UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

FACULDADE DE COMPUTAÇÃO

ARIANE ELIZABETH NUNES GARCIA

MONTAGEM DOS GENOMAS DA CULTURA NÃO-AXÊNICA DE *NOSTOC* SP. CACIAM 19 UTILIZANDO TÉCNICAS DE METAGENÔMICA

Belém - PA

2017

Ariane Elizabeth Nunes Garcia

Montagem dos genomas da cultura não-axênica de *Nostoc* sp. CACIAM 19 utilizando técnicas de metagenômica

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.

Orientador: Profª. Drª. Regiane Silva Kawasaki Francês.

Co-orientador: Msc. Alex Ranieri Jerônimo Lima

Belém - PA

2017

Ariane Elizabeth Nunes Garcia

Montagem dos genomas da cultura não-axênica de *Nostoc* sp. CACIAM 19 utilizando técnicas de metagenômica

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.

Data da defesa: Belém-PA, 24 de Abril de 2017

Conceito: Excelente

_____

Prof.ª Drª Regiane Silva Kawasaki Francês – Orientador
Faculdade de Computação – ICEN/UFPA

_____

Msc. Alex Ranieri Jerônimo Lima – Co-orientador
Programa de pós-graduação em genética e biologia molecular- ICB/UFPA

_____

Profª Drª Danielle Costa Carrara Couto– Membro
Faculdade de Computação- ICEN/UFPA

_____

Bel. Renato Renison Moreira Oliveira- Membro
Faculdade de Computação- ICEN/UFPA

Dedico este trabalho aos meus pais, Valdemir Castro Garcia e Cleide Maria Lima Nunes, que estiveram ao meu lado me incentivando me ajudando e, principalmente, me ensinando a lutar por meus objetivos. Dedico também a minha filha, Maria Eduarda Nunes Garcia Carneiro, que foi a minha maior motivadora a não desistir dos meus objetivos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, Pai, Filho e Espirito Santo, por ter me guiado, me iluminado em todo meu caminho!

Agradeço aos meus pais, Valdemir Garcia e Cleide Nunes, por tudo que fizeram por mim ao longo da vida, por estarem ao meu lado em todos os momentos, por todo amor e carinho que sempre me deram e por não desistirem de mim até mesmo quando eu já havia desistido. Vocês são meus tudo. Agradeço também ao meu irmão (meu gigante), Wallace Gabriel, por todo amor, carinho e palavra de conforto nos meus momentos difíceis, por ser esse irmão-amigo que está comigo em todos meus momentos. Agradeço em especial à minha filha, Maria Eduarda, por ser minha fortaleza, por todo amor que me entrega todos os dias, por ser minha motivação de buscar algo melhor e por encher minha vida de luz quando eu estou desanimada. Vocês são minha base, dedico este trabalho a vocês.

Agradeço aos meus familiares, por toda força e compreensão durante estes anos na graduação, por me desculparem nas faltas das comemorações familiares, por toda ajuda (seja ela qual for) que me deram e por todas as orações destinas a mim. Meu muito obrigada.

Agradeço ao meu namorado, Ivo Reis, uma pessoa especial que apesar de pouco tempo, veio acrescentar muito à minha vida. Agradeço pelas palavras, conselhos, esculhambações, sorrisos, pela atenção que tem para comigo e principalmente por me aguentar nos meus momentos de estresse. Obrigada meu amor.

Agradeço aos meus amigos, por toda força que me deram, por toda palavra de apoio nos momentos difíceis, pelos risos, pelas brigas, pelas conversas e principalmente por estarem comigo me aguentando do jeito que eu sou. Obrigada a todos.

Agradeço aos professores da instituição, em especial a minha Orientadora professora Regiane Kawasaki, por todo apoio ao desenvolvimento deste trabalho, por ter depositado sua confiança em mim para a realização do mesmo e por toda dedicação e atenção que disponibilizou para mim. Obrigada Professora.

Agradeço ao meu co-orientador, Alex Lima, que juntamente com a professora Regiane Kawasaki me auxiliaram no desenvolvimento deste trabalho com dedicação e atenção, sempre se mantendo a disposição de me ajudar. Muito obrigada.

Agradeço à todos os meus amigos da graduação, em especial Alcyr Almeida, Alessandra Araujo, Alexandre Freitas, Arthur Moraes, Bruno Akira, Bleno Vale, Dan Jhonatan, Daniel Henrique, Eric Pereira, Géssica Pinheiro, Moisés Felipe, Paulo Souza, Ronald Souza, que ao longo desses anos enfrentam a luta do cotidiano da graduação com muitas conversas, risos e companheirismo. *Zoeira Never Ends*. Obrigada a todos vocês.

E por último, mas não menos importantes agradeço a todos que contribuíram direta ou indiretamente para a conclusão deste trabalho.

Obrigada a todos!

“ A persistência é o caminho do êxito” (Charles Chaplin)

RESUMO

A bioinformática consiste no estudo computacional que pode obter, organizar e analisar informações biológicas a partir do conhecimento de sequências de biomoléculas. É uma nova ciência que tem raízes em ciências da computação, estatísticas e na biologia molecular. Foi desenvolvida para organizar os resultados obtidos no sequenciamento de genes, que cada vez mais se aprimoram e produzem quantidades cada vez maior de dados sobre sequências. Com o avanço tecnológico de plataformas de sequenciamento genômico, torna-se possível aumentar significativamente o processamento de milhares de bases do DNA em uma única execução, diminuindo os custos e acelerando a velocidade de sequenciamento. Este avanço tecnológico possibilita também obter conhecimento a partir de dados biológicos retirados direto de uma comunidade microbiana, caracterizando os estudos de metagenômica. As técnicas metagenômicas consistem na extração do DNA, sequenciamento, montagem, separação e classificação dos genomas. Junto com avanço tecnológico dos sequenciadores surgiram novas abordagens de montagem que sejam capazes de processar uma grande escala de dados com alta cobertura. Neste trabalho, objetivou-se em montar os genomas presente na amostra de cultura não-axênica de cianobactéria *Nostoc* sp. CACIAM 19, utilizando abordagens metagenômicas. Como resultados, foram separados e classificados genomas de seis gêneros na amostra, com a presença do genoma da cianobactéria em menor abundância.

Palavras- chave: Bioinformática, sequenciamento, metagenômica, montagem.

ABSTRACT

Bioinformatics consists of the computational study that can obtain, organize and analyze biological information from the knowledge of sequences of biomolecules. It is a new science that has roots in computer science, statistics and molecular biology. It was developed to organize the results obtained in gene sequencing, which are increasingly improving and producing increasing amounts of sequence data. With the technological advancement of genomic sequencing platforms, it becomes possible to significantly increase the processing of thousands of DNA bases in a single run, reducing costs and speeding up the sequencing speed. This technological advance also makes it possible to obtain knowledge from biological data taken directly from a microbial community, characterizing metagenomic studies. Metagenomic techniques consist of DNA extraction, sequencing, assembling, separation and classification of genomes. Along with technological advancement of the sequencers have come new assembly approaches that are capable of processing a large scale of data with high coverage. In this work, we aimed to assemble the genomes present in the non-axenic cyanobacteria culture sample Nostoc sp. CACIAM 19, using metagenomic approaches. As results, genomes of six genera were separated and classified in the sample, with the presence of the cyanobacteria genome in smaller abundance.

Keywords: Bioinformatics, sequencing, metagenomics, assembly.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

BLAST	*Basic Local Alignment Search Tool*
CIT	Centro de Inovações Tecnológicas
DNA	Ácido desoxirribonucleico
IEC	Instituto Evandro Chagas
GC	Guanina – Citosina
LABIOCAD	Laboratório de Bioinformática e Computação de Alto Desempenho
LTB	Laboratório de Tecnologia Biomolecular
NGS	*Next Generation Sequencing*
NCBI	*National Center for Biotechnology Information*
OLC	*Overlap-layout-consensus*

SUMÁRIO

1. INTRODUÇÃO

Com o aprimoramento das tecnologias computacionais no século XX, a bioinformática surge com o desenvolvimento de novas abordagens capazes de criar análises de dados biológicos. A princípio, a bioinformática parecia algo contraditório pois não parecia possível empregar ferramentas tecnológicas em meio a dados biológicos. Entretanto, nas últimas décadas, a informática tornou-se essencial para os estudos na área biológica (ARAÚJO et al., 2008). Em termos gerais, pode-se definir a bioinformática como um estudo da aplicação de técnicas computacionais para análise de dados biológicos. Sendo uma união de diversas linhas de conhecimento - ciência da computação, matemática, estatística e a biologia molecular- a bioinformática tem como finalidade o desenvolvimento de programas computacionais que são capazes de reconhecer, desvendar e obter uma grande quantidade de dados genéticos e proteicos. (PROSDOCIMI, 2007).

Seu surgimento teve início especialmente durante a execução dos projetos genomas que geraram grandes informações a partir do sequenciamento de DNA (ARAÚJO et al., 2008). Devido ao aumento na obtenção de sequências genéticas, necessitou-se de algoritmos computacionais eficientes que fossem capazes de compartilhar, analisar e armazenar estas sequências (FARIAS, CHACON & SILVA, 2011). Neste período, surgiram as primeiras plataformas de sequenciamento de DNA que geravam uma grande quantidade de sequências com o custo alto e velocidade relativamente baixa, que posteriormente foram substituídas por plataformas de nova geração capazes de sequenciar milhares de pares de bases e um único processamento.

O avanço tecnológico em plataformas de obtenção de dados em massa facilitou o processo de sequenciamento genômico, pois o surgimento de sequenciadores de nova geração possibilitou o processamento e a obtenção de milhares de bases em um único processamento, diminuindo os custos (VARUZZA,2013). Com esta redução de custos e aumento da velocidade na obtenção de dados tornou-se possível obter dados biológicos retirados direto de comunidades microbianas em seus habitats, adquirindo conhecimento genético de diversas espécies em uma única amostra. Este

tipo de estudo denomina-se metagenômica (KUNIN et al,, 2008), que é um conjunto de técnicas que busca compreender a biologia em nível de comunidade (FALEIRO, ANDRADE, & REIS JUNIOR, 2011). Metagenômica é uma ferramenta útil para a obtenção de conhecimento de amostras ambientais onde se é possível obter a caracterização e a diversidade da amostra genética (PEREIRA**,** 2014).

Em virtude deste avanço tecnológico nas plataformas de sequenciamento, surgiu a necessidade de desenvolvimento de novos montadores de genoma capazes de manipular grandes volumes de dados (MILLER, KOREN & SUTTON, 2010). A montagem de genoma é um processo computacional que reconstrói a sequência original utilizando algoritmos. Esta montagem acontece a partir da junção das leituras resultantes do sequenciamento. Porém estas leituras devem ser pré-processadas, excluindo as sequências com baixa qualidade. Contudo, tratando-se de um metagenoma a montagem não terá apenas uma sequência original, mas diversas e, com isto, faz-se necessário a classificação dos genomas presentes na amostra (PEIXOTO, 2013). Classificar um genoma é encontrar semelhança entre as sequências presentes na amostra, tomando como base a sequência de referência. Com a classificação do genoma é possível obter o detalhamento da diversidade biológica que compõe a amostra (NETO, 2012).

Este trabalho utilizou de técnicas metagenômicas para análise dos genomas presentes na amostra de cianobactéria *Nostoc* sp. CACIAM 19, em virtude de que as cianobactérias têm a facilidade de formar relações mutualísticas com outro organismo dificulta a obtenção de culturas axênicas (RIPPKA, 1988). Realizou-se o pré-processamento, corte e filtragem, das leituras resultantes do sequenciamento de cianobactéria concedido pelo Laboratório de Tecnologia Biomolecular -LTB da Universidade Federal do Pará, verificando a qualidade das sequências e das bases presente no sequenciamento. Após o pré-processamento, que se faz necessário para uma montagem de qualidade, foi realizada a montagem de genoma utilizando duas abordagens diferentes: OLC e grafo de Bruijn. Após a montagem dos genomas, foi feita a separação dos genomas presentes nas sequências montadas, utilizando um algoritmo de *binning*. Por fim, foi realizado a classificação taxonômica dos genomas presentes na amostra.

1.1 JUSTIFICATIVA

O avanço revolucionário de equipamentos tecnológicos utilizados para obtenção de dados em grande escala está crescendo cada vez mais. A biologia é um grande exemplo desse avanço tecnológico, pois, os dados biológicos passaram a ser produzido rapidamente e com um grande volume. Logo, a área de bioinformática está envolvida neste contexto, uma vez que esta utiliza meios computacionais para obtenção e análise dos dados biológicos. Pode-se citar a bioinformática como um sistema de informação para a biologia molecular, que possui diversas aplicações práticas (LUSCOMBE; GREENBAUM & GERSTEIN, 2001).

