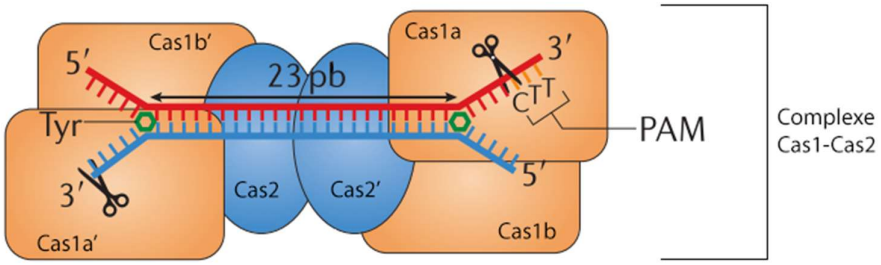


Illustration du complexe Cas1-Cas2 lié à un ADN à double fourche. *Image modifiée à partir d'une illustration par Amitai, Gil et Sorek, Rotem.*

Cas1-Cas2 fonctionne comme une règle moléculaire, la distance entre les deux sites actifs de Cas1 spécifiant la longueur de l'espaceur et coupe probablement le proto-espaceur avant son intégration (Wang et al. 2015)*. Les interactions entre l'espaceur et le complexe Cas1-Cas2 sont indépendantes de sa séquence (celle du duplex), ce qui permet l'intégration de toute séquence d'espacement* (J. Nuñez et al. 2015)*.*

1.3.1.3. L'AAM

Un autre motif de séquence, le AAM, séquence de 2 à 4 nucléotides (le plus souvent dinucléotide), présent dans l'extrémité 5' de l'espaceur (duplex) à un intervalle de 30 pb en aval du motif 5' – AAG – 3', a été proposé comme un motif important pour l'adaptation plutôt que pour l'interférence (Yosef et al. 2013), Cependant, l'importance de l'AAM doit être démontrée davantage par d'autres études.

1.3.1.4. L'intégration de l'espaceur

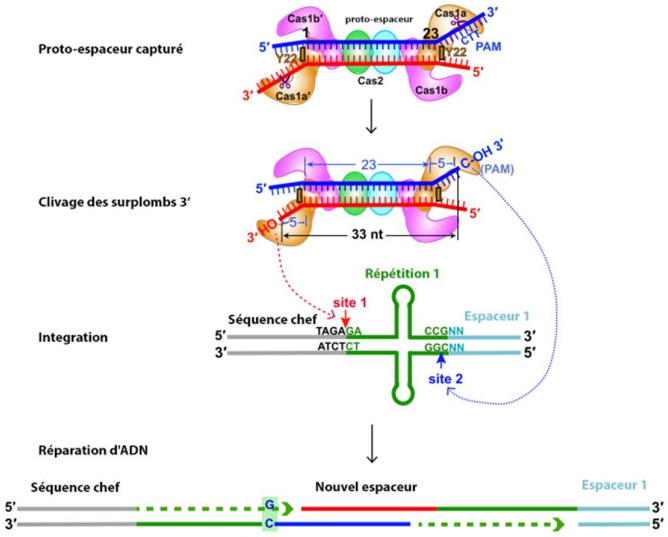


Image modifiée à partir d'une illustration par Wang et al.

Le complexe Cas1-Cas2 clive les surplombs des extrémités 3' du proto-espaceur à 5 nt de la frontière du duplex, ainsi générant un espaceur de 33 nt. Les deux extrémités 3' du proto-espaceur entrant sont impliquées dans l'attaque nucléophile sur le locus CRISPR, comme indiqué par les flèches rouges et bleues en pointillés, respectivement. Le PAM aide à orienter le nouvel espaceur de sorte que tous les espaceurs soient dans la même orientation vis-à-vis de la séquence chef (Mojica et al. 2009; Shmakov et al. 2014). Enfin, le duplex à brèche est réparé par la machinerie de réplication de l'ADN de l'hôte. La paire de bases GC provenant de la séquence PAM est mise en évidence par un fond vert. Il a été suggéré que le complexe Cas1-Cas2 nécessite, dans le cadre de sa liaison au locus, une séquence palindromique pouvant potentiellement former une structure d'ADN cruciforme / « épingle à cheveux », ce qui est une exigence caractéristique de diverses intégrases (Coté and Lewis 2008).

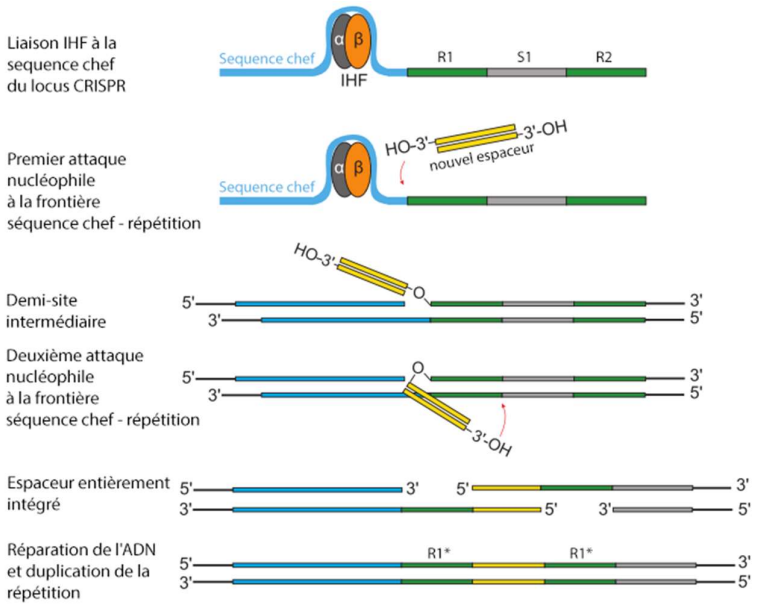


Schéma du processus d'adaptation (Wang et al. 2015) : Le nouvel espaceur est intégré dans le locus grâce à un mécanisme de type intégrase coordonné par Cas1-Cas2 (J.K. Nuñez et al. 2015). La liaison de la protéine IHF au site de liaison présent dans la séquence chef provoque très probablement la flexion de l'ADN du locus, permettant l'accès du complexe Cas1-Cas2 au site d'intégration. Ainsi peut s'expliquer la spécificité de l'extrémité de la séquence chef proche de la suite répétitions-espaceurs du locus comme site d'intégration des espaceurs (Wang et al. 2015).

Image modifiée à partir d'une illustration par Nuñez et al.

2. Notre projet

2.1. L'idée

En 2017, les scientifiques ont fait les premiers pas vers un Internet quantique (Ren et al. 2017), ont créé de l'hydrogène métallique (Silvera and Cole 2010), un traitement pour le cancer qui utilise les cellules T du patient (Smith et al. 2016), un utérus artificiel qui pourrait sauver les bébés nés de façon prématurée (Partridge et al. 2017) et ont modifié génétiquement pour la première fois un embryon humain afin d'éliminer une maladie cardiaque héréditaire (Ma et al. 2017). Ces exemples ne font que partie d'un ensemble plus vaste de travaux importants qui n'ont été réalisés que lors de cette année. Toutes ces découvertes ont un point commun essentiel : afin de se servir des avantages qu'elles offrent, on a besoin de conserver les recherches qui leur sont associées, c'est-à-dire de l'information.

Grâce à la dimension digitale de notre époque, la société moderne s'est permis d'accélérer l'utilisation, la création et le partage des informations, ce qui a conduit au développement de tous les domaines de l'existence humaine et d'arriver au plus haut niveau de qualité de vie mondiale jamais enregistré. Malheureusement, le stockage digital des données devient de plus en plus coûteux et difficile. Un nouveau rapport du Laboratoire National de Lawrence du Département de l'Energie à Berkeley indique que les centres de données des Etats-Unis utilisent énormément d'énergie, soit 70 milliards kWh par an. Cela représente 1,8% de la consommation totale d'électricité américaine (Shehabi et al. 2016). À un coût moyen de 10 cents par kWh, le coût annuel de toute cette électricité est de l'ordre de 7 milliards de dollars. À des fins de comparaison, 1 kWh suffit pour maintenir dix ampoules de 100 watts allumées pendant une heure ou pour qu'un smartphone reste chargé pendant une année entière. Pour produire 70 milliards de kWh par an, il faudrait des centrales électriques d'une capacité de base de 8 000 mégawatts, ce qui équivaut à environ 8 grands réacteurs nucléaires, 16 centrales à base de charbon (de 500 MW) ou le double de la production de tous les panneaux solaires du pays (EIA 2017). Il y a plus de 130 centres de données en France (pour plus de 100 000 serveurs informatiques en leur sein), des milliers sur la planète, et l'on estime qu'un grand centre de données peut consommer à lui seul plus qu'une ville française de 100 000 habitants. Le représentant de GreenPeace, Gary Cook, soulignait dans Times en 2015, que "Si on compte les [centres de données] et les réseaux de connexions, Internet serait le 6ème pays le plus consommateur d'électricité".

L'humanité a un problème de stockage des données : plus de données ont été produites au cours des deux dernières années que dans tout le reste de l'Histoire. IDC (International Data Company) prévoit que d'ici 2025, la sphère de données globale atteindra 163 zettabytes (ZB, soit un milliard de gigabytes) (Reinsel et al. 2017). C'est dix fois plus que les 16.1 ZB de données générées en 2016. Ce torrent d'informations pourrait bientôt dépasser la capacité des disques durs à le capturer. Pour faire face à « l'Internet des objets », on doit séparer les informations à long terme de celles à court terme. Une solution à ce problème doit alors être trouvée pour éviter l'augmentation d'une pollution déjà importante, le gaspillage d'argent et le pire - la perte de ces données (par exemple à cause d'une attaque cybernétique ou même d'une éruption solaire).

Cependant, nous avons à notre disposition une solution qui existe déjà depuis des centaines de millions d'années, un des plus capables, petits, stables et résistants dépôts d'information connu par l'Homme : l'ADN. Pourquoi cette molécule biologique ? À mesure que l'information numérique continue de s'accumuler, des solutions de stockage à plus haute densité et à plus long terme sont nécessaires (Gantz and Reinsel 2011). L'ADN présente de nombreux avantages pour stocker des données numériques (Bancroft et al. 2001). Tout d'abord, il est très dense. Au maximum théorique, l'ADN peut coder deux bits par nucléotide (nt) ou 455 exabytes par gramme d'ADN à simple brin (voir calcul p. 7-8). Ensuite, contrairement à la plupart des supports de stockage numérique, celui-ci n'est pas limité à une couche planaire et est souvent lisible malgré la dégradation dans des conditions non idéales au cours des millénaires (Pääbo et al. 2004; Bonnet et al. 2010). De plus, tant que les sociétés humaines lisent et écrivent l'ADN, elles seront capables de le décoder. « L'ADN ne se dégradera pas au fil du temps comme les cassettes et les CD et il ne deviendra pas obsolète », explique Yaniv Erlich, informaticien à l'Université Columbia. Contrairement à d'autres approches à haute

densité, telles que la manipulation d'atomes individuels sur une surface, les nouvelles technologies peuvent écrire et lire de grandes quantités d'ADN à la fois, ce qui permet d'augmenter l'échelle. Enfin, le rôle biologique essentiel de l'ADN donne accès aux enzymes naturelles de lecture et d'écriture et garantit que l'ADN restera une norme lisible dans un avenir prévisible.

