O DNA como um novo dispositivo de armazenamento de dados

Danyel Fernandes Contiliani[1], Tiago Campos Pereira[2]

[1]Universidade de São Paulo, Programa de Pós-graduação em Genética, Ribeirão Preto, SP
[2]Universidade de São Paulo, Depto. de Biologia. Avenida Bandeirantes, 3900. CEP 14040-901, Ribeirão Preto, SP

Autor para correspondência - tiagocampospereira@ffclrp.usp.br

Palavras-chave: armazenamento de dados, CRISPR, DNA recording, tecnologia da informação

Genética na Escola – ISSN: 1980-3540

A elevada produção global de informação digital torna urgente o aprimoramento das tecnologias atuais de armazenamento de dados, como dispositivos magnéticos, ópticos e eletrônicos, empurrando-as cada vez mais próximas ao abismo da obsolescência. Diante das inúmeras características que inviabilizam a capacidade de armazenamento de dados a longo prazo por estes dispositivos tradicionais, o DNA surge como um material durável, replicável, resistente, recuperável e com alta capacidade de armazenamento. Dados computacionais podem ser traduzidos em um sistema análogo ao código genético e armazenados, no formato de oligonucleotídeos, em células por meio do uso da maquinaria do sistema imune procariótico, CRISPR. A partir do sequenciamento genômico, tais informações podem ser decodificadas novamente à linguagem computacional e, portanto, reproduzidas. Assim, o DNA *recording* é uma solução promissora para problemas atuais da tecnologia da informação e uma aposta para a próxima geração de dispositivos de armazenamento de dados.

O ácido desoxirribonucleico (DNA) consiste em uma macromolécula complexa capaz de armazenar as inúmeras informações essenciais para o desenvolvimento e manutenção de todas as formas de vida. Em geral, sua estrutura possui duas cadeias compostas por nucleotídeos (com as seguintes bases nitrogenadas - adenina, timina, citosina e guanina), os quais realizam pareamentos específicos entre si por meio de ligações de hidrogênio (modelo conhecido como pareamento de bases Watson-Crick). Em poucas palavras, após uma sequência de DNA ser transcrita em RNA mensageiro (RNAm), agrupamentos de três nucleotídeos – chamados de códons – são traduzidos em aminoácidos correspondentes a diferentes combinações destes conjuntos de bases. Dessa forma, o código genético garante instruções às nossas células para construir cada uma das proteínas do nosso organismo.

Curiosamente, um sistema análogo ao código genético (código de interpretação de aminoácidos a partir dos códons) também pode ser utilizado como uma forma de **criptografia**. Em 1999, pesquisadores da Escola de Medicina de Monte Sinai, em Nova Iorque, escreveram uma mensagem secreta baseada em sequências de DNA. Após a conversão de caracteres do alfabeto romano (letras e números) em 64 possíveis combinações de códons (trincas de bases nucleotídicas no RNAm), sintetizaram moléculas de DNA contendo uma mensagem entre um par de **marcadores genéticos** e as armazenaram em uma carta, a qual foi enviada para os próprios autores via correio. Quando eles a receberam, utilizaram a técnica de **PCR** (do inglês, *Polymerase chain reaction*) para amplificar o número das moléculas de DNA escondidas no ponto final do texto redigido. Assim, após o sequenciamento deste DNA, os pesquisadores identificaram a sequência nucleotídica específica e, traduzindo-a novamente para o alfabeto romano, foram capazes de decodificar a mensagem com sucesso (Figura 1).

E, se nós pudéssemos não apenas armazenar informações de texto no DNA, mas também informações digitais (em **código binário**), como fotos, músicas e filmes? Seríamos capazes de reescrever informações tão robustas e complexas em estruturas nanoscópicas como os genomas? Se sim, poderemos tornar o DNA em um novo substituto para os CDs, DVDs e *pendrives* em um futuro próximo. Neste texto, mostraremos como o casamento entre a tecnologia da informação e as ferramentas atuais de biologia molecular podem transformar o DNA em um dispositivo de armazenamento de dados.

Marcadores genéticos: Informações que caracterizam e/ou diferenciam indivíduos ou organismos. Em especial, aqui a expressão "marcadores genéticos" refere-se a sequências nucleotídicas que caracterizam uma determinada informação.