Pós anos 90, os sequenciadores de alto desempenho, denominados também como sequenciadores de nova geração, foram desenvolvidos tornando possível a obtenção de uma grande escala de dados em um único sequenciamento. Com isso deu origem a uma nova geração de algoritmos de montagem que visam reconstruir a sequência original de maneira mais precisa e obtendo um bom desempenho (MILLER, KOREN & SUTTON, 2010).

Contudo, os algoritmos de montagem de genoma têm suas peculiaridades e limitações, fazendo-se necessário realizar estudo de comparação e analise destes algoritmos. Este trabalho busca realizar testes destes algoritmos, utilizando uma amostra de cultura não-axênica de cianobactéria, comparando os resultados obtidos para assim determinar qual algoritmo obteve a montagem mais precisa. A partir dos resultados obtidos, busca-se definir qual montagem foi mais satisfatória para que posteriormente possa ser realizada a classificação taxonômica dos genomas presentes na amostra.

1.2 OBJETIVOS

Objetivo geral:

Montar e separar os genomas presentes na cultura não-axênica da cianobactéria *Nostoc sp.* CACIAM 19 utilizando técnicas metagenômicas.

Objetivos específicos:

- Realizar o tratamento de qualidade nas *reads* resultantes do sequenciamento, definindo os melhores parâmetros;
- Realizar montagem dos dados genômicos utilizando dois tipos de algoritmos;
- Concluir, a partir dos resultados obtidos, qual foi o melhor algoritmo.
- Realizar separação de genoma, utilizando o algoritmo de *binning*;
- Realizar classificação taxonômica dos genomas encontrados.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado em cinco seções:

- **1ª - INTRODUÇÃO**: Apresenta o trabalho, mostrando sua relevância para a área. Assim como os objetivos e justificativa que o cercam.
- **2ª - REFERÊNCIAL TEÓRICO**: Apresenta os conceitos importantes para a realização deste trabalho.
- **3ª - PROCESSOS DE MONTAGEM DE GENOMA**: Apresenta os métodos e materiais utilizados para a realização do trabalho.
- **4ª - RESULTADOS E DISCUSSÃO**: Apresenta a análise quantitativa dos resultados e sua discussão sobre o processo de montagem da cultura não-axênica de *Nostoc* sp. CACIAM 19.
- **5ª - CONSIDERAÇÕES FINAIS**: Apresenta considerações importantes sobre os resultados obtidos neste trabalho.

2. REFERÊNCIAL TEÓRICO

2.1 CIANOBACTÉRIAS.

Cianobactérias são organismos procarionte capazes de realizar fotossíntese oxigênica (REVIERS, 2006). Podem ser unicelulares, coloniais ou filamentosas (CAMPOS & DUARTE, 2011). As cianobactérias podem habitar diversos nichos ecológicos, pois apresentam habilidades de sobrevivência em locais com altas e baixas intensidades luminosas. Isto se dá devido a sua capacidade de produzir pigmentos que são necessários para absorção de luz em qualquer habitat, além da habilidade de armazenarem nutrientes essenciais em seu citoplasma. Portanto, em funções fisiológicas, morfológicas e ecológicas, as cianobactérias têm grande capacidade adaptativa em diversas condições luminosas e de variação quantitativa de nutrientes (MISCHKE, 2003).

Tendo um papel importante entre os microrganismos, principalmente como produtores primários de matéria orgânica e como fixadoras de nitrogênio, as cianobactérias têm grande importância para o meio ambiente e para saúde pública, tendo em vista que são capazes de produzir toxinas durante suas florações, dependendo da linhagem de cianobactéria (Sant'Anna et al., 2008). Nas cianobactérias existem conteúdos proteicos e de carboidratos que podem ser utilizados na suplementação alimentar humana e animal (CAMPOS & DUARTE, 2011). Contudo, ao se comparar com outros microrganismos com fontes de alimentos, como por exemplo as leveduras, existe uma falta de informação detalhada sobre seu cultivo e assim passaram a ser incluídas em programas de monitoramento de cultivo industrial, pois são capazes de produzir diversas substâncias bioativas que podem ser utilizados na indústria (GAULT & MARLER, 2009).

As cianobactérias sofreram alterações nos genomas que geraram genes que desempenham funções importantes à sobrevivência das mesmas em diferentes habitats, com diversas funções que atraem o interesse biotecnológico (LIMA, 2015). Com obtenção do conhecimento sobre os genomas de cianobactérias torna-se possível descobrir produtos naturais. No entanto, o campo de produção de produtos advindos do metabolismo de cianobactérias é pouco

explorado, manifestando assim a necessidade do seu cultivo em alta escala para a obtenção de seus bioativos que possuem uma grande utilidade para biotecnologias (RODRIGUES, 2016).

No entanto, obter a sequência genômica de cianobactérias é dificultoso, pois elas possuem a facilidade de manter relações mutualísticas com outros microrganismos (RIPPKA, 1988). Estas relações de mutualismo beneficiam o crescimento das cianobactérias, mas dificulta a obtenção de culturas puras e o processo de montagem de seus genomas. Este perfil heterogêneo da cultura abre espaço para o uso de técnicas metagenômicas visando o conhecimento do genoma das cianobactérias e da sua interação com outros organismos, verificando assim se a produção de compostos bioativos ocorre de forma conjunta com outros organismos ou de forma isolada (LIMA, 2015).

2.2 SEQUENCIAMENTO E A METAGENÔMICA

O sequenciamento de genoma é o primeiro passo para a obtenção da composição molecular dos organismos, auxiliando na comparação de genomas de diversos indivíduos, permitindo assim relacionar características e o melhor entendimento do mecanismo de evolução. Além do mais, permite um melhor conhecimento do genoma para realizar a manipulação destes com mais facilidade (HORODESKY, 2014).

O sequenciamento genômico teve início, primeiramente, com o desenvolvimento de um método por Frederich Sanger e Alan Coulson em meados da década de 70. Este método de Sanger se firma na utilização de didesoxinucleotídeos, partindo do princípio da terminação controlada da polimerização, possibilitando assim determinar a sequência de DNA. Contudo, inicialmente não era uma ferramenta adequada para sequenciar grandes sequências de genomas. O método de Sanger passou a ser automatizado e integrado a sistemas computadorizados no fim da década de 80 (VENTER,2001).

No início do novo século, houve uma corrida para o sequenciamento do genoma humano (DAVIES, 2001), o que desencadeou no surgimento de novas estratégias para tal finalidade. Após a publicação do sequenciamento do genoma humano em fevereiro de 2001 (VENTER, 2001), ocorreu um avanço nas técnicas de sequenciamento. Plataformas baseadas em tecnologias de sequenciamento de nova geração (*Next Generation Sequencing*- NGS)

começaram a ser comercializadas e passaram a substituir o método Sanger (HORODESKY, 2014). Com o avanço tecnológico nas últimas décadas e com o auxílio de ferramentas da bioinformática, os sequenciadores de nova geração permitiram o sequenciamento de milhares de pares de bases e de diversos microrganismos em um único processo e com alta velocidade, obtendo sequenciamento com mais eficiência e com maior cobertura (VARUZZA, 2013).

Além do baixo custo e o alto desempenho, os sequenciadores de nova geração se diferenciam dos sequenciadores Sanger, pois têm alto rendimento em operação paralela e por operar em com leituras curtas (MILLER, KOREN & SUTTON, 2010). Apesar do grande avanço na forma de analisar genomas, os sequenciadores de nova geração têm suas limitações. Estes são capazes de fornecer grandes quantidades de dados, porém, possuem taxas de erros elevadas. E também, as tecnologias NGS geralmente fornecem leituras menores quando comparadas ao sequenciamento Sanger (CARVALHO & SILVA, 2010). A grande problemática com o processamento de leituras curtas é que resultam na entrega de menos informações. Com o desenvolvimento de sequenciadores de leituras curtas, deu origem a uma nova geração de algoritmos de montagem. No entanto, para realizar uma montagem de leituras curtas é preciso ter uma alta cobertura que possa satisfazer as características mínimas de sobreposição (MILLER, KOREN & SUTTON, 2010).

Os resultados obtidos no sequenciamento de genomas são postos em formatos de arquivo de texto ou binário. Estes arquivos vêm demonstrando informações de maneira padronizada e estruturada sobre as sequências. Sequenciadores de nova geração produzem principalmente FASTA ou FASTQ como formato de arquivos (SILVA & KREMER, 2016). Estes resultados precisam ser pré-processados com o objetivo de remover adaptadores e sequências de baixa qualidade (SOUZA, 2015). Esse pré-processamento se faz necessário para eliminar a maior quantidade possível de erros do sequenciamento. Esta etapa é definida a partir do valor dado a complexidade das sequências considerando tamanho e qualidade das leituras geradas (SILVA, 2013). Esta qualidade é definida por uma variável em escala logarítmica de qualidade denominada *phred*, definida por $\mathbf{Q = -10\ log_{10}\ P}$, onde *P* é a probabilidade de erro para uma determinada base e *Q* é o *Phred Score*. O *Phred* calcula um grau de confiabilidade para cada base identificada, de acordo com o valor definido no *phred score*. Por exemplo, se o *phred score* for igual a 30, terá 99,9% de confiabilidade, logo 0,1% de erro (SILVA & KREMER, 2016).

Phred score	Confiabilidade
10	90%
20	99%
30	99,9%
40	99,99%
50	99,999%

Tabela 1: valores de *phred score* com as confiabilidades correspondentes

Diante da redução de custos e o aumento da velocidade na obtenção de dados é possível adquirir informações biológicas retiradas diretamente de comunidade microbianas, em seu habitat. Ao invés de sequenciar espécies individualmente, é possível analisar de diversas espécies, todas juntas. O estudo desses dados é denominado de metagenômica (KUNIN, et al., 2008). Metagenômica compõe-se de um conjunto de técnicas de pesquisa que contém diferentes abordagens e metodologias, que busca compreender a biologia em nível de comunidades, buscando os efeitos dos genes em comunidade (FALEIRO, ANDRADE, & REIS JUNIOR, 2011). A metagenômica procura adquirir sequências de genomas de diversos microrganismos num habitat, extraindo e analisando o DNA presente em uma amostra ambiental.

Técnicas metagenômicas consistem na extração de DNA diretamente do ambiente para, posteriormente, realizar a construção de uma biblioteca para o sequenciamento. A metagenômica oferece um caminho para descobrir a diversidade microbiana, proporcionando examinar a genômica funcional dos microrganismos presentes na amostra. Isto resulta na prospecção de bibliotecas metagenômicas que identificam novos genes dos mais variados tipos (HENNE et al., 2000; GUPTA et al., 2002). A criação dessas bibliotecas metagenômicas é possível com as novas tecnologias de sequenciamento que resultam na obtenção de grande escala de informações.

A metagenômica se torna uma ferramenta útil para obter conhecimentos de amostras ambientais, tendo como propriedade importante proporcionar a capacidade de caracterização, de maneira eficaz, a diversidade presente nas amostras genéticas (PEREIRA**,** 2014). E isto resulta no enriquecimento do conhecimento e nas aplicações práticas, recorrente a formação

de bibliotecas metagenômicas e no conhecimento de genes ativos capazes de desenvolver produtos biotecnológicos.

2.3 MONTAGEM DE GENOMA

A montagem consiste em reconstruir a sequência original, a partir das leituras resultantes do sequenciamento. Os algoritmos buscam por sobreposições entre duas ou mais leituras e, quando encontram, ocorre a união deles formando os *contigs*. Os *cotings* constituem uma sequência contígua e consenso das leituras (*reads*) alinhadas. Eles podem ser orientados e ordenados com o auxílio de algoritmos e genomas de referência, o que pode gerar sequências maiores apresentando lacunas (*gaps*) entre elas denominadas de *Scaffolds* (MILLER; KOREN & SUTTON, 2010).

