UNIVERSIDADE FEDERAL DO PARANÁ

EDUARDO TIEPPO

MONTAGEM E ANÁLISE PRELIMINAR DO GENOMA DE *Bradyrhizobium elkanii* 587 UTILIZANDO LEITURAS DE SEQUÊNCIAS DE DNA CURTAS

CURITIBA

2011

EDUARDO TIEPPO

MONTAGEM E ANÁLISE PRELIMINAR DO GENOMA DE *Bradyrhizobium elkanii* 587 UTILIZANDO LEITURAS DE SEQUÊNCIAS DE DNA CURTAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de Mestre em Bioinformática.

Orientador:
Emanuel Maltempi de Souza, Dr.
Co-orientador:
Lucas Ferrari de Oliveira, Dr.

CURITIBA

2011

TERMO DE APROVAÇÃO

EDUARDO TIEPPO

Montagem e análise preliminar do genoma de *Bradyrhizobium elkanii* 587 utilizando leituras de sequências de DNA curtas

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Prof. Dr. Emanuel Maltempi de Souza

Coorientador: Prof. Dr. Lucas Ferrari de Oliveira

Prof. Dr. Leonardo Magalhães Cruz
Universidade Federal do Paraná

Prof. Dr. Jackson Antônio Marcondes de Souza
Faculdade de Ciências Agrárias e Veterinárias – FCAV/UNESP

Curitiba, 25 de fevereiro de 2011

Dedico este trabalho a Deus,
por ele ter sido o primeiro cientista a conseguir
publicar os resultados sem revelar a sua
metodologia.

AGRADECIMENTOS

Aos meus orientadores:

Professor Dr. Emanuel Maltempi De Souza,

Pela experiência compartilhada e pelo exemplo que procurarei seguir na carreira científica; e,

Professor Dr. Lucas Ferrari De Oliveira,

Pela sempre sensata opinião e pelos conselhos. Também, muito obrigado pela amizade construída durante este desenvolvimento. Tenho em você um exemplo de caráter no meio profissional; vou me lembrar e seguir suas dicas e atitudes quando orientar meus próprios alunos.

Ao doutorando Vinicius Weiss:

Pela paciência ao responder perguntas nem sempre inteligentes feitas por mim.

Ao Felipe Renó:

Pelo auxílio e disponibilização do sistema GAAT.

Ao sempre mestre e amigo Dieval:

Que sempre esteve disposto a ajudar na busca por soluções ou lamentar a falta delas. Sua mente é a mais brilhante que conheço.

À minha amiga de sempre Michelly:

Pela companhia durante toda esta caminhada até mesmo compartilhando os piores problemas.

Obrigado também, minha amiga, por sempre acreditar na minha capacidade. Você é a responsável por eu estar aqui.

Ao meu amigo Rodrigo:

Pela amizade, acima de tudo.

Também pela companhia durante todo esse percurso, compartilhando os problemas de montagem, e, quase sempre, a falta das soluções; e pelo auxílio incondicional sobre qualquer besteira dita por mim sobre as ciências biológicas.

Obrigado também pelas centenas de horas de *Flash Games*, *Gartic* e futebol.

À equipe de desenvolvimento *Gartic*:
Por proporcionar horas de entretenimento e contribuir para a formação de mentes criativas.

Ao *read beje0167B12.b00*:
Em nome de todos os *reads*, que sozinhos, ancorados apenas por suas extremidades, fecharam grandes *gaps*.

Às secretárias da Bioinformática Suzana, Mariana e Léa:
Pelo auxílio absoluto sempre que necessário. Muito obrigado pelo empenho em resolver as minhas questões sempre com um sorriso no rosto.

Ao programa de Pós-Graduação em Bioquímica:
Pelo acompanhamento e compartilhamento de informações e experiência durante este desenvolvimento.

Ao programa de Pós-Graduação em Bioinformática:
Pelo suporte em todos os assuntos e infraestrutura cedida.

À CAPES:
Pelo auxílio financeiro.

Aos meus amigos:
Lucas,
Pela amizade, compreensão e admiração que sempre me deram força para ir além; e,
Ani,
Pela amizade e pelas sessões gratuitas e mútuas de psicologia.