PCR: Acrônimo para *Polymerase Chain Reaction* (em português, Reação em Cadeia da Polimerase). Consiste em uma técnica de biologia molecular, capaz de amplificar o número de fragmentos específicos de um genoma, gerando milhões de cópias.

Código binário: Linguagem básica de instrução computacional, na qual os valores são representados por 0 e 1.

Criptografia: Técnica de proteção a uma dada informação secreta que apenas o remetente e o destinatário podem acessar.

Genética na Escola – ISSN: 1980-3540

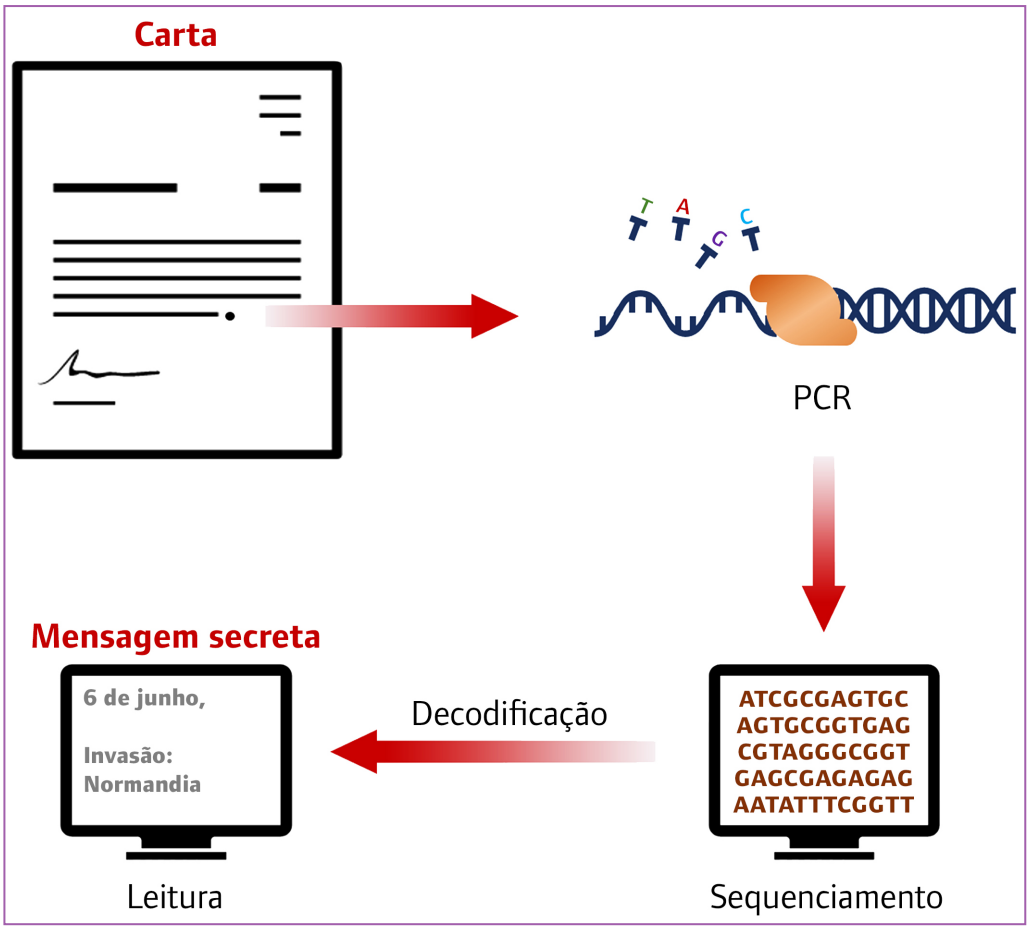

Figura 1.
Criptografia de DNA. De forma esquematizada, a imagem retrata a amplificação das moléculas de DNA (escondidas no ponto final da carta) por meio da técnica de PCR (do inglês, *Polymerase chain reaction*). Em seguida, as sequências nucleotídicas amplificadas são decifradas por sequenciamento de DNA e, por fim, decodificadas para o alfabeto romano, revelando a mensagem secreta. Elementos gráficos da *Noun Project:* DNA – ProSymbols; Monitor – Andy Mc (CC BY 2.0).

Tecnologias de armazenamento de informações

A informação sempre foi importante para a sobrevivência do ser humano e, portanto, a memória escrita serviu como um fator determinante para o progresso cultural da sociedade. Atualmente, a informação não consiste apenas em textos, mas também em imagens, vídeos, áudios e softwares complexos. Portanto, considerando tais mudanças e a elevada produção de dados, tornou-se necessária a criação dos diferentes meios de armazenamento: os meios (i) magnético, (ii) óptico e (iii) eletrônico.