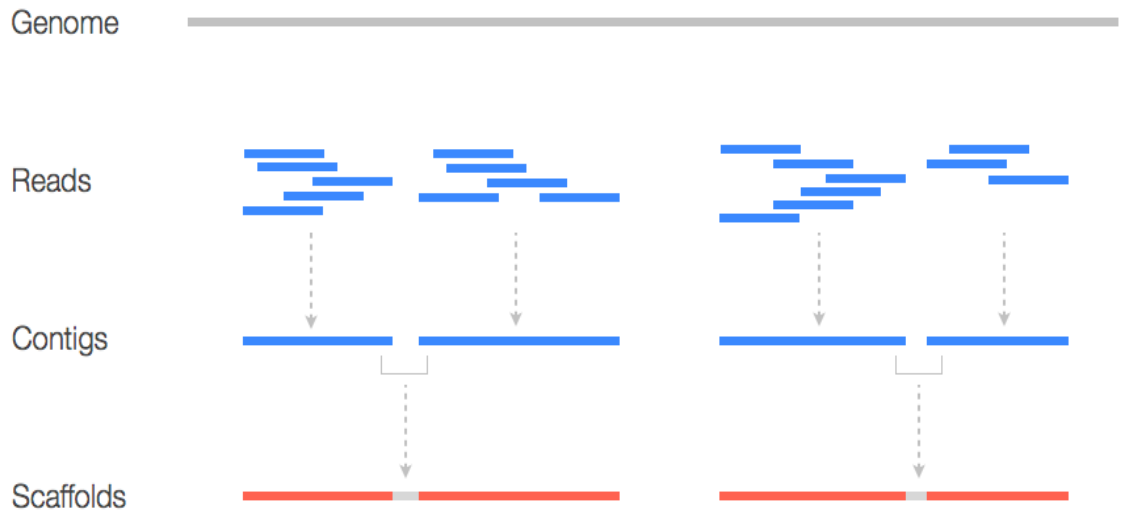

Figura 1: Representação de *reads- contigs- scaffolds* retirada de: *Ecology and Evolution Unit Page*

Durante o processo de montagem deve-se levar em consideração erros no processo de sequenciamento (como bases inseridas, removidas incorretamente detectadas), orientação desconhecida das leituras sequenciadas (já que estes podem ser originários de ambas as fitas de DNA), regiões repetidas e a falta de cobertura (PEREIRA, 2014). Além disso, para se obter bons resultados com a montagem é preciso saber algumas informações da amostra, tais como: que amostra está sendo sequenciada, se utilizou biblioteca pareada ou não, está ciente do sequenciador utilizado, como está a qualidade dos fragmentos gerados no sequenciamento e qual tamanho esperado do genoma (LIMA, 2015). Em posse destas informações é possível

escolher o melhor algoritmo a ser utilizado na montagem e também se há necessidade de realizar tratamento dos dados que contém as leituras.

Os algoritmos de montagem foram surgindo e se aperfeiçoando na medida em que os processos de sequenciamento ficaram mais desenvolvidos. Estes algoritmos podem ser classificados de acordo com três abordagens: algoritmo guloso, algoritmos OLC (*overlap-loyout-consensus*) e grafos de Bruijn.

O algoritmo "guloso" consiste em realizar alinhamentos consecutivos de *reads* com melhor sobreposição, comparando dois *reads* onde o sufixo de um é igual ao prefixo do outro (PIRO, 2014). Baseia-se num conjunto de escolhas que melhor define a solução de montagem, buscando uma solução ótima. Contudo, este algoritmo é utilizado para montagem de genomas pequenos (FRAGA, 2014).

O algoritmo OLC é tradicional de montagem de fragmentos, sendo composto por três passos: *detecção de sobreposição, layout dos fragmentos e decisão da sequência de consenso* (LEMOS, BASILIO & CASANOVA, 2003). A detecção de sobreposição consiste em comparar cada *read* com todas as outras *reads* da amostra, com intuito de detectar e relacionar as sobreposições, criando um grafo onde cada vértice representa um *read* e cada aresta dirigida de um vértice a outro representa a sobreposição. O segundo passo, *layout*, tem como objetivo diminuir o caminho do grafo, ou seja, é feito uma busca por caminho hamiltoniano extraindo o caminho com peso máximo, caminho com maior número de sobreposições. Finalizando o processo, a última etapa consenso consiste em realizar o alinhamento múltiplo de todos as *reads* que fazem parte do caminho hamiltoniano encontrado na etapa *layout*, determinando assim as junções das *reads* gerando sequências consenso (*contigs* ou *scaffolds*) como mostrado na Figura 1 (LEMOS, BASILIO & CASANOVA, 2003). Newbler (454 life Sciences, 2011) e Celera Assembler (MILLER et al., 2008) são exemplos deste tipo de algoritmo.

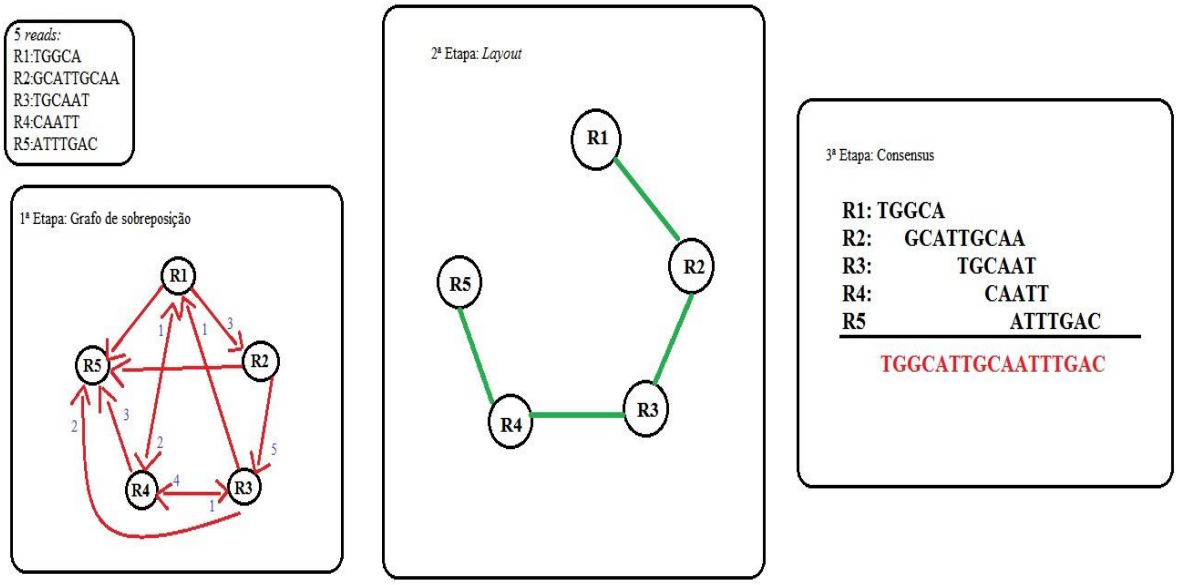

Figura 2: Exemplo das etapas OLC: *Overlap-layout-consensus.* Adaptada de: FRAGA, 2014

Os algoritmos baseados em grafos de Brujin consistem em subdividir as *reads* em fragmentos menores, denominados *k-mer*. O algoritmo funciona construindo um grafo que, para cada *k-mer*, é criado um vértice que possui uma aresta representando uma sobreposição de fragmentos igual a k-1. Nestes algoritmos as arestas equivalem as sobreposições entre o sufixo do vértice de saída e o prefixo do vértice de entrada. Esta abordagem deixa contabilizar a frequência de cada *k-mer* dentro do conjunto (FRAGA, 2014). Estes algoritmos são baseados no caminho Euleriano, o qual determina que todas as arestas sejam visitadas uma única vez, como mostrado na Figura 3. Tem-se como exemplo deste algoritmo o ABySS (SIMPSON et al.,2009) e Velvet (ZERBINO & BIRNEY, 2008).

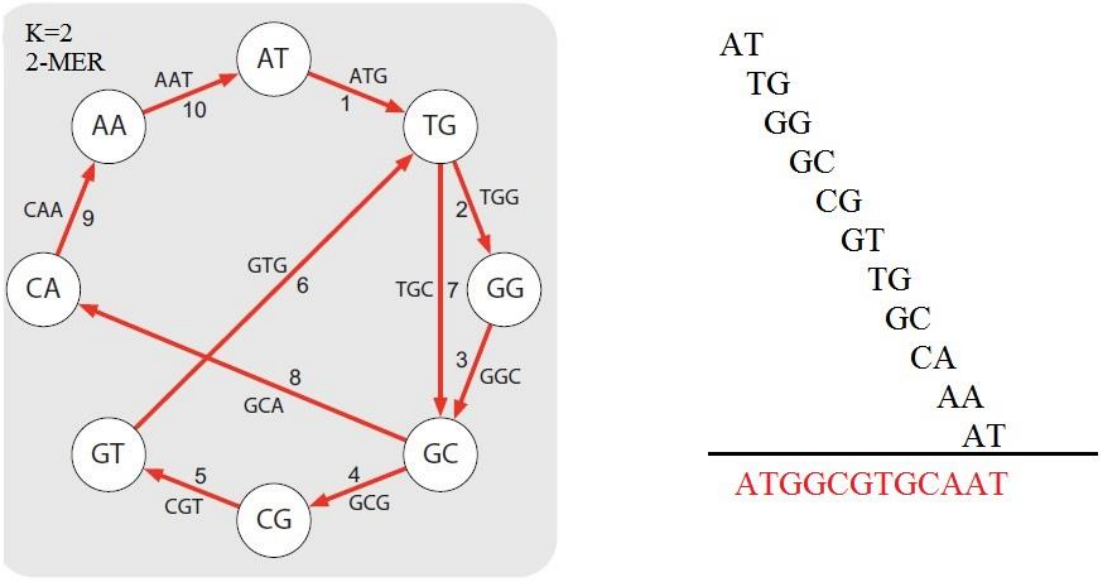

Figura 3: Exemplo do Grafo de Bruijn com k-mer igual a dois. Adaptada de: *Applicazioni biotecnologiche in systems biology*

A qualidade do processo de montagem é avaliada de acordo com a precisão dos *contigs* e *scaffolds*, avaliação de tamanho e pela qualidade das sequências geradas. Por esta razão, são utilizados valores de tamanho máximo e tamanho médio das sequências, número de leituras utilizadas na montagem, tamanho total combinado das sequências e o N50. O N50 refere ao tamanho do menor *contig* dentro de um conjunto onde a somatória de todos os *contigs* represente pelo menos metade do total de bases fornecidas pela montagem. (MILLER; KOREN & SUTTO, 2010).

Pode-se levantar a seguinte consideração sobre o processo de montagem: seu objetivo consiste e montar apenas um genoma. Para se realizar a montagem de um metagenoma é necessário atentar para a existência de múltiplos genomas, o que implica na problemática da diversidade genômica, que é reflexo da variedade de espécies em uma única amostra. Mesmo utilizando os algoritmos avançados, a montagem de metagenomas apresenta limitações como por exemplo: o problema da amostragem ser incompleta resultando em genomas incompletos, a diversidade microbiana e a formação das sequências de espécies diferentes (denominado de quimeras) (PEREIRA, 2014).

2.4 CLASSIFICAÇÃO DE GENOMAS

Com a obtenção das *reads* no sequenciamento do metagenoma e sua montagem posteriormente, resultando na formação dos *contigs/scaffolds* há a necessidade de associar essas sequências com os organismos dos quais elas foram originadas. Deste modo, é possível uma interpretação da amostra a partir do detalhamento da diversidade biológica ali presentes (NETO, 2012). O processo de classificação foi defino por Brady e Salzberg, como sendo atribuição de um rótulo específico aos membros dos conjuntos da amostra. Estes autores conceituaram *binning* como um agrupamento de *reads* do conjunto de dados (DRADY & SALZBERG, 2009). A estratégia de *binning* é utilizada para agrupar sequências tendo como relevância suas características comuns. Essa estratégia tem grande importância para a análise de genomas completos (ou quase completos) na metagenômica, permitindo analisar o conteúdo genômico de microrganismos não-cultiváveis (LEMOS, 2015).

Uma das maneiras de classificar essas sequencias é encontrar semelhanças com a sequência referência, as quais serão usadas para construir uma árvore de espécies (WOOLEY, GODZIK & FRIEDBERG, 2010). Essa técnica se torna útil quando a maioria das sequências da amostra apresenta semelhanças significativas com as sequências de referências conhecidas. Os resultados na árvore de espécies mostram em uma visão geral as espécies dominantes na amostra. Porém, esta abordagem de *binning* tem limitações quanto a base de dados de referência, atualmente incompleta e altamente tendenciosa. Limita-se também quanto aos genes oriundos de dados metagenômicos, principalmente aqueles que obtiveram montagem mínima, que comumente são fragmentados gerando alinhamento incompleto (PEIXOTO, 2013). Além disso, os genes conservados filogeneticamente representam apenas uma pequena parte do total dos metagenomas.

