

ARMAZENAMENTO DE DADOS EM DNA: REVISÃO SISTEMÁTICA DOS MECANISMOS DE CODIFICAÇÃO E DECODIFICAÇÃO DE DADOS

Maria Carolina Azevedo Veiga¹, Fabrício Gustavo Henrique², Anna Patricia Zakem China³

¹Faculdade de Tecnologia de Ribeirão Preto (FATEC)

Ribeirão Preto, SP – Brasil

maria.veiga@fatec.sp.gov.br,
fabricio.henrique@fatec.sp.gov.br,
anna.china@fatec.sp.gov.br

Resumo. *Estima-se que, até 2025, a humanidade terá produzido cerca de 175 zettabytes de dados, mas os dispositivos de armazenamento atuais não acompanham esse crescimento, gerando um déficit iminente. Armazenar dados em DNA surge como uma possível solução, devido à alta densidade e durabilidade da macromolécula. Este artigo revisa sistematicamente os mecanismos de codificação e decodificação no armazenamento de dados em DNA. Observou-se diversas técnicas que apresentam a redução de custos, de tempo de síntese, melhora na correção de erros e na densidade de armazenamento. As taxas de correção de erros podem chegar a 98% e a densidade de 1,98 bits/nt. Perspectivas futuras reforçam a viabilidade do armazenamento de dados em DNA.*

Abstract. *It is estimated that by 2025, humanity will have produced around 175 zettabytes of data, but current storage devices are not keeping pace with this growth, creating an imminent deficit. Storing data in DNA emerges as a potential solution due to the macromolecule's high density and durability. This article systematically reviews the encoding and decoding mechanisms used in DNA data storage. Various techniques have been observed that reduce costs, synthesis time, improve error correction, and increase storage density. Error correction rates can reach up to 98% and a density of 1.98 bits/nt. Future perspectives reinforce the feasibility of DNA data storage.*

1. Introdução

O processo de armazenamento da informação está intimamente relacionado com o desenvolvimento humano. Os mecanismos de registro evoluíram, das cavernas/pedras aos computadores. Atualmente, com a relativa facilidade de acesso e a evolução da tecnologia, a humanidade está produzindo volumes cada vez maiores de dados, como nunca visto.

Somado ao crescente acesso há divergência entre a taxa de crescimento de produção de dados e a capacidade de armazenamento dos dispositivos, tais como HDs, SSDs, entre outros. Assim, surgem estudos que almejam solucionar essa problemática, como, por exemplo, explorar a ideia de armazenamento de dados em DNA. Observando a estrutura e todos os estudos com o DNA ao longo dos anos, é possível ver o quanto ele

é durável e possui uma alta densidade de armazenamento de dados.

Destarte, o presente trabalho realiza uma revisão sistemática da literatura sobre o armazenamento de dados em DNA (*DNA Storage*). O objetivo é investigar e sintetizar as técnicas de codificação e decodificação, buscando comparar os princípios, as vantagens e as desvantagens de cada técnica, além de explorar tendências futuras.

O artigo está estruturado em cinco partes principais: introdução, referencial bibliográfico, metodologia, resultados e conclusão. Especificamente, o referencial teórico busca esclarecer o tema de estudo. Nele, há a seção ‘2. Panorama geral’, ‘2.1. Ácido desoxirribonucleico - DNA’, ‘2.2. Armazenamento de dados em DNA’, ‘2.2.1. Funcionamento geral do armazenamento de dados em DNA’ e ‘2.2.2. Desafios relacionados ao armazenamento de dados em DNA’. No tópico seguinte, temos a metodologia com o intuito de esclarecer todos os passos realizados para a execução da pesquisa. Este tópico está estruturado nas seções ‘3. Metodologia’, ‘3.1. Delimitação do método’ e ‘3.2. Procedimento adotado’. Logo após, há a apresentação dos resultados obtidos no tópico ‘4. Resultados’ e ‘4.1. Técnicas’. Por fim, a seção ‘5. Considerações Finais’, apresentam se os objetivos do artigo foram atingidos e as dificuldades enfrentadas.

2. Panorama geral

De modo geral, um dado pode ser considerado como aquilo que conseguimos registrar, considerando desde pinturas rupestres e escritos em papel aos registros nos computadores (COSTA e PINTO, 2017). Historicamente, os dispositivos de armazenamento de dados foram desde os cartões perfurados, tubo de Williams, tambor de memória, Fita UNISERVO, RAMAC, entre outros que são antigos e pouco conhecidos nos dias de hoje, até os obsoletos disquetes e CDs, chegando aos modernos discos rígidos, SSDs – *Solid-State Drive*, cartões de memória, pen drives e a computação em nuvem (Cloud Computing). Esta evolução ocorreu no período aproximado de 140 anos, considerando o aumento da produção e a importância do tratamento de dados para a sociedade (COSTA e PINTO, 2017).

Estima-se que, até 2025, poderá ser entre 175 zettabytes (BUKO et al., 2023) e 250 ZB de dados no mundo (SAIS et al., 2023). Pressupõe-se que, desde 1956, a quantidade de informações dobre a cada 13 meses, diferentemente da densidade de armazenamento, que desacelerou nas últimas décadas. Por assim ser, estima-se que haja uma lacuna entre o volume de informações produzidas e a capacidade de armazenamento por área - densidade (CAMPBELL, 2020).

Os métodos de armazenamento atuais, tais como os magnéticos HD, SSD e ópticos, possuem, respectivamente, vantagens como confiabilidade (OKUDA, 2015), velocidade de armazenamento (YE et al., 2019) e durabilidade (RUAN, 2016). Porém, todos possuem uma densidade de armazenamento baixa quando analisamos as previsões de crescimento de dados, além das suas limitações específicas.