Dans ce projet, nous allons non seulement étudier les possibilités de stockage d'informations dans l'ADN, mais aussi exploiter le système immunitaire CRISPR de type I-E de Escherichia coli afin de pouvoir stocker des informations au sein du génome de cette bactérie. Pourquoi une bactérie ? La raison est simple : non seulement le génome bactérien est circulaire, améliorant la stabilité à long terme de l'ADN (puisqu'il le protège des exonucléases), mais la bactérie dispose aussi des machineries de défense contre des organismes extra-chromosomiques et de préservation de l'ADN. Aussi, elle peut se multiplier de manière exponentielle, copiant ainsi l'information à travers les générations. Lorsque ces données sont stockées sur un disque dur qui pourrait théoriquement augmenter en taille et donc en capacité de stockage au fil du temps, les scientifiques pourraient les lire en examinant le génome de n'importe quelle bactérie de la colonie. Puis, tous les ordinateurs modernes sont vulnérables aux pannes électriques, au vol de données ou même aux tempêtes solaires qui pourraient détruire tout l'électronique. Quant aux bactéries, celles-ci sont immunisées contre les cyberattaques et l'ADN est résistant aux catastrophes naturelles et à la dégradation causée par le temps. De plus, le système CRISPR-Cas est un moyen très efficace de stockage de données grâce à la manière séquentielle par laquelle sont introduites les données qui permet de récupérer les informations stockées beaucoup plus rapidement (via une analyse génétique appelée génotypage). Grâce aux nouvelles technologies, ce système de stockage d'informations peut être étendu à d'autres types d'organismes en tant que module : celui-ci peut être intégré dans un plasmide contenant les gènes *cas* requis et inséré dans une cellule quelconque. Ainsi, on pourrait créer un disque dur à partir d'un organisme extrêmophile comme l'archée hyperthermophile *Pyrococcus furiosus* et stocker des informations dans des milieux qui autrement détruiraient l'ADN ou même les dispositifs de stockage d'informations actuels. On pourrait donc avoir toute la filmographie de Gérard Depardieu, Alain Delon et Louis de Funès à la surface de roches volcaniques à 120 °C !

Le concept d'utilisation de l'ADN pour le stockage de données numériques n'est pas nouveau, il est apparu il y a quelques dizaines d'années. La première personne à avoir converti du binaire en ADN est l'artiste Joe Davis en 1988, dans une collaboration avec des chercheurs de Harvard. La séquence d'ADN, qu'ils ont insérée dans E. coli, codait seulement 35 bits. Une fois organisées en une matrice 5 × 7, avec les uns correspondant aux pixels sombres et les zéros correspondant aux pixels lumineux, elles formaient l'image d'une ancienne rune germanique représentant la vie et la Terre femelle (« Microvenus »). Une étude récente de l'European Bioinformatics Institute et de Harvard a montré que les progrès des méthodes modernes de manipulation de l'ADN pourraient aujourd'hui rendre le stockage des données numériques dans l'ADN à la fois pratique et possible. Plusieurs groupes de recherche, dont l'Université de l'Illinois à Urbana-Champaign, l'ETH Zurich et l'Université de Columbia travaillent sur ce problème. Les géants technologiques ont aussi montré de l'intérêt dans le stockage biologique des informations. Microsoft, par exemple, prévoit d'ajouter du stockage de données dans l'ADN à son service cloud.

Les périphériques de stockage conventionnels tels que les DVD, les clés USB et les disques durs stockent des données numériques en modifiant les propriétés optiques, électriques ou magnétiques du matériau pour les stocker sous forme de 0s et de 1s. Pour stocker des données numériques dans l'ADN, l'idée est la même, mais un processus différent doit être utilisé. Les molécules d'ADN comprennent de longues séquences d'unités appelées nucléotides. Ces nucléotides sont appelés adénine, thymine, guanine et cytosine, désignés généralement comme A, T, G et C. Ainsi, l'information doit être stockée dans l'ADN sous forme de séquence de nucléotides, au lieu de séquences de 0s et 1s comme dans les médias électroniques. Plusieurs méthodes ont été utilisées afin de réaliser cette conversion, mais la plupart de celles-ci ne réussissent pas à surmonter les exigences imposées par la synthèse sujette à erreurs de l'ADN ou utilisent un schéma de codage non-universel, pertinent que dans l'expérience en question.

Dans ce projet, nous allons proposer plusieurs modèles de codage d'information dans l'ADN qui universalisent la conversion d'une séquence binaire en une séquence de nucléotides, tout en respectant les contraintes de la synthèse d'ADN et de la bactérie hôte. Ensuite, nous allons analyser l'efficacité de notre algorithme dans une expérience *in vivo* où nous allons activer et exploiter le système immunitaire CRISPR-Cas d'une colonie de bactéries Escherichia coli BL21(DE3) afin d'introduire dans leur génome deux fragments d'ADN codant le texte « hello. ».

2.2. Comparaison aux autres technologies

Calcul de la densité d'information dans l'ADN à simple brin :

La densité d'ADN théorique a été calculée en utilisant 2 bits par nucléotide d'ADN à simple brin.

L'ADN ne peut pas exister en tant que molécule isolée en dehors de la cellule d'un être vivant car la répulsion des deux brins, causée par les charges négatives des liaisons phosphodiesters, provoqueraient la dénaturation de celui-ci. En plus d'un pH neutre (7.0) et d'une température faible, des ions Na^+ doivent être ajoutés afin de neutraliser la molécule. Chaque fois que deux nucléotides sont liés ensemble, une molécule d'eau est éliminée (ou "perdue") par un processus appelé synthèse par déshydratation. Cela dit, afin d'obtenir une estimation précise de la masse molaire d'un brin d'ADN, on doit tenir compte de la perte d'eau et de l'ajout de sodium.

$$\begin{aligned} Masse\ molaire\ d'un\ brin\ d'ADN = {} & nA * (mA - mH - mO + mNa) \\ & + nT * (mT - mH - mO + mNa) \\ & + nG * (mG - mH - mO + mNa) \\ & + nC * (mA - mH - mO + mNa) \end{aligned}$$

Ou *nA* est le nombre de bases d'adénine, *mA* est la masse molaire de la désoxyadénosine monophosphate, *nT* est le nombre de bases de thymine, *mT* est la masse molaire de la désoxythymidine monophosphate, *nG* est le nombre de bases de guanine, *mG* est la masse molaire de la désoxyguanosine monophosphate, *nC* est le nombre de bases de cytosine, *mC* est la masse molaire de la désoxycytidine monophosphate, *mH* est la masse molaire de l'hydrogène, *mO* est la masse molaire de l'oxygène et *mNa* est la masse molaire du sodium. C'est-à-dire :

$$Masse\ molaire\ d'un\ brin\ d'ADN\ =\ nA\ *335.2 +\ nT\ *\ 326.2 +\ nG\ *\ 351.2 +\ nC *311.2$$

Pour un brin d'ADN avec des niveaux d'ATGC équilibrés et du sodium associé, la masse molaire moyenne d'un nucléotide serait de :

$$\frac{335.2 +\ 326.2 +\ 351.2 +\ 311.2}{4} = 330.95\ g.mol^{-1}$$

Le nombre de mol dans 1 g d'ADN est égal à $\frac{1}{330.95}\ mol$

Le nombre de nucléotides dans 1g d'ADN est égal à :

$$Constante\ d'Avogadro * nombre\ de\ mol = \frac{6.022 * 10^{23}}{330.95} \text{nucléotides}$$

La masse de 1 nucléotide est de $\frac{1}{\frac{6.022*10^{23}}{330.95}}\ g$ donc la « masse » de 1 bit est de :

$$\frac{1}{2 * \frac{6.022 * 10^{23}}{330.95}} = 2.75 * 10^{-22}\ g$$

Donc, théoriquement, 1 gramme d'ADN pourrait stocker $4.5\ x\ 10^{20}$ bytes (1 byte = 8 bits), soit 450 exabytes (EB) ou pourquoi pas 450 000 000 disques durs de 1 térabyte (TB). Evidemment, les maximums pratiques seraient plus faibles de plusieurs ordres de grandeur en fonction de la méthode de codage utilisée.

Etiquette	Date	Densité	Commentaires
Bande magnétique	1928	256 bits / in^2	Le premier enregistreur à base de bande magnétique, l'Univac Uniservo, enregistrait à une densité de 128 bits / in sur une bande magnétique d'un demi-pouce.
Disque dur	1956	2 000 bits / in^2	L'unité de stockage sur disque IBM 350, le premier lecteur de disque, stocke 5 millions de caractères (6 bits chacun), c'est-à-dire 3,75 MB sur 50 disques de diamètre 24 in avec 100 surfaces dotées de 100 bandes d'enregistrement.
CD	1982	~0.50 Gbit / in^2	CD-ROM de 120 mm diamètre et de capacité 700 MB (en réalité 847 MB).
DVD (SL)	1996	~3.27 Gbit / in^2	DVD-SS-SL de 120 mm diamètre et de capacité 4.7 GB (en réalité 5.5 GB).
Blu-ray (SL)	2002	~14.73 Gbit / in^2	Blu-ray-SL de 120 mm diamètre et de capacité 25 GB.
Disque dur	2012	1 000 Gbit / in^2	Seagate 1 Tb (térabit) / in^2
Mémoire magnétique à base d'atomes	2012	~71 684 Gbit / in^2	12 atomes de fer / bit, 9 nm^2 / bit, mémoire non volatile à basse température
Mémoire à base de positionnement d'atomes	1992	645 160 Gbit / in^2	« IBM » écrit avec des atomes de xénon espacés de 1 nm sur un réseau de 14 * 5 nm^2, 1 bit / nm^2
Holographie quantique	2008	~890 320 Gbit / in^2	Paire d'images 35 bits, atomes aériens 17*17 nm^2 et espace de lecture 4*5 = ((4*5) / (17*17)) * 20 bits / nm^2 = 1.38 bits / nm^2
ADN	2017	~1 897 529 Gbit / in^2 pour ADN à simple brin ou ~3 795 058 Gbit / in^2 (mais plutôt 3.6e12 Gbits / g)	ADN arrangé, diamètre de l'hélice de 2 nm, distance entre nucléotides de 0.34 nm, 10 paires de bases par rotation complète.

2.3. L'algorithme de conversion

Comme l'ADN possède 4 nucléotides possibles (A, T, G ou C), celui-ci est un système de numération de base 4. Le binaire est un système de numération de base 2. En sachant cela, l'algorithme de conversion le plus simple serait une conversion directe de base 2 en base 4 : A correspond à 00, T correspond à 01, G correspond à 10 et C correspond à 11. Cet algorithme pourrait atteindre la limite théorique de densité d'information stockable sur 1 nucléotide, celle de 2 bits / base. Cependant, la synthèse d'ADN est sujette à erreurs lorsque la séquence en question contient des homopolymères longs (répétitions de mononucléotides) ou des pourcentages d'AT ou de GC extrêmes.

Exemple : « *E=mc²* » est :

1. « $01000101001111010110110101100011001100 10$ » en binaire,
2. « 10110331123112030302 » en base 4,
3. « $\mathrm{TA}\mathbf{TT}\mathrm{A}\mathbf{CCTT}\mathrm{GC}\mathbf{TT}\mathrm{GACACAG}$ » en ADN (les homopolymères sont en gras).

Dans cet exemple, les homopolymères sont négligeables. Cependant, si on utilisait cet algorithme pour convertir un fichier d'ordinateur qui peut comprendre des suites de 0s ou de 1s très longues, les homopolymères résultants seraient de $\frac{l}{2}$ en longueur, où l est la longueur de la suite de répétitions de 0s ou de 1s. Cela peut correspondre à des centaines de répétitions.

Ces homopolymères sont non seulement problématiques pour la synthèse de l'ADN, mais interfèrent aussi avec notre objectif de transformer une colonie bactérienne en un disque dur. Ceci nous a conduit à concevoir notre propre algorithme de conversion de binaire en une séquence de nucléotides qui respecte les exigences imposées par la bactérie et par la synthèse sujette à erreurs de l'ADN. Comme nous allons exploiter l'étape d'adaptation du système immunitaire CRISPR-Cas de type I-E de Escherichia coli afin d'insérer nos propres informations, les séquences d'ADN doivent donc posséder les propriétés d'un proto-espaceur.