Recentemente, Wu e colaboradores desenvolveram uma ferramenta computacional para automatizar a identificação e validação de *bins*, o MaxBin (WU et al.; 2014). O MaxBin procura automatizar a busca por *bins* baseando-se no algoritmo de Maximização de expectativas (*expectation-maximazation algorithm*), que inicialmente calcula o nível de cobertura diferencial de cada *contig* e as frequências de tetranucleotídeos (WU et al., 2014). Essas informações são combinadas e usadas para a identificação dos *bins*. Essa abordagem computacional está ligada a reconstrução de genomas microbianos aplicados em dados

metagenômicos. O MaxBin processa, resumidamente, em três etapas (Figura 4): a primeira é a geração de informações de entrada, definindo o nível de cobertura de sequenciamento para cada *contig/scaffolds* calculando a frequências de tetranucleotídeos. A segunda etapa objetiva-se identificar *bins* e validar cada espécie presente na amostra. E por fim, mostrar as espécies que representam genomas individuais encontrados no metagenoma.

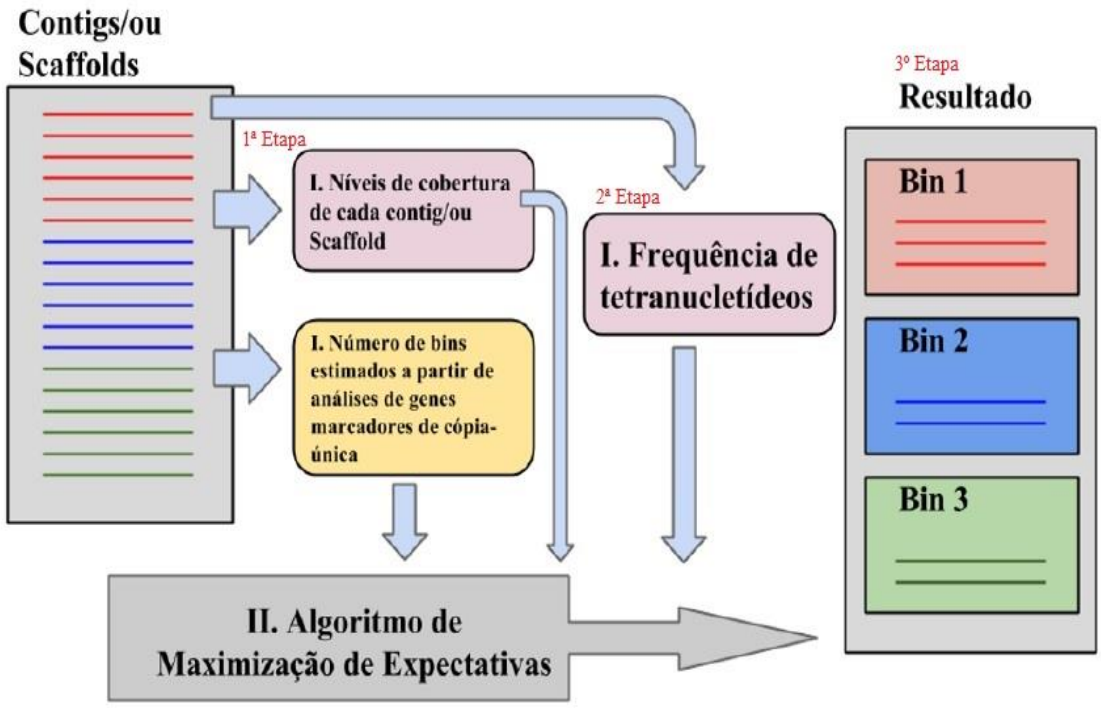

Figura 4:Processamento computacional Maxbin. Adaptada de WU et al., 2014.

Uma outra abordagem de classificação de genoma é a baseada em composição, que utiliza analises estatísticas das sequências (WOOLEY, GODZIK & FRIEDBERG, 2010). Um exemplo dessa abordagem é o modelo de Markov, que se baseia em frequências de *k-mer*, mostrando-se muito efetivos para análises estatísticas (NETO, 2012). Contudo, esta abordagem não está livre de erros, pois, quanto mais numerosas e mais próximas filogeneticamente são as espécies encontradas no metagenoma, maior é a frequência de erros na classificação. Na classificação baseada na frequência de *k-mers* não são necessárias as sequências de referências (PEIXOTO, 2013). Outra abordagem de classificação de genomas é baseada na comparação das sequências, que compara os *contigs* com amostra de bases de dados conhecidas que incluem alinhamentos múltiplos de sequências e anotações. A comparação dos *contigs* com genes conhecidos oriundos das bases de dados é realizada por programas de comparação por

similaridade de sequência. Os *contigs* de entrada são diretamente comparados com as bases de dados e a atribuição em sequências destes *contigs* em suas respectivas espécies ocorre de acordo com o melhor resultado feito na comparação (LEMOS, 2015), um exemplo deste tipo de classificação são os programas BLAST (disponível em https://blast.ncbi.nlm.nih.gov/Blast.cgi).

3. MATERIAL E MÉTODOS

Nesta seção é descrito como se procedeu o processo de montagem de genoma da cianobactéria *Nostoc sp.* CACIAM 19 utilizando técnicas metagenômicas. Serão apresentados os métodos e material utilizado para a realização deste trabalho, explanando um pouco sobre o pré-processamento das sequências advindas do sequenciamento, os algoritmos de montagem e de classificação de genomas que foram utilizados.

3.1 OBJETO DE ESTUDO

O genoma da cianobactéria *Nostoc* sp. CACIAM 19 foi sequenciado pelo Centro de Inovações Tecnológicas (CIT) do Instituto Evandro Chagas (IEC) em parceria com o Laboratório de Tecnologia Biomolecular (LTB), posteriormente disponibilizado os dados para a parceria com o Laboratório de Bioinformática e Computação de Alto Desempenho (LABIOCAD), ambos da Universidade Federal do Pará. A plataforma de sequenciamento utilizada foi o 454 GS FLX *Titanium* (Roche), com o uso de bibliotecas não-pareadas construídas a partir da extração do material genético da amostra de cultura não-axênica da cianobactéria *Nostoc* sp. CACIAM 19.

As *reads* resultantes do sequenciamento da cianobactéria *Nostoc* sp. CACIAM 19 foram verificadas quanto sua qualidade e quantidade, analisando número de bases, tamanho máximo e mínimo. Essa verificação foi feita utilizando FASTX-TOOLKIT (HANNON LAB, 2012) e MOTHUR (SCHLOSS, 2009).

O FASTX-TOOLKIT é um conjunto de ferramentas de pré-processamento de arquivos oriundos do sequenciamento e tem como objetivo manipular as sequências para gerar bons resultados no mapeamento. Foi utilizado os comandos:

- ***fastq_quality_trimmer****:* Realiza cortes, a partir das extremidades, na sequência com base no valor qualidade do valor que fora fornecido. O comando usado foi: ***Fastq_quality_trimmer –Q33 –[t N] –[l N] –i (in).fastq –o (out_trimmed).fastq****,* onde **Q** é o tipo de formatação da qualidade *phred* (mantido por padrão), **t** é o limite de

qualidade de nucleotídeos que corta, a partir das extremidades, as sequências que tiverem valor inferior ao que foi fornecido **l** é o comprimento mínimo da sequência a ser mantida. Sequências que depois os cortes apresentarem valores menores que o valor fornecido, serão descartadas.

- ***fastq_quality_filtrer****:* Realiza filtragem, após o corte, de acordo com a qualidade sugerida. Comando utilizado**:** ***Fastq_quality_filter –Q33 –[q N] –[p N] –i (out_trimmed).fastq –o (out_filtered).fastq***, de **q** é o valor mínimo para manter a qualidade, **p** é o percentual mínimo de bases que devem ter o valor igual ou superior ao valor mínimo. **Q** é o tipo de formatação da qualidade *phred* (mantido por padrão). Observa-se que o arquivo de entrada é o arquivo de saída do processo anterior, pois a filtragem foi realizada após os cortes das sequencias com qualidade inferior.

Assim como o FASTX-TOOLKIT, o MOTHUR foi desenvolvido para analisar as sequências resultantes do sequenciamento. Ambos buscam realizar um pré-processamento das sequências baseando na qualidade das bases. No MOTHUR foi utilizado usando o comando ***trim.seqs***, que realiza o corte e a filtragem. Como demonstrado*:* ***mothur> trim.seqs(fasta=arquivo.fasta, qfile=arq.qual, qavarege=N),*** recebendo os arquivos fasta, com as *reads* do sequenciamento, e o arquivo ".qual" com os valores de qualidades das bases. Usou-se a opção *qaverage,* que calcula a pontuação média da qualidade para cada sequência e remove as sequências que tem a média menor que o valor que foi fornecido na opção.

Em ambas ferramentas foram realizadas seis rodadas com a variação de parâmetros mostrados na Tabela 1:

Rodadas	***fastq_quality_trimmer***	***fastq_quality_filtrer***	**MOTHUR**
Rodada 1	t=20 l=20	q=20 p=20	*qavarege=20*
Rodada 2	t=22 l=22	q=22 p=22	*qavarege=22*
Rodada 3	t=25 l=25	q=25 p=25	*qavarege=25*
Rodada 4	t=28 l=28	q=28 p=28	*qavarege=28*
Rodada 5	t=30 l=30	q=30 p=30	*qavarege=30*
Rodada 6	t=35 l=35	q=35 p=35	*qavarege=35*

Tabela 2: Parâmetros utilizados no pré-processamento das sequências, utilizando as ferramentas FASTX-TOLLKIT e MOTHUR

Com os resultados obtidos nas duas ferramentas, foi realizada análise comparativa dos dados do sequenciamento utilizando a ferramenta FASTQC (ANDREWS, 2010), que retorna um relatório com informações, tais como total de bases, total de *reads*, comprimento máximo e mínimo dos fragmentos, entre outras. Deste modo, foi realizada a verificação do controle da qualidade nos dados sequenciados. A partir da análise comparativa, utilizando o critério de qual ferramenta e qual parâmetro manteve uma melhor qualidade das sequências, foi escolhida a melhor filtragem para que com esta fosse realizado a montagem do genoma da cianobactéria *Nostoc sp*. CACIAM 19.

3.2 MONTAGEM DE GENOMA

Após ter realizado a análise comparativa, a filtragem com a melhor qualidade foi submetida aos algoritmos de montagem. Os algoritmos de montagem utilizado neste trabalho foram: OLC utilizando o Newbler Assembler 2.9 (454 life Sciences), e o grafo de Bruijn com o uso do programa ABySS 2.0 (SIMPSON et al., 2009). Em ambos algoritmos foram realizadas cinco montagens, modificando os parâmetros de sobreposição, Tabela 2. No *software* Newbler os valores modificaram na variável "*Minimum overlap length*" (Sobreposição mínima), mantendo as demais variáveis por padrão, como mostrado na Figura 4. No *software* ABySS os valores modificaram na variável *k* (que representa *k-mer*) no comando *single-end*: ***ABySS –k N aqv.fasta/q –o saída.fa.*** Entretanto, de acordo com a definição do grafo de Bruijn, é preciso adicionar uma unidade a cada valor de *k-mer*, pois estes têm k-1 caminhos para percorrer. Diante disto, o valor da sobreposição se iguala a utilizada no software Newbler

Rodadas	**Newbler**	**ABySS (k + 1)**
Rodada 1	20	21
Rodada 2	25	26
Rodada 3	30	31
Rodada 4	35	36
Rodada 5	40	41

Tabela 3: Valores utilizados nas rodadas realizadas no processo de montagem com os softwares Newbler e ABySS.

