UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Anderson Furtado de Nazaré
Nicoli da Silva Pereira de Souza

MODELOS OCULTOS DE MARKOV PARA PREDIÇÃO DE GENES NIF11 EM CIANOBACTÉRIAS DA REGIÃO AMAZÔNICA

Belém - PA
2017

Anderson Furtado de Nazaré
Nicoli da Silva Pereira de Souza

MODELOS OCULTOS DE MARKOV PARA PREDIÇÃO DE GENES NIF11 EM CIANOBACTÉRIAS DA REGIÃO AMAZÔNICA

Trabalho de conclusão de curso de graduação apresentado para obtenção do título de Bacharel em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Faculdade de Computação. Universidade Federal do Pará.
Orientador Prof. Dra. Regiane Silva Kawasaki Francês.
Coorientador Msc. Alex Ranieri Jerônimo Lima

Belém - PA
2017

MODELOS OCULTOS DE MARKOV PARA PREDIÇÃO DE GENES NIF11 EM CIANOBACTÉRIAS DA REGIÃO AMAZÔNICA

Anderson Furtado de Nazaré
Nicoli Pereira de Souza

Trabalho de conclusão de curso apresentado ao Instituto de Ciências Exatas e Naturais da Universidade Federal do Pará como requisito para a obtenção do título de Bacharel em Ciência da Computação. Julgado em ___ de abril de 2017, com o seguinte conceito _____.

_____

Profa. Dra. Regiane Silva Kawasaki Francês
ORIENTADORA – ICEN – UFPA

_____

Msc. Alex Ranieri Jerônimo Lima
COORIENTADOR-ICB-UFPA

_____

Profa. Dra. Danielle Costa Carrara Couto
CONVIDADA – ICB– UFPA

_____

Msc. Andrei Santos Siqueira
CONVIDADO – ICB – UFPA

Belém - PA
2017

RESUMO

As cianobactérias possuem uma alta variabilidade estrutural e diversidade fenotípica, assim estando presentes em vários habitats. A grande diversidade no bioma da região amazônica somada aos estudos sobre bacteriocinas peptídeos produzidos por vários microrganismos, inclusive cianobactérias são de grande importância para várias indústrias, dentre elas a farmacêutica e alimentícia. Outro fator relevante para pesquisas com bacteriocinas é sua relação com genes do domínio *nif11*, presentes em cianobactérias da região amazônica. O presente trabalha utiliza-se de técnicas e ferramentas da bioinformática como alinhamentos múltiplos, softwares de visualização e a modelagem de perfis por inferência, para este último foram utilizados os Modelos Ocultos de Markov (*Hidden Markov Model* – HMM), onde um perfil foi desenvolvido e treinado para identificação de sequências dos genes *nif11* em sete genomas. Posteriormente a eficácia do modelo desenvolvido foi verificada ao serem comparados os resultados obtidos com os gerados pelo modelo TIGR03798, onde foi constatado um aumento de sensibilidade na identificação das sequências alvos.

Palavras-chaves: *HMM. Cianobactérias. nif11. Bacteriocina.*

ABSTRACT

The cyanobacteria have a high structural variability and phenotypic diversity, because of this reason they are present in several different habitat. The high diversity of Amazonian biome, summed to bacteriocin studies (peptides produced by numerous microorganisms, including cyanobacteria) have a great importance to the pharmaceutical and food industry. Besides that, other relevant point for bacteriocin related research is their relation with genes of the domain nif11, present in cyanobacteria from the Amazon region. In the present study, we used bioinformatics techniques and tools such as multiple sequence alignment, visualization software, and profiling by inference. For the later we used a developed and trained Hidden Markov Model (HMM) to identify sequences of nif11 genes in seven genomes.The efficiency of the developed model was verified by comparing the results obtained with those generated by the TIGR03798 model, where there was an increase on the sensibility of the target sequences identification.

Key words: *HMM, cyanobacteria, nif11, bacteriocin.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

FBN - Fixação Biológica do Nitrogênio.

HMM - Hidden Markov Model (Modelos Ocultos de Markov).

RNA - Ribonucleic Acid (Ácido Ribonucleico).

DNA - Ácido Desoxirribo nucléico.

PH - Potencial hidrogeniônico.

N2 - Nitrogênio.

NCBI - National Center for Biotechnology Information (Centro Nacional de Informação Biotecnológica).

LaBioCAD - Laboratório de Bioinformática e Computação de Alto Desempenho.

ITB - Laboratório de Tecnologia Biomolecular.

SUMÁRIO

1. INTRODUÇÃO

As algas azuis, ou cianobactérias, são microrganismos com características celulares procariontes, contudo, seu sistema fotossintetizante se assemelha ao das algas, tornando-as bactérias fotossintetizantes (XIONG e BAUER, 2002). As cianobactérias têm sua origem estimada há aproximadamente 3,5 bilhões de anos e foram os principais produtores primários da biosfera e ainda hoje tem participação fundamental na oxigenação da atmosfera terrestre (RAYMOND e BLANKENSHIP, 2008). A plasticidade fenotípica e metabólica apresentada por esses organismos permite que este grupo esteja presente em diversos ambientes como sistemas de água doce, salgada e até em habitats extremos de acidez e temperatura (WHITTON e POTTS, 2002).

Uma importante característica das cianobactérias é a Fixação Biológica do Nitrogênio (FBN), sendo os principais responsáveis por esse processo em oceanos abertos (BERMAN-FRANK, LUNDGREN e FALKOWSKI, 2003). Os responsáveis por darem as características FBN a várias espécies são chamados de genes *nif*, sendo que o domínio predominante para esta função nas cianobactérias é denominado *nif11*, que tem por característica a grande variabilidade em suas cadeias de aminoácidos. O estudo de toda essa diversidade bioquímica é de extrema importância, sobretudo as pesquisas relativas às bactericidas, que possuem uma relação ainda pouco conhecida com os genes *nif11* (HAFT, BASU e MITCHELL, 2010).

Visto que tais pesquisas são fundamentais para alguns tipos de indústrias, como a farmacêutica e a alimentícia, e ainda sendo importante na produção de vários produtos como imunossupressores, antibióticos, pigmentos alimentícios, antitumorais, entre outros. Há a necessidade de conhecer melhor as potencialidades das cianobactérias da região amazônica, e uma área que fornece técnicas e ferramentas para melhor compreender esses processos é a bioinformática (PATTANAIK e LINDBERG, 2015).

Por ser um ramo da ciência que surgiu da grande necessidade dos profissionais da biologia em entender as funções biológicas de uma maneira mais rápida e eficiente, a bioinformática se utiliza de métodos como o alinhamento múltiplo de sequências e a modelagem de perfil por inferência, muitas das análises feitas seriam impossíveis sem a ajuda da potente capacidade de processamento dos computadores (FILHO, 2002).

A modelagem por inferência oferece a possibilidade de criar um perfil a partir de características calculadas a partir de um alinhamento, o método é capaz de identificar sequências de nucleotídeos e é baseado no Modelo Oculto de Markov (*Hidden Markov Model* - HMM). Com o perfil desenvolvido HMM, pode-se realizar buscas em genomas ou bancos

de dados de sequências e assim encontrar membros correspondentes ao perfil (KROGH, 1998).

1.1 MOTIVAÇÃO E JUSTIFICATIVA

Segundo (FILHO, 2002), por ser uma área considerada nova nos ramos de pesquisa, a Bioinformática encara como um de seus maiores problemas a existência de certas lacunas nos conhecimentos já adquiridos. Este trabalho será desenvolvido com o intuito de melhorar os conhecimentos e enriquecer a literatura na área da bioinformática e adicionar conhecimentos sobre as cianobactérias da região amazônica, neste sentido, há uma carência no estudo destes organismos (HAGEN, 2002).