O formato mais comum e seguro de armazenamento de informações é o magnético, sendo representado pela fita magnética, pelos disquetes e pelo atual disco rígido (HD). Já o armazenamento por meio óptico, como CDs e DVDs, foram muito úteis para a gravação e reprodução de informações multimídia (*e.g.*, filmes, música), pois a densidade de dados armazenados é muito maior, além da qualidade superior das reproduções. Por fim, representado por *pen drives*, cartões de memória e os SSDs (do inglês, *Solid state drives*), o armazenamento eletrônico é a tecnologia mais recentemente difundida deste campo e possui a vantagem de acesso rápido à informação.

Embora estas tecnologias sejam atuais e continuem em constante aprimoramento, todas já estão atingindo seus limites de densidade de dados (bits para cada unidade de volume físico), uma vez que estamos produzindo mais informações do que podemos armazenar. Além disso, o tempo de vida útil de discos rígidos pode chegar até a 30 anos se armazenados em condições controladas, *i.e.*, inúmeros refrigeradores para a manutenção da temperatura. Por fim, cada um destes dispositivos requer diferentes leitores, os quais são rapidamente substituídos devido à alta obsolescência destas tecnologias. Por exem-

Bits: Um bit (do inglês, *Binary Digit*) consiste em uma unidade binária, correspondendo à menor parcela de informação computacional, assumindo somente 2 valores (0 ou 1).

Genética na Escola – ISSN: 1980-3540

plo, embora os disquetes tenham sido muito utilizados nos anos 1990, nenhum dos computadores atuais possui leitores para este tipo de dispositivo – inclusive muitos *notebooks* já não apresentam leitores de CD e DVD.

Quando o código genético foi reportado pela primeira vez como uma possível forma de criptografia, a comunidade científica e empresas de tecnologia voltaram os olhos para uma iminente solução aos problemas das tecnologias atuais de armazenamento de dados. Afinal, se era possível escrever mensagens de texto em moléculas de DNA, por que não seria possível a inserção de dados cada vez mais robustos? Para isso, assim como aprendemos a traduzir textos de um idioma para outro em uma folha de papel, é necessário aprendermos como traduzir os dados complexos em código genético e escrevê-los em um DNA. De forma intrigante, as bactérias serão as nossas professoras.

1.1 CRISPR

Desde os primórdios, em meio às guerras contra **bacteriófagos**, as bactérias reescreveram os seus próprios genomas como uma estratégia de defesa. Em alguns genomas procarióticos existem lócus (regiões) compostos por sequências nucleotídicas repetidas que se intercalam com sequências espaçadoras. Estes lócus são conhecidos como CRISPR (Repetições Palindrômicas Curtas Agrupadas e Regularmente Interespaçadas) e, junto a um conjunto de genes Cas (genes associados às CRISPR), formam um sistema imune adaptativo de procariotos, o qual se resume em três etapas: (i) adaptação, (ii) biogênese do crRNA e (iii) interferência.

A etapa de **adaptação** ocorre quando uma bactéria é infectada pela primeira vez por um bacteriófago e o complexo enzimático Cas1-Cas2 cliva o material genético viral em pequenos fragmentos (~32 nt), subsequentemente integrando-os no lócus de CRISPR como sequências espaçadoras. Em seguida, a transcrição deste lócus composto por sequências virais leva à **biogênese do crRNA** (RNA derivado de CRISPR), que consiste em pequenas moléculas de RNA correspondentes aos espaçadores. A etapa final – **interferência** - ocorre quando a bactéria é infectada novamente pelo bacteriófago e a enzima Cas9 forma um **complexo ribonucleoproteico** com o crRNA que reconhece e, por fim, destrói a sequência genética viral.

O entendimento do sistema biológico CRISPR-Cas abriu portas para o uso de uma maquinaria no desenvolvimento de ferramentas biotecnológicas com aplicações na medicina, na agricultura, na indústria e na ciência básica. Devido às profundas implicações da técnica de CRISPR-Cas na sociedade, as pesquisadoras Emmanuelle Charpentier e Jennifer A. Doudna foram laureadas ao Prêmio Nobel de Química em 2020 pelo desenvolvimento desta tecnologia. Dentre tantas possibilidades para o uso do sistema CRISPR-Cas, uma nova aplicação tecnológica foi recentemente reportada: o DNA *recording*.