Tendo em vista este desafio, o armazenamento em Ácido Desoxirribonucleico - DNA sintético é como uma possível solução, sendo este inspirado na eficiência da natureza em compactar dados em espaços mínimos dentro de células microscópicas,

oferecendo excelente densidade e durabilidade (SAIS et al., 2023).

2.1. Ácido Desoxirribonucleico - DNA

Ácido Desoxirribonucleico - DNA, é uma macromolécula que detém a função de armazenar todas as informações genéticas (características hereditárias) dos seres vivos (KATSURA, 2018). Ela também possui papel central na síntese de proteínas no interior das células, garantindo e determinando a existência e o desenvolvimento dos seres vivos (LEHNINGER et al., 2014).

Historicamente, há dois marcos sobre o início dos estudos do DNA e da genética. O primeiro é o estudo com ervilhas realizado por Mendel em 1866, marcando o início da área de estudo das características hereditárias, a genética. O segundo é a exploração do DNA pelo físico Miescher em 1868 (BHATTACHARJEE; LEE, 2020).

Com o desenvolvimento da área, descobriu-se que o DNA possui, como é possível visualizar na Figura 1, uma estrutura em formato de dupla hélice, com 2 nm de diâmetro (KATSURA, 2018) e aproximadamente 2 metros de comprimento por cada filamento (NAOUM, 2010). Ele é composto por bases nitrogenadas, açúcar (desoxirribose) e fosfato. As 4 diferentes bases são: (A) adenina, (G) guanina, (T) timina e (C) citosina (KATSURA, 2018). Pelos estudos de Erwin Chargaff, observando as proporções e quantidades de cada base na macromolécula de DNA, foi possível determinar que a base nitrogenada guanina se liga com a citosina e a adenina se liga com a timina na dupla hélice do ácido desoxirribonucleico (BHATTACHARJEE; LEE, 2020).

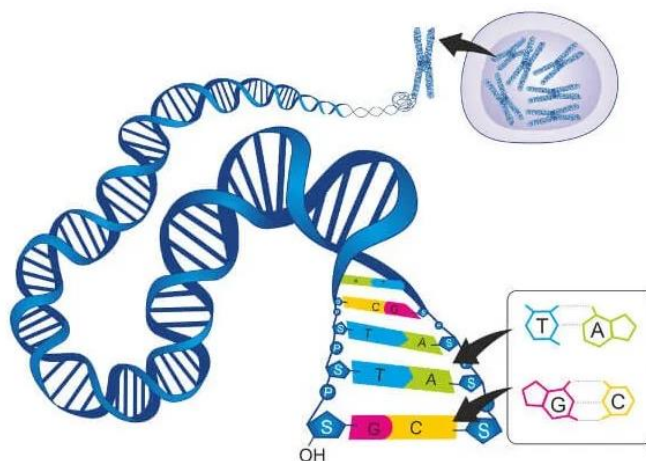


Figura 1. Estrutura da macromolécula de DNA

Fonte: (Shutterstock¹)

Sua organização e estrutura conferem ao DNA uma estabilidade química tão alta que, em muitos aspectos, ele se aproxima da inércia química. Isso resulta em uma longa durabilidade, possibilitando sua preservação por milhares de anos e facilitando sua recuperação e identificação (NAOUM, 2010). Os estudos na área de sequenciamento genético, possibilitaram ampliar a compreensão, a manipulação e a síntese de novas

¹ <https://www.shutterstock.com/pt/image-illustration/schematic-illustration-shows-structure-double-stranded-1112556206>

moléculas de DNA (BHATTACHARJEE; LEE, 2020). Tendo em vista esses fatores, a macromolécula de DNA passa a ser considerada como uma alternativa para o armazenamento de dados.

2.2. Armazenamento de dados em DNA

Em 1959 Feynman teorizou a construção de um computador em nível molecular, especulando sobre a possibilidade de armazenamento de dados em DNA (XU et al., 2020). Estima-se que o primeiro estudo prático na área ocorreu em 1988, codificando um texto por meio do uso de DNA (CHURCH et al., 2012). Com o passar dos anos, os estudos na área vão crescendo, armazenando imagens simples, textos e iniciando processos de decodificação, para ler os dados gravados. Em 2012, em Harvard, Church armazena um livro completo em DNA. Este continha 50.000 palavras e representa o marco que impulsiona a vasta gama de estudos na área de armazenamento de dados em DNA (XU et al., 2020).

Durante 3 bilhões de anos, a natureza tem mostrado como armazenar um volume grande de dados em pouco espaço por meio do DNA (SAIS et al., 2023). Em relação à densidade de armazenamento (quantidade de dados armazenados em uma determinada área), o DNA possui um limite estimado de 1 exabyte por mm^3 , desconsiderando as possíveis perdas (mutação, adição, remoção). Deste modo, não há outros meios de armazenamento que foram desenvolvidos e/ou que estejam em desenvolvimento capazes de alcançar a mesma densidade (BUKO et al., 2023).

Para além da capacidade de armazenamento alta, a durabilidade do DNA é outro fator que desperta interesse. Ele possui um tempo de meia-vida estimado em mais de 500 anos (BORNHOLT et al., 2016), quando armazenado em condições corretas, evitando a interferência de radiação, produtos químicos, temperatura e umidade (XU et al., 2020).

Em termos de consumo energético, os dispositivos de armazenamento atual apresentam grandes gastos, sendo este um fator de preocupação quando abordamos a sustentabilidade e a taxa de emissão de carbono para a atmosfera. Estima-se que cerca de 2% da produção mundial de energia seja destinada aos DataCenters, levando à produção de diversos trabalhos na área com o intuito de reduzir o consumo (FREITAS, 2023). No armazenamento em DNA, o uso de energia para a escrita de dados gira em torno 10^{-19} joules/bit, enquanto as outras alternativas para armazenamento que estão sendo estudadas necessitam de aproximadamente 10^{-15} joules/bit (MANICKA et al., 2023). Este consumo é baixo, ainda mais considerando que a escrita dos dados é uma das etapas que mais consome energia.