Visto que os clusters de genes biossintéticos possuem grande variabilidade (WANG, FEWER e SIVONEN, 2011), e consequentemente seus peptídeos precursores são de difícil padronização (BERMAN-FRANK, LUNDGREN e FALKOWSKI, 2003), fez-se necessário um estudo que contribuísse na identificação desses elementos. Assim, com o objetivo de esclarecer os clusters[1] de genes de bactericida e seus peptídeos precursores e expandir a predição desse tipo de sequência em cianobactérias da região amazônica, foi um desenvolvido um modelo probabilístico baseado em técnicas de inferências conhecidas HMM (BAUM e EAGON, 1967).

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Desenvolver um modelo treinado de HMM a partir de sequências de aminoácidos do peptídeo precursor *nif11* de cianobactérias a fim de expandir a predição deste tipo de sequências presentes em cianobactérias isoladas do ambiente Amazônico.

1.2.2 Objetivos Específicos

- Construir e treinar o modelo HMM utilizando como base as sequências *nif11* de cianobactérias;
- Comparar os resultados do modelo HMM treinado com os resultados provenientes de modelos HMM existentes, quando aplicados em dados genômicos de cianobactérias isoladas do ambiente Amazônico.

---

[1] Conjunto de genes homólogos dentro de um organismo, aglomerados de genes.

2. REVISÃO BIBLIOGRÁFICA

2.1 BIOINFORMÁTICA

O termo bioinformática foi inicialmente utilizado por (HESPER e HOGEWEG, 1970), o nome foi cunhado para designar o estudo de processos computacionais em sistemas biótica sendo a área conhecida por sua interdisciplinaridade, agregando conhecimentos da matemática, tecnologias computacionais e biologia molecular (HAGEN, 2002).

Segundo (SANGER, NICKLEN e COULSON, 1977) e (MOREIRA, 2015), a bioinformática é definida como a pesquisa, desenvolvimento e utilização de ferramentas computacionais para análises e propagação de informações biológicas ou de áreas afins, geralmente com enfoque na biologia molecular. Atuando também na organização, armazenamento e visualização de dados. Assim, a bioinformática veio para suprir a necessidade de manipular um montante cada vez maior de dados tanto qualitativos quanto quantitativos.

Atualmente, segundo (PROSDOCIMI, CERQUEIRA, *et al.*, 2002), a bioinformática atua principalmente na busca de uma maior compreensão dos processos bióticos, coletando e processando dados de genomas para analisar suas funções proteicas, assim também contribuindo na pesquisa de novos compostos farmacêuticos, detalhando estruturas de DNA, afim de facilitar o desenvolvimento de drogas mais eficientes.

2.1.1 Técnicas e Ferramentas

Dentre algumas das principais aplicações computacionais desenvolvidas e utilizadas na bioinformática, temos: o alinhamento de sequências (podendo ser global ou local); a modelagem e simulação de sistemas biofísicos ou bióticos; a construção e organização de bases de dados; anotações, entre outros (LESK, 2008).

As técnicas de alinhamento são utilizadas para identificação de regiões de similaridade entre sequências de DNA, RNA ou proteínas, que podem ser consequências de relações funcionais, estruturais ou evolucionárias. A técnica pode também ser aplicada de forma global ou loca, com a primeira tendo o objetivo de comprar duas sequências por inteiro, para obter a identificação de todas as correspondências entre elas. Já a segunda é utilizada para identificar um conjunto de regiões que apresentam correspondência entre sequências. Os resultados e interpretações dessas aplicações tem grande valor científico e serve de base para a construção de árvores filogenéticas, desenvolvimento de hipóteses sobre relações identificadas, além de outras informações obtidas desse processo (OGATA, 2005). Dentre as principais ferramentas

para aplicação da técnica de alinhamento podemos citar o BLAST, CLUSTALW, MUSCLE, MAFFT e MEGA.

Sobre a modelagem e simulação de sistemas biófitos ou bióticos, a técnica possibilita a criação de modelos tridimensionais sobre moléculas, relação entre elementos proteicos, simulação entre processos espaciais dinâmicos, entre outros.

Já as anotações são utilizadas para o mapeamento ou inclusão de análises, comentários ou referências em regiões de sequências, podendo ser feita de forma manual ou semiautomática. Sendo usualmente aplicadas para a criação de mapas genômicos, evidenciando posições, classificações e funções de determinado gene. Dentre as ferramentas que auxiliam na técnica podemos citar o COG, PFAM, BLSATP e KEGG (FARINA, 2012).

Para técnicas de predição, que objetivam deduzir resultados, criando modelos através de técnicas estatísticas de inferência, podemos citar alguns softwares como GLIMMER, GENEMAK. HMM, PRODIGAL e HMMER (WHEELER, CLEMENTS e FINN, 2014).

É importante esclarecer que as aplicações listadas não se tratam de uma classificação oficial, mas sim uma tentativa de apresentar algumas estratégias utilizadas no campo da bioinformática.

2.2 CIANOBACTÉRIAS

O filo Cianobactéria pertence ao domínio Bactéria e é composto por micro-organismos procariontes, fotossintetizantes e gram-positivos, que tem sua origem datada em aproximadamente 3 bilhões de anos atrás (BLANK e SÁNCHEZ-BARACALDO, 2010). Por se tratar de um grupo muito antigo e possuírem uma alta diversidade ecológica e morfológica, as cianobactérias podem ser encontradas em diversos tipos de ambientes, dentre eles os marinhos e de água doce. Devido à sua alta tolerância a incidência de raios ultravioletas, as cianobactérias podem ser encontradas em lugares considerados inóspitos, como lugares com concentrações elevadas de metais pesados, baixas concentrações de oxigênio, ambientes com temperaturas extremas como em desertos, águas termais e lagos antárticos assim como em lugares que apresentam extremo pH (WHITTON e POTTS, 2002).

Em determinados ecossistemas, as cianobactérias podem chegar a se tornar a espécie dominante (STENICO, 2017). Contudo, a combinação de um conjunto de fatores ambientais, pode desencadear, de forma inesperada, a proliferação massiva de cianobactérias num curto espaço de tempo (florescências), aonde determinados gêneros e espécies de cianobactérias chegam a produzir substâncias tóxicas, também denominadas de cianotoxinas, que podem ter efeitos negativos em mamíferos, com efeitos neurotóxicos e hepatóxicos (MOLICA e

AZEVEDO, 2009). As cianotoxinas são metabolitos secundários quimicamente diversos, e tal diversidade se aplica também aos seus produtos com utilidade para a indústria.

De acordo com (SINGH, KATE e BANERJEE, 2005) as cianobactérias são excelentes fontes de compostos bioquímicos, de combustíveis renováveis e compostos bioativos, tendo assim grande potencial para ser explorado em várias áreas como cosméticos, indústrias alimentícia e farmacêutica (TAN, 2007).

Atualmente verificamos a existência de múltiplos trabalhos sobre cianobactérias, a maioria deles tendo como base amostras de cianobactérias encontradas no hemisfério norte do planeta ou até mesmo em regiões do Brasil, como sul e sudeste (MONDARDO, 2004) (LORENZI, 2004), (FONSECA, 2014), (DA ANUNCIAÇÃO GOMES, SAMPAIO, et al., 2017). E mesmo que as cianobactérias estejam amplamente distribuídas em todas as regiões do nosso país (FONSECA, 2014) poucos estudos são direcionados as cianobactérias encontradas no meio Amazônico, deixando assim essa região com baixa representatividade nas bases de dados internacionais, a exemplo do NCBI.