2. DNA *recording* por CRISPR

O DNA *recording* consiste no armazenamento de informações específicas (digitais ou não) em ácidos nucleicos. Em linhas gerais, de acordo com Shipman et al. (2017), o processo se resume a três etapas centrais: (i) escrita, (ii) armazenamento e (iii) leitura (Figura 2).

2.1 Princípios moleculares

Inicialmente, a **escrita** é feita a partir da codificação da informação digital (em código binário) em sequências de nucleotídeos (código genético). Ao supor que estamos trabalhando com uma imagem a ser codificada, precisamos traduzir a informação de cada **pixel** em código genético. Para isto, Shipman e seus colaboradores (2017) estabeleceram um dicionário atribuindo valores binários a cada base nucleotídica (*i.e.*, C = 00, T = 01, A = 10 e G = 11) e, em seguida, sintetizaram uma cadeia de DNA com 28 nucleotídeos (denominados de **oligonucleotídeos**) que codificam 9 pixels, cada um codificado por uma trinca de bases (Figura 2A). Desta forma, a informação binária de cada conjunto de 9 pixels da imagem foi traduzida em

Complexos ribonucleoproteicos: Agregados formados por RNA e proteínas.

Bacteriófagos: Vírus que infectam bactérias.

Pixel: Aglutinação das palavras em inglês, ***Pic****ture* e ***El****ement*. Consiste no menor elemento de uma imagem digital, composto por 3 pontos (verde, vermelho e azul), os quais formam combinações que definem 256 tonalidades diferentes quando oito bits são usados para a informação de cada cor primária.

Oligonucleotídeos: polímeros curtos de ácidos nucleicos, geralmente contendo entre seis e trinta nucleotídeos.

Genética na Escola – ISSN: 1980-3540

um oligonucleotídeo diferente, cada um especificado com um código de barras – uma pequena sequência inicial de quatro nucleotídeos, chamada de "pixet".

A próxima etapa – **armazenamento** – consiste em guardar estas moléculas em tubos de ensaio (*in vitro*) ou em células (*in vivo*). Em especial, é na abordagem de armazenamento *in vivo* que o sistema imune procariótico – CRISPR – apresenta-se como um protagonista. Neste sentido, os oligonucleotídeos são administrados em uma população de bactérias que apresentam o lócus CRISPR em seus genomas e superexpressam o complexo enzimático Cas1-Cas2. Como espaçadores, estas sequências exógenas são incorporadas ao genoma da bactéria, de maneira similar ao mecanismo de adaptação a infecções virais (Figura 2B).

Por fim, após o crescimento das bactérias, já se torna possível a **leitura** da informação armazenada. Utilizando a técnica de PCR, os arranjos de CRISPR devem ser amplificados e, posteriormente, submetidos ao sequenciamento. Os resultados do sequenciamento são filtrados por uma série de algoritmos, incluindo o isolamento dos espaçadores recém adquiridos pelas bactérias (Figura 2C). Por fim, a decodificação dos oligonucleotídeos novamente à linguagem computacional possibilita a reprodução da informação original.

