DNA-Barcoding e aprendizado de máquina no alinhamento e localização de *Primers* para o reconhecimento de cianobactérias

DNA-Barcoding and machine learning in the alignment and localization of *Primers* for the recognition of cyanobacteria

DOI:10.34117/bjdv8n4-637

Recebimento dos originais: 21/02/2022
Aceitação para publicação: 31/03/2022

Victor Sanmartin
Graduado em Ciência da Computação (UNISC)
Instituição: Universidade de Santa Cruz do Sul (UNISC), Departamento de Engenharias, Arquitetura e Computação
Endereço: Av. Independência, 2293 – Bairro Universitário, Santa Cruz do Sul – RS, CEP: 96815-900
E-mail: vsanmartin67@gmail.com

Rejane Frozza
Doutorado em Ciência da Computação (UFRGS)
Instituição: Universidade de Santa Cruz do Sul (UNISC), Departamento de Engenharias, Arquitetura e Computação, Programa de Pós-Graduação em Sistemas e Processos Industriais, Programa de Pós-Graduação em Letras
Endereço: Av. Independência, 2293 – Bairro Universitário, Santa Cruz do Sul – RS, CEP: 96815-900
E-mail: frozza@unisc.br

Alexandre Rieger
Doutorado em Genética e Biologia Molecular (UFRGS)
Instituição: Universidade de Santa Cruz do Sul (UNISC), Departamento de Ciências da Vida, Laboratório de Biotecnologia e Genética, Programa de Pós-Graduação em Promoção da Saúde, Programa de Pós-Graduação em Tecnologia Ambiental
Endereço: Av. Independência, 2293 – Bairro Universitário, Santa Cruz do Sul – RS
CEP: 96815-900
E-mail: rieger@unisc.br

Alexandro Cagliari
Doutorado em Genética e Biologia Molecular (UFRGS)
Instituição: Universidade Estadual do Rio Grande do Sul (Uergs)
Endereço: Avenida Independência, 2824 – Renascença, Santa Cruz do Sul – RS
CEP: 96816-501
E-mail: alexandro-cagliari@uergs.edu.br

RESUMO

A Bioinformática é uma área que vem ganhando visibilidade para a previsão de doenças através do DNA (Ácido desoxirribonucleico). Esta área sempre esteve diretamente associada à biologia molecular, campo da biologia responsável por estudar a estrutura e as funções do material genético, bem como as proteínas, que são os resultados obtidos em uma síntese de DNA. A área da Genômica é um ramo da bioquímica que estuda o genoma completo de um organismo. Com a evolução das tecnologias de informação, torna-se possível a manipulação de um grande volume de dados e isso tem permitido que os estudiosos da área obtenham cada vez mais rapidamente seus resultados. A genômica e suas derivadas são áreas de investigação alicerçadas na geração de um grande volume dados que tem aumentado exponencialmente ao longo dos anos. As técnicas de DNA-*Barcoding* e Aprendizado de Máquina tendem a auxiliar ainda mais os pesquisadores da área, buscando soluções rápidas e inteligentes. Assim, este trabalho uniu as técnicas de DNA-*Barcoding* e Aprendizado de Máquina, com a técnica de agrupamento, cujo aprendizado é não supervisionado, para o sequenciamento e identificação de Cianobactérias. A metodologia incluiu bibliometria quantitativa e qualitativa para busca e análise de trabalhos relacionados; modelagem e desenvolvimento do sistema de sequenciamento e reconhecimento de cianobactérias; implementação do BLAST (que realiza o alinhamentos de sequência de DNA), do sequenciamento pelo DNA-*Barcoding*, da identificação das regiões de *primers* e do agrupamento das sequências; realização de testes e ajustes no sistema. A partir dos resultados obtidos, destaca-se a velocidade de execução de todo o processo, desde o *BLAST* inicial até o *BLAST* final, bem como a busca pelos *Primers* e o agrupamento.

Palavras-chave: bioinformática, dna, *dna-barcoding*, aprendizado de máquina, cianobactérias

ABSTRACT

Bioinformatics is an area that is gaining visibility for the prediction of diseases through DNA (deoxyribonucleic acid). This area has always been directly associated with molecular biology, the field of biology responsible for studying the structure and functions of genetic material, as well as proteins, which are the results obtained in a DNA synthesis. The Genomics area is a branch of biochemistry that studies the complete genome of an organism. With the evolution of information technologies, it is becoming possible to manipulate a large volume of data, and this has allowed scholars in the field to obtain their results more and more quickly. Genomics area is based on a generation of a large volume of data that has increased exponentially over the years. DNA-Barcoding and Machine Learning techniques tend to help researchers in the field even more, seeking quick and intelligent solutions. Then, this work joined the techniques of DNA-Barcoding and Machine Learning, with the grouping technique, whose learning is unsupervised, for the sequencing and identification of Cyanobacteria. The methodology included quantitative and qualitative bibliometrics for searching and analyzing related works; modeling and development of the cyanobacteria sequencing and recognition system; implementation of BLAST (which performs DNA sequence alignments), sequencing by DNA-Barcoding, identification of primer regions, and grouping of sequences; tests and adjustments to the system. From the results obtained, the speed of execution of the entire process stands out, from the initial BLAST to the final BLAST, as well as the search for Primers and grouping.

Keywords: bioinformatics, dna, *dna-barcoding*, machine learning, cyanobacteria.

1 INTRODUÇÃO

A Bioinformática é o ato de conceituar a biologia em termos moleculares e aplicar técnicas de informática, as quais derivam de áreas como matemática aplicada, ciência da computação e estatística, a fim de entender e organizar as informações associadas a essas moléculas em larga escala (LUSCOMBE, GREENBAUM e GERSTEIN 2001). A Bioinformática integra, essencialmente, o desenvolvimento de programas computacionais para tratar dados biológicos preexistentes e identificar sequências de genes. Caracteriza-se como ferramenta indispensável na ordenação e agrupamento dos resultados gerados pelas análises de sequenciamento de genes, que produzem uma quantidade cada vez maior de dados sobre a composição de DNA (ácido desoxirribonucléico), RNA (ácido ribonucléico) e proteínas.