Até o momento, há relatos na literatura de armazenamento de texto, músicas e imagens já realizados utilizando DNA Storage (armazenamento em DNA) (BUKO, TUCZKO e ISHIKAWA, 2023). É possível entender melhor a evolução de armazenamento de arquivos em DNA, por alguns marcos, tais como os que estão

representados na Figura 2.

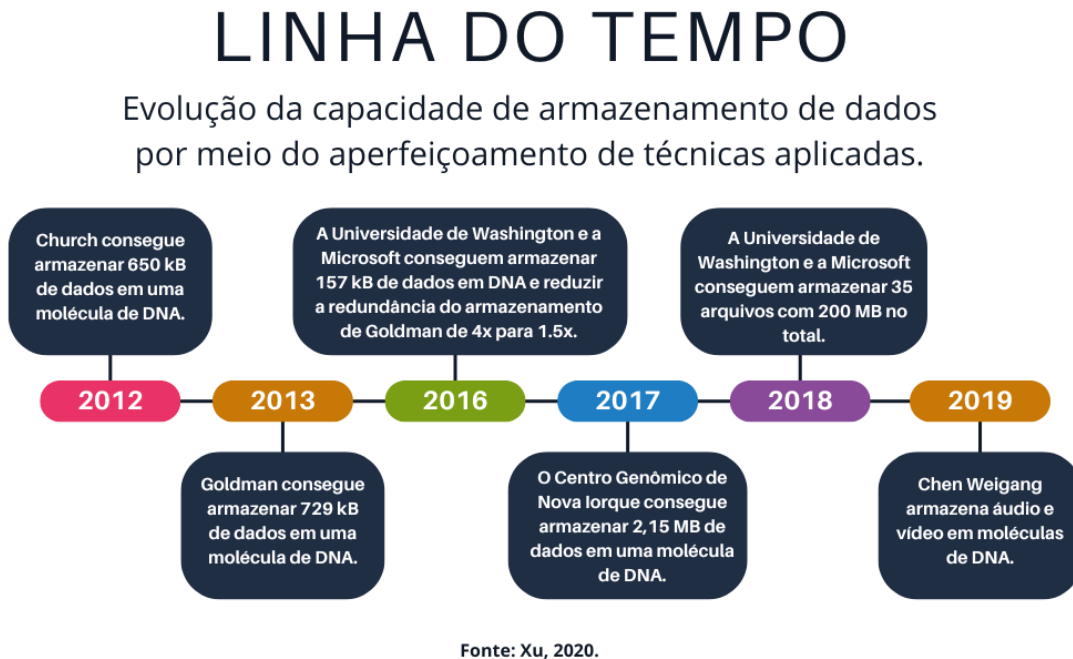


Figura 2. Breve linha do tempo com a evolução do armazenamento de dados em DNA.

Fonte: (feita pela autora com base em Xu, 2020)

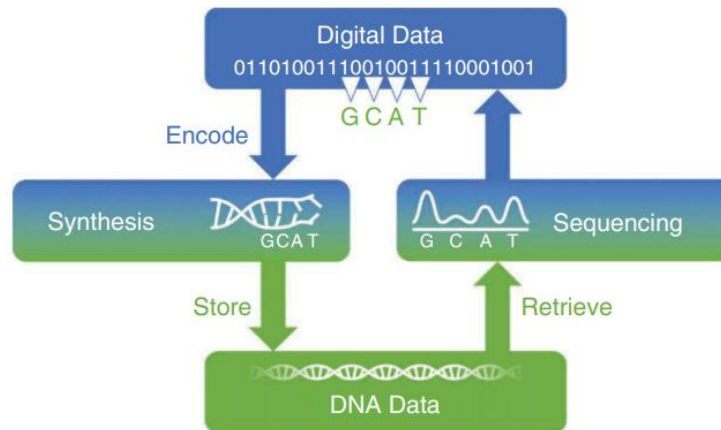
Este conjunto de fatores tornam a tecnologia relevante, atual e promissora. Deste modo, ela tem ganhado cada vez mais espaço, com o desenvolvimento de pesquisas como por exemplo na Big Tech Microsoft, com o grupo *Data Storage Alliance*, além de diversas universidades e feitas por cientistas de diversos países (CONTILIANI e PEREIRA, 2022). Em 2019, ela ficou entre as 10 tecnologias inovadoras no ranking da revista “Scientific American” (XU et al., 2020).

2.2.1. Funcionamento geral

Para Buko, Tuczko e Ishikawa (2023), a técnica de armazenamento em DNA é a mais promissora entre todos os estudos atuais, comparada com pesquisas realizadas com fitas magnéticas. Existem diversas técnicas que podem ser utilizadas, cada uma possuindo, de modo específico, o passo a passo para a sua codificação/decodificação. A ideia é apresentar neste tópico uma abstração do funcionamento; os métodos levantados durante a pesquisa serão detalhados no tópico “4.0 Resultados”.

Vamos imaginar que a imagem simples de uma maçã será armazenada. Inicialmente, a proposta é codificar os dados que estão em código binário (máquina) para cada pixel, em uma base quaternária/código genético (a sequência de nucleotídeos Guanina-G, Timina-T, Citosina-C e Adenina-A). Nesta etapa, cada base nitrogenada recebe um conjunto de bits (CONTILIANI; PEREIRA, 2022), como pode ser visto na

Figura 03.

**Figura 3. Ciclo de vida do armazenamento por DNA****Fonte: (Campbell, 2020)**

Após a atribuição, a ordem das bases nitrogenadas corresponderá à sequência de bits que formam a imagem. A partir desse ponto, é realizada a síntese do DNA, formando a exata molécula que corresponde à ordem de armazenamento de informação do conjunto de bits (CAMPBELL, 2020), no nosso caso hipotético, os dados do armazenamento da imagem da maçã.