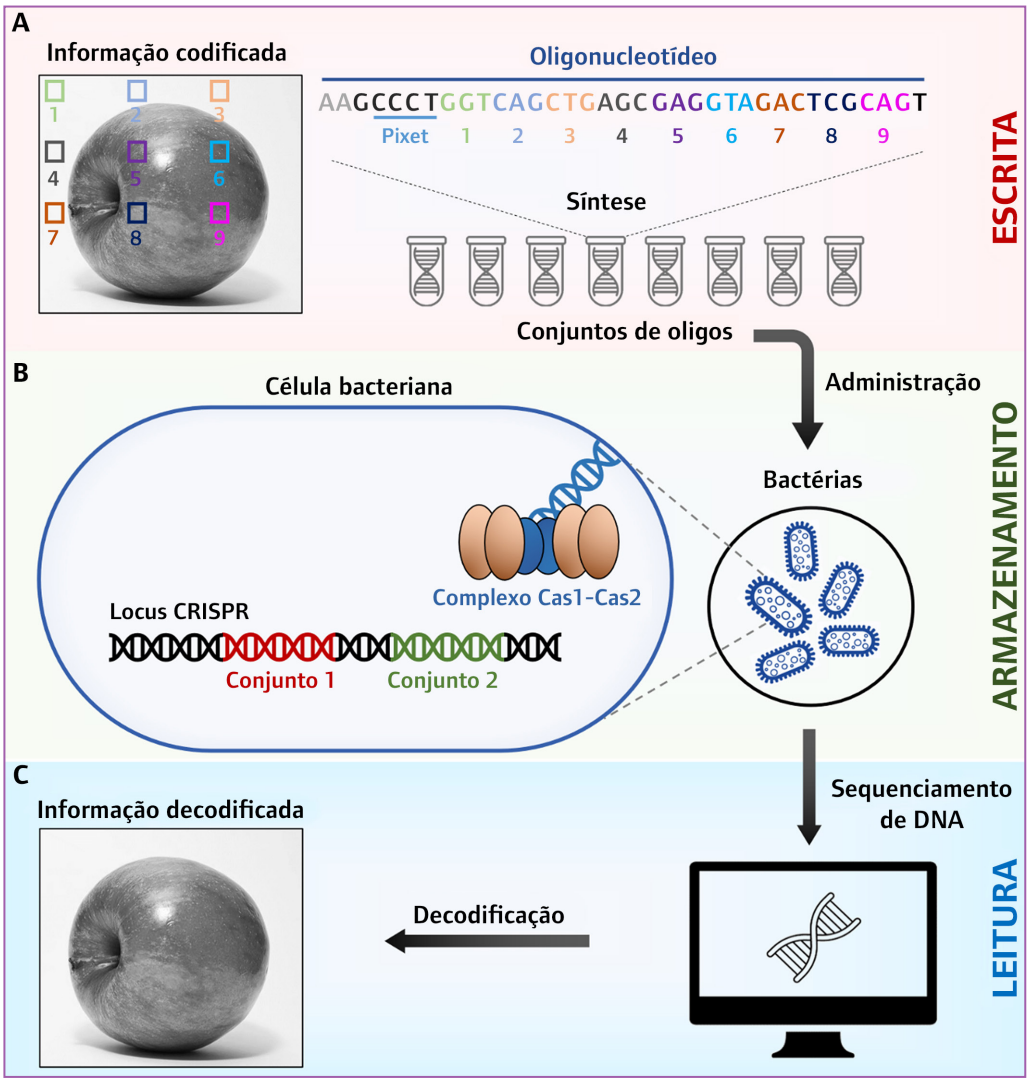

Figura 2.
Esquema do processo de DNA *recording* por CRISPR. Em **A**, conjuntos de pixels de uma imagem são geneticamente codificados e sintetizados em oligonucleotídeos. Em **B**, as moléculas são introduzidas na célula bacteriana, cujo sistema CRISPR trabalha na incorporação das oligonucleotídeos no genoma. Em **C**, o DNA bacteriano é sequenciado e decodificado em linguagem computacional para a reconstrução da informação armazenada. Elementos gráficos da *Noun Project:* DNA – Olivia; Vectors Market (CC BY 2.0).

Genética na Escola – ISSN: 1980-3540

Aplicações

CRISPR não se limita apenas ao armazenamento de informações estáticas, como textos e imagens, no DNA. De maneira similar ao exemplo anterior, Shipman et al. (2017) também executaram o armazenamento de um GIF adaptado de "O Cavalo em Movimento" por Eadweard Muybridge (1878), que consiste em uma sequência de 5 fotografias. Por conta de ser uma informação dinâmica, cada quadro do GIF foi codificado em um conjunto de 104 oligonucleotídeos espaçadores. Assim, cada célula bacteriana foi capaz de armazenar um conteúdo de 2.6 quilobytes (veja mais em: https://www.youtube.com/watch?v=yWqlwYjpc1A).

Um fato curioso é que, conforme as moléculas exógenas de DNA são introduzidas à célula bacteriana, estas são integradas como espaçadores no lócus CRISPR de maneira sequencial, como um registro cronológico de eventos biológicos. Desta forma, DNA *recording* por CRISPR também possibilita o monitoramento de atividades biológicas. Em um estudo recente, um plasmídeo responsivo a estímulos foi introduzido na bactéria *Fusicatenibacter saccharivorans*, que naturalmente apresenta a expressão do complexo Cas-1-Cas2 fusionado à enzima **transcriptase reversa** (TR). Conforme a bactéria sofre um determinado estímulo, a transcrição do plasmídeo exógeno é desencadeada. Subsequentemente, as moléculas de RNA produzidas são reversamente transcritas em DNA pela TR e, por fim, integradas ao genoma bacteriano pelo complexo Cas1-Cas2. Assim, os pesquisadores reportaram a tecnologia baseada em CRISPR, Record-seq, como uma ferramenta promissora para o monitoramento transcricional *in vivo*.