Já o conceito de DNA-*Barcoding* está diretamente relacionado com o que se conhece como código de barras, como, por exemplo, a identificação do código de barras de um produto no mercado. É com esse mesmo intuito que foi sugerido por Paul Hebert, juntamente com seus colaboradores, a criação de um código de barras molecular que identificasse espécies conhecidas, diminuindo a necessidade de utilizar-se de métodos mais complexos para a sua identificação (HEBERT et al., 2003). DNA-*Barcoding* é uma ferramenta para rápida identificação de espécies com base em sequências de DNA e toda a estrutura do genoma (KRESS e ERICKSON, 2008).

Para que as sequências mantenham uma integridade de informação, deve-se realizar o BLAST, que é a ferramenta de pesquisa e alinhamento de sequências que realiza a comparação de nucleotídeos ou proteínas com bancos de dados de sequências e calcula a significância estatística das correspondências. O BLAST pode ser usado para inferir relações funcionais e evolutivas entre sequências, bem como auxiliar a identificar membros de famílias de genes (NCBI, 2020).

Após ser realizado o sequenciamento das moléculas de DNA, identifica-se o que se chama de Primer, que é uma sequência curta de ácido nucléico, de até 60 nucleotídeos semelhantes em todas as sequências, que fornece um ponto de partida para a síntese de DNA. A área entre os Primers servirá para a identificação das moléculas (DELONG e ZHOU, 2015). Em relação às ferramentas e métodos computacionais nesta área, o campo do Aprendizado de Máquina está relacionado ao desenvolvimento e aplicação de algoritmos de computador que melhoram com a experiência (MITCHELL, 1997). Assim, o Aprendizado de Máquina de genomas pode ser usado para "aprender" a reconhecer padrões nas sequências de DNA. Uma grande variedade de métodos de Aprendizado de

Máquina foi desenvolvida para auxiliar no entendimento dos mecanismos subjacentes à expressão gênica. Algumas técnicas têm como objetivo prever a expressão de um gene baseado apenas na sequência de DNA (LIBBRECHT e NOBLE, 2015).

O objetivo principal é desenvolver uma aplicação para identificar regiões de *Primers* em sequências de DNA de cianobactérias, utilizando a técnica de DNA-Barcoding e Aprendizado de Máquina não supervisionado, com uso de agrupamento. O sistema de sequenciamento e reconhecimento precisa ser inteligente o suficiente para aprender durante a interação com diversos genomas e com a identificação rápida e eficaz dos *Primers*, em todas as sequências. E utilizando os dados de sequências curadas, que são sequências confiáveis que já foram verificadas e validadas por profissionais, provenientes da Base de Dados Genômicos do NCBI, o sistema traz um resultado confiável e rápido.

O problema de pesquisa definido foi "De que forma é possível localizar regiões de *Primers* em sequências de DNA para identificação de organismos, utilizando DNA-*Barcoding* e técnica de agrupamento?

O artigo está organizado nas seguintes seções: a seção 2 apresenta a metodologia adotada para desenvolvimento da pesquisa; a seção 3 descreve o desenvolvimento e os resultados atingidos; a seção 4 destaca as conclusões.

2 METODOLOGIA

A bibliometria quantitativa, definida por Araújo (2006) como uma técnica quantitativa de produção científica, foi realizada a fim de encontrar os artigos científicos que abordam temas relacionados ao tema de pesquisa deste trabalho, os quais foram publicados entre o período de janeiro de 2015 a janeiro de 2020. Para isso, foram utilizados os termos de busca "DNA-*Barcoding"*, "*Cyanobacteria*", "*Metabarcoding*", "*Machine Learning*" e "*Genetics*". As bases de dados escolhidas foram Scopus e PubMed, filtrando apenas por artigos científicos. Foi encontrado apenas 1 trabalho relacionado com os termos de busca em conjunto. Como a quantidade de artigos encontrados através da combinação de todos os termos foi pequena, foi realizada uma outra pesquisa combinando os seguintes termos: i) "DNA-Barcoding" AND "Metabarcoding" AND "Machine Learning", sendo selecionados 3 artigos científicos relacionados; ii) "Machine Learning" AND "Genetics", sendo selecionado 1 artigo científico relacionado.

Os quatro artigos selecionados foram estudados e contribuíram para o desenvolvimento da pesquisa, sendo eles: i) Libbrecht e Noble (2015), o qual tem o objetivo de descrever como o aprendizado de máquina auxilia no entendimento de dados genéticos; ii) Beckers et al. (2016), que busca avaliar o desempenho de pares de *Primers* de 16s rDNA, na análise de comunidades bacterianas presentes no solo e raiz, caule e folhas da rizosfera; iii) Cordier et al. (2018), que utiliza o aprendizado de máquina supervisionado, aliado ao DNA-*Barcoding*, para realizar o biomonitoramento marinho; iv) Gerhard e Gunsch (2019), que realiza uma pesquisa microbiológica da água de lastro de navios cargueiros em portos utilizando aprendizagem de máquina.

A Quadro 1 apresenta um comparativo das principais características dos trabalhos selecionados. Os critérios definidos para comparação foram: objetivos e técnicas utilizadas.