2.2.1 Bacteriocina e Genes *nif11*.

Metabolitos secundários foi um termo cunhado a mais de cem anos atrás por (KOSSEL, 1897), se referindo a compostos orgânicos produzidos pelos seres vivos que não se encaixavam na definição de metabolitos primários (essenciais para o desenvolvimento e reprodução), porém estudos mais recentes tem mostrado que esses compostos podem ser de grande auxilio para sucesso reprodutivo, promover mecanismos de defesa e auxiliar na competição interespecífica do mesmo (DEMAIN e FANG, 2000), (VINING, 1990), (SINGH, TIWARI, *et al.*, 2011) e (PENN, WANG, *et al.*, 2014). Na natureza são encontrados exemplos de metabolitos secundários como agentes inibidores de crescimento (AERTS, SNOEIJER, *et al.*, 1991), antibactericidas e antifúngicos (CROTEAU, KUTCHAN e LEWIS, 2000)e protetores contra foto-destruição (GRACE e LOGAN, 2000) e (SCHREINER, MARTÍNEZ-ABAIGAR, *et al.*, 2014), dentre essa variabilidade de metabólitos secundários, as bacteriocinas ganham destaque.

As bacteriocinas são peptídeos com atividade bactericidas ou bacteriostática contra outros organismos, sendo da mesma espécie ou de gênero diferente (VELHO, 2010). Para

biossíntese dessas substâncias, os genes geralmente são agrupados em *operons*[2] que incluem o gene estrutural da bacteriocina (COTTER, HILL e ROSS, 2005). Estes curtos genes possibilitam a codificação de polipeptídios incialmente como precursores e posteriormente os mesmos sofrem modificações químicas durante sua maturação (JACK, TAGG e RAY, 1995) e (COTTER, EARL, *et al.*, 2015). Contudo, foi identificada uma peculiaridade nas bacteriocinas específicas da região amazônica, onde as mesmas apresentam uma diferença em relação as demais bacteriocinas presentes nos bancos de dados. A bacteriocina possui uma região conservada e uma região variável (HAFT, BASU e MITCHELL, 2010), e foi averiguado que as cianobactérias analisadas possuíam um marcador inicial para a região variável diferente das demais cianobactérias nos bancos de dados, essa característica está relacionada com genes fixadores de nitrogênio (*nif*) (HAFT, 2010).

O Nitrogênio, assim como o carbono, hidrogênio e o oxigênio é um elemento de grande importância para a vida, pois está presente na maior parte dos componentes celulares. Mesmo sendo o elemento em maior quantidade na nossa atmosfera, não é aproveitável para a maioria dos organismos em sua forma molecular (N2) (FARINA, 2012), tornando assim a FBN em uma importante etapa no ciclo do nitrogênio. Alguns organismos possuem a capacidade de executarem essa fixação do nitrogênio na forma de substâncias utilizáveis para a maioria dos organismos como amônia e nitratos. Sendo uma característica de organismos procariotos, os genes *nif* são domínios de sequência geralmente encontrados numa grande variedade organismos tais como bactérias e cianobactérias (CHAI, 2007).

Os genes *nif* são responsáveis pela produção das enzimas responsáveis pela fixação do nitrogênio atmosférico em outras formas de nitrogênio. Dentre as enzimas produzidas pelos genes *nif* estão a nitrogenase, que ficam organizadas em clusters dentro dos genes *nif*. (DOS SANTOS, 2010.). Dos vários tipos de genes *nif* existentes, foram encontradas em algumas cianobactérias uma família de genes *nif11*, que tem sua função desconhecidas e sua principal característica é o ponto de clivagem sendo uma dupla glicina (HAFT, BASU e MITCHELL, 2010).

---

[2] É um conjunto de genes nos procariontes e em alguns eucariontes que se encontram funcionalmente relacionados, contíguos e controlados coordenadamente, sendo todos expressos em apenas um RNA mensageiro. Ou seja, é constituído pelo promotor, o operador e os genes estruturais.

2.3 MODELO OCULTO DE MARKOV

Grandes partes dos processos que envolvem sistemas reais se apresentam de forma tão complexa que mesmo que haja uma forma analítica para resolvê-los, há casos no qual acaba sendo mais vantajoso lançar mão do uso da teoria de probabilidade. É fundamental que, para aplicar a teoria da probabilidade ao mundo real, seja introduzido o conceito de "variável estocástica". Assim uma variável X é dita variável estocástica se, dentro de um conjunto {xi} de possíveis realizações, seu valor é determinado pelo resultado de um experimento (REICHL, 1998).

Dentre os mais variados modelos probabilísticos estocásticos existentes, temos os Modelos Markovianos e os Modelos Ocultos de Markov.

Os HMM's foram descritos pela primeira vez no final da década de 1960 (BAUM e EAGON, 1967), ele trata de uma variação dos Modelos de Markov, descritas pelo matemático Andrey Markov em 1906. Inicialmente utilizados para o reconhecimento da fala na década de 1970 (BAUM, 1972) e (BAKER, 1975) e posteriormente adotadas para aplicações nas áreas de reconhecimento de padrões temporais como gestos, voz e palavras escritas. Por permitir uma melhor análise em sequências biológicas, principalmente do DNA e cadeias de proteínas (AQUINO, 2012).

Inicialmente, antes de conceituarmos HMM, temos que explicar o conceito de Modelos Marvokianos. Um processo de Markov nada mais é do que um processo que aplica a propriedade de Markov, possuindo a propriedade a seguir (ATUNCAR, 2011).

$$P\left(X_{t+1} \in A \vee X_1 = x_1, X_2 = x_2 \ldots, X_t = x_t\right) = P\left(X_{t+1} \in A \vee X_t = x_t\right)$$

De forma geral, o HMM é definido como um modelo onde as observações das relações do sistema é feita de forma indireta, agindo na forma de uma função probabilística da mudança entre estados conhecidos (ESPINDOLA, 2010). Apesar dos modelos de Markov poderem ser discretos e contínuos, o problema abordado no trabalho é discreto no espaço de dados e no tempo.

Segundo (AQUINO, 2012), p. 8 um HMM por definição é composto por:

- Um conjunto de Q estados distintos{ $E_1, E_2, \ldots, E_Q$ };
- Um conjunto de probabilidades de transição Pt, de tal maneira que Pt_{E_i,E_j} indica a probabilidade de estar no estado E_i e transitar para o estado E_J . O valor Pt_{E_i,E_j} depende exclusivamente do estado anterior E_i , de acordo com a propriedade de Markov;
- Um conjunto de símbolos de saída Σ = { $a_1, a_2, \ldots, a_\Sigma$ };

- Um conjunto de probabilidades de emissão Pe associado a cada estado, de tal maneira que $Pe_{E_2}(a_1)$ indica a probabilidade de emitir o símbolo de saída no estado E_j ;
- Um conjunto de probabilidades de iniciação π, tal que indica a π_{E_2}, probabilidade da emissão da sequência de observação iniciar no estado E_j.

Para ilustrar como funciona a cadeia de Markov, é apresentado o exemplo citado (ESPINDOLA, 2010) apud (RABINER, 1989).