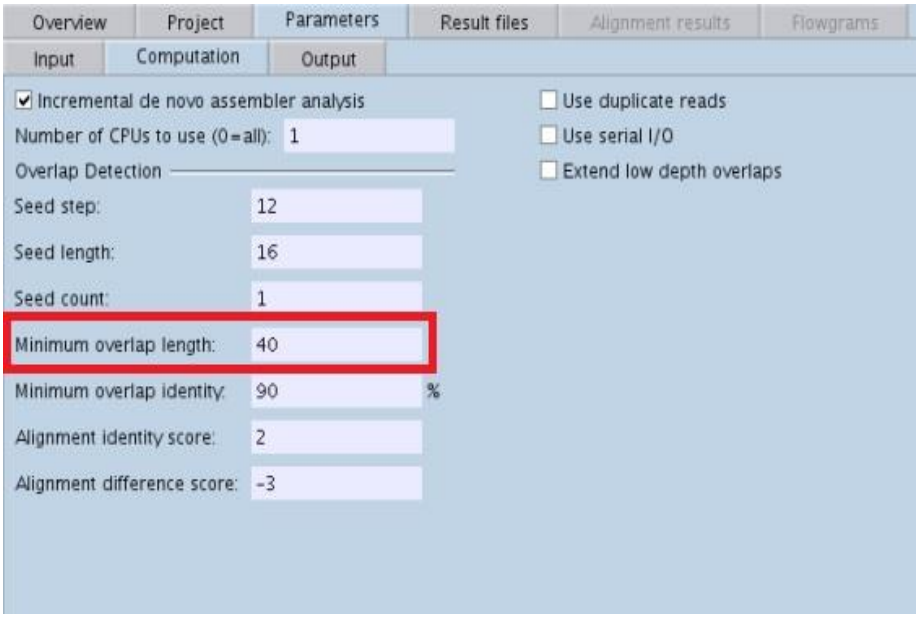

Figura 5:Tela de parâmetros da ferramenta Newbler, mostrando onde variou os valores. Fonte: 454 sequencing, 2011

Com as montagens feitas nas duas ferramentas, os resultados foram comparados e definido a melhor montagem para cada montador. Os critérios para escolha das melhores montagens foram: o tamanho de maior *contig*, média e quantidade total de *contigs* produzidos, tamanho total de bases e o N50.Após a definição da montagem mais precisa em cada montador, foi feita a comparação entre as duas abordagens para a definição de melhor montador, o qual obteve a montagem de genoma com melhor qualidade.

3.3 CLASSIFICAÇÃO DOS GENOMAS

A partir do resultado da melhor montagem do melhor montador foi utilizado o algoritmo de *binning*, utilizando a ferramenta MaxBin 2.2(Wu et al,2010). No MaxBin foi utilizado o comando: ***Run_MaxBin.pl – contig mycontig,fasta –reads – myreads.fasta –out myout1***, que compara os *contigs* resultantes da montagem com os *reads* do sequenciamento. Depois de classificar os ***contigs/scaffolds*** em diferentes blocos, o MaxBin relata características genômicas, como o tamanho do genoma, conteúdo GC (Guanina-Citosina) e níveis de cobertura do genoma, em arquivos em formato tabular. Fornece também arquivos dos *contigs* com semelhanças, com total de *contigs*, tamanho máximo e mínimo, total de bases e N50.

Os *bins* resultantes do processo de classificação com o MaxBin, foram submetidos a classificação taxonômica utilizando o programa BLASTp (disponível em https://blast.ncbi.nlm.nih.gov/Blast.cgi). O BLAST está hospedado no site do *National Center for Biotecnology Information* (NCBI) e realiza buscas das sequências por similaridade no banco de dados. Neste trabalho as buscas foram feitas pela comparação sequência de proteínas. Os resultados do BLASTp mostram as sequências que tiveram similaridades com a sequência de entrada.

Os resultados do BLASTp foram postos no MEGAN 6 (HUSON et al., 2007). O MEGAN permite analisar grandes conjuntos de dados, este processa os resultados do BLAST realizando estimativas e explorações interativas para obter a possível classificação taxonômica da sequência (PEIXOTO, 2013). MEGAN resulta em uma árvore taxonômica que é possível descobrir as categorias taxonômicas das sequências.

4. RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os resultados obtido em todo o processo de tratamento de qualidade dos dados sequenciados, montagem e a classificação dos genomas. Além dos resultados quantitativos, serão apresentadas discussão em torno dos mesmos. E com as análises realizadas definir os melhores resultados.

4.1 TRATAMENTO DE QUALIDADE

Os resultados do pré-processamento, corte e filtragens, das *reads* estão representadas graficamente abaixo. Havendo a comparação dos valores obtidos nas ferramentas utilizadas, FASTX-TOOLKIT e MOTHUR, e com estes valores realiza-se a análise qualitativa entre as

ferramentas através das métricas: tamanho máximo de *reads*, quantidade total de *reads* e total de bases.

Comparando os resultados obtidos nas ferramentas, observa-se que a ferramenta MOTHUR elimina uma grande quantidade de *reads*, não garantido que todas estas *reads* eliminadas sejam de baixa qualidade, pois o comando utilizado calcula uma média para cada *read* e a compara com o valor fornecido. Enquanto a ferramenta FASTX-TOOLKIT utiliza um percentual mínimo de bases que devem ser mantidas, considerando que estas bases devem ter o valor igual ou superior ao valor mínimo de qualidade. Isto pode ser observado nas métricas Total de *Reads* e Total de Bases, mostrados nas Figuras 5 e 6, qual mostram que na ferramenta FASTX-TOOLKIT manteve uma grande quantidade de *reads*, eliminando somente *reads* inferiores ao valor mínimo de qualidade, obtendo um valor maior de *reads* e de bases.

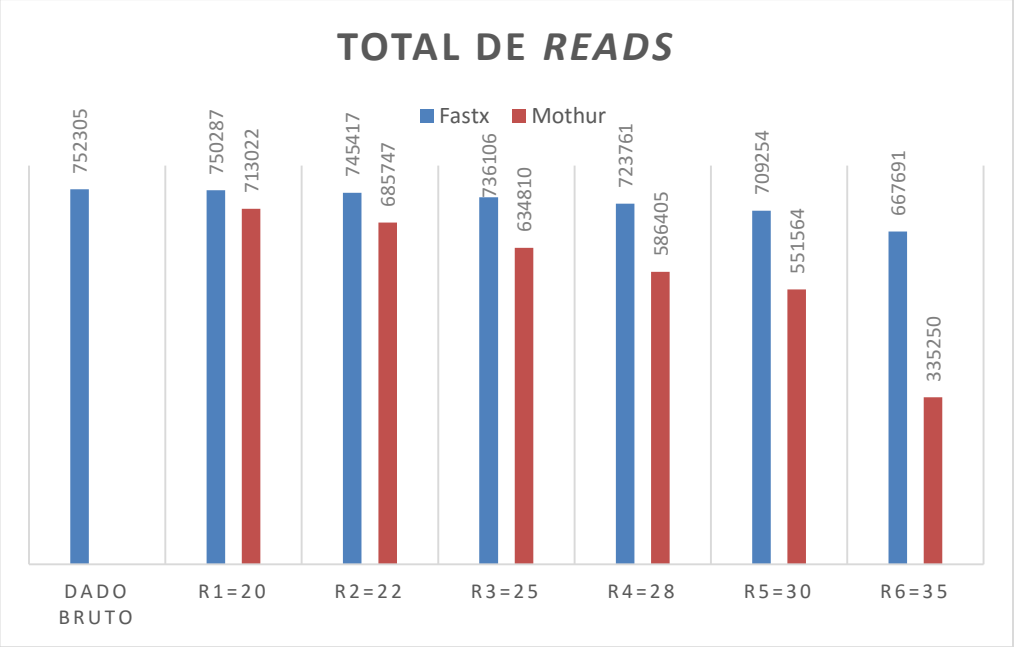

Figura 6: Comparativo do resultado do pré-processamento das *reads* referente a métrica "total de *reads*". Onde R(1,...,6) significa rodadas de pré-processamento igualando aos valores utilizados.

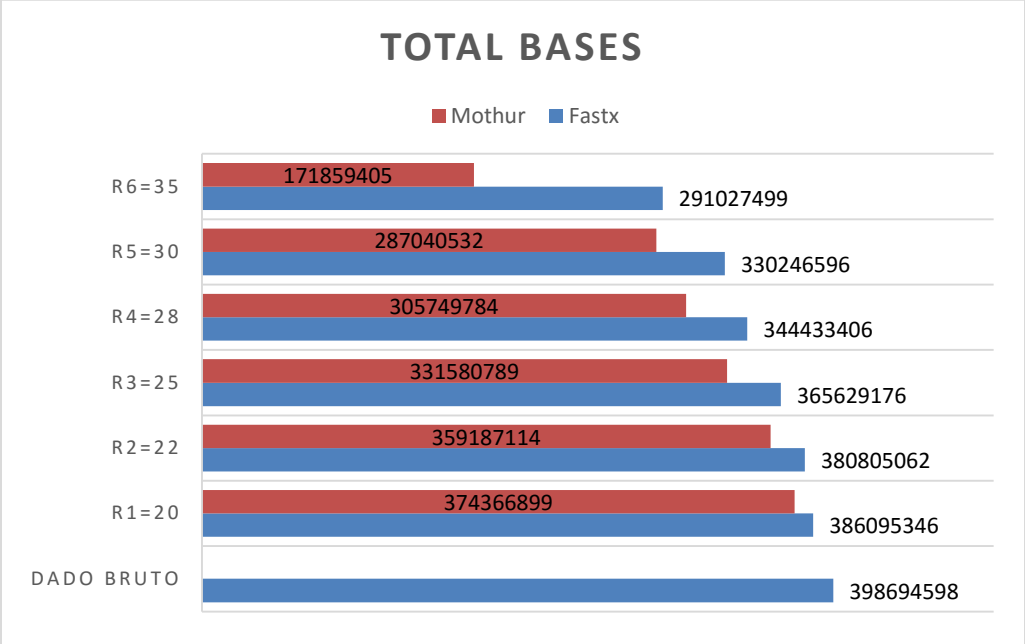

Figura 7: Resultado comparativo do pré-processamento das *reads* - "total de bases". Onde R(1...6) representa rodada.

Diante dos resultados obtidos nas métricas anteriores, o maior tamanho de *read* foi obtido na ferramenta FASTX-TOOLKIT, pois nesta grande parte das *reads* foram mantidas.

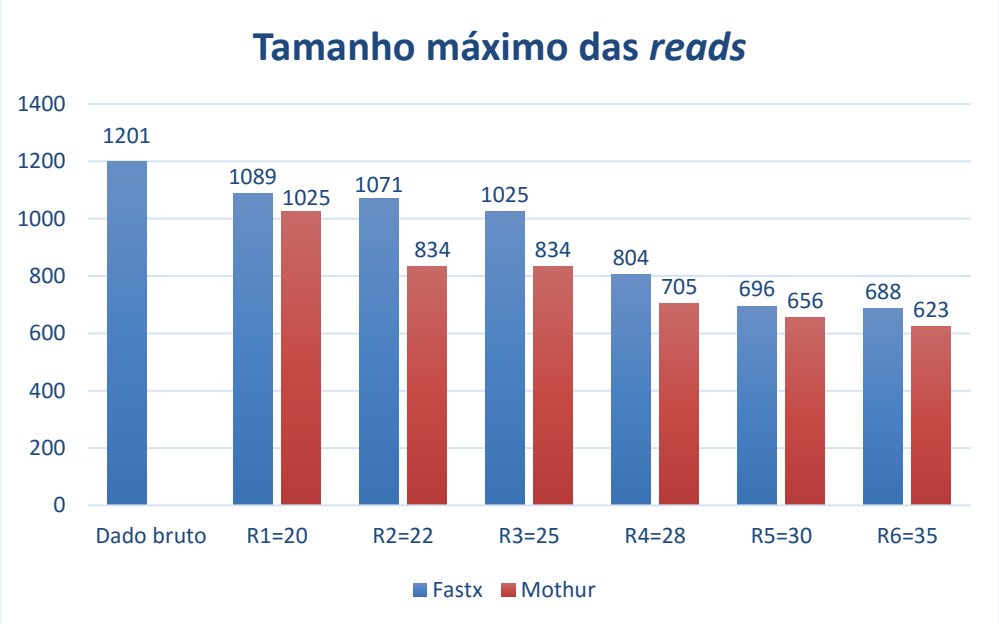

Figura 8: Resultado do pré-processamento de *reads* -"tamanho máximo das *reads*". Onde R(1,...,6) representa rodada de pré-processamento igualado com os valores fornecidos.

Dentre os valores dos parâmetros utilizados, é preciso definir qual deles se obteve o melhor resultado. Analisando os gráficos acima é possível notar que em ambas ferramentas o valor que mantem grande quantidade de *reads* é a rodada com valor igual a vinte, pois, nesta rodada poucas *reads* foram descartadas mantendo assim uma grande parte da amostra. Dentre as duas ferramentas utilizadas, FASTX-TOOLKIT foi a ferramenta que mais conservou a amostra, eliminando somente *reads* de baixa qualidade. Portanto, afirma-se que o melhor resultado foi encontrado no parâmetro igual a vinte da ferramenta FASTX-TOOLKIT.