Quadro 1 – Comparativo dos trabalhos relacionados

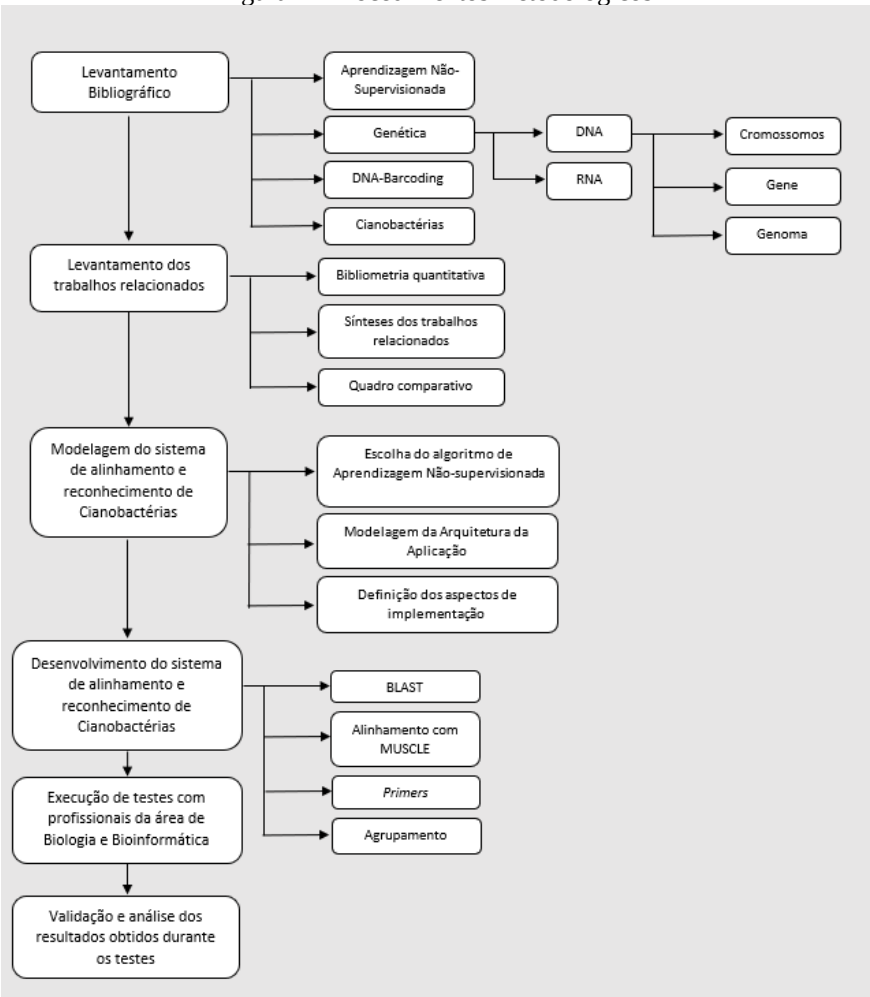
Artigo	Objetivo	Técnicas utilizadas
LIBBRECHT E NOBLE (2015)	Descrever como o Aprendizado de Máquina auxilia no entendimento de dados genéticos.	Aprendizagem de máquina.
GERHARD E GUNSCH (2019)	Realizar uma pesquisa microbiológica da água de lastro de navios cargueiros em portos.	Aprendizagem de Máquina e análise de dados com algoritmo em R, utilizando a biblioteca *phyloseq*.
CORDIER *et al.* (2018)	Utilizar Aprendizado de Máquina supervisionado, aliado ao DNA-*Barcoding*, para realizar o biomonitoramento marinho.	Aprendizagem de máquina supervisionado, aliada ao DNA-*Barcoding*, e *Random Forest.*
BECKERS *et al.* (2016)	Avaliar o desempenho de pares de *Primers* de 16s rDNA, na análise de comunidades bacterianas presentes no solo e raiz, caule e folhas da rizosfera.	Não foi utilizado método computacional, mas utilizaram a metodologia de identificação de *Primers.*
Este Trabalho	Desenvolver uma aplicação para identificar regiões de *Primers* em sequências de DNA de cianobactérias, utilizando a técnica de *Barcoding* e Aprendizado de Máquina, não supervisionado, com uso de agrupamento.	Técnica de Aprendizado de Máquina do tipo não-supervisionada (agrupamento), em conjunto com DNA-*Barcoding*

A partir deste estudo, foi possível observar que algoritmos de aprendizagem supervisionada e não supervisionada aliados ao DNA-Barcoding vêm sendo propostos e testados como forma de se criar uma identificação mais rápida e eficaz de organismos.

Conforme a Figura 1, este trabalho envolveu uma pesquisa bibliográfica, com o objetivo de obter aprofundamento nos assuntos relacionados ao tema de pesquisa. Os tópicos explorados dizem respeito à Bioinformática, sua definição e as técnicas utilizadas, como o DNA-Barcoding, Aprendizagem de Máquina, seus tipos e sua definição. Também foi realizado o levantamento dos trabalhos relacionados, iniciando pela bibliometria

quantitativa, com o intuito de analisar o número de trabalhos existentes nas áreas relacionadas a este trabalho, nos últimos anos. Na bibliometria qualitativa, foram sintetizados os trabalhos relacionados e comparados quanto a seus objetivos e técnicas utilizadas. Foi definida a modelagem com os passos de desenvolvimento da aplicação, assim como também os aspectos de implementação. Seguindo, foi desenvolvido o sistema de sequenciamento e reconhecimento de Cianobactérias, utilizado DNA-Barcoding e Aprendizado de Máquina não-supervisionado. A partir desse sistema desenvolvido, foram realizados testes e ajustes de parâmetros.