Assumindo que uma dada variável estocástica X representa o tempo cujos estados estão definidos pelo conjunto {S1 = Chuvoso; S2= nublado; e S3 = ensolarado}. Para efeito de estudo, as observações do clima são realizadas uma vez ao dia, sempre no mesmo horário, e o resultado é necessariamente um dos elementos do conjunto. Assim, a probabilidade de transição dos estados é retratada pela matriz a seguir.

$$\hat{A} = \{a_{ij}\} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}$$

Tendo constatado que o tempo no primeiro dia de verificações foi ensolarado (X0 = S3), tem-se o objetivo de saber qual a probabilidade que o clima nos próximos sete dias seja retratado por uma dada sequência, tal como: ensolarado, ensolarado, chuvoso, chuvoso, ensolarado, nublado e ensolarado. Com o cenário preparado, a sequência de observações o descreve é :

$$O=\{X_0=S_3, X_1=S_3, X_2=S_3, X_3=S_1, X_4=S_1, X_5=S_3, X_6=S_2, X_7=S_3\}$$

Assim, a probabilidade de O é dado por:

$$\begin{aligned} P(O|Model) &= P(S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3|Model) \\ &= P(S_3)P(S_3|S_3)P(S_3|S_3)P(S_1|S_3)P(S_1|S_1)P(S_3|S_1)P(S_2|S_3)P(S_3|S_2) \\ &= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23} \\ &= 1 \cdot (0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2) \\ &= 1.536 \times 10^{-4} \end{aligned}$$

Onde a notação verifica a probabilidade inicial de cada estado é dada por:

$$\pi_i = P(X_0 = S_i), \quad 1 \leq i \leq N$$

2.3.1 Elementos de um HMM

Um elemento HMM para as observações de símbolos discretos é caracterizado da seguinte forma (ESPINDOLA, 2010):

- N, o número de estados do modelo. Os estados individuais são rotulados como { 1, 2, …, N} e o estado no tempo t como .

- M, o número de símbolos de observações distintas por estado, por exemplo, o tamanho do alfabeto discreto.

- A distribuição de probabilidade da transição do estado $A=\{a_{ij}\}$ onde:

$$a_{ij}=P[q_{t+1}=j \mid q_t=i], 1\le i, j\le N$$

Para o caso especial onde qualquer estado pode alcançar qualquer outro estado em uma simples etapa, tem-se $a_{ij}>0$ para todo i, j. Para outros tipos de HMM, podem-se ter $a_{ij}=0$ para um ou mais pares (*i, j*).

A distribuição de probabilidade de símbolos de observações, $B=\{b_j(k)\}$ define a distribuição de símbolos no estado $j, j=1,2,...,N$ onde:

$$b_j(k)=P[O_t=v_k \mid q_t=j], 1\le i\le M$$

A distribuição do estado inicial $\pi=\{\pi_i\}$, onde:

$$\pi_i=P[q_1=i], 1\le i\le N$$

Pode-se observar que uma completa especificação de um HMM requer especificação de dois parâmetros do modelo, N e M, especificação da observação de símbolos, e a especificação de três conjuntos de medidas de probabilidade $A, B e \pi$. Por conveniência será utilizada a notação compacta $\lambda=(A,B,\pi)$ para indicar o completo conjunto de parâmetros do modelo. Este conjunto de parâmetros, naturalmente, define a medida de probabilidade para 0, por exemplo, o qual será visto nas seções seguintes.

2.3.2 Perfil HMM para Representação de Família de Sequências Genômicas

A partir do alinhamento múltiplo de sequencias genômicas, um perfil HMM pode representar a família destas sequências. O modelo HMM representa de forma bem específica as informações de cada coluna, sendo assim capaz de demonstrar quais são as regiões conservadas e quais nucleotídeos tem a maior probabilidade de aparecer naquela posição (DURBIN, 1998).

De acordo com (AQUINO, 2012), é possível construir um perfil HMM que modele as principais características das famílias das sequências utilizadas é possível a partir das seguintes etapas:

1. Para cada coluna *j* do alinhamento múltiplo, um estado *M*j é criado no HMM;
2. Para cada estado *Mj*, uma transição de *Mj* para o estado *Mj+1* é criada, com probabilidade de transição P tMj ,Mj+1 = 1;
3. O conjunto Σ de símbolos de saída é definido como o conjunto de aminoácidos ou bases nitrogenadas que formam as sequências;
4. Para cada estado *Mj* e cada símbolo si de Σ, a probabilidade de emissão *P* e *Mj* (*si*), é determinada por:
5. As probabilidade de iniciação são definidas tal que $\pi M1 = 1$ e $\pi Mj = 0$, ∀j = 2...Q. Para exemplificar, serão levadas em consideração as sequências mostradas na Figura 2.1.

		colunas do alinhamento				
		1	2	3	4	5
S_1	=	A	C	G	T	A
S_2	=	A	C	G	A	A
S_3	=	A	C	T	T	A
S_4	=	A	C	T	C	A
S_5	=	A	C	C	C	G

Figura 2.1 *Alinhamento de sequências.*

Fonte: *Aquino(2012).*

Após concluir as cinco etapas apresentadas anteriormente, podemos representar nosso modelo na forma de um autômato, como mostra a Figura 2.2, onde podemos observar os estados chamados de Match (M) modelando cada coluna do alinhamento. É importante ressaltar que qualquer sequência de saída gerada pelo HMM deve começar em M_1, por ser o estado inicial do autômato que representa este modelo. As probabilidades de transição entre estados está representada por P_t sobre as setas que se encontram entre os estados. E por fim, dentro de cada estado, estão sendo indicadas as probabilidades de emissão de cada símbolo de saída P_e.

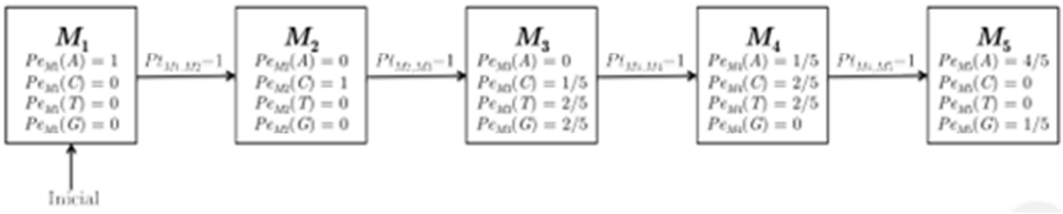

Figura 2.2 *HMM representante de uma família de sequências a partir do alinhamento da figura 2.1.*

Fonte: *Aquino(2012).*

A partir de um alinhamento múltiplo, tem-se a possibilidade de otimizar os parâmetros do HMM com métodos interativos, para isso é necessária a inclusão de novas sequências de observação. Esse processo é chamado de treinamento do HMM (PORITZ, 1988), e é realizado pelo algoritmo Baum-welch (SCHUSTER-BÖCKLER, 2004). O algoritmo atua na análise o comportamento do HMM durante a emissão de novas sequências assim podendo modificar as probabilidades de transição e emissão, deste modo novas sequências podem ser incluídas na família modelada pelo HMM.

Portanto, para calcular a probabilidade de uma sequência ser gerada por um perfil HMM, é necessário levar em consideração todas as transições que alcançam cada estado e do HMM. Essa operação, denominada avaliação da sequência, é realizada por dois algoritmos diferentes, porém equivalentes: o algoritmo *Forward* e o algoritmo *Backward* (FINK, 2008). Nesses algoritmos, todos os caminhos no HMM que podem ser tomados para gerar um conjunto de sequências são contabilizados para calcular o *score* [3]de dada sequência em relação ao HMM.

[3] Pontuação obtida de um alinhamento

3. MATERIAS E MÉTODOS

Este trabalho faz parte do Laboratório de Bioinformática e Computação de Alto Desempenho (LaBioCAD-UFPA) em colaboração com o Laboratório de Tecnologia Biomolecular (LTB-UFPA), que nos cedeu as 6 sequências genômicas das cianobactérias, das quais três estão disponíveis no (NCBI): *Cyanobium* sp. CACIAM 14 (GenBank: JMRP00000000.1) (LIMA, SIQUEIRA, *et al.*, 2014), *Limnothrix* sp. CACIAM 69d (GenBank: MKGP00000000.1), *Mycrocystis aeruginosa* CACIAM03 (GenBank: MCIH00000000.1) (CASTRO, LIMA, *et al.*, 2016), *Synechococcus* sp. CACIAM 66, *Synechocystis* sp. CACIAM 05, *Tolypothrix* sp. CACIAM 22. Adicionalmente foi utilizada a sequência genômica de *Synechocystis* sp. PCC6803 (GenBank: U67397.1).