À minha namorada Hellen:
Por transformar os últimos meses deste trabalho, normalmente estressantes e nervosos, em meses sempre leves, alegres e apaixonados.
Obrigado, meu amor!

Aos meus familiares:

Leonardo,

Simplesmente pela amizade absoluta; e,

Nani,

Pela amizade e pela companhia nas horas de distração.

À minha irmã Paula:

Por ser meu exemplo em toda minha vida acadêmica. Obrigado por, desde sempre, me ensinar como eu deveria caminhar.

Aos meus pais, Claudio e Berenice:

Que fizeram sempre todo o possível para que eu pudesse ser o que quisesse. Sem eles nada seria possível.

A Deus:

Por tudo.

“E depois de tantos anos
Só decepções, desenganos
Dizem que sou um burguês
Muito privilegiado
Mas burgueses são vocês
Eu não passo
De um pobre coitado
Mas quem quiser ser como eu
Vai ter é que penar um bocado...”

O Pequeno Burguês
Martinho da Vila

RESUMO

A fixação biológica de nitrogênio é um processo essencial para a vida na Terra e é realizada apenas por procariotos. Entre os fixadores de nitrogênio, as bactérias da ordem *Rhizobiales* são capazes de formar simbiose com leguminosas utilizadas na alimentação humana e animal, como soja e feijão. Nesta associação, a bactéria sofre diferenciação para uma forma bacterióide e ocupa um órgão específico, chamado de nódulo, onde ocorre a fixação de nitrogênio fornecendo os nutrientes necessários à planta. Em termos econômicos, associação de soja e *Bradyrhizobium* é a mais importante deste tipo de simbiose, representando uma economia estimada para os agricultores brasileiros de cerca de 1,7 bilhão de dólares por ano. Neste estudo, foi obtida a sequência preliminar do genoma de *Bradyrhizobium elkanii* estirpe 587, utilizada como inoculante para soja no Brasil. Cerca de 6 milhões de leituras Illumina-Solexa, representando uma profundidade de cobertura entre 28x e 32x, foram montadas usando o montador Velvet. Além disso, aproximadamente 40 mil leituras Sanger (profundidade de cobertura entre 4x e 5x) foram montadas com o montador Phrap. Os *scaffolds* produzidos pelo montador Velvet foram alinhados utilizando o programa MUMmer e o genoma de *Bradyrhizobium japonicum* USDA 110 como referência. Finalmente, as falhas foram fechadas manualmente utilizando Consed.

Palavras-chave: Sequenciamento de DNA, montagem genômica, *Bradyrhizobium elkanii*, fixação biológica de nitrogênio.

ABSTRACT

The biological nitrogen fixation is an essential process for life on earth and is carried out only by prokaryotes. Among the nitrogen-fixers, bacteria of the order *Rhizobiales* are able to form symbiosis with legume plants used as human and animal food such as soybean and common bean. In this association, the bacteria suffers differentiation to a bacterioid form and occupy a specific organ called nodule, where it fixes nitrogen supplying the plant needs. The association soybean-*Bradyrhizobium* is the most important of such symbioses economically, representing an estimated economy for Brazilian farmers of approximately 1.7 billion dollars per year. In this study we obtained the draft genome sequence of *Bradyrhizobium elkanii* strain 587 which is used as soybean inoculant in Brazil. Approximately 6 million Illumina-Solexa reads, representing a coverage depth between 28x and 32x, were assembled using Velvet. Additionally, approximately 40.000 Sanger reads (coverage depth between 4x and 5x) were assembled with Phrap. The scaffolds produced by Velvet were aligned using the program MUMmer 3.0 and the genome of *Bradyrhizobium japonicum* USDA 110 as reference. Finally, gaps were manually closed using Consed.