Figura 1 - Procedimentos metodológicos



3 DESENVOLVIMENTO E RESULTADOS OBTIDOS

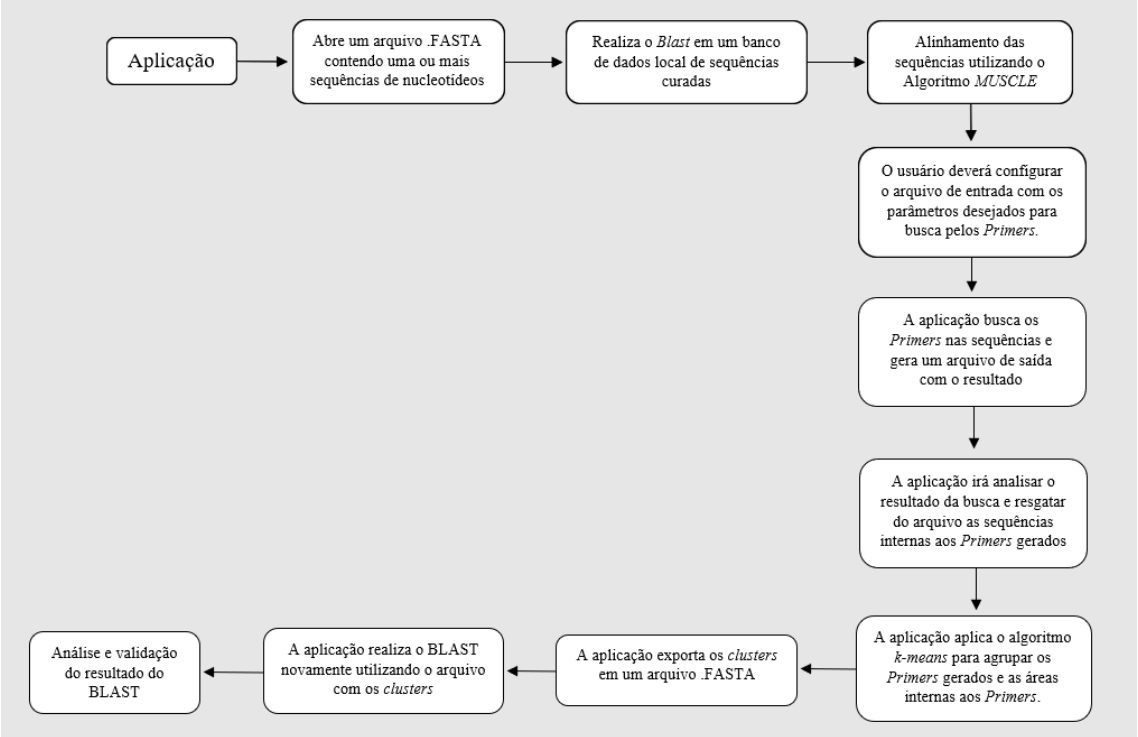
Esta seção apresenta as características do desenvolvimento do sistema para sequenciamento e reconhecimento de cianobactérias.

3.1 MODELO DESENVOLVIDO

O sistema foi modelado como apresentado na Figura 2, com o usuário abrindo um arquivo com a extensão ".FASTA", que contém as sequências utilizadas no processo de identificação de cianobactérias. Em seguida, o sistema realiza o BLAST destas sequências com um banco de dados local. Após o BLAST ser realizado, o sistema alinha as sequências, utilizando o algoritmo MUSCLE (algoritmo de alinhamento múltiplo de sequências), para depois ser realizada a identificação das possíveis regiões de Primers em todas as sequências alinhadas. Para realizar a busca dos Primers, o usuário deve configurar um arquivo de entrada com as sequências resultantes do alinhamento e configurar os parâmetros necessários de acordo com a necessidade. Após a busca pelos Primers, o sistema gera um arquivo com os Primers e as sequências localizadas entre os Primers e utiliza o algoritmo k-means para realizar o agrupamento destas sequências.

Por último, o sistema gera um novo arquivo .FASTA com os clusters encontrados, e, assim, realizar novamente o BLAST das regiões internas aos Primers, a fim de encontrar diferenças que possam distinguir uma sequência da outra e validar a integridade dos Primers gerados para identificar a espécie de Cianobactéria.

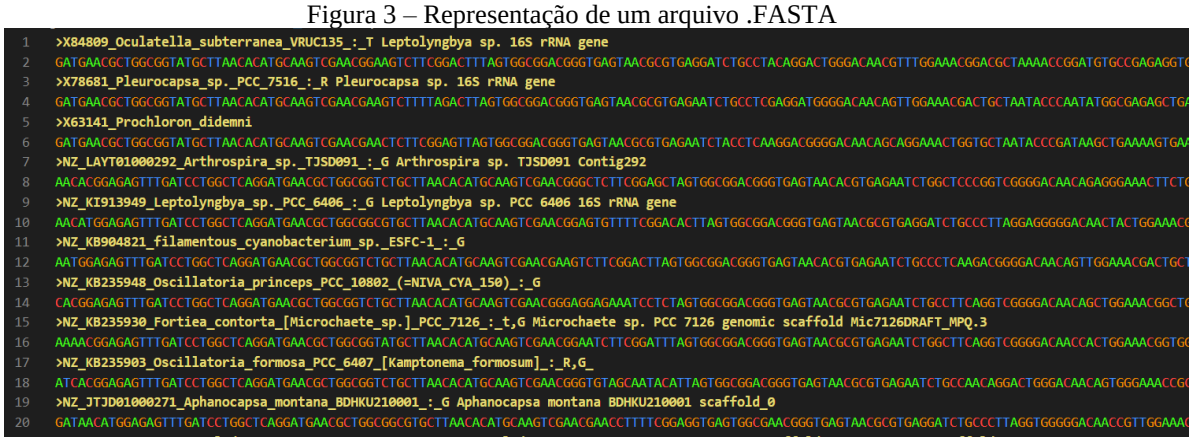
Figura 2. Modelo do Sistema



O formato do arquivo .FASTA é baseado em texto para representar tanto sequências de nucleotídeos como sequências de peptídeos, com códigos de uma única

letra (PEARSON e LIPMAN, 1988) . O conteúdo do arquivo é iniciado com o símbolo de “>”, seguido pela identificação da sequência genética. Na linha seguinte consta a sequência de nucleotídeos referente à identificação anterior, como ilustrado na Figura 3.