As análises computacionais a seguir foram realizadas a partir de um computador com Intel® Core™ i3-4130 CPU com 3.40GHz × 4, 64-bit e 8 gigas de memória Kingston HyperX Gamer, utilizando o sistema operacional Ubuntu 16.04 LTS.

3.1 ALINHADOR.

Foram utilizadas no alinhamento seis sequências de aminoácidos do peptídeo precursor *nif11* dos clusters de genes de bacteriocina preditos pelo antiSMASH a partir do genoma da cianobactéria *Cyanobium* sp. CACIAM 14, Além de 1228 sequências de cianobactérias obtidas a partir do banco de dados do NCBI (NCBI, 2016), as sequências foram selecionadas com a utilização do filtro para buscar somente cianobactérias com o domínio de sequências *nif11*.

Para o alinhamento, foi utilizado o software MAFFT v7.310 (KATOH e STANDLEY, 2013), versão para desktop, utilizando como parâmetro de *threshold*[4] o *score* igual a 39, com *e-value*[5] sendo 8.4e-11, que gerou um arquivo no formato FASTA.

A partir do alinhamento, foi plotada a árvore filogenética das sequências devida o grande número de amostras, foi feito um recorte para a visualização somente nos clusters mais próximos das cianobactérias objetos de estudo do presente trabalho. A partir deste ponto, foram comparadas a eficácia e a sensibilidade do nosso perfil em relação ao perfil disponibilizado no TIGRFAM. O processo foi feito para todas as cianobactérias disponíveis e a comparação entre elas é apresentada na Figura 3.1.

---

[4] Limiar de identificação de sequências.
[5] Significância estatística da sequência.

O *e-value* foi ajustado para menor que um *e-value* << 1 para ter resultados mais significativos, mesmo modelo de corte utilizado pelo perfil TIGR0398, nosso modelo mostrou valores similares para os algoritmos MSV, Vitterb e Forward, com ambos os perfis aplicado para o algoritmo MSV, para o Vitterb, e para o Forward, todos com variância menor que 0.7 no valor final, assim demostrando que o diferencial do modelo desenvolvido é maior abrangência obtida através do treinamento feito com as cianobactérias da Amazônia.

3.2 CRIAÇÃO DO PERFIL - HMMBUILD.

A partir do arquivo resultante do alinhamento das sequências, com o auxílio da ferramenta HMMER-3.1b2 (EDDY, 1998), que se utiliza de técnicas Markovianas, foi criado um Perfil HMM, o mesmo foi utilizado para o próximo passo do processo de análise.

3.3 UTILIZAÇÃO DA TÉCNICA DE BUSCA – HMMBUILD

Ainda com a ferramenta HMMER-3.1b2, o perfil criado no passo anterior foi confrontado com sete genomas de cianobactérias, sendo eles: *Cyanobium* sp. CACIAM 14, *Limnothrix* sp. CACIAM 69d, *Mycrocystis aeruginosa* CACIAM03, *Synechococcus* sp. CACIAM 66, *Synechocystis* sp. CACIAM 05, *Synechocystis* sp. PCC6803 e a *Tolypothrix* sp. CACIAM 22. Em seguida, foi confrontado contra os mesmos sete genomas o modelo TIGR03798, disponível no banco de dados do TIGRFAM (HAFT, SELENGUT e WHITE, 2003).

4. RESULTADOS E DISCURSSÃO

Ao compararmos de forma quantitativa as sequências que os diferentes modelos foram capazes de detectar, como mostrados na Figura 4.1, pode-se notar que em alguns casos, o perfil desenvolvido a partir do modelo treinado neste estudo obteve uma sensibilidade maior na busca por sequências *nif11*, em relação ao modelo disponível no banco de dados do TIGRFAM. É importante esclarecer que os dados da Figura 4.1 não levam em consideração as identificações obtidas através do *threshold*.

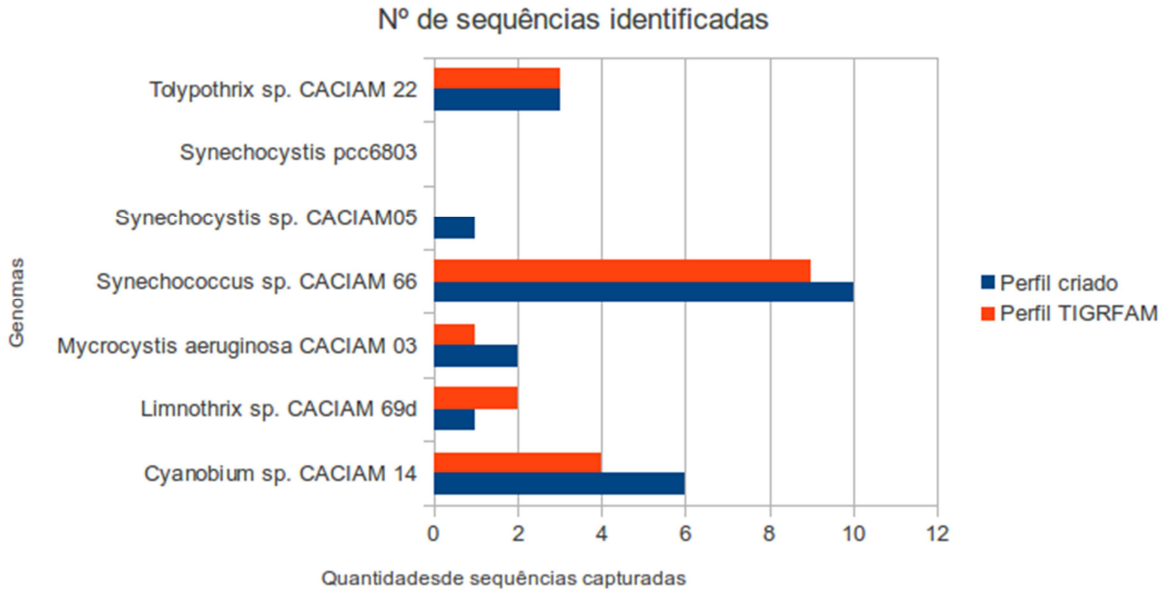

Figura 4.1 *Comparação quantitativa de sequências do domínio nif11 identificadas pelos modelos.*

Nota-se que para o genoma da *Synechocystis PCC* 6803, não houve identificação por nenhum modelo, isso corrobora com trabalhos atuais sobre a cianobactéria em questão, como os estudos de (BERMAN-FRANK, LUNDGREN e FALKOWSKI, 2003) e (WANG, FEWER e SIVONEN, 2011), onde mostraram que a mesma não apresenta indícios do domínio *nif11*. Cada análise comparativa será discutida mais à frente.