Keywords: DNA Sequencing, genomic assembly, *Bradyrhizobium elkanii*, biological nitrogen fixation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 *Bradyrhizobium elkanii*

Atualmente, o termo rizóbio é empregado para designar bactérias capazes de fixar nitrogênio em associação com leguminosas através de nodulação (WILLEMS, 2006). A bactéria *Bradyrhizobium elkanii* pertence à Ordem *Rhizobiales*, caracterizada por bactérias aeróbias, gram-negativas e diazotróficas. Em um nível taxonômico inferior, *B. elkanii* pertence à Família *Bradyrhizobiaceae* e ao Gênero *Bradyrhizobium*, e tem como característica crescimento lento (JORDAN, 1982; GIONGO, 2007).

A taxonomia de *B. elkanii* apresenta-se da seguinte maneira (KUYKENDALL *et al.*, 1992; RUMJANEK *et al.*, 1993; GARRITY *et al.*, 2005):

- Domínio: *Bacteria*
- Filo: *Proteobacteria*
- Classe: *Alphaproteobacteria*
- Ordem: *Rhizobiales*
- Família: *Bradyrhizobiaceae*
- Gênero: *Bradyrhizobium*
- Espécie: *Bradyrhizobium elkanii*

Bactérias do gênero *Bradyrhizobium* são comumente encontradas no solo e realizam a fixação biológica de nitrogênio principalmente em simbiose com plantas da Família *Fabaceae* (SPRENT, 1995). Tal gênero possui ainda um interesse agronômico, pois associa com diversas leguminosas, incluindo a soja, contribuindo para o crescimento da planta e aumentando a produtividade de grãos. Além do interesse agronômico, tais bactérias são alvos comerciais, já que algumas estirpes de *Bradyrhizobium japonicum* e *B. elkanii*, por exemplo, funcionam como inoculantes para a soja, favorecendo o desenvolvimento das plantas (JORDAN, 1982; ALBERTON *et al*., 2006; GIONGO, 2007).

1.1.1 Fixação Biológica de Nitrogênio

O processo de fixação biológica de nitrogênio é realizado por organismos chamados diazotróficos, e consiste na redução de nitrogênio atmosférico (N_2) em amônia (NH_3), reação catalisada pela enzima nitrogenase. A amônia produzida pode então ser utilizada pelo metabolismo dos organismos (GIONGO, 2007).

Os organismos diazotróficos mais estudados são aqueles capazes de estabelecer simbiose com plantas leguminosa, e, entre estes, os mais importantes são os da ordem *Rhizobiales*. Os rizóbios ganham atenção especial por fixar nitrogênio associados a plantas e são localizados em estruturas diferenciadas chamadas nódulos. A amônia produzida pelo rizóbio é exportada para a planta que não depende mais da disponibilidade deste nutriente no ambiente (RHIJN *et al*., 1995; SADOWSKY *et al*., 1995; GIONGO, 2007).

As culturas de plantas precisam de fontes de nitrogênio para o seu desenvolvimento. A simbiose entre rizóbios e leguminosas provê o nitrogênio requerido pela planta. Para a cultura de soja, por exemplo, são necessários 80 quilogramas de nitrogênio por hectare para a produção de uma tonelada de grãos (HUNGRIA *et al*., 2005). A fixação biológica de nitrogênio e os fertilizantes são as duas fontes de nitrogênio disponíveis para a soja (HUNGRIA *et al*., 2001), mas a primeira dispensa o uso de fertilizantes, não agride o meio ambiente, e ainda resulta na economia de quase R$ 3 bilhões anualmente no Brasil (MENNA *et al*., 2006; GIONGO, 2007).

1.2 Sequenciamento e montagem genômica

Através do sequenciamento e montagem genômica é possível determinar a sequência de bases do genoma de uma bactéria, e determinar características relevantes para a compreensão das ações biológicas (STEIN, 2001).

O processo de sequenciamento e montagem genômica, em termos gerais, refere-se ao conjunto de instrumentos, dispositivos, protocolos e métodos direcionados à determinação da sequência de DNA componente dos organismos, desde a coleta das amostras e obtenção de dados, até o tratamento destes dados e a finalização da sequência (CHAN, 2005).