Le mécanisme suivant a été conçu pour stocker des informations dans la bactérie Escherichia coli BL21(DE3) qui possède un système immunitaire CRISPR-Cas de type I-E. Ce mécanisme peut être adapté à n'importe quel autre système CRISPR-Cas grâce à la variabilité de l'algorithme dont le code peut être trouvé sur GitHub (lien à la fin du projet). Cela dit, toutes les variables qui seront mentionnées sont strictement utilisées pour ce type de CRISPR. Ce système impose trois exigences (Shipman et al. 2017) :

1. Le PAM ne doit pas exister au sein de l'espaceur

Le PAM est comme un signal pour Cas1-Cas2 de cliver le proto-espaceur, c'est pourquoi il ne doit apparaitre qu'au début de celui-ci. La séquence du contenu ne doit donc pas comprendre la séquence 5' – AAG – 3' ou son complémentaire 5' – CTT – 3'.

2. Le proto-espaceur ne doit pas avoir des répétitions de nucléotides trop longues

La séquence du proto-espaceur ne doit pas comprendre des homopolymères (répétitions de mononucléotides) plus grands ou égaux à 4 (ex : AAAA) puisque ceux-ci peuvent provoquer des erreurs de synthèse d'ADN.

3. Le taux de GC de l'espaceur doit être supérieur ou égal à celui d'AT

A cause du fait que les séquences riches en AT sont moins stables thermodynamiquement que celles riches en GC et que souvent elles remplissent des fonctions spéciales dans Escherichia coli, (par exemple, elles correspondent aux origines de réplication (Rajewska et al. 2012)), le pourcentage de GC du proto-espaceur doit être supérieur à celui d'AT.

L'information que l'on souhaite stocker dans le génome de la bactérie correspond à des espaceurs de 33 nt séparés par des séquences répétitives au sein du locus CRISPR 1. Chaque espaceur stocke 24 bits d'information (l'organisation de l'information dans celui-ci sera détaillée un peu plus tard). Donc, tout d'abord, la séquence binaire qui correspond à l'information numérique que l'on souhaite stocker (étape assez simple) doit être divisée en sous-séquences de 24 bit chacune.

Exemple : « *Newton* »

$[010011100110010101110111][011101000110111101101110]$

Ensuite, les apparitions de 0s et de 1s de chaque sous-séquence sont comptées. Le chiffre binaire qui apparait le plus sera associé aux nucléotides G et C, l'autre sera associé aux nucléotides A et T. Cette étape garantit que le taux de GC de la séquence finale sera plus grand ou égal à celui d'AT, respectant ainsi la règle 3.

$[010011100110010101110111] \rightarrow$ 10 zéros, 14 uns $\rightarrow$ 0 correspond à A et T, 1 correspond à G et C

Dans la dernière étape, la séquence d'ADN est construite. Chaque chiffre binaire est remplacé par le nucléotide correspondant diffèrent de celui employé la dernière fois.

$[010011100110010101110111] \rightarrow [\mathrm{AGTAGCGTACGTACTGACGCTGCG}]$

Les répétitions de mononucléotides sont alors impossibles ; la règle 2 est donc respectée. Par conséquent, la règle 3 est aussi respectée, puisque le PAM canonique contient des répétitions.

Après avoir construit la séquence d'ADN, on rajoute un « bit marqueur » qui indique le chiffre binaire auquel correspondent les nucléotides G et C. Si G et C correspondent à 0, on rajoute un A ou un T à la fin de la séquence. Sinon, on rajoute un G ou un C. La règle de l'alternance doit être respectée partout.

G et C correspondent à 1 → [AGCTGCGATCGATCAGTCGCAGCG][C]

Ceci est l'essence de l'algorithme. Evidemment, comme cette séquence ne code qu'une partie du contenu, on doit rajouter un index, c'est-à-dire un nombre qui représente la position de l'information codée dans le flux de bits. Dans ce projet, nous avons choisi une longueur de 4 nucléotides pour l'index, mais celle-ci est variable. On peut numéroter sur 4 bits $2^4 = 16$ séquences, ce qui est largement suffisant pour notre expérience. Comme nous allons insérer seulement 2 proto-espaceurs, nous aurions même pu utiliser 1 seul nucléotide pour représenter l'index, mais nous allons poursuivre avec 4 à des fins de démonstration.

En plus de l'index, on va inclure une méthode de détection d'erreurs à travers un bit de parité représenté par un nucléotide. Si lors du décodage celui-ci équivaut à 0, le nombre de nucléotides G et C de l'index et du contenu doit être pair. Sinon, celui-ci doit être impair. Au cas où la parité du nombre d'apparitions de G et de C est en contradiction avec le bit de parité, cela signifie que l'espaceur a subi une mutation et que l'information est corrompue. L'ensemble constitué du bit de parité et de l'index est appelé « l'entête ». L'ensemble constitué de l'entête et du contenu est appelé « paquet ».

Enfin, comme cette séquence fait partie d'un proto-espaceur, on doit rajouter le PAM et l'AAM, respectivement au début et à la fin de la séquence construite. En pratique, on rajoute le PAM au début du processus afin de respecter la règle de l'alternance de nucléotides et l'AAM à la fin du processus de codage. Le PAM, placé avant le contenu, possède la séquence 5' – AAG – 3'. L'AAM possède la séquence 5' – GA – 3' et est placé après le contenu. Bien que son importance ne soit pas encore établie, nous avons choisi de l'inclure aussi afin de maximiser la probabilité de réussite du projet.

Les étapes sont donc les suivantes :

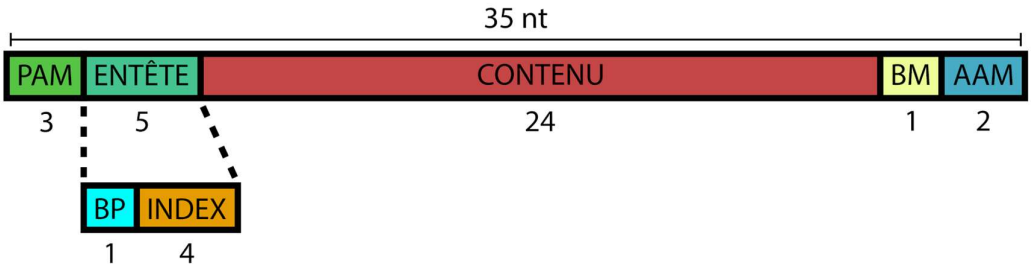
1. Rajouter le PAM.
2. Compter le nombre d'apparitions de chaque chiffre binaire (0 et 1) dans l'index et le contenu.
3. Associer chaque chiffre binaire à une paire de nucléotides par rapport au nombre d'apparitions de chacun dans l'index et le contenu.
4. Rajouter le bit de parité.
5. Rajouter l'index converti en nucléotides.
6. Rajouter le contenu converti en nucléotides.
7. Rajouter le bit marqueur.
8. Rajouter l'AAM.

Exemple : « *Newton* »

Proto-espaceur #1 :
5' – [AAG][T][ATAT][ACTAGCGTACGTACTGACGCTGCG][C][GA] – 3'

Proto-espaceur #2 :
5' – [AAG][T][ATAC][TGCGACTATGCAGCGCTGCAGCGT][G][GA] – 3'

Les proto-espaceurs auront donc la structure suivante :



Où BP correspond au bit de parité et BM correspond au bit marqueur.

Au cas où la longueur de la séquence binaire n'est pas un multiple de 24, on doit rajouter des nucléotides de remplissage et aussi une façon dont les détecter. Comme une des exigences du proto-espaceur est un taux élevé de GC, la redondance qu'on ajoute sera constituée de répétitions de GC. Afin de détecter la redondance (c'est-à-dire la fin du contenu valide), nous allons utiliser une méthode qui évite toute séquence de type PAM (5' – AAG – 3' ou 5' – CTT – 3') au sein du proto-espaceur. Nous avons conçu alors la règle suivante :

1. Si les deux derniers nucléotides du contenu sont complémentaires, on répète la dernière apparition d'un nucléotide différent de ces deux.

2. Sinon, on répète l'avant-dernier nucléotide.

Exemple :

5′ – [(…)AGT**C**A][**C**][GCGC(…)] – 3′ → Les deux derniers nucléotides ne sont pas complémentaires (CA), on répète alors l'avant-dernier nucléotide du contenu (**C**).

5′ – [(…)A**G**TAT][**G**][CGCG(…)] – 3′ → Les deux derniers nucléotides sont complémentaires (AT), on répète alors la dernière apparition d'un nucléotide différent (**G**).

Cas extrême : Si la configuration du proto-espaceur résulte en l'utilisation d'une seule paire de bases tout au long de l'espaceur, la redondance est introduite par la répétition du dernier nucléotide (ex : 5′ – [(…)GCGC**G**][**G**][CGCGC(…)] – 3′).

Au décodage, notre algorithme détecte que la règle d'alternance de nucléotides est rompue et détecte ainsi le début de la redondance. L'ensemble de ces règles assurent la détection de celle-ci et la non apparition d'un PAM au sein du proto-espaceur.

Remarque : Bien que la fonctionnalité de l'algorithme puisse être étendue à d'autres systèmes CRISPR, c'est-à-dire la configuration des différents secteurs du proto-espaceur, au cas où le PAM diffère de 5′ – AAG – 3′ ou 5′ – CTT – 3′, la garantie que celui-ci n'apparaitra pas dans le proto-espaceur n'est plus valable. Nous n'assumons aucune responsabilité de tout dommage causé par une mauvaise utilisation de la fonctionnalité de l'algorithme. En revanche, nous garantissons une fonctionnalité propre de l'algorithme dans les systèmes CRISPR de type I-E comme celui trouvé dans Escherichia coli.

2.4. L'expérience

2.4.1. Informations préalables

Pour transformer des bactéries en des disques durs moléculaires, nous allons exploiter l'étape d'adaptation du système CRISPR-Cas de type I-E d'une colonie d'Escherichia coli BL21(DE3). Cela consiste en l'introduction de proto-espaceurs dans la cellule qui vont coder pour le texte « hello. », converti du binaire en nucléotides grâce à notre algorithme. Les bactéries traiteront ces proto-espaceurs en tant que séquences d'ADN viral et déclencheront l'étape d'adaptation de leur système immunitaire CRISPR-Cas. Le complexe Cas1-Cas2, engin d'adaptation, intègrera ces proto-espaceurs en tant qu'espaceurs dans le locus CRISPR et ainsi, stockera nos données.

a) Le système CRISPR-Cas est naturellement inactif

Le système CRISPR-Cas est inactif dans la plupart des bactéries Escherichia coli à cause du facteur de répression H-NS (codé par le gène *hns*), protéine qui se lie à des centaines de promoteurs parmi lesquels se trouvent ceux des gènes *cas* et bloque leur transcription. Par conséquent, le système n'acquiert pas de nouveaux espaceurs naturellement et même si c'était le cas, ces espaceurs ne protégeraient pas la bactérie contre des phages car les protéines d'interférence importantes ne sont pas exprimées. La résistance contre les phages dans Escherichia coli sera ainsi beaucoup plus facile à réaliser par d'autres moyens, en particulier par des mutations d'adsorption (par exemple dans LamB). Ainsi, jusqu'à présent il n'y a pas de Escherichia coli sauvage que l'on peut « entraîner » pour protéger des phages en utilisant le système CRISPR-Cas. En utilisant un mutant *Δhns* (sans le gène *hns*), on pourrait, au cours de deux semaines, obtenir une protection contre les plasmides et potentiellement contre des phages persistants tels que M13. Cependant, la probabilité que l'on effectue de telles expériences sur une période si longue, dans le mutant spécifique, sans connaissance préalable du CRISPR-Cas est faible. En revanche, dans d'autres bactéries, comme dans Streptococcus thermophilus, le système sert en fait de système de défense dominant (avec presque toutes les résistances aux phages qui sont dues à ce système).

b) L'activation du système immunitaire

Notre expérience n'aurait pas de succès sans l'activation du système CRISPR-Cas qui est naturellement inactif. De plus, le génome de la souche de bactéries que nous avons à notre disposition ne comporte aucun gène *cas* – il ne possède que deux locus CRISPR. Ainsi, nous avons eu deux choix possibles : soit activer le système par insertion d'un plasmide contenant les gènes *cas1* et *cas2* qui codent pour le complexe d'adaptation Cas1-Cas2, soit acheter une souche de bactéries K12-Δhns qui contient tous les gènes *cas* et ne possède pas le gène hns qui code pour le facteur de répression H-NS. Nous n'avons pas fait le choix du mutant pour deux raisons : premièrement, la suppression du gène *hns* n'a pas un impact local et réduit la compétence des bactéries (Solomon 2012) et deuxièmement, Escherichia coli K12 contient aussi les gènes qui codent pour la machinerie d'interférence Cascade qui pourrait détruire nos proto-espaceurs.