Figura 3 – Representação de um arquivo .FASTA



3.2 BASE DE DADOS

A base de dados utilizada foi organizada a parti do banco de dados *Cyanotype*, que é uma base com genomas de Cianobactérias curadas, ou seja, são sequências genéticas que possuem confiabilidade (RAMOS, MORAIS e VASCONCELOS, 2017). O banco de dados foi criado localmente dentro de um arquivo .FASTA e utilizando o software disponibilizado pelo NCBI, BLAST+, com 341 sequências parciais do gene rDNA 16S de diversas espécies de Cianobactérias.

3.3 ALGORITMOS DE PESQUISA DE SEQUÊNCIAS

A Bioinformática usa diferentes algoritmos que variam de simples linhas de comandos a programas gráficos mais complexos e serviços da Web independentes. Uma das ferramentas mais conhecidas e que foi utilizada neste trabalho é o BLAST (*Basic Local Alignment Search Tool*), amplamente utilizada para fazer o alinhamento de sequências.

O BLAST realiza um algoritmo de alinhamento local para comparar informações de sequências biológicas primárias, como sequências de DNA (KORF, YANDELL e BEDELL, 2003). Permite a comparação de uma sequência fornecida a partir de consulta em uma biblioteca ou base de dados de sequências. O BLAST reduz o tempo necessário para identificar regiões conservadas, usando estratégias de pesquisa rápida.

3.4 ALINHAMENTO DAS SEQUÊNCIAS

Múltiplos alinhamentos de sequências de nucleotídeos são importantes em muitas aplicações, incluindo a visualização da árvore filogenética, previsão de estrutura secundária e identificação de resíduos críticos (EDGAR, 2004). Os algoritmos de alinhamento múltiplo de sequências precisam ter uma precisão biológica e complexidade computacional, ou seja, requisitos de tempo e memória.

MUSCLE é um dos programas de alinhamento múltiplo de melhor desempenho de acordo com os testes de *benchmark* publicados, com precisão e velocidade consistentemente melhores do que CLUSTALW (EDGAR, 2004). Por isso, foi escolhido para ser utilizado no desenvolvimento deste trabalho.

3.5 GERAÇÃO DE *PRIMERS*

PRIMER3 é um programa amplamente utilizado para projetar *Primers* de PCR (*Polymerase Chain Reaction*), além de projetar sondas de hibridização e *Primers* de sequenciamento. A PCR é uma ferramenta essencial e onipresente em genética e biologia molecular, sendo totalmente customizável devido ao grande número de parâmetros editáveis da ferramenta, o que proporciona uma independência para o usuário julgar a qualidade do *Primer* gerado.

A geração dos *Primers* na aplicação desenvolvida é realizada através do software PRIMER3, que é executado a partir das sequências que foram alinhadas. O usuário deverá customizar o arquivo "primer3-run.txt", passando as sequências alinhadas e os parâmetros necessários para a geração dos *Primers*.

A aplicação cria um *data frame* com os *Primers* gerados no lado esquerdo (*PRIMER_LEFT)* e no lado direito (*PRIMER_RIGHT)* e as regiões internas aos *Primers*, como mostra a Figura 4. Este *data frame* é utilizado para que a aplicação possa realizar o agrupamento dos *Primers* gerados.

Figura 4 – Data frame com os Primers gerados

	SEQUENCE_ID	PRIMER_LEFT	PRIMER_RIGHT	PRIMER_INTERNAL
0	1	TGCAACTCGCCTGCATGA	GACGGGCGGTGTGTACAA	CGGTGAATACGTTCCCGGGCC
1	2	TGCAACTCGCCTGCATGA	ACGGGCGGTGTGTACAAG	GCGGTGAATACGTTCCCGGGC
2	3	TAACTCCGTGCCAGCAGC	GCCACCTACGGACGCTTT	CGCGGTAATACGGGGGATGCA
3	4	GCTAACTCCGTGCCAGCA	GCCACCTACGGACGCTTT	CGCGGTAATACGGGGGATGCA
4	5	AACGCTGGCGGTATGCTT	CACGCGTTACTCACCCGT	GCAAGTCGAACGGGATCTTTCGGG

Após o *dataframe* estar completo, a aplicação executa o algoritmo de agrupamento *k-means* para agrupar os *Primers* e validar a semelhança entre eles.

K-means é um dos algoritmos mais frequentemente utilizados para realizar o agrupamento é o *K-means*, proposto por James MacQueen em 1967, no qual o usuário/especialista é responsável por definir o número de partições, ou *clusters* a serem encontradas (JAIN e DUBES, 1988).

3.6 TESTES

Os testes foram realizados iniciando a aplicação com um arquivo contendo 4 sequências, com 621 nucleotídeos cada uma, para a realização do BLAST, conforme a Figura 5.

Figura 5 – Sequências utilizadas para teste

```
>seq1
GATGAACGCTGGCGGTATGCTTAACACATGCAAGTCGAACGAACTCTTCGGAGTTAGTGGCGGACGGGTGAGTAACGCGTGAGAATCTACCTCAAGGACGGGGACAACAGCAGGAAAC
>seq2
AAGATTTATTGCCGAGAGAGGAGCTCGCGTCCGATTAGCTAGTTGGAAGGGTAAAAGCCTACCAAGGCGACGATCGGTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGA
>seq3
GCAGTGGGGAATTTTCCGCAATGGGCGAAAGCCTGACGGAGCAATACCGCGTGAGGGAAGAAGGCTTTTGGGTTGTAAACCTCTTTTGTCAGGGAAGAAGAATTGACGGTACCTGAAG
>seq4
CCGCGGTAATACGGAGGATGCAAGCGTTATCCGGAATCATTGGGCGTAAAGGGTCCGCAGGTGGCGAATAAAGTCTGCTGTTAAAGACCGGGGCTAAACTCCGGAAAGGCAGTGGAAA

```

Ao iniciar a aplicação, o primeiro passo é ler o arquivo com as sequências e realizar o BLAST, que é retornado ao usuário via console com todos os alinhamentos encontrados, como mostra a Figura 6.