O DNA sintetizado é armazenado. Quando houver a necessidade de recuperar a imagem armazenada, o DNA que está guardado é sequenciado de modo a estabelecer a ordem de bases nitrogenadas que o compõem. Nesse ponto, há novamente técnicas de decodificação que buscam minimizar os erros da síntese e facilitar a leitura dos dados. Os métodos que aparecem nos artigos analisados serão apresentados detalhadamente no tópico “4.0 Resultados”. Após o sequenciamento do DNA, cada base nitrogenada recebe os bits correspondentes e o processo inverso é realizado. A ordem de bits é convertida na imagem (CAMPBELL, 2020), como é possível observar na Figura 3.

2.2.2. Desafios relacionados ao armazenamento de dados em DNA

Apesar de ser uma técnica promissora, há pontos de atenção que os pesquisadores buscam melhorar para viabilizar sua aplicação em massa. Aqui serão citados os principais desafios enfrentados. O primeiro deles é o custo para a produção do DNA. Em 2017, o valor estimado para a síntese e armazenamento de 1 MB de dados era de US\$ 3.500,00 (CAMPBELL, 2020). Essa estimativa leva em conta que, para realizar manutenções nos dados, o valor não é alto, porém, a síntese eleva o preço de uso da técnica (CHOI et al., 2019).

Além do alto custo, o tempo de processamento da codificação e decodificação é relativamente lento quando comparado às tecnologias em uso atualmente. Esse ponto é de extrema importância, considerando a atual velocidade de acesso e armazenamento de dados. Outro fator que interfere na aplicação atual da técnica são as imperfeições na síntese do DNA, provocando erros que podem levar à perda de dados (CAMPBELL, 2020).

Por fim, dois fatores que despertam a atenção são os metadados do armazenamento e a *biocybersecurity*. Em relação aos metadados, há uma dificuldade na

gestão e recuperação dos dados armazenados, já que não há o estabelecimento de um processo de coleta. O segundo ponto de alerta, *biocybersecurity*, tem por base a relevante e atual importância de garantir a segurança no armazenamento de dados. Há registros, desde 2017, da inserção de malware na sequência da codificação que desencadeia erros de leitura e/ou sequenciamento. Isso pode levar ao comprometimento e/ou vazamento dos dados (GERVASIO et al., 2024).

Gervasio et al. (2024) apresentam uma previsão futurista sobre a evolução da técnica de armazenamento de dados em DNA. Para eles, nos próximos 10-20 anos, estarão bem estabelecidos os protocolos de segurança e armazenamento de metadados. Outros fatores que serão aprimorados a curto prazo são a correção de erros, a latência e a redução de custos para viabilizar a produção em massa. Por um longo período, o foco será na padronização dos métodos utilizados para codificação, decodificação e leitura. Por fim, a preocupação para mais de 1000 anos a partir de hoje, será a identificação do DNA natural versus não natural, sinalizando o que possui informações digitais para diferenciá-lo do que não possui.

3. Metodologia

Dado que a fase de experimentação no armazenamento em DNA é relativamente nova, o presente artigo propõe investigar o assunto, de modo a proporcionar uma melhor compreensão deste campo promissor e em expansão. Por assim ser, esta pesquisa tem abordagem qualitativa, utilizando o método da revisão sistemática, com objetivo descritivo e uma natureza básica.

3.1. Delimitação do método

O método utilizado para a formulação do presente trabalho consiste na Revisão Sistemática - RS. Considerando o grande volume de pesquisas publicadas, a revisão sistemática emerge como um recurso valioso (BRIZOLA, 2016). O principal objetivo é sintetizar as informações produzidas por pesquisas de fonte primária (realizadas de modo prático), de acordo com a delimitação do tema de estudo (CORDEIRO, 2007). Desse modo, é possível mapear a produção de dados na área, avaliar criticamente e apresentar os resultados de modo a agregar e facilitar novos estudos na área (BRIZOLA, 2016).

A RS deve ser imparcial e apresentar o passo a passo para que outros possam reproduzi-la. Para garantir o rigor dos resultados e evitar que a análise tenda para o ponto de vista do autor sobre o tema, são adotados critérios rigorosos. Dentre esses, podemos citar: estabelecer de modo claro a pergunta de pesquisa; os termos de busca devem possuir alta sensibilidade; mais de uma base de dados deve ser utilizada; toda a literatura pertinente encontrada deve ser analisada; e todo o processo de busca deve ser devidamente documentado (DONATO; DONATO, 2019).

3.2. Procedimento adotado

A presente pesquisa foi elaborada com base na seguinte questão:

O 'DNA Storage' surge como uma solução para a atual problemática da densidade de armazenamento de dados. Há um consenso sobre os pontos positivos e negativos da utilização desta técnica. Um dos principais desafios da aplicação desta é a escalabilidade, que envolve, dentre outros fatores, a leitura e escrita de dados. Desse modo, a presente

pesquisa tem como problemática a compreensão das atuais técnicas em desenvolvimento para solucionar os desafios de codificação e decodificação envolvendo o armazenamento de dados em DNA.

Partindo desse questionamento inicial, com o intuito de alcançar o objetivo da pesquisa e englobar todos os resultados desejados, o seguinte termo de busca foi estabelecido:

("DNA storage" OR "DNA data storage" OR "DNA-based storage") AND ("encoding methods" OR "decoding methods" OR "encoding techniques" OR "decoding techniques" OR "encoding technologies" OR "decoding technologies").

Para a realização da pesquisa, foram selecionadas duas bases de dados relevantes na área do tema pesquisado, a *PubMed* e a *Scopus*. A primeira base de dados tem acesso à MEDLINE, *life science journals* e livros on-line, com mais de 37 milhões de citações na área da biomedicina². Ela foi escolhida devido aos estudos que englobam a manipulação de DNA. Já a segunda base de dados, *Scopus*, é renomada e confiável, apresenta resumos multidisciplinares e possui mais de 91 milhões de produções³.