Pour activer le système CRISPR-Cas, nous avons choisi le plasmide pCas1+2 déposé par Udi Qimron (plasmide Addgene #72676) qui comporte les gènes *cas1* et *cas2* extraits d'une souche d'Escherichia coli K12. Comme nos bactéries ne pourront qu'intégrer les espaceurs et non pas les détruire à cause de l'absence des gènes qui codent pour la machinerie d'interférence Cascade, la probabilité d'intégration de notre ADN est maximisée.

Les gènes *cas1* et *cas2* sont placés sous un opérateur lac et un promoteur T7. Ainsi, leur transcription n'est permise qu'en présence d'ARN polymérase T7 et d'allolactose ou d'IPTG. Afin d'éviter une « fuite d'expression » (au cas où la bactérie ne possède pas le gène *lacI* ou ne l'exprime pas assez - ce qui peut résulter en une expression involontaire de la protéine d'intérêt), le plasmide contient aussi le gène *lacI* qui code pour le répresseur de l'opéron lac. Le plasmide possède CloDF13 pour ORI (origine de réplication) ce qui résulte en 20-40 copies du plasmide dans une cellule. Enfin, il contient aussi un gène de résistance à la streptomycine et à la spectinomycine (SmR) pour aider à la sélection des cellules qui ont reçu le plasmide.

Comme les gènes *cas1* et *cas2* sont placés sous un promoteur T7, nos bactéries doivent synthétiser de l'ARN polymérase T7 puisqu'autrement nos gènes d'intérêt ne seraient pas transcrits. L'ARN polymérase T7 est issue du

bactériophage T7, ce qui signifie qu'elle n'est pas naturellement présente à l'intérieur d'une bactérie, sauf dans le cas d'une transduction. Notre souche de bactéries Escherichia coli BL21(DE3) contient le prophage λDE3 qui abrite le gène de l'ARN polymérase T7 régulé par un promoteur lacUV5 (similaire au promoteur lac). Ainsi, on peut induire l'expression de l'ARN polymérase T7 et la transcription de nos gènes avec de l'IPTG.

Ces gènes sont régulés par un operateur lac dont la répression n'est pas très stricte, c'est-à-dire qu'ils sont transcrits à un niveau basal même en absence d'IPTG. Cela serait un problème si les protéines d'intérêt étaient toxiques, mais comme Cas1 et Cas2 ne le sont pas, l'expression faible de ceux-ci n'est pas problématique pour notre projet même si le complexe Cas1-Cas2 formé commence à intégrer quelques proto-espaceurs présents à l'intérieur de la cellule (ce qui est très peu probable sauf au cas d'une contamination des bactéries par de l'ADN extra-chromosomique).

c) La conception des proto-espaceurs

Pour cette démonstration de faisabilité (« proof of concept », POC) nous avons choisi de coder le texte « hello. » dans deux proto-espaceurs dont la séquence a été établie par notre algorithme de conversion.

Proto-espaceur #1 :

5' – [AAG][T][CGCG][CATGACGCGTACGTCAGTACTAGC][A][GA] – 3'

Proto-espaceur #2 :

5' – [AAG][C][TATG][ACGTCGATACGTCGCGATCAGCGT][G][GA] – 3'

Comme l'ADN ne peut être synthétisé que dans la direction 5' – 3', il est impossible de synthétiser de l'ADN double brin à cause de la propriété antiparallèle de l'ADN. Nous avons eu alors deux choix possibles : soit synthétiser deux oligonucléotides complémentaires pour chaque proto-espaceur (qui se seraient associés), soit synthétiser un long oligonucléotide qui aurait formé une « épingle à cheveux » (c'est-à-dire que la deuxième moitié serait le complémentaire inverse de la première moitié afin de pouvoir s'associer). Nous avons trouvé que nous pouvions utiliser des proto-espaceurs en « épingle à cheveux » constitués de 23 bases d'ADN duplex au cœur du proto-espaceur (J. Nuñez et al. 2015; Wang et al. 2015) avec 7 bases non complétées à l'extrémité 5' du brin inférieur (qui comprend le PAM) et 5 bases non complétées à l'extrémité 5' du brin supérieur formant un lieur en boucle. *Rappel : la partie Cas1a du complexe Cas1-Cas2 reconnait le PAM 5' – CTT – 3' du surplomb 3' de 7 nt du proto-espaceur.* Aussi, l'épingle à cheveux évite la dissociation des brins complémentaires et leur ségrégation dans des cellules différentes lors d'un choc thermique ou d'une électroporation. Enfin, les données structurales du complexe Cas1-Cas2 et les expériences de Seth L. Shipman et al. (Shipman et al. 2017) nous indiquent que le PAM 5' – AAG – 3' est dispensable dans le brin supérieur alors que son complémentaire 5' – CTT – 3' dans le brin inférieur est indispensable (voir la partie d'adaptation du système CRISPR-Cas, section 1.1. du projet). La conception finale de nos proto-espaceurs est basée sur celle de Shipman et al (Shipman et al. 2017).

Pour obtenir l'épingle à cheveux, on doit rajouter à la séquence issue de l'algorithme son inverse-complémentaire dont les 5 premiers nucléotides sont supprimés (ceux de la boucle). Ensuite, pour avoir le surplomb 3', on doit supprimer les 7 premiers nucléotides de la séquence finale (c'est-à-dire de l'extrémité 5').

Exemple :

Séquence :

5' – AAGTCGCGCATGACGCGTACGTCAGTACTAGCAGA – 3'

Inverse-complémentaire :

5' – ~~TCTGC~~TAGTACTGACGTACGCGTCATGCGCGACTT – 3'

Séquence :

5' – ~~AAGTCGC~~GCATGACGCGTACGTCAGTACTAGCAGA
TAGTACTGACGTACGCGTCATGCGCGACTT – 3'

Séquence :

5' – GCATGACGCGTACGTCAGTACTAGCAGA
TAGTACTGACGTACGCGTCATGCGCGACTT – 3'

Illustration des deux proto-espaceurs en « épingle à cheveux » qui codent pour « hello. ».

Les séquences des proto-espaceurs en « épingle à cheveux » sont alors les suivantes :

Proto-espaceur #1 :

5' – [GCATGACGCGTACGTCAGTACTA][GCAGA] [TAGTACTGACGTACGCGTCATGC][GCGACTT] - 3'

Proto-espaceur #2 :

5' – [GACGTCGATACGTCGCGATCAGC][GTGGA] [GCTGATCGCGACGTATCGACGTC][ATAGCTT] - 3'

2.4.2. Protocole

A. Isolation du plasmide pCas1+2

Inoculation d'une culture liquide pendant la nuit

1. Après avoir reçu le récipient de bactéries DH5 alpha contenant le plasmide pCas1+2, on le range à 4˚C (les bactéries peuvent résister pendant ~ 2 semaines).

2. On prépare un stock liquide de spectinomycine de concentration 50 mg / mL dissoute dans de l'eau (la streptomycine pouvait être utilisée aussi mais nous ne disposions que de la spectinomycine).

3. Dans un tube Corning de 50 mL, on ajoute 10 mL de milieu de culture liquide (LB, Luria-Bertani) et un volume de spectinomycine prélevé du stock pour obtenir une concentration de 50 µg / mL.

   Nous avons préparé 2 tubes de 50 mL pour avoir de meilleurs rendements. Chacune de ces étapes doit être répétée ou faite en parallèle pour le deuxième tube.

4. On perce l'agar du récipient de bactéries avec un embout de pipette stérile et on le jette dans le tube de 50 mL.

5. On couvre légèrement le tube avec une feuille d'aluminium stérile ou un bouchon qui n'est pas étanche.

6. On incube la culture bactérienne pendant 16 heures à 37 ° C dans un incubateur à agitation à 120 rotations par minute.

7. Après l'incubation, on vérifie la croissance des bactéries qui est caractérisée par un trouble dans le milieu de culture.

Isolation du plasmide

Chaque étape décrite ci-dessous doit être répétée ou faite en parallèle pour le deuxième tube de 50 mL.

1. On centrifuge le tube de bactéries pendant 30 secondes à 11000 g puis, à l'aide d'une pipette, on enlève le surnageant.

2. On ajoute 1.5 mL de tampon de resuspension PBS puis, après avoir homogénéisé la solution, on transfère les bactéries dans un tube Eppendorf de 1.5 mL.

3. On centrifuge le tube Eppendorf pendant 30 secondes à 11000 g puis, à l'aide d'une pipette, on enlève le surnageant.

4. On ouvre le NucleoSpin® Plasmid pour procéder à l'isolation de plasmide.

Lyse des bactéries

5. On remet en suspension les bactéries dans 250 µL de Tampon A1 de resuspension (« resuspension buffer ») du kit et on vortexe. Ce tampon contient de la RNAse afin de dégrader l'ARN qui se trouve au sein des bactéries.

6. On ajoute 250 µL de Tampon A2 de lyse (« lysis buffer ») pour dénaturer membrane plasmique des bactéries. On mélange doucement en inversant le tube 6-8 fois sans vortexer jusqu'à ce que tout le mélange soit bleu.

7. On incube pendant 5 minutes à température ambiante.

8. On ajoute 300 µL de Tampon A3 de neutralisation pour renaturer les bactéries. On mélange bien en retournant le tube 6-8 fois jusqu'à ce que les échantillons bleus deviennent complètement incolores. Cette étape diminue le pH provoquant la précipitation des protéines et de l'ADN génomique qui a une couleur blanche, tout en laissant les plus petits plasmides libres en solution.

Clarification du lysat

9. On centrifuge les bactéries pendant 5 minutes à 11000 g. Pendant la centrifugation, l'ADN génomique (lié à des protéines histones) forme une boulette au fond du tube tandis que l'ADN plasmidique reste soluble dans le surnageant.

Liaison de l'ADN à la membrane de silice

10. On place une colonne NucleoSpin® Plasmid dans un tube de prélèvement (2 mL) et on pipette un maximum de 750 µL du surnageant sur la colonne. Le plasmide se liera par affinité à la membrane de silice.

11. On centrifuge le tube de prélèvement avec la colonne pendant 1 min à 11 000 g.

12. On jette le liquide résiduel et on replace la colonne NucleoSpin® Plasmid dans le tube de prélèvement. On répète cette dernière étape pour charger le lysat restant.

Lavage de la membrane de silice

13. On ajoute 600 µL de Tampon A4 de lavage (« washing buffer », contient aussi de l'éthanol) pour précipiter le plasmide et laver la colonne afin de supprimer les liaisons non-spécifiques.

14. On centrifuge pendant 1 min à 11 000 g. On jette l'écoulement et on place la colonne NucleoSpin® Plasmid dans le tube de prélèvement vide.