A FIGURA 1.1 mostra as etapas de sequenciamento e montagem genômica. Duas fases podem ser observadas: (i) a fase experimental e (ii) a análise computacional. Na primeira, as bactérias são cultivadas e seus DNAs extraídos; o genoma alvo do sequenciamento é fragmentado e os fragmentos são clonados; por fim, ocorre o sequenciamento destes. Na segunda fase, as sequências individuais ou leituras são ordenadas e montadas por sobreposição parcial de suas extremidades; estratégias de finalização destas sequências são executadas a fim de obter uma sequência única representando o genoma; e, finalmente, genes e regiões de interesse são buscados e anotados para poderem servir de base para novos estudos.

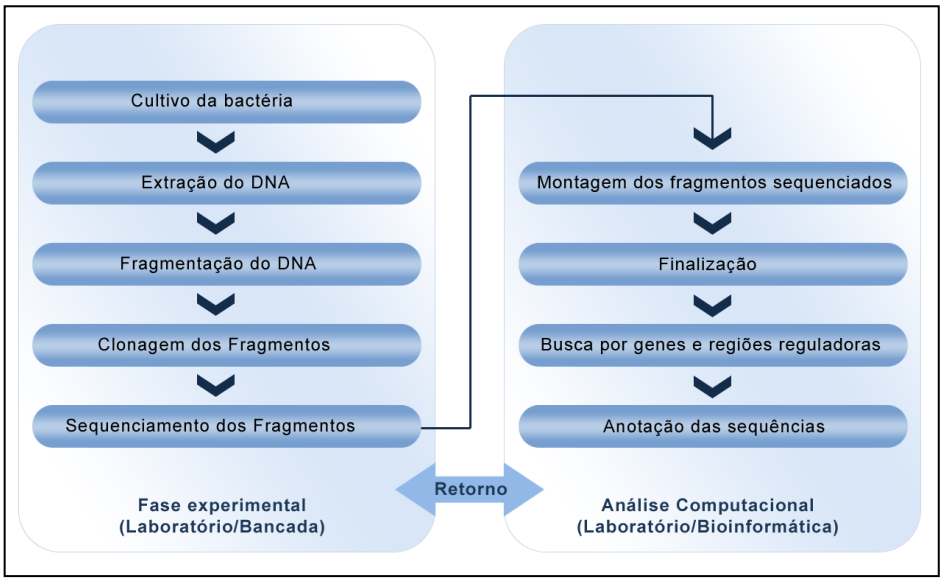

FIGURA 1.1 - ETAPAS DE UM PROCESSO DE SEQUENCIAMENTO E MONTAGEM GENÔMICA
FONTE: Adaptada de GENOPAR. Disponível em: <nfn.genopar.org/nfn/genopar>. Acesso em: 29/11/2010.

1.2.1 Sequenciamento genômico

O sequenciamento genômico e a extração de características provenientes de um conjunto de genes são fatores essenciais da pesquisa biológica (STEIN, 2001). Esses estudos possuem diversos objetivos tais como descobertas de rearranjos genômicos (MOROZOVA *et al*., 2008); análise de transcriptomas (ASMANN *et al*., 2008); testes de diagnósticos moleculares (VOELKERDING *et al*., 2009); variabilidade genética em diversos organismos (IMELFORT *et al*., 2009), descoberta

de vacinas (DHIMAN *et al*., 2009); melhoramento genético de espécies (VARSHNEY *et al*., 2009), *etc*. Estes estudos só são realizáveis devido às tecnologias de sequenciamento de DNA, que vêm sendo aprimoradas nas últimas décadas (SHENDURE *et al*., 2008; FLEISCHMANN *et al*., 1995).

Allan Maxam e Walter Gilbert introduziram em 1976 um método de sequenciamento baseado em modificação química e subsequente clivagem de bases específicas (MAXAM *et al*., 1977; CHEN, 2008).

Um ano depois, Sanger apresentou um método de sequenciamento enzimático baseado na interrupção do crescimento da cadeia polinucleotídica na presença de análogos dos nucleotídeos. Tal método vem sendo aprimorado desde a sua criação e a maioria das sequências de DNA produzida até hoje é oriunda do método Sanger original ou com algumas variações (SANGER *et al*., 1977; SWERDLOW *et al*., 1990; HUNKAPILLER *et al*., 1991).