A pesquisa foi realizada no dia 30 de maio de 2024, buscando pelos termos no título, resumo e nas palavras-chave dos artigos. Retornaram três resultados na *PubMed* e 23 resultados na *Scopus*. Porém, os três artigos que apareceram na *PubMed* também foram encontrados na *Scopus*, resultando em 23 estudos. Para conferir todos os artigos encontrados, acesse o GitHub, disponível em: <https://github.com/carolav13/WorkTec.git>.

Abaixo é possível ver algumas características dos artigos encontrados na busca da revisão sistemática:

- a) **Ano de publicação:** entre 2017 e 2024, sendo no ano atual o maior pico de produtividade (06 até o momento).
- b) **Tipos de documento:** artigo científico (14), *papers* de conferências (6), revisões (2) e revisões de conferências (1).
- c) **Áreas em que os artigos foram publicados:** as 3 mais expressivas são, Ciência da Computação (34,5%); Bioquímica, Genética e Biologia Molecular (16,4%); e Engenharia (14,5%).
- d) **Países que mais desenvolveram estudos na área:** em primeiro lugar a China (9), seguido da Alemanha e de Israel, ambos com 3.⁴

Alguns artigos não seguiam os critérios necessários para a análise, ou seja, apresentar os métodos de codificação e decodificação do armazenamento em DNA. Deste modo, eles foram excluídos da análise. Abaixo é possível ver os critérios de exclusão de cada um:

- a) Não foi possível localizar um estudo completo⁵, portanto, ele não pode ser analisado;
- b) Três resultados (1º HEINIS et al., 2023; 2º WANG et al., 2022; 3º SWATI et

² <https://pubmed.ncbi.nlm.nih.gov/>.

³ <https://www.elsevier.com/pt-br/products/scopus>.

⁴ Resultados fornecidos pela ferramenta “Analyze search results” da *Scopus* de acordo com a busca pelos termos de pesquisa.

⁵ “Proceedings - 2020 8th International Conference on Advanced Cloud and Big Data, CBD 2020”

al., 2017) abordam uma revisão sobre aspectos dos métodos de codificação e decodificação do armazenamento de dados em DNA. Eles não foram incluídos, já que são similares à proposta do presente estudo e o método da revisão sistemática prioriza a análise de artigos de fontes primárias, com o intuito de melhorar a confiabilidade e diminuir a redundância nos resultados;

- c) O artigo de Lenz, et al. 2023 apresenta um estudo sobre canais de comunicação do armazenamento em DNA, abordando um aspecto da codificação e decodificação de dados, mas não os métodos para análise;
- d) Chee et al. (2020), focam exclusivamente no erro de duplicação do DNA que ocasiona em múltiplas cópias em sequência de uma mesma sequência de bases (duplicações de Tandem). Deste modo ele foi excluído por ser restrito e específico.
- e) O estudo de Lenz et al. (2019), busca, por meio de cálculos, determinar a quantidade máxima de armazenamento e recuperação de dados com precisão utilizando armazenamento em DNA. Deste modo, apresenta um foco diferente da revisão, ao não propor métodos de codificação e decodificação.

Assim, totalizam 16 estudos para análise.

O resultado da análise realizada com base no procedimento descrito pode ser encontrado no tópico '4. Resultados'.

4. Resultados

O crescente interesse na área de armazenamento de dados em DNA evidencia sua relevância e sua perspectiva futura no armazenamento de informações. Isso ocorre principalmente devido à sua alta densidade e durabilidade (WANG et al., 2019).

No entanto, a aplicação prática dessa técnica enfrenta várias problemáticas se síntese, armazenamento e sequenciamento, como: síntese limitada na capacidade de bits armazenados por bases nitrogenadas; limitação na quantidade máxima de sequência de nucleotídeos por cadeia; alto custo de produção; redundância de dados que é inserida facilitar o processo de codificação, mas pode ocasionar em erros como substituição, deleção e inserção de bases nitrogenadas; erros de mutação; entre outros (HONG et al., 2024).

Deste modo, há o desenvolvimento, combinação e adaptação de diversas técnicas de correção de erros, codificação e decodificação. Alguns métodos foram destacados na revisão sistemática realizada, os quais estão listados na Tabela 01 'Resultados', disponível no GitHub através do link de acesso: [<https://github.com/carolav13/WorkTec.git>].

De acordo com a Tabela 1, é possível visualizar diferentes objetivos em relação à técnica de armazenamento de dados em DNA, visando ao seu aprimoramento. Especificamente, eles buscam:

- a) Correção de erros em dados corrompidos;
- b) Padronização das técnicas para futura aplicação em massa;

- c) Melhorar a síntese de DNA para atender às principais restrições biológicas (como homopolímeros⁶ e 50% de conteúdo de GC⁷);
- d) Aumentar a densidade de armazenamento, ou seja, a quantidade de bits armazenados em cada base nitrogenada (bits/nt);
- e) Reduzir os custos de codificação e decodificação por meio da combinação de diferentes técnicas;
- f) Corrigir erros de armazenamento por meio da codificação;
- g) Desenvolver novas técnicas que agreguem valor, como por exemplo, reduzir a quantidade de ciclos realizados na síntese;
- h) Melhorar a codificação de diversos tipos de dados (textos, imagens, vídeos, áudios e melodias/notas musicais);
- i) Reduzir o tempo de codificação dos dados;
- j) Aprimorar os métodos de codificação existentes, como os códigos *fountain*.

Durante a análise do levantamento bibliográfico, surgiram diversas técnicas. As que apresentaram maior expressividade e, portanto, maior frequência foram: Códigos *Fountain*: Utilizados na codificação de dados; Códigos *Reed-Solomon*: Utilizados para codificação, decodificação e correção de erros em dados incompletos ou danificados; e Códigos de *Levenshtein*: Utilizados para correção de erros por meio da comparação de dados.