15. On centrifuge pendant 2 min à 11 000 g pour sécher la membrane de silice et on jette le tube de prélèvement.

Elution de l'ADN

16. On place la colonne NucleoSpin® Plasmid dans un tube Eppendorf de 1,5 ml et on ajoute 50 µl de Tampon AE d'élution.

17. On incube pendant 1 min à température ambiante.

18. On centrifuge pendant 1 min à 11 000 g.

19. La solution obtenue contient nos plasmides pCas1+2.

B. Quantification du plasmide pCas1+2

Par spectrophotométrie NanoDrop™ pour estimer la concentration en plasmides

1. On allume le NanoDrop™ et on configure le logiciel et l'appareil pour la mesure.

2. On nettoie les deux socles d'échantillonnage avec 2 µL d'eau déminéralisée (optionnellement sans nucléases).

3. On ouvre le bras supérieur et on éponge fermement les deux socles avec une lingette de laboratoire sèche.

4. On ouvre le logiciel d'exploitation. On clique sur le bouton « Nucleic Acid Quantification ».

5. On ajoute 2 µL d'eau déminéralisée sur le socle d'échantillonnage.

6. On abaisse le bras et on clique sur F3 ou sur le bouton « Blank ».

7. Lorsque la mesure est terminée, on nettoie fermement les deux socles avec une lingette de laboratoire sèche.

8. On prend le tube avec les plasmides et on pipette une goutte du contenu sur le socle d'échantillonnage.

9. On appuie sur le bouton « Measure » et on lit les mesures. On a maintenant la concentration en plasmides de notre solution et la pureté des plasmides (pourcentage de surenroulement défini par le rapport de l'absorbance entre 260 / 280 nm qui doit être aux alentours de 1.8).

Résultats : Le test NanoDrop™ indique une concentration en plasmides de 55 ng / µL et une pureté de 1.8.

Par électrophorèse dans du gel d'agarose pour vérifier la pureté et la masse des plasmides

1. On prépare les échantillons en mélangeant pour chacun 1 µL de la solution avec les plasmides, 9 µL d'eau et 2 µL de colorant de chargement (« loading dye ») de concentration 6X. Ce colorant contient aussi du glycérol qui augmente la densité de l'échantillon d'ADN et le dépose au fond du gel pour éviter sa diffusion dans le tampon.

2. On prépare le gel en mélangeant 50 mL de TAE de concentration 1X et 0.5 g d'agarose qu'on chauffe pendant 45 secondes dans un four à micro-ondes. Cela aboutit à un gel de concentration 1% d'agarose. On laisse le mélange se refroidir jusqu'à environ 50°C.

3. On ajoute dans le mélange du bromure d'éthidium pour obtenir une concentration de 0.5 µg / mL et on le laisse se solidifier dans un bac de coulée pendant 20 minutes avec un peigne (pour créer les puits).

   Le bromure d'éthidium se lie à l'ADN et lui confère une fluorescence forte sous lumière UV, ainsi permettant sa visualisation. La limite de détection par le bromure d'éthidium est aux alentours de 20 ng d'ADN mais cela dépend de la densité et de l'épaisseur du gel et des dimensions du puit.

4. Après la solidification, on enlève le peigne et on met le gel dans la cuve d'électrophorèse, veillant à l'orienter afin que l'ADN puisse migrer vers l'électrode + (l'ADN est chargé négativement).

5. On remplit la cuve avec du TAE de concentration 1X, jusqu'à ce que le liquide dépasse légèrement la surface du gel.

6. On remplit les puits avec nos échantillons et un marqueur de masse

   Nous avons utilisé 1 kb DNA Ladder.

7. On met le couvercle de la cuve et on allume le générateur de courant, le réglant à 120V.

8. Quand on observe une migration suffisante des échantillons dans le gel (par exemple jusqu'à la moitié du gel), on éteint l'appareil et on enlève le gel.

9. On analyse le gel à l'aide d'une lampe UV. Les différentes bandes du marqueur de masse nous donnent une idée de la longueur du plasmide qui doit être aux alentours de 4000-5000 pb (la longueur exacte est de 4711 pb).

   De gauche à droite : 2 échantillons qui contiennent nos plasmides, un échantillon qui contient un plasmide avec le gène efgp et un marqueur de masse 2-Log de 1 kb.

   Remarque : Plus les traits visibles sont proches de la ligne de dépôt, plus l'ADN est long ou impur. Aussi, plus les traits sont éloignés de la ligne de dépôt, plus les plasmides sont purs, c'est-à-dire sous forme surenroulée, puisqu'ils migrent plus vite dans le gel. Si on observe des traits épais et fortement blancs, alors la solution contient aussi de l'ARN, ce qui signifie une purification échouée.

Résultats : Lors de la première électrophorèse avec des échantillons contentant 1 µL de la solution de plasmides, aucun trait n'était visible. Nous avons fait alors une deuxième électrophorèse, mais avec des échantillons contenant 10 µL de la solution de plasmides, 2 µL de colorant et sans eau, accompagnés d'un marqueur de masse 2-Log de 1 kb. Deux traits furent visibles aux alentours de 4000 – 5000 pb (qui correspondent aux deux traits à gauche sur la photo ci-dessus). Nous pouvons voir deux traits dans l'échantillon qui contient le plasmide avec le gène egfp : cela montre qu'au trait le plus proche de la ligne de dépôt se trouvent des plasmides impurs, c'est-à-dire coupés ou détendus. Comme les traits des échantillons avec nos plasmides se retrouvent au même front que celui de l'échantillon avec le plasmide qui contient le gène egfp, nous pouvons déduire de nouveau que nos plasmides sont purs (leurs masses sont similaires).

C. Transformation n°1 - Insertion via choc thermique du plasmide pCas1+2 dans la culture d'Escherichia coli BL21(DE3)

1. On sort de -80°C deux tubes Eppendorf de 1.5 mL contenant 30 µL de bactéries Escherichia coli BL21(DE3) chimiquement compétentes et on les décongèle sur de la glace (environ 20-30 minutes).

2. Dans le premier tube on ajoute 3 µL de la solution de plasmides (165 ng de plasmides dans notre cas) et dans l'autre on met 10 µL (550 ng de plasmides dans notre cas). On note sur chaque le volume de la solution de plasmides prélevé et on les dépose sur de la glace.

Remarque : Suite au résultat très faible de la concentration en plasmides de la solution, nous avons décidé d'utiliser deux quantités différentes de plasmides afin d'assurer l'apparition d'au moins quelques colonies uniques.

3. On prend les tubes et on les chauffe pendant 45 s à 42°C.

4. On met les tubes sur de la glace.

5. En gardant les tubes sur la glace, on ajoute dans chacun 1 mL de milieu de culture liquide LB sans antibiotique.

Remarque : Comme nos bactéries compétentes sont poreuses et viennent de recevoir le plasmide qui leur confère une résistance à la spectinomycine, on n'ajoute pas de l'antibiotique pour permettre aux bactéries de refaire leur membrane et de synthétiser les protéines nécessaires à la destruction de l'antibiotique.

6. On prend les tubes et on les incube pendant 1 heure à 37°C dans un incubateur à agitation à 140 RPM.

7. On prend 4 boîtes de Petri qui contiennent du milieu LB-Agar solide et de la spectinomycine.

8. Pour chacun des deux tubes, on réalise les étapes suivantes :

- On prélève 100 µL de la solution du tube.

- On la dépose en petites gouttes sur une boîte de Petri et, à l'aide d'un épandeur en forme de L, on ensemence sur le milieu solide jusqu'à ce que le liquide sèche et qu'on ressente une forme de résistance.

- Du même tube, on prélève 900 µL qu'on centrifuge pendant 2 min à 2000 g.

- On jette 800 µL du surnageant et on remet en suspension les bactéries afin d'augmenter la concentration et réduire le volume de la solution à 100 µL. Cela évite la diffusion des bactéries dans le liquide qui empêcherait la sélection de colonies uniques.

- On la dépose en petites gouttes sur une boîte de Petri et, à l'aide d'un épandeur en forme de L, on ensemence sur le milieu solide jusqu'à ce que le liquide sèche et qu'on ressente une forme de résistance.

  Remarque : On prend deux volumes différents de bactéries (respectivement de 100 µL et de 900 µL) pour obtenir suffisamment de bactéries transformées et pouvoir sélectionner des colonies uniques.

9. On incube les 4 boîtes pendant 16 heures dans un incubateur à agitation à 37°C.

Remarque : Nous avons aussi fait un contrôle en insérant uniquement un plasmide contenant le gène de fluorescence egfp afin de tester la compétence de nos cellules.

Comme nous allons réaliser une deuxième transformation pour introduire les proto-espaceurs codant pour « hello. », nous devons induire la synthèse des protéines Cas1 et Cas2 responsables de l'intégration de ceux-ci. Cependant, nos cellules ne sont plus compétentes et doivent subir un traitement avec du chlorure de calcium pour être de nouveau compétentes. Pour la suite, nous avons eu alors deux choix possibles : soit induire la culture bactérienne avec de l'IPTG pendant 2 heures avant le traitement, soit l'induire après la deuxième transformation. Nous avons choisi la deuxième variante car nous avons pensé qu'une culture de densité optique d'environ 1.5 serait trop vieille pour le traitement.

D. Traitement des bactéries transformées avec du chlorure de calcium pour les rendre compétentes

1. On choisit une colonie unique d'une boîte de Petri avec de la spectinomycine. Les bactéries d'une colonie sont issues d'une seule bactérie qui a reçu le plasmide pCas1+2 lui conférant une résistance à cet antibiotique.

2. On perce la colonie avec un embout de pipette stérile qu'on dépose dans un erlenmeyer contenant 100 mL de milieu de culture LB liquide et de la spectinomycine.

3. On incube la culture pendant 2 heures dans un incubateur à 37°C.

4. On vérifie la DO600 (densité optique à 600 nm) des bactéries. Si elle n'est pas comprise entre 0.5-0.6, on laisse les bactéries dans l'incubateur jusqu'à ce que la culture atteigne cette densité optique en vérifiant leur densité à des intervalles de 30 min.

5. Après avoir atteint une densité optique adéquate, on prélève 45 mL de la culture dans 2 tubes Corning de 50 mL et on les laisse sur de la glace pendant 10 minutes.

6. On centrifuge les tubes pendant 5 minutes à 2000 RPM à 4°C et on jette le surnageant.

7. Pour chaque tube, on remet en suspension les bactéries sur de la glace dans 10 mL de $CaCl_2$ (chlorure de calcium) de concentration 0.1 M.

8. On centrifuge pendant 5 minutes à 2000 RPM et on jette le surnageant.

9. On remet en suspension les bactéries sur de la glace dans 4 mL de $CaCl_2$ (chlorure de calcium) de concentration 0.1 M.

10. On remet en suspension les bactéries de chaque tube dans 700 µL de glycérol. Le glycérol diminue la température de congélation des bactéries.

11. On aliquote les tubes Corning dans des tubes Eppendorf de 1.5 mL en prélevant 100 µL de la solution dans chaque tube qu'on place dans de l'azote liquide pour une congélation instantanée.

12. On range les tubes Eppendorf à -80°C.

Ces bactéries qui portent le plasmide pCas1+2 sont à nouveau compétentes.

E. Transformation n°2 - Insertion via choc thermique des proto-espaceurs en « épingle à cheveux » codant pour « hello. » dans la culture de bactéries Escherichia coli BL21(DE3) transformées

Pour sélectionner les colonies ayant reçu les proto-espaceurs, on introduit simultanément avec les proto-espaceurs un plasmide pHAT2 contenant le gène de fluorescence egfp (« Enhanced Green Fluorescent Protein ») et un gène de résistance à l'ampicilline. Les bactéries qui grandissent dans un milieu de culture contenant de l'ampicilline et qui sont fluorescentes seront sélectionnées pour la vérification de l'intégration des espaceurs. Pour avoir une sélection stricte, on va introduire 20 fois plus de proto-espaceurs que de plasmides. A cause de la différence en taille et en quantité du plasmide et

des proto-espaceurs, la probabilité qu'une bactérie a reçu le plasmide mais pas nos proto-espaceurs est très faible.

1. On sort de -80°C 2 tubes de bactéries récemment rendues compétentes.

2. Afin de minimiser les erreurs de pipetage, on réalise un « master mix » dans un tube contenant assez de plasmides et de proto-espaceurs pour 3.5 tubes Eppendorf de 1.5 mL (on n'en utilisera que 3 mais le surplus évite les erreurs de pipetage).

3. Pour chaque tube contenant les proto-espaceurs lyophilisés, on ajoute de l'eau pour obtenir une concentration en ADN de 1 µg / µL.