Figura 6 – Console em tempo real mostrando o resultado do BLAST

```
****Alignment****
sequence: gnl|BL_ORD_ID|195 CAIO01000252_Microcystis_aeruginosa_PCC_9809_:_G Microcystis aeruginosa 9809 WGS project CAIO01000000 data, contig AAI_H_2202_325
length: 1477
e value: 2.07603e-25
TTGGGAACTCTAGGGAGACTGCCGGGGACAACTCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTA...
||||| |||||| |||||||||||| |||||  ||||||||||||||||||||||||||||||||||||||||||||...
TTGGGGACTCTAAGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTA...
match: gnl|BL_ORD_ID|194 CAIP01000212_Microcystis_sp._T1-4_:_G Microcystis sp. T1-4 WGS project CAIP0100000
****Alignment****
sequence: gnl|BL_ORD_ID|194 CAIP01000212_Microcystis_sp._T1-4_:_G Microcystis sp. T1-4 WGS project CAIP01000000 data, contig AAI_I_2203_29
length: 1477
e value: 1.07696e-127
ACTCTAGGGAGACTGCCGGGGACAACTCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTACGTCCTGGGCGACACACGTACTACAATGGTTGGGACAAAGGGCAGCAAACTCGCGAGAGTAAGCGAATCTCATCAAACCCAGCCACAGTTCAGATTGCAGGCTGCAACTC
|||||| |||||||||||| |||||  |||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||||||||| |||||||||||||| ||||||||  |||||||| || ||||||| ||| ||||||||||||||||||||||||||||
ACTCTAAGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTACGTCTTGGGCGACACACGTACTACAATGGTCGGGACAAAGGGCAGCGAACTCGCGAGAGCCAGCGAATCCCAGCAAACCCGGCCTCAGTTCAGATTGCAGGCTGCAACTC
match: gnl|BL_ORD_ID|194 CAIP01000212_Microcystis_sp._T1-4_:_G Microcystis sp. T1-4 WGS project CAIP0100000
****Alignment****
sequence: gnl|BL_ORD_ID|194 CAIP01000212_Microcystis_sp._T1-4_:_G Microcystis sp. T1-4 WGS project CAIP01000000 data, contig AAI_I_2203_29
length: 1477
e value: 2.07603e-25
TTGGGAACTCTAGGGAGACTGCCGGGGACAACTCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTA...
||||| |||||| |||||||||||| |||||  |||||||||||||||||||||||||||||||||||||||||||...
TTGGGGACTCTAAGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTA...
match: gnl|BL_ORD_ID|193 CAIQ01000358_Microcystis_aeruginosa_PCC_9701_:_G Microcystis aeruginosa 9701 WGS p
****Alignment****
sequence: gnl|BL_ORD_ID|193 CAIQ01000358_Microcystis_aeruginosa_PCC_9701_:_G Microcystis aeruginosa 9701 WGS project CAIQ01000000 data, contig AAI_K_2204_420
length: 1477
e value: 1.07696e-127
ACTCTAGGGAGACTGCCGGGGACAACTCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTACGTCCTGGGCGACACACGTACTACAATGGTTGGGACAAAGGGCAGCAAACTCGCGAGAGTAAGCGAATCTCATCAAACCCAGCCACAGTTCAGATTGCAGGCTGCAACTC
|||||| |||||||||||| |||||  |||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||||||||| |||||||||||||| ||||||||  |||||||| || ||||||| ||| ||||||||||||||||||||||||||||
ACTCTAAGGAGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCAGCATGCCCCTTACGTCTTGGGCGACACACGTACTACAATGGTCGGGACAAAGGGCAGCGAACTCGCGAGAGCCAGCGAATCCCAGCAAACCCGGCCTCAGTTCAGATTGCAGGCTGCAACTC
match: gnl|BL_ORD_ID|193 CAIQ01000358_Microcystis_aeruginosa_PCC_9701_:_G Microcystis aeruginosa 9701 WGS p
```

Em seguida, a aplicação monta um arquivo .FASTA com todas as sequências resultantes do BLAST para que possa ser aplicado o algoritmo de alinhamento MUSCLE. A Figura 7 mostra o arquivo resultante do alinhamento realizado.

Figura 7 – Arquivo com as sequências alinhadas com o MUSCLE

```
>gnl|BL_ORD_ID|313 AB039012_Nodosilinea_nodulosa_PCC_7104_(Leptolyngbya_sp.)_:_R,G Leptolyngbya PCC71
------ACTCTAAAGAGACTGCCGGGGAC-AACTCGGAGGAAGGTGGGGACGACGTCAAG
TC-ATCATGCCCCTTACGTCTTGGGCTACACACGTCCTACAA--TGC---TACAGACAGA
GGG-CAGCAAGCGCGCGAGTGCAAGCAAATCCCAT-AAACTGTGGCTCAGTTCAGATTGC
AGGCTGCAACTCGCCTGCATGAAGGCGG-AATCGCTAGTAATCGCAGGTCAG-CA-TACT
GCGGTGAATACGTTCCCGGGCCTTGTACACACC-GCCCGTCACACCATG-GGAGT-TGGC
CACGCCCGAAGTCGTTACTCTAACCG----------
>gnl|BL_ORD_ID|80 KF307598_Nodosilinea_nodulosa_UTEX_2910_:_T Nodosilinea nodulosa UTEX 2910 16S ribo
------ACTCTAAAGAGACTGCCGGGGAC-AACTCGGAGGAAGGTGGGGACGACGTCAAG
TC-ATCATGCCCCTTACGTCTTGGGCTACACACGTCCTACAA--TGC---TACAGACAGA
GGG-CAGCAAGCGCGCGAGTGCAAGCAAATCCCAT-AAACTGTGGCTCAGTTCAGATTGC
AGGCTGCAACTCGCCTGCATGAAGGCGG-AATCGCTAGTAATCGCAGGTCAG-CA-TACT
GCGGTGAATACGTTCCCGGGCCTTGTACACACC-GCCCGTCACACCATG-GGAGT-TGGC
CACGCCCGAAGTCGTTACTCTAACCG----------
>gnl|BL_ORD_ID|313 AB039012_Nodosilinea_nodulosa_PCC_7104_(Leptolyngbya_sp.)_:_R,G Leptolyngbya PCC71
TTGGGCACTCTAAAGAGACTGCCGGGGAC-AACTCGGAGGAAGGTGGGGACGACGTCAAG
TC-ATCATGCCCCTTA----------
----------
```