Os estudos ocorreram de modo prático, por meio de simulações em softwares específicos, como o DR4DNA, simulando *In vitro* e *In silico*, e simulações com cálculos matemáticos e/ou códigos desenvolvidos. Esses artigos apontam que houve aumento na taxa de codificação e decodificação bem-sucedidas; a grande maioria dos estudos se preocupam e atendem às principais restrições biológicas, possibilitando sua replicação; houve melhor recuperação de erros (exemplo: em 150 nt com 5 nt de erro, a correção de dados chegou a 96% (ZHANG; ZHOU, 2024); correção de 98% dos erros quando a taxa de erro é de 0,05 (ZAN et al., 2022)); redução nos custos; redução no tempo de codificação; e melhora da densidade de armazenamento de dados (codificação de 1,98 bits/nt (LIU et al., 2020)).

Com base no levantamento de dados, é possível ver que há investimento na área, seu desenvolvimento é relevante e necessário, e as perspectivas futuras são promissoras. Vários pontos necessitam de atenção, porém, a evolução para superar os desafios é clara.

5. Considerações Finais

O presente estudo investigou e sintetizou as técnicas de codificação e decodificação para armazenamento de dados em DNA. Os resultados obtidos evidenciam que a técnica de armazenamento de dados em DNA apresenta um potencial significativo para resolver a

⁶ Homopolímeros: pequenas sequências de nucleotídeos iguais que se repetem (ex: 'AAAA', 'GGGG', 'TTTT', 'CCCC'). Elas podem ocasionar em erros de síntese.

⁷ Restrição de GC: para que a molécula de DNA seja estável quando sintetizada, é necessário que haja uma proporção em torno de 50% das bases nitrogenadas Guanina e Citosina. Ela é justificada devido às propriedades das ligações realizadas na macromolécula de DNA.

problemática da densidade de armazenamento de dados, destacando-se pela alta densidade e durabilidade oferecidas pelo DNA. No entanto, enfrenta desafios complexos, como a síntese limitada, o alto custo de produção e a ocorrência de erros durante o armazenamento e sequenciamento.

Diversas técnicas de correção de erros, codificação e decodificação foram exploradas e analisadas, com destaque para os códigos Fountain, Reed-Solomon e Levenshtein, cada um com suas aplicações específicas e vantagens. A padronização das técnicas para futura aplicação em massa, a melhoria da síntese de DNA para atender às restrições biológicas e o aumento da densidade de armazenamento foram objetivos centrais observados na revisão sistemática.

Para encaminhamentos futuros, é crucial continuar desenvolvendo e aprimorando técnicas que reduzam custos, melhorem a eficiência na codificação e decodificação de diversos tipos de dados e expandam a capacidade de armazenamento do DNA. Considerando novas revisões sistemáticas, apesar dos critérios de execução terem sido atingidos, recomenda-se que sejam incluídas mais bases de dados, de modo a obter um acervo maior para análise.

Assim, a pesquisa reforça a importância contínua deste campo inovador, destacando suas potencialidades e desafios, e delineia um caminho promissor para avanços futuros na área de armazenamento de dados em DNA."

6. Referências

ANAVY, Leon et al. Data storage in DNA with fewer synthesis cycles using composite DNA letters. **Nature Biotechnology**, v. 37, p. 1229-1236, 2019. DOI: 10.1038/s41587-019-0240-x.

BHATTACHARJEE, P.; LEE, S.-W. *DNA Nanomaterials*. World Scientific Series in Nanoscience and Nanotechnology, v. 20, n. 3, p. 23-42, 2020. Disponível em: <<https://www.scopus.com/record/display.uri?eid=2-s2.0-85151271503>>. Acesso em: 1 jun. 2024.

BIA, Kun; WAN, Yu; ZHAO, Xiangwei. FCBC hybrid encoding method based on fountain code and Base64 code. In: 2023 3rd International Conference on Electronic Information Engineering and Computer Communication (EIECC). IEEE, 2023. DOI: 10.1109/EIECC60864.2023.10456627.

BORNHOLT, J. et al. DNA-based archival storage system. ACM SIGPLAN Notices, 2016.

BOULIC, R. AND RENAULT, O. (1991) D Hierarchies for Animation, In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons Ltd., England.

BRIZOLA, Jairo; FANTIN, Nádia. Revisão da literatura e revisão sistemática da literatura. **Revista de Educação do Vale do Arinos**, Juara, v. 3, n. 2, p. 23-39, jul./dez. 2016.

BUKO, T.; TUCZKO, N.; ISHIKAWA, T.. "DNA Data Storage." *BioTech*, vol. 12, no. 2, 2023, p. 44, <https://doi.org/10.3390/biotech12020044>.