4. Dans le tube « master mix », on ajoute $3.5 * 0.1 = 0.35$ µg de plasmides contenant le gène *egfp.*

Dans notre cas, la solution de plasmides contenant le gène a une concentration de 0.64 µg / µL, donc on doit tout d'abord la diluer 10 fois pour obtenir une concentration de 64 ng / µL. Pour ce faire, on prélève 1 µL et on complète avec 9 µL d'eau déminéralisée et sans nucléases. Pour obtenir 0.1 µg, on prélève 1.56 µL de la solution diluée. Pour le « master mix », on prélève $3.5 * 1.56 = 5.46$ µL.

5. Pour avoir 20 fois plus de proto-espaceurs que de plasmides, on ajoute dans le même tube $3.5 * 20 = 70$ ng de proto-espaceurs. Pour ce faire, on doit diluer chaque solution de proto-espaceurs 100 fois pour obtenir une concentration de 0.01 µg / µL. On prélève de chaque solution diluée $3.5 * 2 = 7$ µL et on l'ajoute au « master mix ».

La masse molaire du plasmide contenant le gène egfp est de 2378426.95 Da. Celle du premier proto-espaceur est de 17961.61 Da et celle du deuxième est de 17993.61 Da.

6. Le « master mix » a alors 19.46 µL. On prélève 5.56 µL de celui-ci dans trois tubes Eppendorf de 1.5 mL contenant 50 µL de bactéries compétentes portant le plasmide pCas1+2.

7. On prend les tubes et on les chauffe pendant 45 s à 42°C.

8. On met les tubes sur de la glace.

9. En gardant les tubes sur de la glace, on ajoute dans chacun 950 µL de milieu de culture liquide LB sans antibiotique. Le volume final sera d'environ 1 mL.

10. Dans un tube, on ajoute 2 µL d'IPTG de concentration 500X pour obtenir une concentration finale de 1 mM.

11. Dans l'autre tube, on ajoute 20 µL d'IPTG de concentration 500X pour obtenir une concentration finale de 10 mM.

12. Dans le dernier tube, on n'ajoute pas d'IPTG. On utilise cet échantillon comme contrôle négatif de la synthèse des protéines.

13. On incube les cultures pendant 2 heures à 37°C.

14. On prend 6 boîtes de Petri qui contiennent du milieu LB-Agar solide et de l'ampicilline.

Rappel : Le plasmide contenant le gène egfp contient aussi un gène de résistance à l'ampicilline. On l'utilise pour sélectionner les bactéries dont la probabilité d'avoir intégré aussi les proto-espaceurs est haute.

10. Pour chacun des trois tubes, on réalise les étapes suivantes :

- On prélève 100 µL de la solution du tube.
- On la dépose en petites gouttes sur une boîte de Petri et, à l'aide d'un épandeur en forme de L, on ensemence sur le milieu solide jusqu'à ce que le liquide sèche et qu'on ressente une forme de résistance.
- Du même tube, on prélève 900 µL qu'on centrifuge pendant 2 min à 2000 g.
- On jette 800 µL du surnageant et on remet en suspension les bactéries afin d'augmenter la concentration et réduire le volume de la solution à 100 µL. Cela évite la diffusion des bactéries dans le liquide qui empêcherait la sélection de colonies uniques.
- On la dépose en petites gouttes sur une boîte de Petri et, à l'aide d'un épandeur en forme de L, on ensemence sur le milieu solide jusqu'à ce que le liquide sèche et qu'on ressente une forme de résistance.

F. Vérification de l'intégration des proto-espaceurs

Par amplification avec PCR et électrophorèse

Une première vérification peut être faite à travers l'amplification de la région proximale de la séquence chef du locus CRISPR II. Si la longueur du produit de PCR est plus grande que celle

de la région habituelle flanquée par nos amorces (120 pb vs 242 pb), cela signifie que de l'ADN a été intégré.

1. On sort de -80ºC les tubes contenant les amorces sens et antisens.

2. Pour chaque tube, on ajoute de l'eau déminéralisée et sans nucléases pour obtenir des stocks de concentration 50 µM.

3. Des bactéries cultivées précédemment sur du milieu solide contenant de l'ampicilline, on sélectionne les colonies uniques les plus vertes et fluorescentes.

Rappel : Chez Escherichia coli BL21(DE3), seul le locus II est fonctionnel.

Remarque : A cause du fait que les bactéries transformées qui ont été déposées sur des boîte de Petri n'ont pas assez grandi, nous avons sélectionné 15 colonies uniques vertes et fluorescentes que nous avons cultivées dans du milieu liquide avec de l'ampicilline et de la spectinomycine. 11 cultures ont survécu et ont été envoyées vers la PCR. Une des 11 colonies n'a pas été induite avec de l'IPTG et sert comme contrôle.

4. On prépare le mélange de PCR. Dans notre cas, nous préparons une solution de 300 µL pour 12 tubes de 25 µL chacun qui contient : 60 µL de tampon réactionnel Q5 (5X), 6 µL de mélange de dNTPs (10mM), 3 µL de solution d'amorce sens, 3 µL de mélange d'amorce antisens, 213 µL d'eau déminéralisée et sans nucléases et 3 µL de Q5 ADN polymérase (2 unités / µL).

5. On ajoute 24 µL de cette solution dans 11 tubes de PCR Eppendorf de 200 µL.

Remarque : De nouveau, c'est un « master mix ». Le fait de pipetter moins et d'avoir un surplus de solution évite les erreurs de pipetage.

6. A chaque tube, on ajoute 1 µL de la culture bactérienne correspondante d'Escherichia coli BL21(DE3) cultivée précédemment pour obtenir un volume final de 25 µL.

7. On place les 11 tubes dans la couronne de la PCR.

Remarque : Comme le mélange est chauffé à 98 ºC, on peut introduire les bactéries directement sans faire une purification d'ADN génomique puisqu'à cette température les bactéries sont de toute façon lysées et libèrent leur ADN.

8. On configure le cycle de la PCR. Tout d'abord, la dénaturation pendant 15 secondes à 98ºC. Ensuite, l'association des amorces à l'ADN pendant 30 secondes à 55ºC. Enfin, l'élongation de l'ADN pendant 45 secondes à 72ºC. Ces trois étapes se répètent pendant 35 cycles.

9. On sort les tubes de la PCR et on les pose sur de la glace.

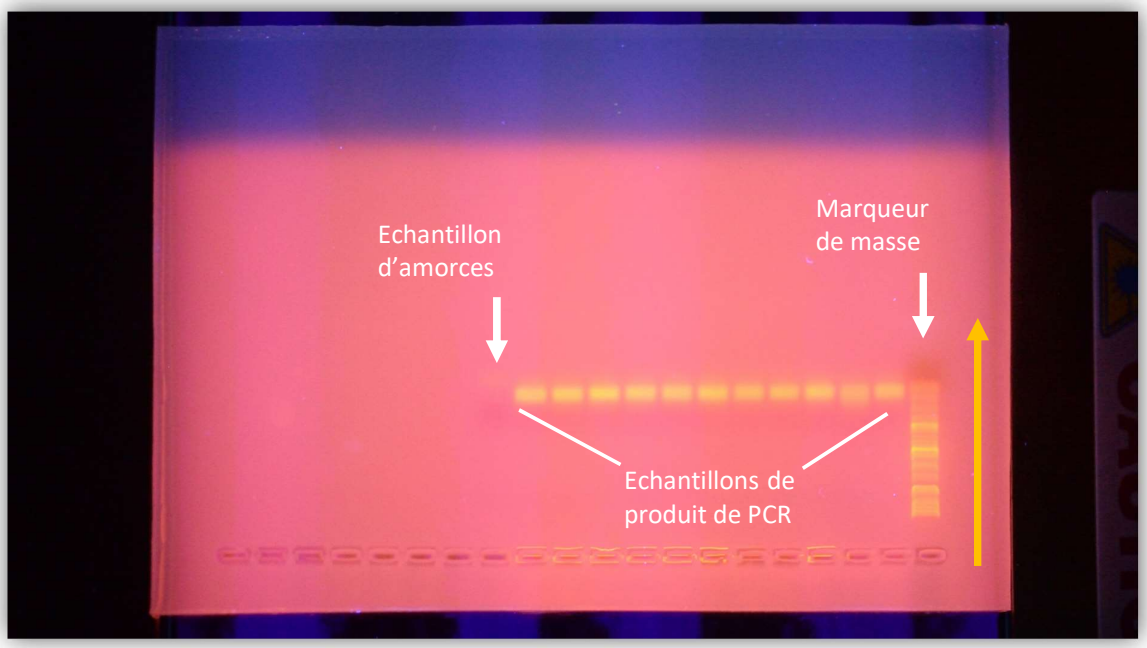
10. On réalise 11 échantillons d'électrophorèse dans des tubes Eppendorf de 1.5 mL. Pour chaque échantillon, on ajoute 10 µL du produit de PCR correspondant et 2 µL de colorant de chargement.

11. On réalise aussi un échantillon qui contient uniquement des amorces sens et antisens, pour pouvoir distinguer le produit de PCR des amorces non-consommées. Pour celui-ci, on respecte la concentration en amorces du mélange de PCR mais on n'inclut pas les autres composants – on remplit avec de l'eau déminéralisée et sans nucléases.

12. On prépare un gel d'agarose de concentration 1.2% et du TAE de concentration 1X puis on prépare les appareils d'électrophorèse (voir le protocole de la partie B).

13. On charge les échantillons dans les puits du gel. On ajoute aussi un marqueur de masse 2-Log de 0.1 kb à 10 kb.

14. On laisse migrer jusqu'à environ ½ du gel, puis on analyse les résultats sous une lampe UV.



RESULTAT POSITIF !

De droite à gauche : marqueur de masse, échantillons de produit de PCR, échantillon d'amorces.

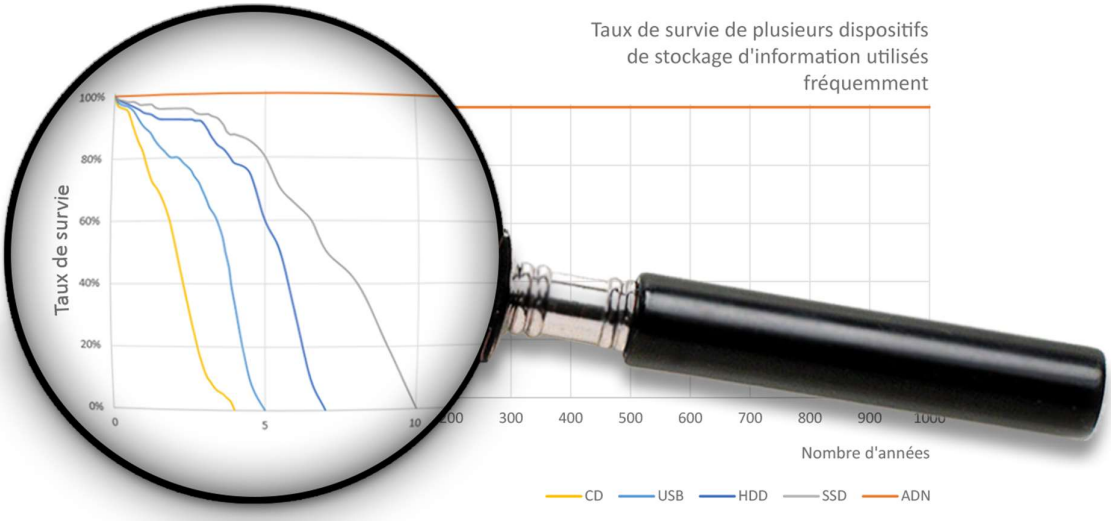
Nous pouvons voir des traits très bien définis pour 10 des échantillons aux alentours de 100 pb, ce qui signifie une amplification réussite du locus CRISPR mais pas d'intégration d'ADN. Cependant, un des échantillons (le deuxième, n°4) présente un trait plus allongé qui s'étend aussi aux alentours de 200 pb (on prend le marqueur de masse comme référence). Cet allongement signifie une intégration d'espaceurs dans le locus. Comme il se trouve près du trait de 200 pb du marqueur de masse, il est clair qu'au moins deux proto-espaceurs ont été intégrés (une unité de répétition-espaceur comprend 61 nucléotides). Pour vérifier que ce n'est pas une simple coïncidence, nous avons répété la PCR et l'électrophorèse après avoir cultivé 4 cultures de bactéries dont la culture n°4 et le résultat fut le même. Nous allons faire une dernière vérification à l'aide de deux enzymes de restriction qui reconnaissent chacune respectivement une sous-séquence de l'un des deux proto-espaceurs puis nous allons envoyer l'échantillon n°4 vers le séquençage. Nous considérons l'expérience comme une réussite grâce aux circonstances qui ne sont pas favorables à l'intégration de tout autres proto-espaceurs que nous avons introduit.

