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

Victor Sanmartin[1], Rejane Frozza[1,2]

Universidade de Santa Cruz do Sul (UNISC) Av. Independência, 2293 – Bairro Universitário Santa Cruz do Sul – RS – Brasil

[1]Departamento de Engenharias, Arquitetura e Computação

[2]Programa de Pós-graduação em Sistemas e Processos Industriais

vsanmartin@mx2.unisc.br, frozza@unisc.br

Abstract. *Bioinformatics is an area that has great prominence and has been gaining even more visibility for prediction of diseases through DNA. Since it appeared, bioinformatics 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 techniques of DNA-Barcoding and Machine Learning tend to help even more researchers in the area, seeking to bring quick and intelligent solutions. And it is with this aim that the objective of this work is based, which seeks to unite the techniques of Barcoding and Machine Learning for the sequencing and identification of Cyanobacteria.*

Resumo. *A bioinformática é uma área que possui um grande destaque e vem ganhando ainda mais visibilidade para a previsão de doenças através do DNA. Desde que surgiu, a bioinformática 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. As técnicas de DNA-Barcoding e Aprendizado de Máquina tendem a ajudar ainda mais os pesquisadores da área, buscando por soluções rápidas e inteligentes. Assim, o objetivo desta pesquisa ainda em andamento, busca unir as técnicas de Barcoding e Aprendizado de Máquina para o sequenciamento e identificação de Cianobactérias.*

1. Introdução

A bioinformática é o ato de conceituar a biologia em termos moleculares e aplicar técnicas de informática, as quais derivam de disciplinas como matemática aplicada, ciência da computação e estatística, a fim de entender e organizar as informações associadas a essas moléculas em larga escala (Luscombe, 2001). A bioinformática integra essencialmente o desenvolvimento de programas computacionais para tratar dados biológicos preexistentes e identificar sequências de genes.

Já o conceito de DNA-*Barcoding* está diretamente relacionado com o que se conhece hoje 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 outros métodos mais complexos para a sua identificação. DNA-*Barcoding* é uma ferramenta

para rápida identificação de espécies com base em sequências de DNA e de 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 a bancos de dados de sequências e calcula a significância estatística das correspondências (BLAST, 2020). O *BLAST* pode ser usado para inferir relações funcionais e evolutivas entre sequências, bem como ajudar a identificar membros de famílias de genes (NCBI, 2020). Após ser realizado o sequenciamento das moléculas de DNA, deverá ser identificado o que se chama de *Primer*, que é uma sequência curta de ácido nucleico, de até 60 nucleotídeos semelhantes em todas as sequências, que fornece um ponto de partida para a síntese de DNA. Eles atuam como delimitadores, sendo que 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 foram desenvolvidos para ajudar a entender os 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). Como a área da Computação está diretamente ligada a este assunto, ela pode ser utilizada para o desenvolvimento de um *software* que realize o alinhamento das sequências de DNA e utilize DNA-*Barcoding* e Aprendizado de Máquina para fazer a identificação de cianobactérias, a partir da identificação de regiões de *Primers*.

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. E o problema de pesquisa refere-se a: É possível, de uma forma mais otimizada, localizar regiões de Primers em sequências de DNA para identificação de organismos do tipo cianobactérias?

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 trará um resultado confiável e rápido.

O artigo está organizado nas seguintes seções: a seção 2 apresenta uma breve fundamentação teórica, a seção 3 descreve a metodologia, na seção 4 são abordados os aspectos referentes à proposta do trabalho e a seção 5 apresenta as considerações finais.

2. Fundamentação teórica

O DNA *Barcoding* ou código de barras de DNA objetiva sistematizar e “etiquetar” as sequências de DNA para otimizar o tempo de identificação de espécies que já estejam catalogadas e acelerar processos de descobertas (Herbert *et al.*, 2003).

Os *Primers*, que são sequências curtas de DNA, podem variar de 20 até 60 nucleotídeos semelhantes, fornecem um ponto de partida para a síntese de DNA. Os primers são desenhados baseando-se no código genético (DNA molde) da espécie que se deseja estudar, sendo capazes de se anelar às suas sequências alvo de DNA durante a reação em cadeia da DNA polimerase (PCR) (Giegerich, Meyer & Schleiermacher, 1996).