Em seguida, a aplicação utiliza o arquivo resultante do alinhamento e retira somente as sequências alinhadas, escrevendo-as no arquivo que será executado pelo PRIMER3, que possui os parâmetros customizáveis, que devem ser alterados de acordo com as necessidades do usuário. Após a execução do PRIMER3, o algoritmo gera um arquivo com o resultado da busca dos Primers, como mostra a Figura 8.

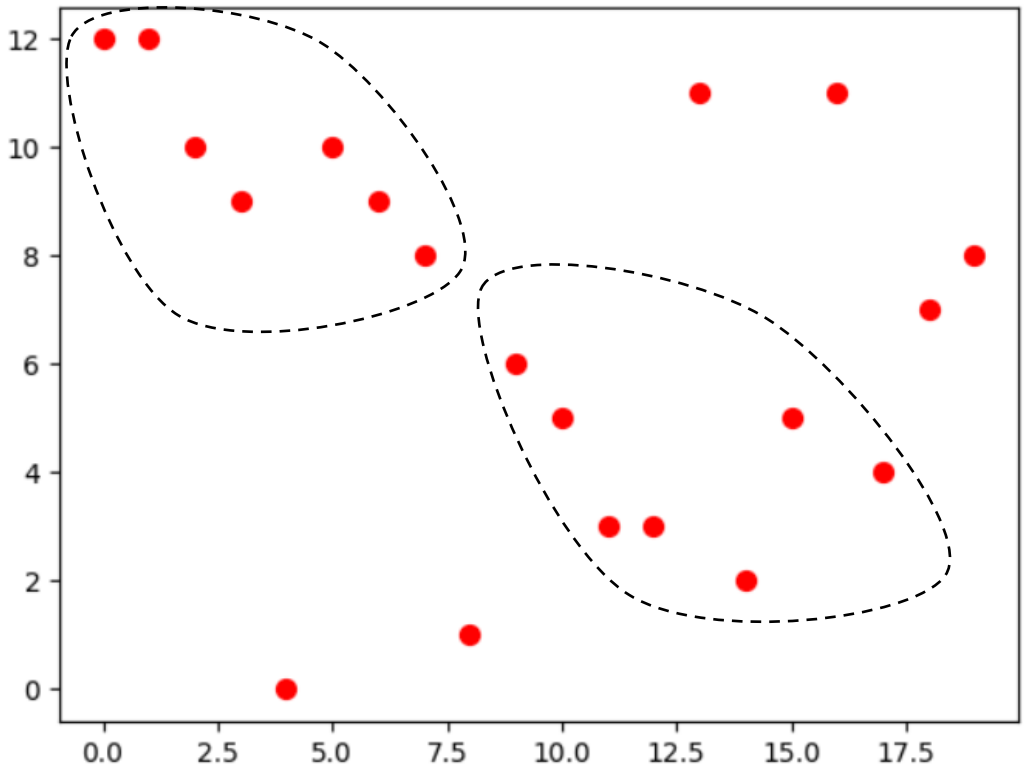
Figura 8 – Arquivo resultante do PRIMER3

```
SEQUENCE_ID=seq1
SEQUENCE_TEMPLATE=GATGAACGCTGGCGGTATGCTTAACACATGCAAGTCGAACGGGATCTTTCGGGATCTAGTGGCGGACGGGTGAGTAACCCTTGAGTAACATGTACTCATGTAATAATTGGGGACAACAGTTGGA
PRIMER_LEFT_EXPLAIN=considered 8331, GC content failed 23, low tm 5332, high tm 611, high any compl 1, high hairpin stability 852, ok 1512
PRIMER_RIGHT_EXPLAIN=considered 8334, GC content failed 23, low tm 5305, high tm 627, high hairpin stability 773, ok 1606
PRIMER_INTERNAL_EXPLAIN=considered 14415, GC content failed 8, low tm 7236, high tm 1877, high hairpin stability 1999, ok 3295
PRIMER_PAIR_EXPLAIN=considered 625, unacceptable product size 618, ok 7
PRIMER_LEFT_NUM_RETURNED=5
PRIMER_RIGHT_NUM_RETURNED=5
PRIMER_INTERNAL_NUM_RETURNED=5
PRIMER_PAIR_NUM_RETURNED=5
PRIMER_PAIR_0_PENALTY=0.061745
PRIMER_LEFT_0_PENALTY=0.032376
PRIMER_RIGHT_0_PENALTY=0.029369
PRIMER_INTERNAL_0_PENALTY=1.148357
PRIMER_LEFT_0_SEQUENCE=TGCAACTCGCCTGCATGA
PRIMER_RIGHT_0_SEQUENCE=GACGGGCGGTGTGTACAA
PRIMER_INTERNAL_0_SEQUENCE=CGGTGAATACGTTCCCGGGCC
PRIMER_LEFT_0=1240,18
PRIMER_RIGHT_0=1332,18
PRIMER_INTERNAL_0=1294,21
PRIMER_LEFT_0_TM=59.968
PRIMER_RIGHT_0_TM=59.971
PRIMER_INTERNAL_0_TM=59.852
```

Como é possível observar na Figura 8, o arquivo mostra os Primers encontrados do lado esquerdo, no campo PRIMER_LEFT_0_SEQUENCE, o Primer do lado direito, no campo PRIMER_RIGHT_0_SEQUENCE, e ainda a área conservada aos Primers, no campo PRIMER_INTERNAL_0_SEQUENCE.