Além disso, o desenvolvimento de novas tecnologias de sequenciamento tem sido incentivado por diversos fatores, como redução de custos, maior potencial de sequenciamento, necessidade de estratégias mais práticas e realizáveis para a obtenção das sequências de DNA, entre outros (SHENDURE *et al*., 2008; SHENDURE *et al*., 2004). Diferentes métodos de sequenciamento de DNA foram descritos: através de eletroforese, hibridização, observação em tempo real de moléculas e sequenciamento por vetor cíclico (SHENDURE *et al*., 2004).

Os dois métodos de sequenciamento utilizados para obtenção dos dados deste trabalho utilizaram uma abordagem chamada W*hole-Genome Shotgun* (WGS), também introduzida por Sanger, em 1977. Esta estratégia consiste no sequenciamento de fragmentos estocásticos do DNA, resultando em leituras desordenadas e necessitando de técnicas computacionais para montagem e reconstrução da sequência inicial de DNA (BATZOGLOU *et al*., 2002).

1.2.1.1 Sequenciamento pelo método Sanger

O método Sanger foi introduzido na década de 70 por Sanger e colaboradores.

A FIGURA 1.2 ilustra uma variação do método Sanger utilizada por modernos sequenciadores automáticos. Em um pré-processo, a bactéria é cultivada e seu DNA extraído. Na FIGURA 1.2 (1) o DNA é fragmentado através de um processo de

nebulização. Em (2), os fragmentos de DNA são clonados *in vivo* em um vetor de plasmídeo, resultando em uma biblioteca de fragmentos de DNA. Em (3), os clones são individualmente sequenciados utilizando anelamento e extensão dos *primers*, DNA polimerase, dideoxinucleotídeos, e didesoxinucleotídeos marcados com fluoróforos. A DNA polimerase utiliza como DNA molde o fragmento clonado e ao incorporar um didesoxinucleotídeo marcado, a replicação do DNA é interrompida, resultando em fragmentos parcialmente polimerizados representando partes do fragmento inicial. Em (4), os fragmentos derivados da polimerização parcial são separados e ordenados por tamanho através de uma eletroforese capilar, onde o DNA migra de uma ponta capilar a outra de acordo com seu tamanho e passa por um detector a laser que excita o fluoróforo marcador do didesoxinucleotídeo, resultando na emissão de um sinal e gerando um arquivo de faixas de sequenciamento de acordo com o tipo de sinal emitido, possibilitando a identificação das bases componentes do fragmento inicial e a criação de uma sequência ou leitura (SANGER *et al*., 1977; CHEN, 2008; SHENDURE *et al*., 2008).