Figura 4.2 *Representação do alinhamento e conservação de colunas, obtida através do software Geneious 10.1.3.*

Na figura 4.2, podemos ver como ficaram as sequências *nif11* após os alinhamentos, os gap[6]s foram ignorados para melhorar a visualização. Percebemos de forma definida, pelas diferentes cores dadas para representar diferentes aminoácidos, algumas das regiões altamente conservadas. A incidência de conservação se inicia no aminoácido 115, representado pelo logo "M", em roxo, a cadeia conservada se estende até o aminoácido 604, representado pela logo "L", em alaranjado, a partir deste aminoácido a região variável do alinhamento se inicia. Um importante ponto a ser citado é que o motivo de dupla glicina (GG), característico nos genes *nif11* e presente nos estudos de (HAFT, BASU e MITCHELL, 2010) e Letzel (2014), não foi identificado no neste trabalho

Na figura 4.3, temos a sequência consenso que obtivemos a partir do alinhamento das sequências de cianobactérias. A sequência consenso nos mostra, nesse caso, o aminoácido que aparece com maior frequência em cada posição do alinhamento. Uma outra forma de visualização está apresentada no Apêndice A, no qual temos um recorte da região melhor conservada da sequência consenso em forma de *Sequence logo*. Na *Sequence logo*, as letras coloridas representam os aminoácidos que dominam aquela posição e quanto maior for a representação, maior a conservação do aminoácido para determinada coluna.

Com o perfil HMM modelado a partir do alinhamento, foi verificada uma diferença entre tamanho do modelo desenvolvido e do modelo TIGR03798, com o modelo desenvolvido neste trabalho tendo um comprimento de 74, enquanto o modelo TIGR03798 possui comprimento igual a 68. Tais valores refletem o tamanho da sequência consenso em cada perfil, representando a capacidade de cada modelo em detectar suas sequências alvo.

O incremento, atingido é resultado direto da qualidade das amostras utilizadas, sobretudo as adquiridas em nossa região, visto que a variabilidade dos genes *nif11* conseguiu ser mapeada com eficácia, gerando um modelo mais abrangente.

[6] Sequênciade espaços ou apenas um nó no alinhamento.

Figura 4.3 *HMM logo feita a partir da sequência consenso, obtida através do software Geneious 10.1.3.*

Nas tabelas apresentadas a seguir, estão expostos dados provenientes da comparação feita através do comando hmmsearch, os dados do perfil criado para este trabalho foram posicionados à esquerda e os do perfil disponibilizado pelo TIGRFAM à direita. O *score* é basicamente a pontuação de Match da sequência encontrada, sendo assim a representação bruta se a sequência foi pareada com qualidade ou não. Já o *e-value* representa a significância estatística da sequência e, ao contrário do *score*, quanto menor for seu valor, mais significativo é. Tem-se também o *threshold*, que representam sequências identificadas abaixo do limite de corte, mas com proximidade o suficiente para serem relevantes. Tendo esses conceitos em mente, apresentamos a seguir os resultados das comparações feitas.

A primeira comparação feita entre perfis foi contra a cianobactéria *Cyanobium* sp. CACIAM 14. Na Tabela 4.1, observamos que o perfil desenvolvido obteve um acréscimo em sua sensibilidade, identificando um total de seis sequências, possuindo duas a mais que as encontradas pelo perfil TIGR03798. Nota-se também que nas sequências que foram encontradas por ambos os perfis, o *score* e o *e-value* possuem, em sua maioria, melhores valores para as sequências descobertas pelo perfil apresentado neste trabalho. Vale também ressaltar, que além de possuir maior número de sequências encontradas, o perfil incluiu uma sequência no *threshold*.

Tabela 4.1 *Perfis criados a partir do modelo treinado e do modelo TIGR03798 para a cianobactéria Cyanobium sp. CACIAM 14.*

Cyanobium sp. CACIAM 14				
Perfil Treinado		Perfil TIGR03798		
E-Value	Score	E-Value	Score	
5.9^{-16}	56.5	3.8^{-15}	53.4	Contig00012
2.7^{-11}	41.6	1.3^{-14}	51.7	Contig00012
7.3^{-8}	30.6	7.6^{-6}	23.6	Contig00090
7.3^{-6}	24.2	3^{-5}	21.7	Contig00135
0.00063	18.0			Contig00067
0.0062	14.8			Contig00059
Incluídos no threshold				
0.13	10.6			Contig00379

Em seguida, a Tabela 4.2 mostra os perfis criados para a *Limnothrix* sp. CACIAM 69d. Para tal organismo, o perfil criado a partir do modelo treinado teve menor sucesso ao identificar sequências, possuindo apenas uma sequência encontrada, uma a menos que as

encontradas pelo modelo do TIGR03798. Apesar de uma quantidade menor de sequências, o valor do *e-value* é relativamente melhor, assim como o *score*, que é melhor cotado comparando a mesma sequência encontrada pelo outro modelo. O modelo desenvolvido foi capaz de incluir no *threshold* duas sequências não encontradas pelo modelo do TIGR03798.

Tabela 4.2 *Perfis criados a partir do modelo treinado e do modelo TIGR03798 para a cianobactéria Limnothrix sp. CACIAM 69d.*

Limnothrix sp. CACIAM 69d				
Perfil Treinado		Perfil TIGR03798		
E-Value	Score	E-Value	Score	
1.4^{-11}	43.0	6.5^{-9}	33.9	Contig00026
		0.0015	16.8	Contig00003
Incluídos no threshold				
0.018	13.8			Contig00118
0.021	13.6			Contig00019

A Tabela 4.3 mostra o comparativo dos perfis criados para a *Mycrocystis* sp. CACIAM 03 e invertendo o caso da *Limnothrix* sp. CACIAM 69d, o perfil apresentado se mostra mais eficaz encontrando mais sequências com valores melhores que as encontradas pelo modelo TIGR03798, porém com menos inclusões no *threshold*. O mesmo pode ser percebido na Tabela 4.4, cujo genoma confrontado pertence à *Mycrocystis aeruginosa* CACIAM 03, onde os perfis mostraram-se equivalentes, no entanto, com o perfil treinado identificando uma sequência a mais e tendo valores de *e-value* e *score* melhores. O genoma em questão obteve a maior quantidade de sequências identificadas, em comparação com os demais aqui utilizados.

Tabela 4.3 *Perfis criados a partir do modelo treinado e do modelo TIGR03798 para a cianobactéria Mycrocystis Aeruginosa CACIAM 03.*

Mycrocystis Aeruginosa CACIAM 03				
Perfil Treinado		Perfil TIGR03798		
E-Value	Score	E-Value	Score	
5.6^{-13}	47.6	1.7^{-8}	32.7	MCIH01000184
0.0043	15.9			MCIH01000084
Incluídos no threshold				
2.6	13.8	0.017	13.4	MCIH0100075 / MCIH01000187
		0.02	13.2	MCIH0100075

Tabela 4.4 *Perfis criados a partir do modelo treinado e do modelo TIGR03798 para a cianobactéria Synechococcus sp. CACIAM 66.*

Synechococcus sp. CACIAM 66				
Perfil Treinado		Perfil TIGR03798		
E-Value	Score	E-Value	Score	
1.1^{-14}	56.	1.6^{-14}	51.4	Scaffold00041
5.4^{-14}	50.2	1.9^{-12}	44.7	Scaffold00035
9.8^{-12}	42.9	1.2^{-9}	35.7	Scaffold00010
1.6^{-10}	39.1	2.3^{-8}	31.6	Scaffold00044 / Scaffold00027
2.6^{-9}	35.2	6.4^{-7}	27.0	Scaffold00019 / Scaffold00018
8.5^{-9}	33.5	1.7^{-6}	25.6	Scaffold00018 / Scaffold00044
1.9^{-8}	32.4	3.7^{-5}	21.3	Scaffold00027 / Scaffold00019
7.2^{-8}	30.6	0.00024	18.7	Scaffold00035 / Scaffold00035
0.0017	16.5	0.004	14.8	Scaffold00005 / Scaffold00034
0.0057	14.9			Scaffold00016
Incluídos no threshold				
0.012	13.8	0.014	13.0	Scaffold00034 / Scaffold00005
0.067	11.4	0.03	12.0	Scaffold00045 / Scaffold00016
		0.071	10.8	Scaffold00010

Nos testes aplicados à *Synechococcus* sp. CACIAM 05 (Tabela 4.5), ambos os perfis apresentaram dificuldades na identificação de sequências, o perfil desenvolvido identificou *nif11* em um *scaffold* com baixa qualidade tanto no *e-value* quanto no *score,* enquanto o perfil TIGR03798, apenas uma sequência no limiar de corte.