4.2 MONTAGEM DE GENOMA

Com a definição do melhor resultado do pré-processamento, foi realizada a montagem de genomas utilizando as abordagens OLC e grafo de Buijn. Os resultados obtidos estão representados graficamente abaixo. Com estes valores foi possível o melhor parâmetro para cada abordagem e posteriormente estabelecer qual montador obteve a melhor montagem.

O estudo quantitativo e qualitativos dos resultados se dá a partir da análise e dos estudos de suas métricas, pois, neste caso nem sempre o melhor valor será o melhor resultado. Faz-se necessário levar em consideração alguns fatores, tais como: o algoritmo utilizado, definição de tamanho de sobreposição e precisão da sobreposição. As métricas utilizadas para esta análise são: total de *contigs*, tamanho máximo de *contigs*, total de bases e N50.

A primeira métrica utilizada na análise dos resultados se refere ao total de *contigs* resultantes da montagem do genoma da cianobactéria *Nostoc* sp. CACIAM19. Como mostrado no gráfico da Figura 8 e Figura 9, sobre as métricas Total de *contigs* e total de bases, observa-se que em ambos softwares o resultado maior foi na primeira rodada quando o parâmetro de sobreposição é igual a vinte (vinte e um na abordagem grafo de Bruijn). Entretanto, este resultado não significa melhor resultado de montagem, pois com este valor há uma cobertura de sobreposição pequena que não garante uma montagem precisa, tendo em vista que quanto maior for o valor de sobreposição mais confiável será. Pode-se afirmar que, uma sobreposição de quarenta bases é possível realizar uma montagem mais precisa dos genomas e assim reduzir erros por repetição.

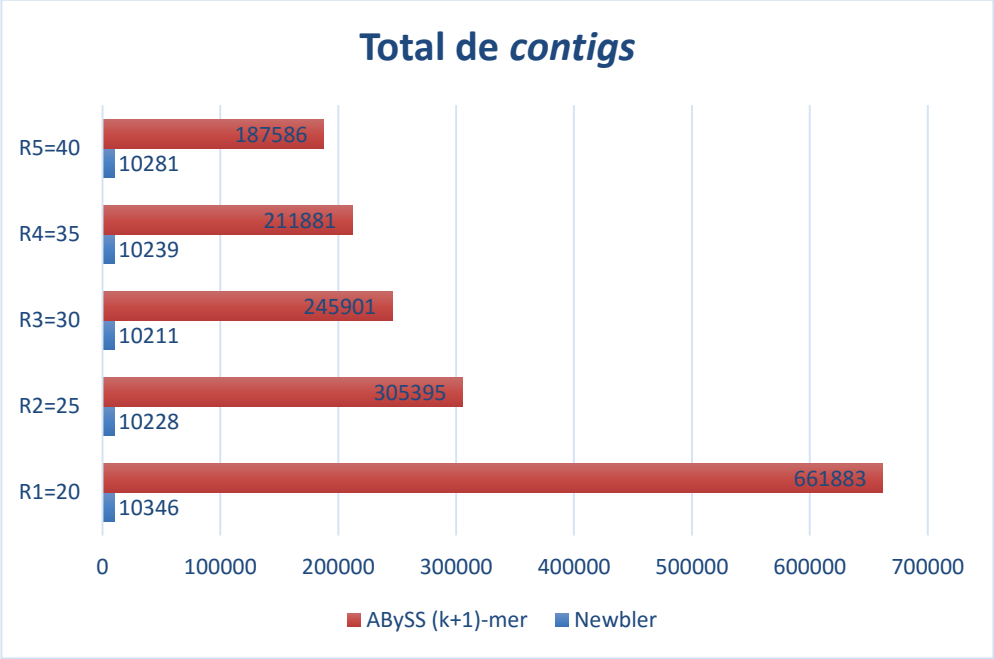

Figura 9:Total de contigs resultante da montagem de genoma. R (1...5) representa rodada de montagem com os valores de sobreposição utilizado a cada rodada, levando em consideração que no ABySS adiciona uma unidade

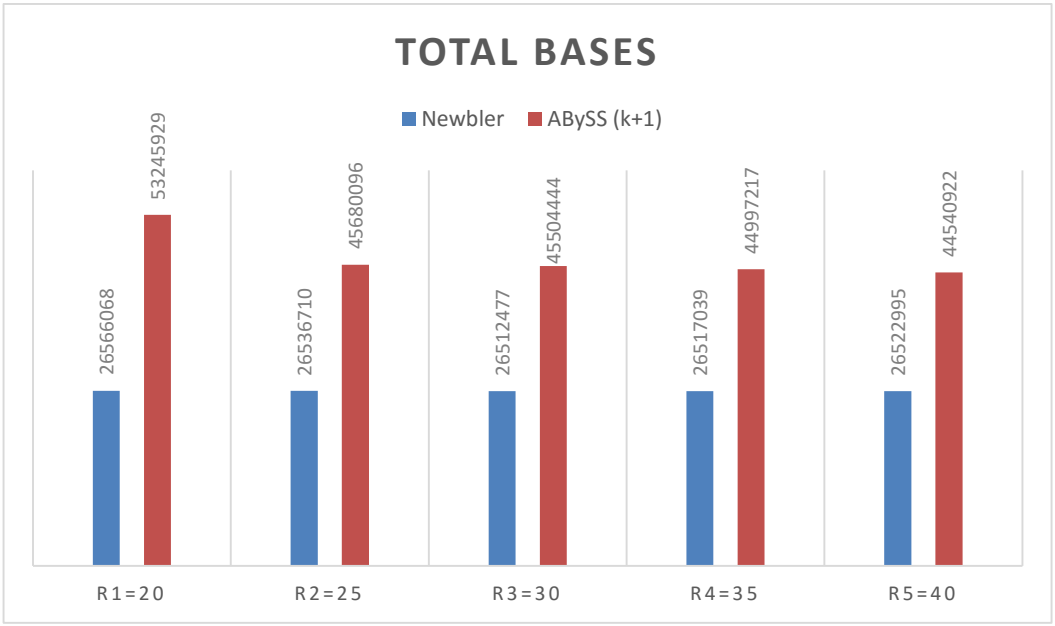

Figura 10: Total de bases presente na montagem. R (1...5) representa rodada de montagem igualando aos valores de sobreposição utilizados, levando em consideração que no ABySS adiciona uma unidade

O N50 (Figura 10) representa o menor *contig* de um conjunto cuja a somatória é igual ou maior a 50% das bases formadas. Nota-se que no *software* ABySS o melhor resultado foi encontrado quando a sobreposição é igual a quarenta, enquanto no *software* Newbler os resultados se mantiveram próximos ou iguais. O N50 está ligado ao tamanho máximo de *contigs (*Figura 11) onde é possível perceber que no software ABySS obteve-se menores valores de tamanho máximo, em virtude de que, a montagem é realizada com fragmentos de *reads*.

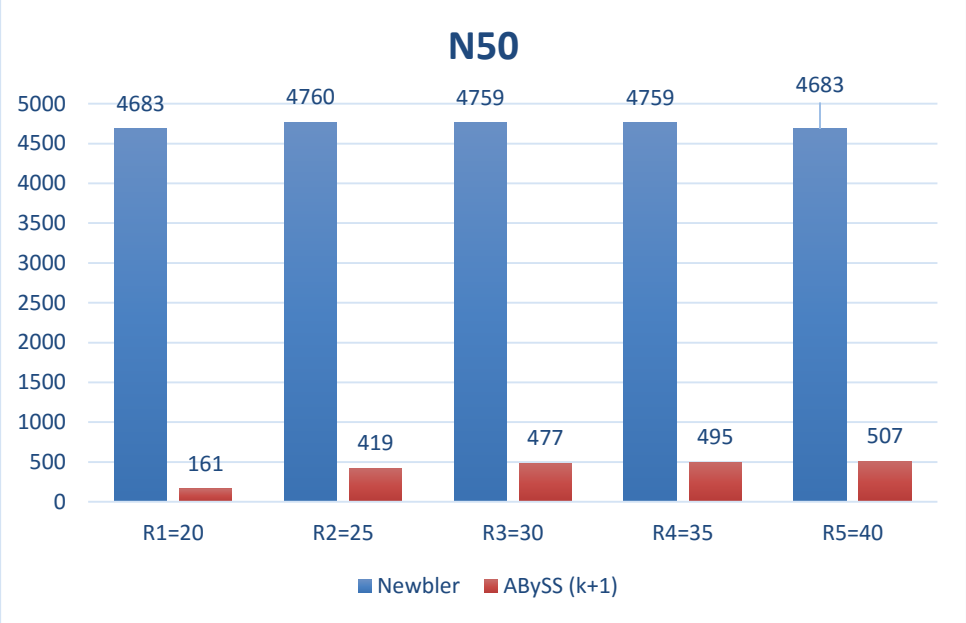

Figura 11: Valor N50 da montagem de genoma. R (1...5) representa rodada de montagem igualados com valores de sobreposição definidos, levando em consideração que no ABySS ganha uma unidade

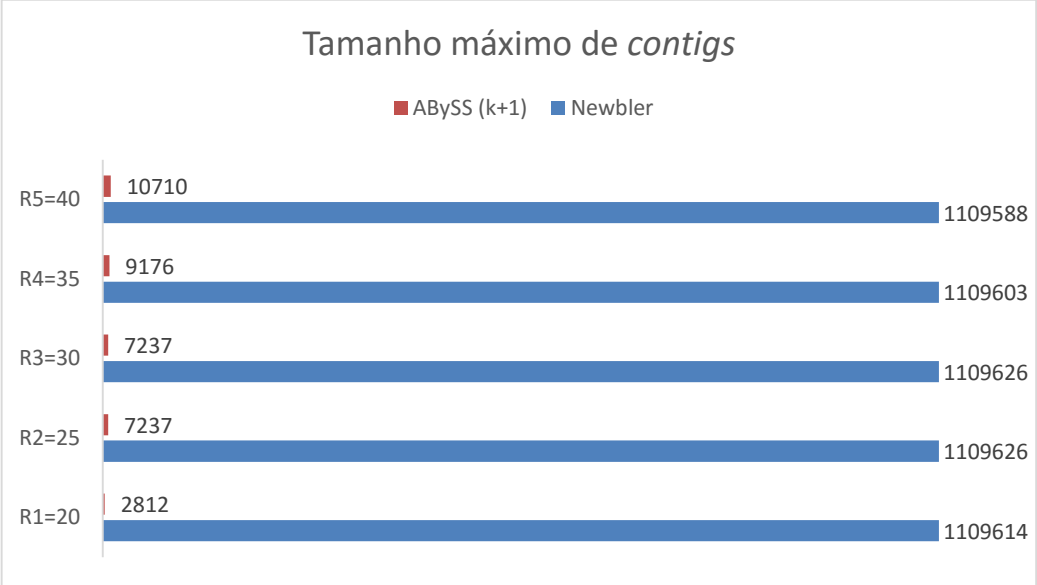

Figura 12: Tamanho máximo de *contigs*. R (1...5) significa rodada de montagem sendo igualado com valores de sobreposição definidos, levando em consideração que no ABySS adiciona uma unidade

Com os resultados obtidos e com as análises feitas, é possível definir qual a montagem mais precisa para cada algoritmo. Observou-se que em ambos os casos o valor de sobreposição igual a quarenta é mais satisfatório, pois há um alinhamento de bases maior. Comparando os dois resultados e comparando as abordagens utilizadas pode-se afirmar que os melhores resultados foram obtidos no *software* Newbler, pois com o algoritmo OLC é possível ter maior cobertura. Apesar da abordagem de grafos de Bruijn ter uma grande facilidade em lidar com uma grande quantidade sequências, a questão de fragmentar as *reads* em *k-mers* pode gerar problemas de repetições de bases na montagem. Enquanto a abordagem do algoritmo OLC comprar todas as *reads* entre si, buscando similaridades para realizar a montagem (Li; Chen; et al., 2011). Portanto, pode-se afirmar que para a montagens de genomas da cianobactéria *Nostoc* sp. CACIAM 19 o melhor resultado foi obtido na abordagem do algoritmo OLC com o montador Newbler, com o valor de parâmetro de sobreposição igual a quarenta.

4.3 CLASSIFICAÇÃO DE GENOMA

Os *contigs* resultantes da montagem pelo Newbler foram submetidos ao software MaxBin para a classificação do genoma. O software MaxBin foi utilizado para descobrir as espécies abundantes na amostra. Foram encontradas seis espécies na amostra da cianobactéria *Nostoc sp*. CACIAM 19. Após a separação dos genomas, os resultados foram submetidos ao BLASTp e posteriormente ao MEGAN para a classificação taxonômica. A Tabela 3 mostra a classificação taxonômica, abundância, tamanho do genoma, total de *contigs*, total de bases e tamanho máximo de *contigs* de cada espécie encontrada.