As cianobactérias (também chamadas de algas verde-azuladas ou cianofíceas) são um grupo antigo de micróbios fotossintéticos que ocorrem na maioria das águas interiores e que podem ter grandes efeitos na qualidade da água e no funcionamento dos ecossistemas aquáticos (Vincent, 2009). Estes organismos realizam fotossíntese capturando a luz solar para obter energia e são encontrados em lagos, lagoas, córregos, rios e mares e representam um papel muito importante na fixação de nitrogênio, carbono e oxigênio em ambientes aquáticos.

Segundo Rezende (2003), "Aprendizado de Máquina é uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento de forma automática".

3. Metodologia

A bibliometria quantitativa (Araújo, 2006) foi realizada, com o objetivo 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 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.

Então, foram escolhidos quatro artigos científicos, após leitura dos resumos para serem estudados e contribuírem 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.

Os procedimentos metodológicos definidos são: i) levantamento bibliográfico, para obter aprofundamento nos assuntos da pesquisa; ii) levantamento de trabalhos relacionados, com realização da bibliometria; iii) modelagem do sistema a ser desenvolvido; iv) desenvolvimento do sistema de sequenciamento e reconhecimento de Cianobactérias, utilizado DNA-*Barcoding* e aprendizado de máquina; v) realização de testes e ajustes de parâmetros; vi) análise do desempenho alcançado utilizando o método proposto.

4. Proposta do sistema de alinhamento e identificação de cianobactérias utilizando DNA-Barcoding e Aprendizado de Máquina

O sistema será modelado como apresentado na Figura 1, com o usuário abrindo um arquivo com a extensão ".FASTA", o qual terá as sequências que deverão ser pesquisadas. Em seguida, deverá ser realizado um *BLAST* dessa sequência com um banco de dados a ser escolhido. Após o *BLAST* ter sido realizado, o sistema deverá alinhar as sequências e buscar possíveis regiões de *Primers* em todas as sequências, podendo variar o tamanho e a sequência. Por último, o sistema deverá gerar um novo arquivo .FASTA com os *Primers* gerados, para que possa ser realizado novamente o *BLAST* e buscar por sequências semelhantes para validar a integridade dos *Primers* gerados.