Remarque : En effet, nos rendements furent pauvres : seulement 1 des 15 cultures a intégré nos proto-espaceurs. Cependant, cela est probablement dû non seulement à la chronologie de l'induction des bactéries avec de l'IPTG (à cause de la toxicité de l'IPTG pour des cultures jeunes), mais aussi à la quantité trop petite de plasmides contenant le gène egfp et celui de résistance à l'ampicilline. Il est très possible que la plupart des bactéries ayant intégré les proto-espaceurs furent tuées par l'ampicilline à cause du fait qu'elles n'ont pas intégré le plasmide qui les aurait sauvées. Il est aussi possible que la plupart des cellules aient perdu le plasmide pCas1+2 à cause de l'absence d'un stress provoqué par la présence de spectinomycine ou bien à cause de la dégradation des proto-espaceurs par des endonucléases, toujours lié à la chronologie de l'induction.

2.5. Conclusion

A l'issue de nos expériences, nous pouvons conclure ainsi sur notre projet : la réussite de l'intégration et de la récupération du message montre que le stockage de données dans le génome d'un organisme est possible.

Le stockage d'informations dans de l'ADN est durable grâce à la stabilité de la molécule. En effet, celle-ci peut résister des dizaines, même des centaines de milliers d'années dans des conditions appropriées. Par exemple, le plus ancien ADN séquencé a pu être daté entre -450 000 et -800 000 années après sa découverte au Groenland (Willerslev et al. 2007) et on a réussi à séquencer le génome complet du mammouth (Palkopoulou et al. 2015). Théoriquement, l'ADN pourrait résister 6.8 millions d'années puisque sa demi-vie est de 521 années, mais, selon les scientifiques, il ne serait plus lisible après 1.5 millions d'années.



De plus, cette méthode est durable grâce aux mécanismes de défense de la cellule hôte. Non seulement l'ADN peut résister des milliers d'années, mais il est aussi protégé des organismes exogènes au sein de la bactérie. Par exemple, celui-ci est notamment protégé des exonucléases grâce à la forme plasmidique du génome. De plus, même si la bactérie meurt par exemple à cause du manque de nourriture, l'ADN génomique ne subit aucun dommage (les organismes procaryotes ne contiennent pas de lysosomes) et peut facilement être récupéré. Par conséquent, nous n'aurions pas besoin d'infrastructures quelconques pour protéger l'ADN.

Ensuite, stocker des données dans de l'ADN est très efficace grâce à la compacité de celui-ci. Comme 1 gramme d'ADN peut théoriquement stocker 450 exabytes d'informations, quelques grammes d'ADN pourraient stocker l'entièreté des données produites par l'humanité tout au long de l'histoire. Un exemple plus concret serait le stockage des 100 pétabytes d'un centre de données dans 220 µg d'ADN. Ainsi, nous n'aurions plus besoin d'utiliser des centaines d'hectares pour construire des centres de données puisque nous pourrions archiver toute l'information dans de l'ADN.

Par ailleurs, la quantité d'énergie quasi-nulle requise par les bactéries montre l'efficacité de cette méthode. Bien que les bactéries nécessitent en effet des nutriments pour subsister, la quantité de milieu de culture nécessaire pour atteindre une colonie assez grande est minuscule. Pour éviter trop de mutations dans les données stockées, à un moment donné, la réplication des bactéries doit être stoppée soit par congélation de celles-ci, soit par famine. Quant à l'ADN, celui-ci ne nécessite aucune source d'énergie pour son existence en tant que sel. Par conséquent, on réduirait énormément les émissions de carbone dues à la consommation d'énergie par les dispositifs de stockage de données actuels.

Enfin, cette méthode est efficace à travers la réplication exponentielle des bactéries. En effet, en se répliquant, les bactéries font plusieurs copies du même ADN génomique ou plasmidique. Avoir plusieurs copies du même ADN peut servir à corriger les mutations que l'ADN peut subir sans avoir besoin de synthétiser d'autres oligonucléotides. Ainsi, nous pourrions avoir une méthode de correction d'erreurs issue de la réplication naturelle des bactéries, ce qui signifierait à nouveau une diminution des coûts de stockage des données.

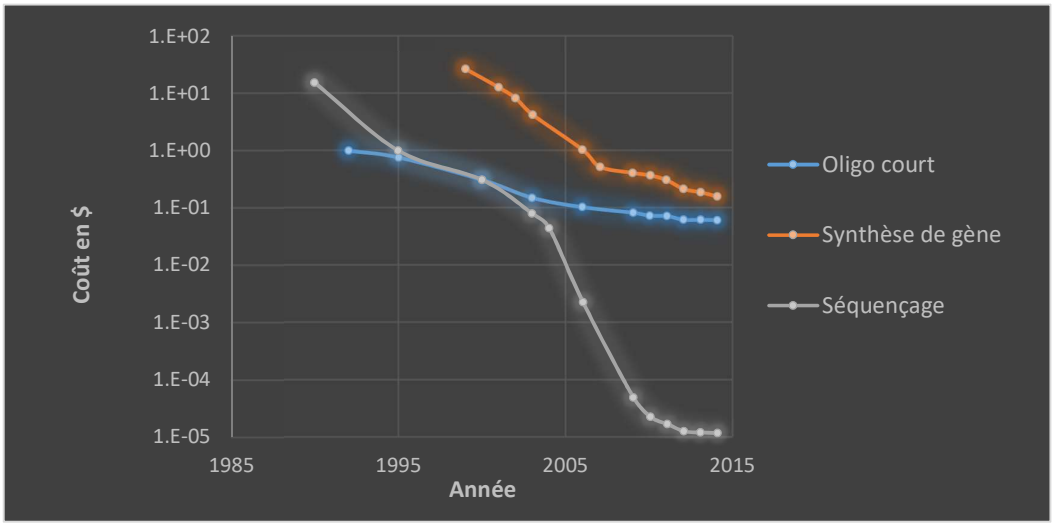
Passant au processus d'intégration de l'ADN dans le génome bactérien, quoique celui-ci peut paraitre long, complexe et exhaustif, les avantages de durabilité que celui-ci présente ne doivent pas être ignorés dans un monde où l'électronique est si fragile face aux catastrophes naturelles qui d'ailleurs sont imprévisibles.

Par ailleurs, notre algorithme peut aussi être utilisé pour coder des informations dans de l'ADN isolé sous forme de sel. En effet, les propriétés de celui-ci sont favorables à une synthèse d'ADN non-sujette à erreurs. Par exemple, l'algorithme construit des séquences qui ne contiennent pas des homopolymères et dont la paire de bases dominante peut être contrôlée. L'algorithme présente aussi l'avantage de pouvoir décoder le message en utilisant soit le brin sur lequel l'espaceur est introduit, soit son complémentaire dans la direction 5' – 3'. Si l'on souhaite avoir un taux équilibré d'AT et de GC, on peut simplement encoder le binaire à l'aide d'un schème Huffman pour équilibrer le nombre de 0 et de 1 binaires.

Ensuite, le système de stockage d'informations peut non seulement être utilisé dans de multiples colonies de bactéries, mais aussi peut être étendu à d'autres souches : grâce aux nouvelles technologies, il est possible d'intégrer dans un plasmide un locus CRISPR artificiel accompagné des gènes codant pour les protéines nécessaires à l'adaptation CRISPR-Cas. En insérant ce plasmide, il serait désormais possible de stocker des données par exemple dans des organismes extrêmophiles afin de garantir la survie des informations dans des milieux hostiles. La versatilité de notre algorithme ajoute donc à l'efficacité de ce système de stockage d'informations, devenant modulaire et accessible à tous les organismes (même à des eucaryotes).

Cependant, bien que cette méthode soit en effet durable, son efficacité peut encore être débattue à cause du prix de l'ADN et de la vitesse de synthèse et de séquençage de celui-ci. Aujourd'hui, il est impossible de remplacer un disque dur d'un ordinateur ordinaire par un disque à ADN ou bactéries puisqu'il ne pourrait pas répondre aux exigences en temps réel de l'utilisateur. La synthèse d'un nucléotide coûte environ $0.50, ce qui est trop cher pour pouvoir stocker domestiquement des données telles que des images numériques. Par conséquent, l'usage pratique actuel se retrouve dans l'archivage de données et pas dans le stockage à court terme.

Néanmoins, ce processus peut certainement être amélioré dans le futur pour répondre aux problèmes d'inefficacité. Cette démonstration de faisabilité est destinée à paver une nouvelle voie de stockage d'informations pour répondre aux besoins d'archivage des données permanentes telles que les recherches scientifiques qui sont essentielles à la survie intellectuelle de notre espèce. D'ailleurs, les données statistiques actuelles prévoient une augmentation exponentielle de la vitesse de synthèse et de séquençage d'ADN, accompagnée d'une diminution exponentielle de leurs coûts.



Ainsi, nous venons de montrer que l'encapsulation dans un organisme d'un ADN modifié est faisable. Bien qu'aujourd'hui elle ne soit qu'une une alternative efficace et durable à l'archivage de données et non pas au stockage d'informations quotidiennes, cette méthode de stockage d'informations numériques dans de l'ADN montre toujours son potentiel d'extension et de distribution en masse (« scalable ») pour pouvoir être mise à disposition du public, dépendant des investissements et de l'avancement des technologies qui semblent être favorables.

Remerciements :

Nous remercions Dr. Andrei-José Petrescu, PhD, Chef du Département de Bio-informatique et Biochimie Structurale de l'Institut de Biochimie de l'Académie Roumaine, Dr. Catalin Lazar, PhD, chercheur du Departement de Glycoprotéines Virales de l'Institut de Biochimie de l'Académie Roumaine et leurs collègues de l'Institut de Biochimie de l'Académie Roumaine de nous avoir aidé tant matériellement qu'intellectuellement tout au long du projet, Dr. Seth L. Shipman, PhD, chercheur postdoctoral du Département de Génétique de Harvard Medical School pour les conseils offerts et nos professeurs de sciences pour l'aide, les encouragements et les conseils qu'ils nous ont accordé.

Liens utiles :

Algorithmes du projet : https://github.com/alexbrt/DNAdrive

Glossaire :

A

- **Acide nucléique :** Polymères dont l'unité de base, ou monomère, est le nucléotide. Ces nucléotides sont reliés par des liaisons phosphodiesters. Un acide nucléique peut être soit de l'ADN, soit de l'ARN.
- **Allolactose :** Isomère de lactose qui induit l'opéron lac.
- **Arabinose :** Sucre qui est utilisé pour induire l'expression de gènes sous contrôle d'un opéron araBAD.
- **Archée :** Organisme vivant procaryote.

B

- **Bactérie :** Organisme vivant procaryote.
- **Complexe :** Ensemble de protéines. Les protéines peuvent s'associer pour former des complexes multiprotéiques.

C

- **Choc thermique :** Stress subi par un organisme vivant, suite à une augmentation soudaine de sa température au-delà d'une certaine valeur. Celui-ci facilite l'insertion d'ADN exogène dans une cellule mais son fonctionnement précis est inconnu.
- **Compétente :** Se dit d'une cellule capable d'intégrer de l'ADN extra-chromosomique tels que des plasmides à travers une électroporation (électrocompétence) ou un choc thermique (chimio-compétence).

D

- **dsDNA :** ADN à double brin.