CAMPBELL, M.; Trace3. "DNA Data Storage: Automated DNA Synthesis and

- Sequencing Are Key to Unlocking Virtually Unlimited Data Storage." *Revista IEEE Computer Society*, 2020, DOI: 10.1109/MC.2020.2967908.
- CHOI, Yeongjae et al. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. **Scientific Reports**, [S.l.], v. 9, n. 1, art. 6582, Dec. 2019. DOI: 10.1038/s41598-019-43105-w. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85065076517&doi=10.1038%2fs41598-019-43105-w&partnerID=40&md5DOI>. Acesso em: 07 jun. 2024.
- CHRUCH, G. M., GAO, Y., & KOSURI, S. (2012). Next-Generation Digital Information Storage in DNA. *Science*, 337, 1628. Retrieved from <https://www.science.org>.
- CONTILIANI, Anyel Fernandes; PEREIRA, Tiago Campos. O DNA como um novo dispositivo de armazenamento de dados. *Genética na Escola*, v. 17, n. 1, p. 11-16, 2022. ISSN 1980-3540. Sociedade Brasileira de Genética.
- CORDEIRO, Alexander Magno; OLIVEIRA, Glória Maria de; RENTERÍA, Juan Miguel; GUIMARÃES, Carlos Alberto; GRUPO DE ESTUDO DE REVISÃO SISTEMÁTICA DO RIO DE JANEIRO (GERS-Rio). Revisão sistemática: uma revisão narrativa. **Comunicação Científica**, Rio de Janeiro, v. 34, n. 6, p. 428-431, nov./dez. 2007.
- COSTA, Israel Reis; PINTO, Liliane Faria Corrêa. A evolução dos dispositivos de armazenamento de dados na perspectiva da história. *Curso de Licenciatura em Informática, Universidade Federal do Maranhão, Campus Codó VII (UFMA)*.
- DIMOPOULOU, Melpomeni; SAN ANTONIO, Eva Gil; ANTONINI, Marc. A JPEG-based image coding solution for data storage on DNA. In: EUSIPCO 2021. DOI: 10.1109/EUSIPCO54536.2021.9687997.
- DONATO, Helena; DONATO, Mariana. *Acta Med Port. Revista Científica da Ordem dos Médicos*, Coimbra, v. 32, n. 3, p. 227-235, mar. 2019. Disponível em: <https://doi.org/10.20344/amp.11923>. Acesso em: 04 jun 2024.
- DYER, S., MARTIN, J. AND ZULAUF, J. (1995) Motion Capture White Paper, http://reality.sgi.com/employees/jam_sb/mocap/MoCapWP_v2.0.html, Dezembro.,
- FREITAS, Adriel de Oliveira; MADEIRO, João Paulo do Vale. Desenvolvimento de uma plataforma IoT móvel para aquisição, processamento e compartilhamento de sinais cardiológicos para fins de pesquisa em ambiente colaborativo. *Revista de Engenharia da Universidade Católica de Petrópolis*, v. 17, n. 1, 2023.
- GERHARD, T E; SILVEIRA, D. T. Método de Pesquisa. Rio Grande do Sul. *Editora UFRGS*. 2009.
- GERVASIO, J.H.D.B.; DA COSTA OLIVEIRA, H.; DA COSTA MARTINS, A.G.; PESQUERO, J.B.; VERONA, B.M.; CERIZE, N.N.P. How close are we to storing data in DNA? **Trends in Biotechnology**, v. 42, n. 2, p. 156-167, 2024. DOI: 10.1016/j.tibtech.2023.08.001. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85169891387&doi=10.1016%2fj.tibtech.2023.08.001&partnerID=40&md5>.
- HAGHIGHAT, Javad; DUMAN, Tolga M. A Practical Concatenated Coding Scheme

- for Noisy Shuffling Channels with Coset-based Indexing. In: 2023 IEEE Global Communications Conference (GLOBECOM). IEEE, 2023. DOI: 10.1109/GLOBECOM54140.2023.10437131.
- HEINIS, T., SOKOLOVSKII, R., & ALNASIR, J. J. (2023). Survey of Information Encoding Techniques for DNA. *ACM Computing Surveys*, 56(4), Article 107. <https://doi.org/10.1145/3626233>.
- HOLTON, M. AND ALEXANDER, S. (1995) Soft Cellular Modeling: A Technique for the Simulation of Non-rigid Materials, *Computer Graphics: Developments in Virtual Environments*, R. A. Earnshaw and J. A. Vince, England, Academic Press Ltd., p. 449-460.
- HONG, Jingwei et al. VSD: A Novel Method for Video Segmentation and Storage in DNA Using RS Code. **Mathematics**, v. 12, n. 1235, 2024. Disponível em: <https://doi.org/10.3390/math12081235>.
- JEONG, Jaeho et al. Cooperative sequence clustering and decoding for DNA storage system with fountain codes. **Bioinformatics**, v. 37, n. 19, p. 3136-3143, 2021. DOI: 10.1093/bioinformatics/btab246.
- JEONG, Jaeho et al. Iterative Soft Decoding Algorithm for DNA Storage Using Quality Score and Redecoding. **IEEE Transactions on Nanobioscience**, v. 23, n. 1, p. 81-90, janeiro 2024.
- KATSURA, Shinji. Application 5 - Addressing of Nanoparticles by Using DNA Molecules. In: NAITO, Makio; YOKOYAMA, Toyokazu; HOSOKAWA, Kouhei; NOGI, Kiyoshi (Ed.). *Nanoparticle Technology Handbook* (Third Edition). Elsevier, 2018. p. 423-426. ISBN 9780444641106. Disponível em: <<https://doi.org/10.1016/B978-0-444-64110-6.00012-3>>. Acesso em: 1 jun. 2024.
- KIRYANOVA, Olga Yu et al. A novel approach to encode melodies in DNA. **BioSystems**, v. 237, 2024. Disponível em: <https://doi.org/10.1016/j.biosystems.2024.105136>.
- KNUTH, D. E. (1984), *The TeXbook*, Addison Wesley, 15th edition.
- LEHNINGER, T. M., NELSON, D. L. & COX, M. M. *Princípios de Bioquímica*. 6ª Edição, 2014. Ed. Artmed.
- LIMA JUNIOR, Eduardo Brandão; OLIVEIRA, Guilherme Saramago de; SANTOS, Adriana Cristina Omena dos; SCHNEKENBERG, Guilherme Fernando. Análise documental como percurso metodológico na pesquisa qualitativa. *Cadernos da Fucamp*, v. 20, n. 44, p. 36-51, 2021.
- LIMA, Telma Cristiane Sasso de; MIOTO, Regina Célia Tamaso. Procedimentos metodológicos na construção do conhecimento científico: a pesquisa bibliográfica. *Revista Katál*. Florianópolis, v. 10, n. esp., p. 37-45, 2007. Páginas 37-45. Universidade Federal de Santa Catarina (UFSC). Recebido em 20.02.2007. Aprovado em 03.04.2007.
- LIU, Q.; WANG, P.; CUI, J.; QI, H. MRC: A High Density Encoding Method for Practical DNA-based Storage. In: *2020 Eighth International Conference on Advanced Cloud and Big Data (CBD)*. IEEE, 2020. p. 13-19. DOI: 10.1109/CBD51900.2020.00012. Disponível em:

<https://doi.org/10.1109/CBD51900.2020.00012>.