Classificação	**Abundância**	**Tamanho do genoma**	**Total de *contigs***	**Total de bases**	**Tamanho máximo de *contigs***
Fluviicola	5.46	3771010	62	3757986	1109588
Hyphomonas polymorpha1109	4.33	4751682	273	4715530	788769
Rhizobiales Rhosospirillaceae	1.72	4091458	756	4100066	31308
Acetobacteraceae	0.88	1230544	792	1237291	8507
Belnapia sp. f-4-1	0.86	3114916	1668	3135557	8541
Nostoc calcícola	0,80	6507864	2488	6514325	20752

Tabela 4: Classificação dos genomas e suas especificações.

Os resultados do MEGAN são fornecidos em árvores de classificação taxonômicas. Nesta arvore é possível descobrir qual(is) espécie(s) se encontram para cada classificação (ou *bin*) obtido pela ferramenta MaxBin, como pode ser observado nos genomas presentes na amostra utilizada neste trabalho. Dentre os seis genomas encontrados apenas dois estão sem

contaminação, Figura 13 e Figura 17. Os outros quatro genomas (Figuras 13,14,15 e 17) possuem mais de uma espécie, por conter semelhança nos *contigs* presentes.

Na figura 13, mostra a classificação taxonômica para o gênero *Fluviicola*, sem haver contaminações. Este gênero contém 62 *contigs* com total de bases igual 3757986. Este genoma está em maior abundância presente na amostra.

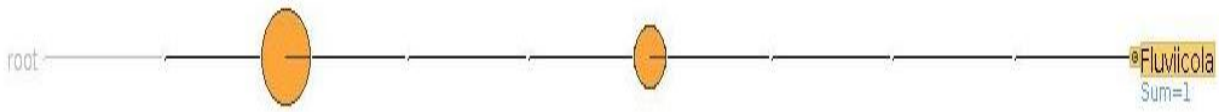

Figura 13: MEGAN: classificação taxonômica *bin*1, gênero: *Fluviicola.*

O gênero *Hyphomonas*, Figura 14, apresenta-se em maior evidência no segundo *bin*, se dividindo em três espécies: *Hyphomonas* hirschiana, *Hyphomonas* jannaschiana e em maior quantidade *Hyphomonas* polymorpha. Contudo, o gênero tem uma pequena contaminação da espécie Chloroflexi bacterium OLB13.

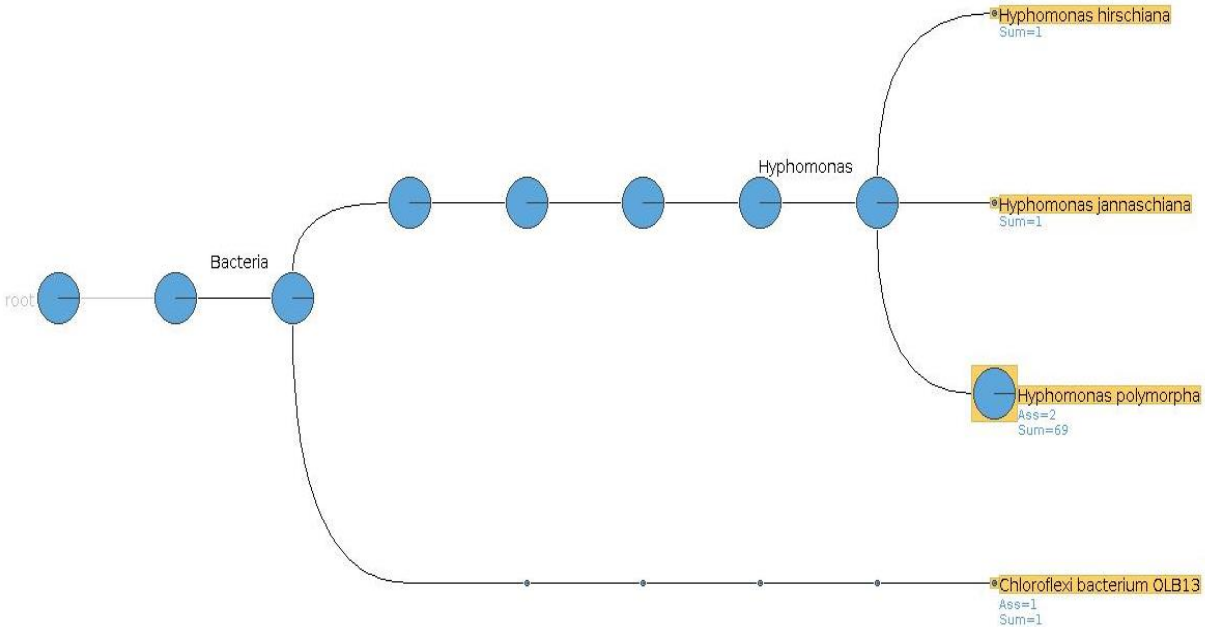

Figura 14: MEGAN: classificação taxonômica *bin2,* espécie em evidência: *Hyphomonas polymorpha*.

A Figura 15, mostra o gênero *Alphaproteobacteria*, que se ramifica em duas espécies *Rhizobiales* e *Rhodospirillaceae* em igual quantidades. Este encontra-se com um total de contigs igual a 756.

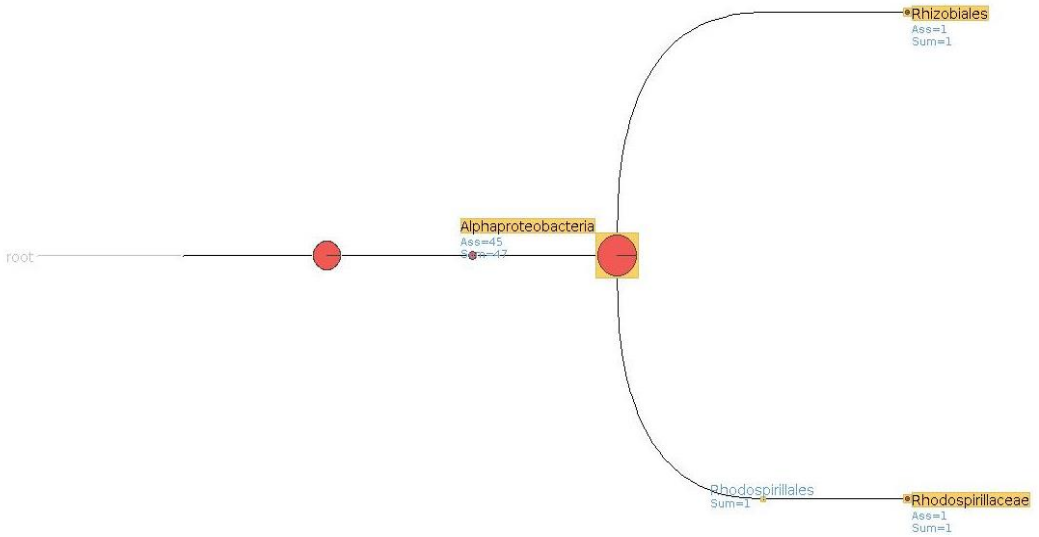

Figura 15: MEGAN: Classificação taxonômica *bin3,* gêneros: *Rhizobiales* e *Rhodospirillaceae*

Dentre todos os resultados encontrados na classificação, o filo *Proteobacteria* foi o mais ramificado (Figura 16) mostrando em quais gêneros o genoma pode ser classificado. Entre estes gêneros observa-se uma maior evidência para *Acetobacteria.*

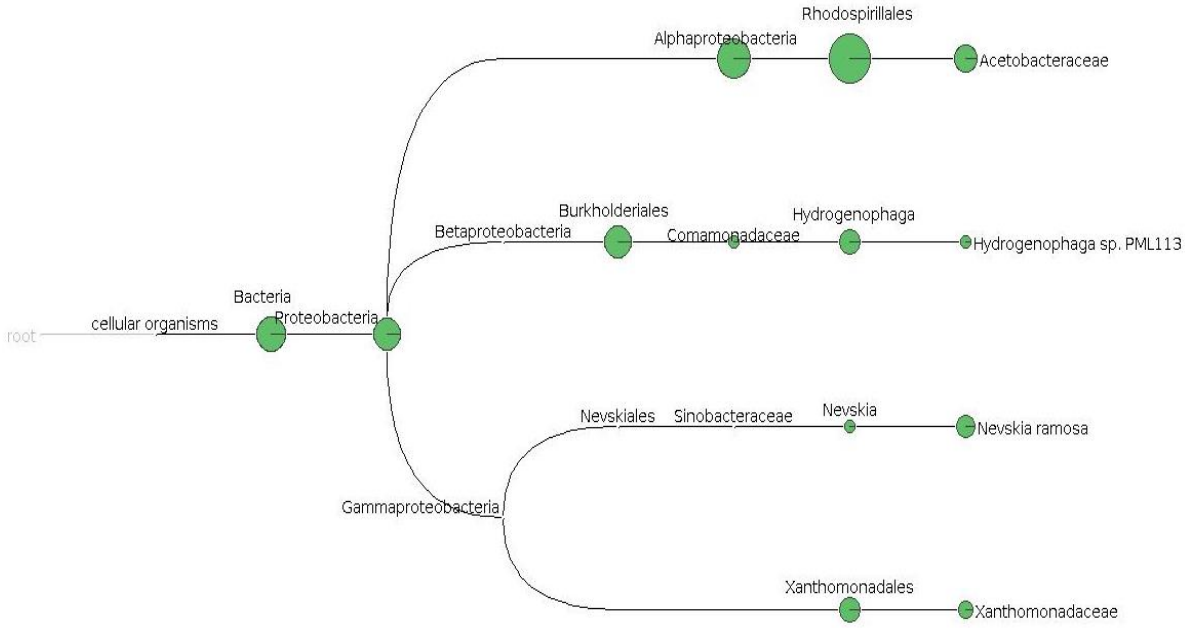

Figura 16: MEGAN: classificação taxonômica *bin4*, gênero em evidência: *Acetobacteraceae*

A Figura 17, mostra classificação taxonômica para espécie *Belnapia* sp. F-4-1, sem contaminações. Com um total de *contigs* igual a 1668.

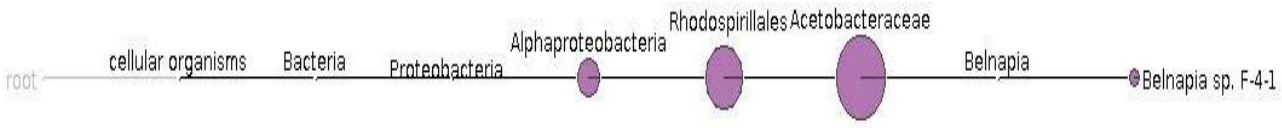

Figura 17: Classificação taxonômica do *bin5* MEGAN, espécie: *Belnapia* sp. f-4-1

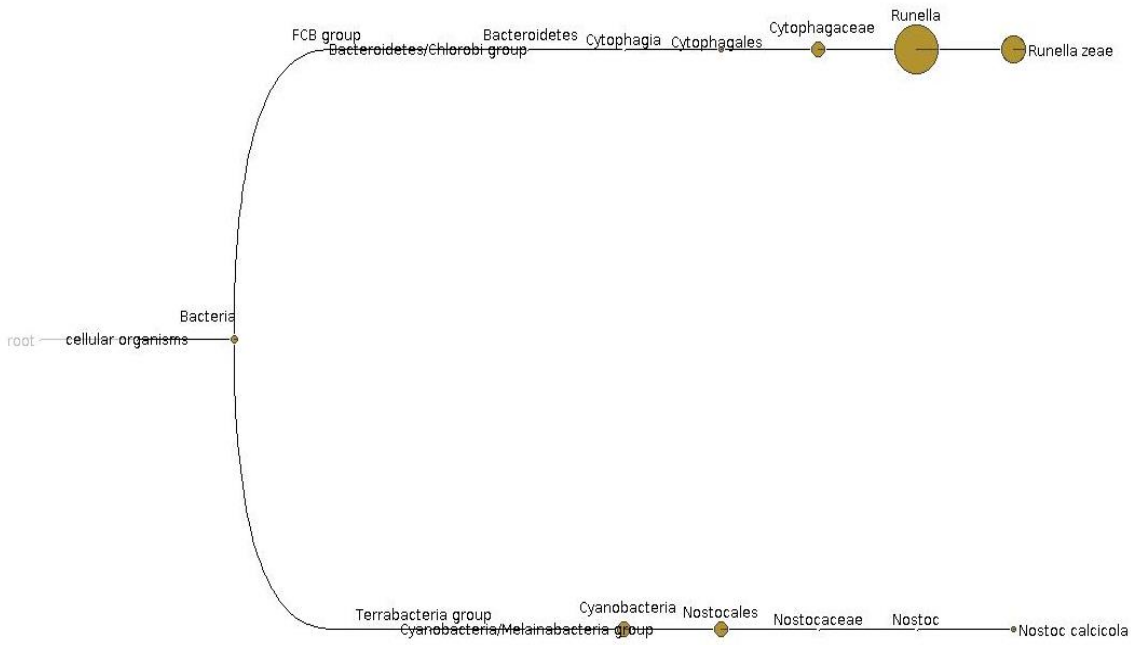

Figura 18: Classificação taxonômica resultante do MEGAN, presença do genoma de cianobactéria: *Nostoc calcícola*.