E

- **Electroporation :** Technique microbiologique qui consiste à appliquer un champ électrique sur les membranes cellulaires pour les déstabiliser et augmenter leur perméabilité afin d'introduire de l'ADN exogène. Le fonctionnement précis de cette méthode est inconnu.
- **Endogène / endo- :** Qui provient de l'organisme.
- **Exogène / exo- :** Qui ne provient pas de l'organisme.
- **Expression hétérologue :** Expression d'un gène dans un organisme hôte qui ne le possède pas naturellement.
- **Extra-chromosomique :** Acide nucléique qui n'appartient pas au chromosome de l'organisme en question.

F

G

- **Gène :** Segment d'A.D.N. conditionnant la synthèse d'une ou de plusieurs protéines et, donc, la manifestation et la transmission d'un caractère héréditaire déterminé.
- **Gène régulateur :** Gène qui régule l'expression d'autres gènes.

H

- **Histone :** Protéine qui s'associe à l'ADN génomique pour l'empaqueter, le replier et le neutraliser.
- **Homopolymère :** Polymère dont la séquence est un monomère répété.

I

- **Intégrase :** Protéine qui catalyse l'étape dite d'intégration du cycle réplicatif des agents infectieux comme le rétrovirus. L'ADN étranger est ainsi « intégré » dans le génome de l'organisme hôte.
- **IPTG :** Sucre similaire à l'allolactose mais qui ne peut pas être hydrolysé par β-galactosidase (métabolisé), ce qui permet d'avoir une concentration constante d'agent inducteur tout au long de l'expérience.

J

K

- **Knockout :** Suppression de l'expression d'un gène.

L

- **Locus :** Région du chromosome.
- **Locus CRISPR :** Région d'ADN qui correspond au système immunitaire CRISPR-Cas.

M

- **Monomère :** Composé chimique pouvant subir une polymérisation, c'est-à-dire pouvant s'associer à une chaine de monomères.

N

- **Nucléophile :** Réactif dont la caractéristique principale est son attraction pour les espèces aux charges positives, contrairement au réactif électrophile.

O

- **Oligonucléotide :** ADN à simple brin court.

P

- **PCR :** Réaction en Chaîne par Polymérase (PCR est l'abréviation anglaise de « polymerase chain reaction »). Celle-ci est une méthode d'amplification génique qui permet d'amplifier / dupliquer exponentiellement une séquence d'ADN connue à l'aide d'amorces sens et antisens qui flanquent la région d'ADN ciblée.
- **Phage (bactériophage) :** Virus infectant exclusivement les bactéries.
- **Plasmide** : Molécule d'ADN circulaire distincte de l'ADN chromosomique, capable de réplication autonome et non essentielle à la survie de la cellule.
- **Polymère :** Chaine de monomères.
- **Primer (anglais) / Amorce (français) :** Courte séquence d'ARN ou d'ADN servant de point de départ à la synthèse d'un brin complémentaire à ce dernier modèle par une ADN polymérase.

Q

R

- **Régulation de l'expression des gènes :** Ensemble des mécanismes de régulations mis en œuvre pour passer de l'information génétique incluse dans une séquence d'ADN à un produit de gène fonctionnel (ARN ou protéine). Elle a pour effet de moduler, d'accroître ou de décroître la quantité des produits de l'expression des gènes (ARN, protéines).
- **Répresseur :** Protéine régulant négativement un ou plusieurs gènes en se liant à une séquence spécifique sur l'ADN appelée opérateur.

S

- **Surexpression génétique :** Expression augmentée d'un ou plusieurs gènes spécifiques.
- **ssDNA :** ADN à simple brin.

T

U

V

- **Virus :** Agent infectieux très petit qui possède un seul type d'acide nucléique (ADN ou ARN) et qui ne peut se reproduire qu'en parasitant une cellule.

W

X

Y

Z

Références et bibliographie :

Amitai G, Sorek R. 2016. CRISPR-Cas adaptation: Insights into the mechanism of action. Nat. Rev. Microbiol. 14:67–76. doi:10.1038/nrmicro.2015.14.

Bancroft C, Bowler T, Bloom B, Clelland CT. 2001. Long-term storage of information in DNA. Science (80-.). 293:1763–1765.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science (80-.). 315:1709–1712. doi:10.1126/science.1138140.

Barrangou R, Marraffini LA. 2014. CRISPR-cas systems: Prokaryotes upgrade to adaptive immunity. Mol. Cell 54:234–244. doi:10.1016/j.molcel.2014.03.011.

Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151:2551–2561. doi:10.1099/mic.0.28048-0.

Bonnet J, Colotte M, Coudy D, Couallier V, Tuffet S. 2010. Chain and conformation stability of solid-state DNA : implications for room temperature storage. 38:1531–1546. doi:10.1093/nar/gkp1060.

Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin E V., van der Oost J. 2008. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. 321:960–964.

Coté AG, Lewis SM. 2008. Mus81-Dependent Double-Strand DNA Breaks at In Vivo-Generated Cruciform Structures in S. cerevisiae. Mol. Cell 31:800–812. doi:10.1016/j.molcel.2008.08.025.

Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat. Commun. 3:945–947. doi:10.1038/ncomms1937.

Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. 2008. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. 190:1390–1400. doi:10.1128/JB.01412-07.

Díez-Villaseñor C, Guzmán NM, Almendros C, García-Martínez J, Mojica FJM. 2013. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biol. 10:792–802. doi:10.4161/rna.24023.

EIA. 2017. Electric Power Monthly with Data for November 2016. U.S. Energy Inf. Adm.

Fineran PC, Charpentier E. 2012. Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. Virology 434:202–209. doi:10.1016/j.virol.2012.10.003.

Gantz BJ, Reinsel D. 2011. Extracting Value from Chaos State of the Universe : An Executive Summary. :1–12.

Grissa I, Vergnaud G, Pourcel C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8:1–10. doi:10.1186/1471-2105-8-172.

Hayes RP, Xiao Y, Ding F, van Erp PBG, Rajashankar K, Bailey S, Wiedenheft B, Ke A. 2016. Structural basis for promiscuous PAM recognition in Type I-E Cascade from E. coli. Nature 530:499–503. doi:10.1038/nature16995.

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J. Bacteriol. 169:5429–5433. doi:10.1128/jb.169.12.5429-5433.1987.

Künne T, Kieper SN, Bannenberg JW, Vogel AIM, Miellet WR, Klein M, Depken M, Suarez-Diez M, Brouns SJJ. 2016. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol. Cell 63:852–864. doi:10.1016/j.molcel.2016.07.011.

Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature 520:505–510. doi:10.1038/nature14302.

Ma H, Marti-Gutierrez N, Park S-W, Wu J, Lee Y, Suzuki K, Koski A, Ji D, Hayama T, Ahmed R, et al. 2017. Correction of a pathogenic gene mutation in human embryos. Nature 548:413–419. doi:10.1038/nature23305.

Makarova KS, Craft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, et al. 2011. Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. 9:467–477. doi:10.1038/nrmicro2577.Evolution.

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. 2015. An updated evolutionary classification of CRISPR–Cas systems. Nat. Rev. Microbiol. 13:722–736. doi:10.1038/nrmicro3569.

Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155:733–740. doi:10.1099/mic.0.023960-0.

Nuñez J, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. 2015. Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature 527:535–538. doi:10.1038/nature15760.Foreign.

Nuñez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. 2016. CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol. Cell 62:824–833. doi:10.1016/j.molcel.2016.04.027.

Nuñez JK, Kranzusch PJ, Noeske J, Wright A V., Davies CW, Doudna JA. 2014. Cas1 – Cas2 complex formation mediates spacer acquisition during CRISPR – Cas adaptive immunity. Nat. Struct. Mol. Biol. 21:528–534. doi:10.1038/nsmb.2820.

Nuñez JK, Lee ASY, Engelman A, Doudna JA. 2015. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature 519:193–198. doi:10.1038/nature14237.

van der Oost J, Westra ER, Jackson RN, Wiedenheft B. 2014. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nat. Rev. Microbiol. 12:479–492. doi:10.1038/nrmicro3279.Unravelling.

Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M. 2004. Genetic

Analyses From Ancient DNA. Annu. Rev. Genet. 38:645–679. doi:10.1146/annurev.genet.37.110801.143214.

Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar H, Götherström A, et al. 2015. Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth. Curr. Biol. 25:1395–1400. doi:10.1016/j.cub.2015.04.007.

Partridge EA, Davey MG, Hornick MA, McGovern PE, Mejaddam AY, Vrecenak JD, Mesas-Burgos C, Olive A, Caskey RC, Weiland TR, et al. 2017. An extra-uterine system to physiologically support the extreme premature lamb. Nat. Commun. 8:15112. doi:10.1038/ncomms15112.

Pougach K, Semenova E, Bogdanova E, Datsenko KA, Djordjevic M, Wanner BL, Severinov K. 2010. Transcription, processing and function of CRISPR cassettes in Escherichia coli. Mol. Microbiol. 77:1367–1379. doi:10.1111/j.1365-2958.2010.07265.x.

Rajewska M, Wegrzyn K, Konieczny I. 2012. AT-rich region and repeated sequences - the essential elements of replication origins of bacterial replicons. FEMS Microbiol. Rev. 36:408–434. doi:10.1111/j.1574-6976.2011.00300.x.

Reinsel D, Gantz J, Rydning J. 2017. Data Age 2025 : Don't Focus on Big Data; Focus on the Data That's Big. :1–25.

Ren J, Xu P, Yong H, Zhang L, Liao S. 2017. Ground-to-satellite quantum teleportation. doi:10.1038/nature23675.

Shehabi A, Smith SJ, Sartor DA, Brown RE, Herrlin M, Koomey JG, Masanet ER, Horner N, Azevedo IL, Lintner W. 2016. United States Data Center Energy Usage Report. Lawrence …. doi:LBNL-1005775.

Shipman SL, Nivala J, Macklis JD, Church GM. 2017. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature 547:345–349. doi:10.1038/nature23017.

Shmakov S, Savitskaya E, Semenova E, Logacheva MD, Datsenko KA, Severinov K. 2014. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res. 42:5907–5916. doi:10.1093/nar/gku226.

Silvera IF, Cole JW. 2010. Metallic hydrogen: The most powerful rocket fuel yet to exist. J. Phys. Conf. Ser. 215. doi:10.1088/1742-6596/215/1/012194.

Smith AJ, Oertle J, Warren D, Prato D. 2016. Chimeric antigen receptor (CAR) T cell therapy for malignant cancers: Summary and perspective. J. Cell. Immunother. 2:59–68. doi:10.1016/j.jocit.2016.08.001.

Solomon M. 2012. Deleting the hns Gene for the H-NS Protein in Escherichia coli Reduces the Transformation Efficiency Following the Heat Shock Transformation Protocol. J. Exp. Microbiol. Immunol. 16:119–122.

Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. 2010. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet. 26:335–340. doi:10.1016/j.tig.2010.05.008.

Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ. 2012. CRISPR Interference Directs Strand Specific Spacer Acquisition. PLoS One 7:1–7. doi:10.1371/journal.pone.0035888.

Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. 2015. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell 163:840–853. doi:10.1016/j.cell.2015.10.008.

Westra ER, van Erp PBG, Künne T, Wong SP, Staals RHJ, Seegers CLC, Bollen S, Jore MM, Semenova E, Severinov K, et al. 2012. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol. Cell 46:595–605. doi:10.1016/j.neuron.2009.10.017.A.

Westra ER, Semenova E, Datsenko KA, Jackson RN, Wiedenheft B, Severinov K, Brouns SJJ. 2013. Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. PLoS Genet. 9. doi:10.1371/journal.pgen.1003742.

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, et al. 2007. Ancient Biomolecules from Deep Ice Cores Reveal a Forested Southern Greenland. Science (80-.). 317:111–114. doi:10.1126/science.1141758.

Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. 40:5569–5576. doi:10.1093/nar/gks216.

Yosef I, Shitrit D, Goren MG, Burstein D, Pupko T, Qimron U. 2013. DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. Proc. Natl. Acad. Sci. 110:14396–14401. doi:10.1073/pnas.1300108110.