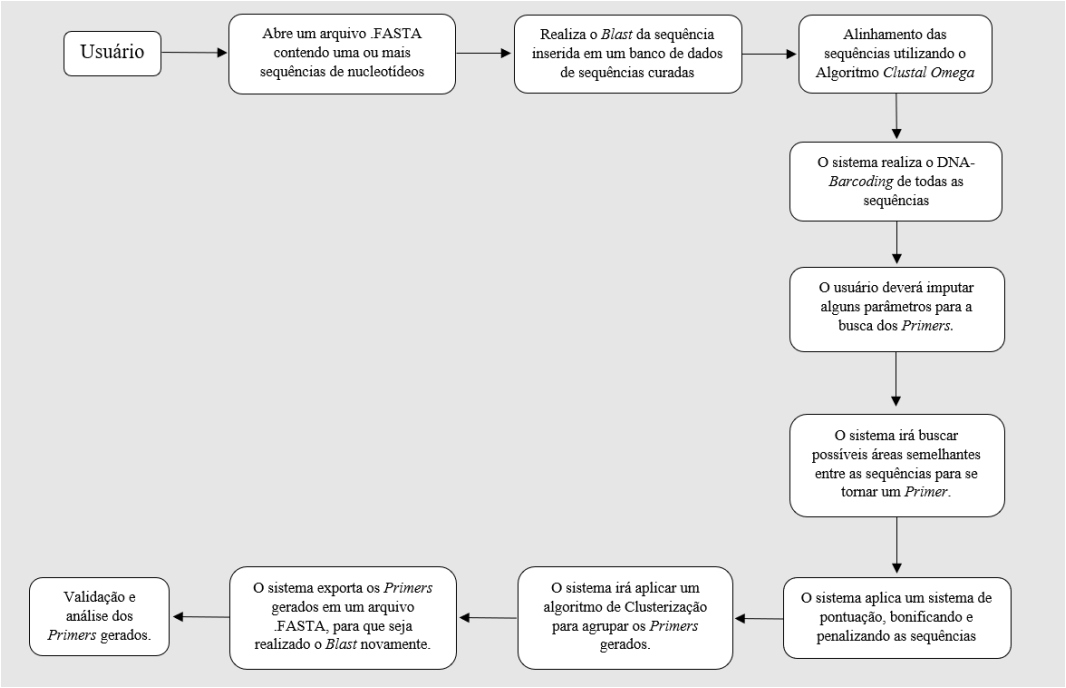

Figura 1. Modelo do Sistema proposto

Após o usuário realizar o upload do arquivo .FASTA com as sequências a serem pesquisadas, o sistema irá se conectar com o banco de dados GenBank, do site do NCBI e hospedado nos servidores da UNISC, para realizar o *BLAST* e buscar por sequências semelhantes. Em seguida, o sistema irá realizar o alinhamento de todas as sequências retornadas após o *BLAST*, utilizando um algoritmo chamado Clustal Omega (Clustal Omega, 2020).

Cada alinhamento possui um valor de pontuação que indica o quanto o alinhamento tem posições similares/idênticas. Essa pontuação é calculada por algoritmos que computam o valor do pareamento de cada letra no alinhamento e depois somam esses valores, com bonificação se for encontrada a letra similar ou penalização se for uma letra diferente. Para comparações de nucleotídeos, são utilizados os valores +1 para as combinações de letras no alinhamento (*match*) e -2 para não combinações (*mismatch*). Os *gaps* também têm um custo negativo (penalidade) para abertura (*gap-creation ou gap-opening*) e outro custo para extensão de um *gap* existente (*gap-extension*). Os valores de criação (penalidade) do *gap* são maiores que os valores de extensão do *gap* (Genebio, 2020).

Após o alinhamento realizado, a partir da pontuação calculada pelo algoritmo, para a localização dos *Primers*, será necessário que o usuário impute alguns parâmetros necessários, como tamanho, temperatura de anelamento, os quais são necessários para que o sistema realize a busca de acordo com a necessidade do usuário. Assim que os parâmetros estiverem definidos, o sistema irá buscar as áreas semelhantes em todas as sequências, e destacar as sequências que possivelmente pode se tornar um *Primer*. Assim como é realizado no algoritmo de alinhamento, a busca de *Primers* é realizada baseada no sistema de pontuação comentado anteriormente, cujas sequências que possuírem uma pontuação maior terão maior possibilidade de se tornar um *Primer*.

Em seguida, assim que os *Primers* forem definidos, o sistema aplicará um algoritmo de Clusterização para agrupar os *Primers* gerados. As técnicas de clusterização têm como objetivo buscar padrões em dados desordenados, classificando e organizando os objetos de acordo com suas semelhanças, distância ou densidade em um espaço multidimensional (Jain e Dubes, 1988).

5. Considerações finais

A partir do estudo realizado no referencial teórico, é visível que a bioinformática é uma área que vem crescendo exponencialmente e deve-se ao crescimento das tecnologias de informação, que estão em constante avanço, a cada dia que passa. É uma área que está ligada às áreas da Biologia e da Saúde, buscando resolver problemas e identificar estes problemas mais rapidamente. Trabalhos recentes mostram como 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, nota-se que os resultados obtidos são muito bons, o que torna o estudo ainda mais interessante e atual.

Para a próxima etapa deste trabalho, será desenvolvido um sistema de alinhamento e reconhecimento de sequências de DNA em Cianobactérias, utilizando as técnicas de DNA-*Barcoding* e Aprendizado de Máquina. Também serão executados testes e validações com profissionais da área da Biologia e Bioinformática para avaliar o desempenho do sistema desenvolvido.

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.

CLUSTAL OMEGA. Disponível em: https://www.ebi.ac.uk/Tools/msa/clustalo/. 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.

GENEBIO, BLAST. 2020. Disponível em http://www.genebio.ufba.br/BLAST/. Acessado em: 16/06/2020.

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.

GIEGERICH, R.; MEYER, F; SCHLEIIERMACHER, C. GeneFisher: software support for the detection of postulated genes. In: INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS FOR MOLECULAR BIOLOGY, 4.; 1996. BethesdaMD: NCBI. p. 68-77.

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.

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.

REZENDE, Solange Oliveira. Sistemas Inteligentes – Fundamentos e Aplicações. São Paulo: Manole, 2003.

VINCENT, W. F. *Cyanobacteria. Encyclopedia of Inland Waters,* 2009. 226–232.