3. Vantagens e desvantagens

Como uma iminente solução para os problemas atuais de armazenamento de dados em dispositivos magnéticos, ópticos e eletrônicos, o DNA apresenta propriedades altamente relevantes: **a alta densidade de informações** (215 petabytes para cada 1 grama de material genético), em que 1 grama de DNA é capaz de armazenar cerca de 215 petabytes; a **elevada durabilidade** (de séculos a milênios), visto que é possível recuperar informações genéticas de fósseis e/ou microrganismos preservados em geleiras; a **alta resistência a condições ambientais extremas**, considerando o uso da bactéria *Deinococcus radiodurans* como uma proteção biológica, devido à sua capacidade de tolerar dessecação, altas temperaturas, luz ultravioleta e doses de radiação ionizante mil vezes maiores do que a tolerada por seres humanos; a **facilidade de replicação**, uma vez que a técnica de PCR pode criar milhões de cópias de uma determinada informação em poucas horas; por último, a **baixa obsolescência**, visto que, em cerca de 4 bilhões de anos de vida na Terra, o DNA é a forma mais dispersa e antiga de armazenamento de informações.

Petabytes: Um **byte** consiste em um conjunto de 8 bits. Um **petabyte** consiste em 10^{15} bytes ou 1.000 terabytes.

Entretanto, a tecnologia de DNA *recording* ainda é recente, pouco investigada e requer aperfeiçoamentos. Enquanto isso, alguns fatores ainda inviabilizam seu uso comercial e em larga escala, como: o **custo** da síntese e do sequenciamento de DNA ainda é altamente elevado, de tal forma que o armazenamento de 1 MB (megabyte) no DNA custaria em torno de 3.500 dólares; o **ruído** gerado pelas imperfeições na síntese de DNA e pelos erros de sequenciamento; o **acesso direcionado** a uma determinada informação no DNA ainda é prejudicado, devido à necessidade de uma nominalização para cada informação dentro de um imenso conjunto de dados; por fim, o **desempenho** da tecnologia ainda é muito baixo, uma vez que a velocidade de síntese química/enzimática de DNA é muito inferior à velocidade de armazenamento de dados em dispositivos eletrônicos.

Transcriptase reversa: Enzima que realiza o processo de síntese de um DNA complementar (cDNA) a partir de um molde de RNA. Este processo é conhecido como transcrição reversa pois é feito em sentido contrário à transcrição de DNA para RNA.

4. Perspectivas

A quantidade de informação gerada nos últimos anos tem crescido descontroladamente, excedendo os limites físicos para o seu armazenamento. O advento da tecnologia de DNA *recording*, que utiliza ácidos nucleicos para o armazenamento de dados, abre portas para a resolução deste importante problema,

Genética na Escola – ISSN: 1980-3540

possibilitando o armazenamento de toda a informação estimada a ser produzida até 2025 (175 **zettabytes**) em uma única sala. Entretanto, o importante gargalo da performance de síntese e sequenciamento de DNA ainda precisa ser resolvido. Para isso, a automação do armazenamento de dados em DNA é um dos focos principais do *DNA Data Storage Alliance*, um grupo formado por mais de dez organizações importantes, como a Microsoft e a Universidade de Washington. Assim, nos próximos anos, os avanços destes grupos poderão viabilizar o uso de DNA como uma solução para lacunas atuais das tecnologias de armazenamento de dados.

Para saber mais

CAMPBELL, Mark. DNA data storage: automated DNA synthesis and sequencing are key to unlocking virtually unlimited data storage. *Computer*, v. 53, n. 4, p. 63-67, 2020.

DE CARLI, Gabriel J. et al. A revolucionária técnica de edição genética "CRISPR". *Genética na escola*, v. 12, n. 2, 114-123.

SHIPMAN, Seth L. et al. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*, v. 547, n. 7663, p. 345-349, 2017.

Zettabytes: Unidade de medida equivalente a 10^{21} bytes ou 10^{9} terabytes.