Após a geração dos Primers, novamente as sequências de Primers são repassadas para um arquivo, o qual será utilizado para a criação de um data frame, que será base do algoritmo k-means, que é o algoritmo utilizado para o agrupamento dos Primers. O resultado do agrupamento leva em consideração o tamanho e a quantidade dos Primers gerados, e como resultado, foram gerados 4 clusters. É possível observar que dois clusters estão bem definidos e dois clusters um pouco mais distantes, como mostra a Figura 9.

Figura 9 – Plotagem dos agrupamentos dos Primers gerados



Para finalizar a execução, a aplicação cria um arquivo com as regiões conservadas dos Primers e realiza novamente o BLAST, para validar a identidade da espécie de Cianobactéria pesquisada, aplicando a técnica de DNA-Barcoding e a confiabilidade do banco de dados.

4 CONCLUSÃO

A partir do estudo realizado, é visível que a Bioinformática é uma área que vem crescendo exponencialmente devido também ao aprimoramento das tecnologias de informação, que estão em constante avanço. Está ligada às áreas da Biologia e da Saúde, buscando identificar problemas e resolvê-los mais rapidamente. Trabalhos recentes mostram que estão sendo utilizadas as técnicas de DNA-Barcoding e Aprendizado de Máquina juntas e comprovam os benefícios que estas tecnologias refletem diretamente nos resultados obtidos.

Após a pesquisa de trabalhos relacionados, foi possível encontrar diversos estudos que estão utilizando a técnica de Barcoding para o sequenciamento de DNA e a identificação mais rápida e confiável dos organismos. E aliando-se de técnicas de Aprendizagem de Máquina, os resultados obtidos são promissores.

Esta pesquisa trabalhou com uma base de dados relativamente grande, em termos de tamanho de armazenamento, aproximadamente 600 Gigabytes, e deixa a possibilidade de aumentar esta base futuramente. Ainda, a possibilidade de armazenar uma maior diversificação de espécies, sendo possível adaptar a base para a utilização de espécies de vírus e bactérias. Além disso, sugere-se a utilização de um banco de dados na nuvem, o que talvez impactasse um pouco no tempo de execução, mas abriria um leque de possibilidades de espécies a utilizar, levando em consideração que os bancos de dados de vírus, bactérias e protozoários podem ultrapassar 1 Terabyte de armazenamento.

Os testes realizados se destacaram pela velocidade de execução e pelo resultado de cada processo dentro da aplicação.

Em relação ao problema de pesquisa definido “De que forma possível localizar regiões de Primers em sequências de DNA para identificação de organismos utilizando DNA-Barcoding e técnica de agrupamento?”, os resultados demonstram que a solução desenvolvida otimiza o processo realizado, trazendo a confiabilidade e redução de tempo e recursos para os especialistas que utilizam este tipo de aplicação.

REFERÊNCIAS

ARAÚJO, C. A. (2006) Bibliometria: evolução histórica e questões atuais. Em Questão, Porto Alegre, v. 12, n. 1, p. 11-32.

BECKERS, B., BEECK, M. O., THIJS, S., TRUYENS, S., WEYENS, N., BOERJAN, W., VANGRONSVELD, J. Performance of 16s rDNA Primer Pairs in the Study of Rhizosphere and Endosphere Bacterial Microbiomes in MetaBarcoding Studies. Frontiers in Microbiology, 2016, v. 7, p. 1-15.

BLAST. Disponível em: https://blast.ncbi.nlm.nih.gov/Blast.cgi. Acessado em: 14/10/2020.

CORDIER, T., FORSTER, D., DUFRESNE, Y., MARTINS, C. I. M., STOECK, T., PAWLOWSKI, J. Supervised machine learning outperforms taxonomy-based environmental DNA metaBarcoding applied to biomonitoring. Molecular Ecology Resources. 2018, v. 18, p. 1381-1391.

DELONG, R. K., ZHOU, Q. Polymerase Chain Reaction (PCR). Introductory Experiments on Biomolecules and Their Interactions, 2015, 59–66.

EDGAR, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113. 2004.

GERHARD, W. A., GUNSCH, C. K. MetaBarcoding and machine learning analysis of environmental DNA in ballast water arriving to hub ports. Environment International, 2019, v. 124, p. 312-319.

HEBERT, P. D. N.; CYWINSKA, A.; BALL, S. L.; DEWAARD, J. R. Biological identifications through DNA barcodes. Proceedings. Royal Society Biological Sciences Meeting. 2003, v. 270, 313-321.

JAIN, Anil K.; DUBES, Richard C. Algorithms for Clustering Data. Prentice Hall. New Jersey. 1988.

KORF, Ian & YANDELL, Mark, BEDELL, Joseph. BLAST - an essential guide to the basic local alignment search tool. 2003.

KRESS, W. J., ERICKSON, D. L. DNA barcodes: genes, genomics, and bioinformatics. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105, 2761–2762.

LIBBRECHT, Maxwell W., NOBLE, William S. Machine learning in genetics and genomics. HHS Public Access, 2015, v. 16, p. 321-332.

LUSCOMBE, N. M., GREENBAUM, D., GERSTEIN, M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf. Med. 2001, 40, 346-358.

MITCHELL, Tom. *Machine Learning*, McGraw-Hill; 1997.

NCBI, Blast. 2021. Disponível em https://blast.ncbi.nlm.nih.gov/Blast.cgi. Acessado em: 21/01/2021.

PEARSON, William. LIPMAN, David J. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America. 85. 2444- 8. 1988

RAMOS, V., MORAIS, J., VASCONCELOS, V. A curated database of cyanobacterial strains relevant for modern taxonomy and phylogenetic studies. Scientific Data 4. 2017.