- MANICKA, A. et al. Automated routing of droplets for DNA storage on a digital microfluidics platform. *Digital Discovery*, v. 2, n. 5, pp. 1436-1451, 2023. Document Type: Article. Publication Stage: Final. Source: Scopus. DOI: 2-s2.0-85170690637.
- NAOUM, P. C. O DNA, seus incríveis números e seus segredos. **LAES&HAE**, n. 000, p. 01-02, 2010. *Biologia Molecular*. Disponível em: <a.c.t@terra.com.br>. Acesso em: 1 jun. 2024.
- OKUDA, H. (2015). Vicissitude of magnetic tape data storage: comparison with video tape recording. In 2015 ICOHTEC/IEEE International History of High-Technologies and their Socio-Cultural Contexts Conference (HISTELCON) (pp. 1-5). Tel-Aviv, Israel. doi: 10.1109/HISTELCON.2015.7307305.
- RUAN, H. (2016). Ultra-high density optical storage technologies for big data center. In Asia Communications and Photonics Conference, ACP, 2016-November (Art. No. 8680160). <https://doi.org/2-s2.0-85091234293>.
- SAES, Keylla Ramos. *Abordagem para a integração automática de dados estruturados e não estruturados em um contexto big data*. 2019. Dissertação (Mestrado em Sistemas de Informação) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2019.
- SAIS, M.; RAFALIA, N.; ABOUCHABAKA, J.. "DNA technology for big data storage and error detection solutions: Hamming code vs Cyclic Redundancy Check (CRC)." In: E3S Web of Conferences, v. 412, 2023, p. 01090.
- SCHWARZ, Peter Michael; FREISLEBEN, Bernd. Data recovery methods for DNA storage based on fountain codes. **Computational and Structural Biotechnology Journal**, v. 23, p. 1808-1823, 2024. Disponível em: <https://github.com/umr-ds/DR4DNA>.
- SILVA DE SOUSA, Angélica; OLIVEIRA, Guilherme Saramago de; ALVES, Laís Hilário. A pesquisa bibliográfica: princípios e fundamentos. *Cadernos da Fucamp*, v. 20, n. 43, p. 64-83, 2021.
- SMITH, A. AND JONES, B. (1999). On the complexity of computing. In *Advances in Computer Science*, pages 555–566. Publishing Press.
- SWATI, A., MATHURIA, F., BHAVANI, S., MALATHY, E., & MAHADEVAN, R. (2017). A review on various encoding schemes used in digital DNA data storage. *International Journal of Civil Engineering and Technology (IJCET)*, 8(12), 108–114. Retrieved from <http://iaeme.com/Home/issue/IJCET?Volume=8&Issue=12>. (Article ID: IJCET_08_12_013).
- WANG, C., MA, G., WEI, D., ZHANG, X., WANG, P., LI, C., XING, J., WEI, Z., DUAN, B., YANG, D., WANG, P., BU, D., & CHEN, F. (2022). Mainstream encoding–decoding methods of DNA data storage. *CCF Transactions on High Performance Computing*, 4, 23–33. <https://doi.org/10.1007/s42514-022-00094-z>.
- WANG, Yixin et al. Construction of Bio-Constrained Code for DNA Data Storage. **IEEE Communications Letters**, v. 23, n. 6, p. 963-966, junho 2019.
- XU, P., FANG, G., SHI, X., LIU, W. (2020). DNA Storage and Its Research Progress

- [DNA存储及其研究进展]. Dianzi Yu Xinxu Xuebao/Journal of Electronics and Information Technology, 42(6), 1326-1331. Recuperado de 2-s2.0-85087525973.
- YE, X., ZHAI, Z., & LI, X. (2019). A multi-channel redundant check storage method for solid-state disks based on cold and hot data stripes. *Journal of Engineering Science and Technology Review*, 12(2), 46-52. <https://doi.org/2-s2.0-85068782649>.
- ZAN, Xiangzhen et al. A Hierarchical Error Correction Strategy for Text DNA Storage. **Interdisciplinary Sciences: Computational Life Sciences**, v. 14, p. 141-150, 2022. Disponível em: <https://doi.org/10.1007/s12539-021-00476-x>.
- ZAN, Xiangzhen et al. An Efficient Bucket-allocation Decoding Method Based on Forward Error Correction Codes for Deoxyribo Nucleic Acid Storage. **Journal of Electronics & Information Technology**, v. 44, n. 10, p. 3650-3656, 2022.
- ZHANG, X.; ZHOU, F. An Encoding Table Corresponding to ASCII Codes for DNA Data Storage and a New Error Correction Method HMSA. *IEEE Transactions on Nanobioscience*, v. 23, n. 2, p. 344-354, abr. 2024.
- ZHAO, Yunzhu et al. DBTRG: De Bruijn Trim rotation graph encoding for reliable DNA storage. **Computational and Structural Biotechnology Journal**, v. 21, p. 4469-4477, 2023. Disponível em: <https://doi.org/10.1016/j.csbj.2023.09.004>.
- ZHENG, Yanfen et al. DNA-QLC: an efficient and reliable image encoding scheme for DNA storage. **BMC Genomics**, v. 25, n. 266, 2024. Disponível em: <https://doi.org/10.1186/s12864-024-10178-5>.