Tabela 4.5 *Perfis criados a partir do modelo treinado e do modelo TIGR03798 para a cianobactéria Synechocystis sp. CACIAM 05.*

Synechocystis sp. CACIAM 05				
Perfil Treinado		Perfil TIGR03798		
E-Value	Score	E-Value	Score	
0.0081	14.6			Scaffold v.6 2-isoproylmalate
Incluídos no threshold				
		0.06	11.3	Scaffold v.6 hypothetical p.

Confrontados com a cianobactéria *Tolypothrix* sp. CACIAM 22, apresentada na Tabela 4.6, há uma similaridade na qualidade de captura para esse organismo. O modelo TIGR03798 localizou três sequências, a mesma quantidade do perfil desenvolvido, além de ter identificado três sequências incluídas no *threshold*.

Tabela 4.6 *Perfis criados a partir do modelo treinado e do modelo TIGR03798 para a cianobactéria Tolypothrix sp. CACIAM 22.*

Tolypothrix sp. CACIAM 22				
Perfil Treinado		Perfil TIGR03798		
E-Value	Score	E-Value	Score	
1.7^{-17}	62.7	8.9^{-17}	59.8	Scaffold00422
3.2^{-8}	32.9	8^{-6}	24.7	Scaffold03150
0.00018	21.0	0.01	14.8	Scaffold01704
Incluídos no threshold				
0.027	14.0			Scaffold01097
0.086	12.4			Scaffold01121
0.94	9.0			Scaffold00355

5. CONCLUSÃO

Os dados obtidos no estudo, juntamente com a análise comparativa entre o modelo desenvolvido e o TIGR03798, mostraram-se promissores, visto melhora para predição e identificação das variações de aminoácidos característicos das sequências de *nif11* de clusters de genes de bactericidas em genomas de cianobactérias amazônicas.

Os estudos mais aprofundados dos mecanismos de funcionamento dos algoritmos utilizados nos Modelos Ocultos de Markov nos dão a possibilidade de treinar e ajustar modelos desenvolvidos para abranger diversas situações, o que contribuiu de forma significativa para os resultados deste trabalho. Assim, este presente estudo apresentou:

- O treinamento e construção de um modelo HMM para identificar sequências de genes *nif11* em cianobactérias da região amazônica;
- A comparação com o modelo TIGR03798, onde foi verificado uma diferença entre os modelos, no comprimento de suas sequências consenso e na sensibilidade para identificação de genes *nif11*;

Claramente, os resultados deste trabalho indicam a necessidade de estudos complementares sobre as cianobactérias da região amazônica, visto que o sítio de clivagem característico dos genes *nif11* não foi identificado nas sequências de cianobactérias analisadas, por tanto, a necessidade de novos estudos que visem a modelagem de novos perfis para melhor caracterizar este motivo com ausência de dupla glicina.

Também, a continuidade dessa pesquisa com a inserção de mais amostras de cianobactérias da nossa região tendência um aumento na sensibilidade do modelo, podendo assim obter resultados mais expressivos na identificação de genes do domínio *nif11*.

REFERÊNCIAS

AERTS, R. J. et al. Allelopathic inhibition of seed germination by Cinchona alkaloids? Phytochemistry, v. 30, p. 2947-2951, 1991. ISSN 9.

AQUINO, S. B. F. Estratégias de Otimização em GPU para Análise de Sequências Biológicas (Master's thesis), 2012.

ATUNCAR, G. S. **Conceitos básicos de processos estocásticos**. Universidade Federal de Minas Gerais. Belo Horizonte. 2011.

BAKER, J. K. **Stochastic Modeling as a Means of Automatic Speech Recognition.** Departamento de Ciência da Computação Carriegie-Mellon University. Pittsburgh. 1975. (Tese (Doutorado)).

BAUM, L. E. An Inequality and Associated Maximisation Technique in Statistical Estimation for Probabilistic Functions of a Markov Process. **Inequalities**, Los ANgeles, v. 3, p. 1-8, 1972.

BAUM, L. E.; EAGON, J. A. An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology. **Bull. Amer. Math. Soc.**, v. 73, p. 369-363, Maio 1967. ISSN 3.

BERMAN-FRANK, I.; LUNDGREN, P.; FALKOWSKI, P. Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria.. **Research in microbiology**, v. 154, p. 157-164, 2003. ISSN 3.

BLANK, C. E.; SÁNCHEZ-BARACALDO, P. Timing of morphological and ecological innovations in the cyanobacteria - a key to understanding the rise in atmospheric oxygen. **Geobiology**, v. 8, p. 1-23, Janeiro 2010. ISSN 1.

CASTRO, W. O. et al. Draft genome sequence of Microcystis aeruginosa CACIAM 03, a cyanobacterium isolated from an Amazonian freshwater environm, 2016.

CHAI, Y. H. **Characterization of Nitrogen Fixation (nif) Genes From Paenibacillus polymyxa.** Universiti Sains Malaysia. [S.l.], p. 95. 2007. (Dissertação (Mestrado)).

COTTER, C. A. et al. Preparation of cell cultures and vaccinia virus stocks. **Current protocols in microbiology**, 2015.

COTTER, P. D.; HILL, C.; ROSS, R. P. Bacteriocins: developing innate immunity for food. **Nature Reviews Microbiology**, v. 3, p. 777-788, 2005. ISSN 10.

CROTEAU, R.; KUTCHAN, T. M.; LEWIS, N. G. (Eds.) Biochemistry & Molecular Biology of Plants. Rockville: American Society of Plant Physiologists. **Natural Products (Secondary Metabolites). In: Buchanan B., Gruissem W., Jones R.**, p. 1250-1318, 2000.

DA ANUNCIAÇÃO GOMES, A. M. et al. FLORAÇÕES DE CIANOBACTÉRIAS TÓXICAS EM UMA LAGOA COSTEIRA HIPEREUTRÓFICA DO RIO DE JANEIRO/RJ (BRASIL) E SUAS CONSEQUÊNCIAS PARA SAÚDE HUMANA. **Oecologia Australis**, v. 13, p. 322-345, 2017. ISSN 2.

DEMAIN, A. L.; FANG, A. The natural functions of secondary metabolites. Advances in Biochemical Engineering/Biotechnology, v. 69, p. 1-39, 2000.

DOS SANTOS, H. R. M. **Diversidade de Bactérias em Nódulos de Inga vera Willd. (LEGUMINOSAE-MIMOSOIDEAE) do Sul da Bahia.**. Universidade Estadual de Santa Cruz. Ilhéus, p. 69. 2010. (Dissertação (Mestrado)).

DURBIN, R. E. A. **Biological sequence analysis: probabilistic models of proteins and nucleic acids**. Cambridge university press. [S.l.]. 1998.

EDDY, S. R. Profile hidden Markov models. Bioinformatics, v. 14, p. 775-763, Julho 1998. ISSN 9.

ESPINDOLA, L. D. S. **Um estudo sobre modelos ocultos de markov hmm-hidden markov model**. [S.l.]. 2010.