Observa-se que a cianobactéria *Nostoc* foi encontrada no ultimo *bin* (Figura 18), contendo um total de *contigs* igual a 2488. Contudo, o genoma encontra-se contaminado pela espécie *Runella* zeae em grande quantidade. Diante dos resultados obtidos, é possível comprovar a heterogeneidade da amostra de cianobactéria *Nostoc* sp. CACIAM 19 caracterizando-a como uma cultura não-axênica.

5. CONSIDERAÇÕES FINAIS

O desenvolvimento de sequenciadores de alto desempenho possibilitou descobrir, em larga escala e em processo único, informações estruturais, metabólicas e evolutiva da amostra. Em virtude deste avanço se fez necessário a utilização de programas computacionais capazes de processar sequenciamentos de grandes volumes de dados e assim realizar o processo de montagem, classificação e anotação de genomas.

O processo de tratamento de qualidade, que é necessário para a obtenção de uma montagem mais eficiente, das *reads* obteve o melhor resultado utilizando a ferramenta FASTX-TOOLKIT, pois manteve a maior parte da amostra pelo fato de utilizar um percentual mínimo para manter de bases. Enquanto o MOTHUR calcula uma média para cada *reads* comparando-a com o valor definido e isto não garante que apenas *reads* de baixa qualidade estejam sendo eliminadas. Entre os valores determinados para qualidade, o que obteve melhor resultado foi igual a vinte, pois este manteve a amostra eliminando apenas *reads* de baixa qualidade.

Dentre as abordagens de montagem utilizados para a realização deste trabalho, o algoritmo OLC mostrou-se mais efetivo que o grafo de Bruijn, pois neste há um alinhamento de *reads* explicito enquanto no grafo de Bruijn a relação de sobreposição por *k-mer* torna o alinhamento implícito gerando uma quantidade maior de *contigs*. Deste modo, o montador Newbler 2.9, com o parâmetro de sobreposição igual a quarenta, revelou-se mais eficiente para a montagem de uma cultura não-axênica.

Desta forma, com a separação dos genomas foram encontrados seis genomas em abundância, sendo um deste genoma de cianobactéria. Isto mostra a importância da utilização de técnicas metagenômicas para a reconstrução de genomas em cultura não-axênica.

Como trabalhos futuros, pretende-se realizar a anotação dos genomas encontrados, com intuito de identificar e caracterizar as regiões funcionais presentes na amostra, realizar a ordenação dos *contigs* a partir da comparação com as sequências de referência. Pretende-se também verificar se os contaminantes definidos na classificação taxonômica são de fato contaminantes, identificando os genes presentes para ver a possibilidade de não serem e se forem confirmados como contaminantes, identificar os *contigs*.

6. REFERÊNCIAS

454 LIFE SCIENCES; Newbler Assembler [página da internet]; Disponível em http://www.454.com/products/analysis-software

ALVARENGA, D. Análise Genômica e funcional da cianobactéria *Nostoc* sp. CENA67 e caracterização da sua comunidade microbiana associada. USP, 2015

ARAÚJO, N. et al.; Era da Bioinformática: Seu potencial e suas implicações para as ciências da sáude. PUCPR, 2008.

ANDREWS, S.; Fastqc: a quality control tool for high throughtput sequence data. Babraham bioinformatics. 2010.

ARUMUGAM, M. et al. Smashcommunity: a metagenomic annotation and analysis tool. Bioinformatics; 2010.

Brady, A. & SALZBERG, S. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nature Methods, 2009.

CAMPOS, F. & DUARTE, L.; Cianobactérias: um panorama geral; EcoDebate, 2011.

CARMICHAEL, W.. Cyanobacteria secondary metabolites-the cyanotoxins. Journal of Applied Bacteriology, USA, 1992

CARVALHO, M & SILVA, D.; Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas; Ciência Rural vol.40 no.3 Santa Maria Mar. 2010.

CODD, G. & BELL, S. Eutrophication and toxic cyanobacteria in freshwater. Walter Pollut, 1985.

Davies K. Decifrando o genoma: a corrida para desvendar o DNA humano. Companhia das Letras, 2001

DITTMANN, E. & WIEGAND, C.; Cyanobacterial toxins-occurrence, biiosynthesis and impact on human affairs. Mol Nutr Food Res. Vol.50, p.7-17, 2006.

FALEIRO, F. G.; ANDRADE, S. R. M. & REIS JUNIOR, F. B.; Biotecnologia estado da arte e aplicações na agropecuária. Embrapa. 1ª edição. Cap. 6. p. 175-193, , 2011.

FARIAS, A.; CHACON, P. & SILVA, N. A bioinformatica como ferramenta de formação de recursos humanos no IFRN. IFRN- HOLOS 2011;

FRAGA, J.S.; Algoritmos genéticos e o problema da montagem de reads; UFMS;2014.

GAULT, P.M. & MARLER, H.J. 2009. Handbook on Cyanobacteria: biochemistry, biotechnology, and applications. Nova Science Publishers, New York, 538 p.

GERLACH, W.; STOYE, J. Taxonomic classification of metagenomic shotgun sequences with carma3. Nucleid Acids Research; v. 39; n. 14; 2011.

GUPTA, R.; BEG, Q. K.; LORENZ, P. Bacterial alkaline proteases: molecular approachesand industrial applications. Applied Microbiology and Biotechnology, 2002.

HANDELSMAN, J. Metagenomics: Aapplication of genomics to uncultured microorganisms.Microbiology and molecular biology reviews; v. 68; n. 4; p. 669–685; 2004.

HANNON LAB. FASTX TOOLKIT [pagina da internet]. Disponível em http://hannonlab.cshl.edu/fastx_toolkit/download.html, 2012.

HAQUE, M. M. et al. Sort-items: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics; v. 25; p. 1722–1730; 2009.

HORODESKY, A; Sequenciamento de nova geração para analises metagenômicas: enfoque ao uso do sequenciador ILLUMINA. GIA, novembro 2014.

HUSON, D. et al. MEGAN analysis of metagenomic data. *Genome research*, 17(3): 377-386, 2007

KUNIN, V.; COPELAND, A.; LAPIDUS, A.; MAVROMATIS, K. AND HUGENHOLTZ, P. A Bioinformatician's Guide to Metagenomics. Microbiology and Molecular Biology Reviews, 2008.

LEMOS, L; Reconstrução e análise de genomas de bactérias de compostagem a partir de dados metagenômicos; USP, 2015.

LEMOS, M., BASÍLIO,A. &CASANOVA, M.A.; Um estudo dos algoritmos de montagem de fragmentos de DNA; PUC-RJ, 2003.

Li, Z.; Chen, Y.; et al. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-buijn-graph. Briefings in Functional Genomics. Vol II Nª I. p 25-37. 2011.

LIMA, A.R.J.; Potencial biotecnológico de cianobactérias amazônicas para a produção de hidrocarbonetos: da montagem de genomas à modelagem comparativa. UFPA, 2015.

LUSCOMBE; GREENBAUM &GERSTEIN; What is Bioinformatics? A proposed Definition and Overview of the Field, 2001.

MICALLEF; Exploring Cyanobacterial genomes for natural product biosyntheis pathways. Mar Genomics, Junho 2015.

MILLER, J. R. et al. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics, v. 24, p. 2818–2824, 2008.

MILLER, J., KOREN, S. & SUTTON, G.; Assembly Algorithms for next-generation sequencing Data. Genomics v 95. p. 315-327, 2010.

MISCHKE, U. Cyanobacteria associations in shallow polytrophic lakes: influence of environmental factors. Acta Oecologica, 2003.

NETO, H. Classificação de sequências metagenômicas. UFMT 2012.

PEIXOTO, B.; Bioinformática aplicada a um projeto de metagenômica. UNICAMP, 2011.

PEIXOTO, B. Classificação de sequências e análise de diversidade em metagenômica. UNICAMP, 2013

PEREIRA, V. Montagem e análise de genomas a partir de metagenomas. Universidade de São Paulo. Novembro,2014

PESSOA FILHO, M. A. C. de P. Metagenômica e sua aplicação no estudo de Diversidade e função de microrganismos de solos do cerrado. EMPRABA, 2010.

PINOTTI, M. & SEGATO, R. cianobactérias importância econômica; 1991

PIRO, V. Desenvolvimento da ferramenta para finalização de montagensde genomas *in silico* – FGAP. 2014.

PROSDOCIMI, F. Introdução à bioinformática. Biotecnologia ciência e Desenvolvimento, 2007.

RAJENDHRAN,J. E GUNASEKARAN,P.; Strategie for acesseing soil metagenome for desired applications, Biotechnol Adv. 2008.

REVIERS, B. Biologia e filogenia das algas. Artmed, Porto Alegre, p. 21. 2006.

RIPPKA, R. Isolation and purification of cyanobacteria. Methods in enzymology, v. 167, p. 3–27, 1988.

RODRIGUES, J. AS algas, as cianobactérias e a biotecnologia. FCiências, 2016.

SANGER, F. & COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology, v.94, n.3, p.441-448, 1975.

SANT'ANNA, C.L, AZEVEDO, M.T.P., WERNER, V.R., DOGO, C.R., RIOS, F.R. &

CARVALHO, LR. 2008. Review of toxic species of Cyanobacteria in Brazil. Algological Studies.

SCHBATH, S., PRUM, B. & TURCKHEIM, E. . Exceptional motifs in different Markov chain

models for a statistical analysis of DNA sequences. Journal of computational biology.

SCHLOSS, P.D. & HANDELSMAN, J. Biotechnological prospects from metagenomics. Curr Opin Biotechnol. p.303-310. 2003.

SCHLOSS, P; Mothur [página da internet]. Disponivel em http://www.mothur.org/.; 2009

SCHLOSS, P.; WESTCOTT, S. et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2009, p. 7537–7541

SHARON, I.; BANFIELD, J. F. Genomes from metagenomics. Science; v. 342; p. 1057–1058; 2013

SILVA, L. & KREMER, F. Plataformas de sequenciamento de nova geração e pré-processamentos de dados. UFPel, 2016.

SIMPSON, J.; WONG, K.; JACKMAN, S.; SCHEIN, J.; JONES, S. & BIROL,I.; ABySS: A parallel assembler for short read sequence data.; Genome Research; 2009.

SIQUEIRA, D. & OLIVEIRA-FILHO, E. Cianobactérias de água doce e saúde pública: uma revisão. *Universitas* Ciências da Saúde- Vol.03 n.01- pp. 109-1272005

SOUZA, R. Análise de expressão diferencial em transcriptomas. Curso de Introdução à bioinformática aplicada a genômica, UTFPR 2015

Universidade Federal da Bahia- GENEBIO: Genética e Bioinformática [página da internet]. BLAST. Disponível em http://www.genebio.ufba.br/?page_id=260

VARUZZA, L; Introdução à análise de dados de sequenciadores de nova geração. Abril 2013

VENTER, J. C., ADAMS; M. D.; MYERS, E.; LI, P.W.; MURAL, R. J.; SUTTON, G. G.; SMITH, H. O.;et al. The sequence of the human genome. Science, 2001.

WOOLEY, J.C.; GODZIK,, A. & FRIEDBERG, I.; A Primer on Metagenomics. PLoSComput Biol, 2010.

WU, Y.W., TANG,Y.H., TRINGE,S.G., SIMMONS,B.A. & SINGER,S.W.; Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome, 2014.

ZERBINO, D. R. & BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research, v. 18, p. 821–829, 2008.