FARINA, R. **Diversidade de bactérias promotoras do crescimento vegetal associadas à cultura de canola (Brassica napus L.) cultivada no município de Vacaria.** Instituto de Biociências, Universidade Federal do Rio grande do sul. Porto ALegre , p. 104. 2012. (Tese (Doutorado)).

FILHO, F. C. P. F. C. G. Bioinformática: Manual do Usuário. **Biotecnologia Ciência e Desenvolvimento**, Brasília, v. 5, p. 12-25, Janeiro 2002. ISSN 29. Disponivel em: <http://www.biotecnologia.com.br/revista/bio29/bioinf.pdf>. Acesso em: 11 Fevereiro 2017.

FINK, G. A. Markov Model for Pattern Recognition – From Theory to Applications. **Springer**, 2008.

FONSECA, B. M. . E. A. Biovolume de cianobactérias e algas de reservatórios tropicais do Brasil com diferentes estados tróficos. **Hoehnea**, v. 41, p. 9-30, Março 2014. ISSN 1.

GRACE, S. C.; LOGAN, B. A. Energy dissipation and radical scavenging by the plant phenylpropanoid pathway. **Philosophical Transactions of The Royal Society B** , v. 355, p. 1499-1510, Outubro 2000. ISSN 1402.

HAFT, D. H.; BASU, M. K.; MITCHELL, D. A. Expansion of ribosomally produced natural products: a nitrile hydratase- and Nif11-related precursor family. **BMC Biology**, v. 8, Maio 2010. ISSN 70.

HAFT, D. H.; SELENGUT, J. D.; WHITE, O. The TIGRFAMs database of protein families. Nucleic acids research, v. 31, p. 371-373, 2003. ISSN 1.

HAGEN, J. B. The origins of bioinformatics. **Nature Reviews Genetics**, v. 1, p. 231-236, Dezembro 2002. ISSN 3.

HESPER, B.; HOGEWEG, P. Bioinformatica: een werkconcept. Kameleon 1. **Leiden: Leidse Biologen Club**, Dutch, n. 6, p. 28-29, 1970.

JACK, R. W.; TAGG, J. R.; RAY, B. Bacteriocins of gram-positive bacteria. **Microbiological reviews**, v. 59, p. 171-200, 1995. ISSN 2.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular Biology and evolution**, n. 30, p. 772-780, 2013.

KOSSEL, A. Über die chemische zusammensetzung der zelle. Archiv für Physiologie, p. 181-186, 1897.

KROGH, G. V. Care in knowledge creation. **California management review University of California Press Journals**, California, n. 3, p. 133-153, 1998.

LESK, A. M. **Introdução à Bioinformática**. 2ª. ed. Porto Alegre: Artmed, 2008.

LIMA, A. R. J. et al. Draft genome sequence of the Brazilian Cyanobium sp. strain CACIAM 14. **Genome announcements**, n. 2(4), 2014. ISSN e00669-14.

LORENZI, A. S. **Abordagens moleculares para detectar cianobactérias e seus genótipos produtores de microcistinas presentes nas represas Billings e Guarapiranga**. Doctoral dissertation, Centro de Energia Nuclear na Agricultura, Universidade de São Paulo. São Paulo. 2004.

MOLICA, R.; AZEVEDO, S. Ecofisiologia de cianobactérias produtoras de cianotoxinas. **Oecologia Australis**, v. 13, p. 229-246, Junho 2009. ISSN 2.

MONDARDO, R. I. Influência da pré-oxidação na tratabilidade das águas via filtração direta descendente em manancial com elevadas concentrações de microalgas e cianobactérias., 2004.

MOREIRA, L. M. **Ciências Genômicas:** Fundamentos e Aplicações. Ribeirão Preto, Brasil: Sociedade Brasileira de Genética, 2015.

NCBI. **NCBI - National Center for Biotechnology Information.** Disponivel em: <https://www.ncbi.nlm.nih.gov/>. Acesso em: 29 Desembro 2016.

OGATA, S. O. E. A. **Alinhamento de seqüências biológicas com o uso de algoritmos genéticos**. [S.l.]. 2005.

PATTANAIK, B.; LINDBERG, P. Terpenoids and their biosynthesis in cyanobacteria. **Life**, n. 5, p. 269-293, Janeiro 2015.

PENN, K. et al. Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. **The ISME Journal**, v. 8, p. 1866-1878, Março 2014.

PORITZ, A. B. Hidden Markov Models: a Guided Tour. **In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)**, p. 7-13, 1988.

PROSDOCIMI, F. et al. Bioinformática: manual do usuário. **Biotecnologia Ciência & Desenvolvimento**, v. 29, p. 12-25, 2002.

RABINER, L. R. . E. A. HMM clustering for connected word recognition. Acoustics, Speech, and Signal Processing. **1989 International Conference on. IEEE ICASSP-89**, 1989.

RAYMOND, J.; BLANKENSHIP, R. The origin of the oxygen-evolving complex. **Coord Chem Rev 252**, p. 377–383, 2008.

REICHL, L. E. **A Modern Course in Statistical Physics.** 2ª. ed. Nova York: John Wiley-Interscience, 1998.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors.. **Proceedings of the National Academy of Sciences**, v. 74, p. 5263-5467, Dezembro 1977. ISSN 12.

SCHREINER, M. et al. UV-B Induced Secondary Plant Metabolites. Optik & Photonik, v. 9, p. 34-37, Maio 2014. ISSN 2.

SCHUSTER-BÖCKLER, B. . S. J. . &. R. S. HMM Logos for visualization of protein families. **BMC bioinformatics**, 2004.

SINGH, R. K. et al. Cyanobacteria: an emerging source for drug discovery. **The Journal of Antibiotics**, v. 64, p. 401-412, Abril 2011.

SINGH, S.; KATE, B. N.; BANERJEE, U. C. Bioactive Compounds from Cyanobacteria and Microalgae: An Overview. Critical Reviews in Biotechnology, v. 25, p. 73-95, 2005. ISSN 3.

STENICO, M. E. E. A. INIBIDORES DE PROTEASES PRODUZIDOS POR CIANOBACTÉRIAS: UMA REVISÃO. **Oecologia Australis 16.2**, p. 183-209, 2017.

TAN, L. T. Bioactive natural products from marine cyanobacteria for drug discovery. Phytochemistry, v. 68, p. 954-979, Abril 2007. ISSN 7.

VELHO, R. V. Identificação de genes de bacteriocinas produzidas por diferentes linhagens de Bacillus isolados da região amazônica, 2010.

VINING, L. C. Functions of secondary metabolites. Annual Review of Microbiology, v. 44, p. 395-427, Outubro 1990.

WANG, H.; FEWER, D. P.; SIVONEN, K. Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. **PloS one**, v. 6, p. e22384, 2011. ISSN 7.

WHEELER, T. J.; CLEMENTS, J.; FINN, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. **BMC Bioinformatics**, v. 15, Janeiro 2014. ISSN 7.

WHITTON, B. A.; POTTS, M. **The ecology of cyanobacteria:** their diversity in time and space. 1ª. ed. Nova Iorque: Kluwer Academic Publishers, 2002.

WHITTON, B. A.; POTTS, M. **The ecology of cyanobacteria:** their diversity in time and space. 1ª. ed. Nova Iorque: Kluwer Academic Publishers, 2002.

XIONG, J.; BAUER, C. E. A cytochrome b origin of photosynthetic reaction centers: an evolutionary link between respiration and photosynthesis. **Journal of molecular biology 322.5** , p. 1025-1037, 2002.

APÊNDICE A – RECORTE DA REGIÃO MELHOR CONSERVADA DA SEQUÊNCIA CONSENSO EM FORMA DE SEQUENCE LOGO.