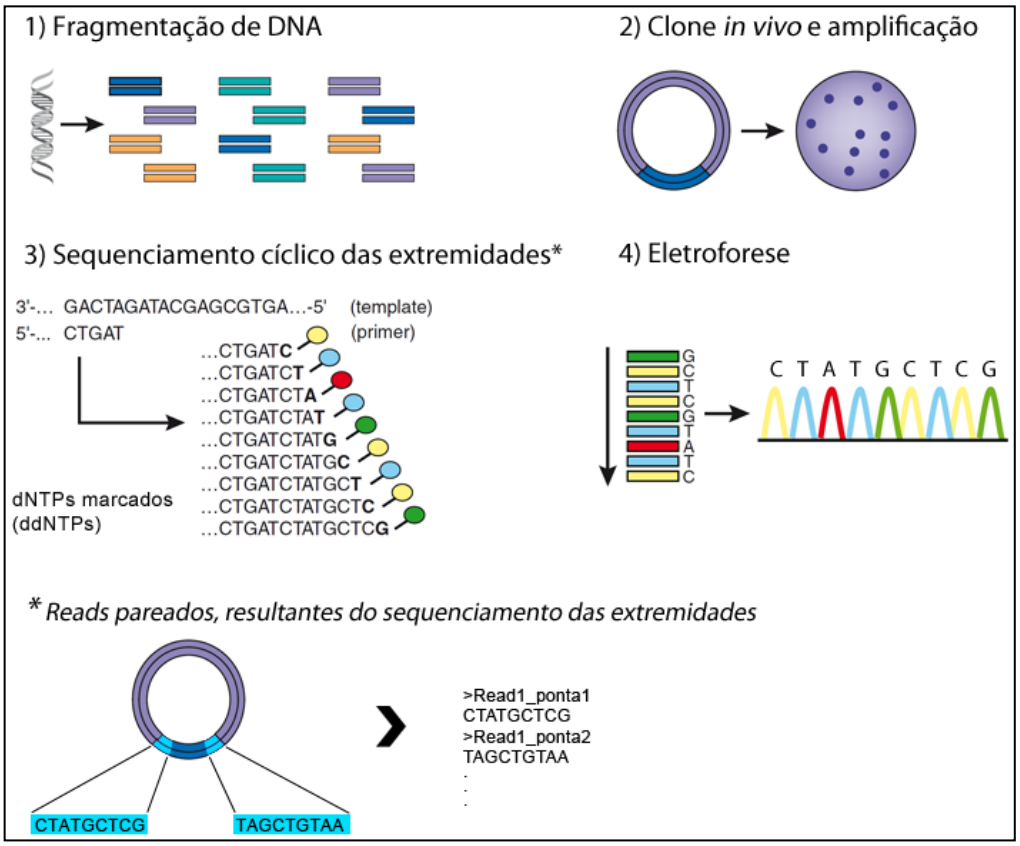

FIGURA 1.2 - SEQUENCIAMENTO PELO MÉTODO SANGER EM SEQUENCIADORES AUTOMÁTICOS

FONTE: Adaptação (SHENDURE *et al*., 2008)

1.2.1.2 Sequenciamento pelo método de vetor cíclico

O sequenciamento através do método de vetor cíclico foi descrito em 2005 por Jay Shendure e Margulies (SHENDURE *et al*., 2005; MARGULIES *et al*., 2005), e na sequência por Bentley e Harris com trabalhos de ressequenciamento genômico e sequenciamento genômico viral, respectivamente (BENTLEY, 2006; HARRIS *et al*., 2008).

O método consiste, basicamente, em ciclos de manipulação enzimática e aquisição de dados baseada em imagem; e, apesar desta estratégia ter diferenças em relação ao método Sanger, o sequenciamento e a obtenção das sequências em ambas as estratégias são similares conceitualmente já que ambas se baseiam na síntese de DNA através do processo de replicação, sendo este interrompido com a incorporação de didesoxinucleotídeos marcados possibilitando o mapeamento de tais bases (SHENDURE *et al*., 2008).

A FIGURA 1.3 ilustra o processo. Em (1), o DNA é fragmentado através de um processo de nebulização. Em (2), os fragmentos de DNA são submetidos a uma PCR chamada *Bridge PCR* (ADESSI *et al*., 2000; FEDURCO *et al*., 2006), onde adaptadores são ligados aos fragmentos de DNA e então presos a uma placa; o processo de replicação resulta em cópias dos fragmentos localizadas próximas ao fragmento original, gerando, ao fim do processo de PCR, “colônias” de fragmentos contendo aproximadamente 1.000 cópias de um fragmento inicial. São incorporados à placa, em (3), *primers*, DNA polimerase, e didesoxinucleotídeos marcados com fluoróforos; em (4), ocorrem ciclos de síntese pela injeção de bases, uma a uma, em uma colônia de fragmentos, resultando na emissão de um sinal luminoso diferente de acordo com a base incorporada. Por fim, em (5), as imagens adquiridas a cada ciclo são analisadas para obtenção da sequência de cada “colônia” de DNA. Assim, a análise de tais imagens, através da comparação de posição das colônias de fragmentos e do sinal emitido, resulta na identificação das bases de cada fragmento e na consequente criação de uma leitura (SHENDURE *et al*., 2008). A maior diferença entre este método e o anterior é que agora não é mais necessário realizar a clonagem de DNA, representando uma grande redução da quantidade de trabalho e custo envolvidos no sequenciamento.

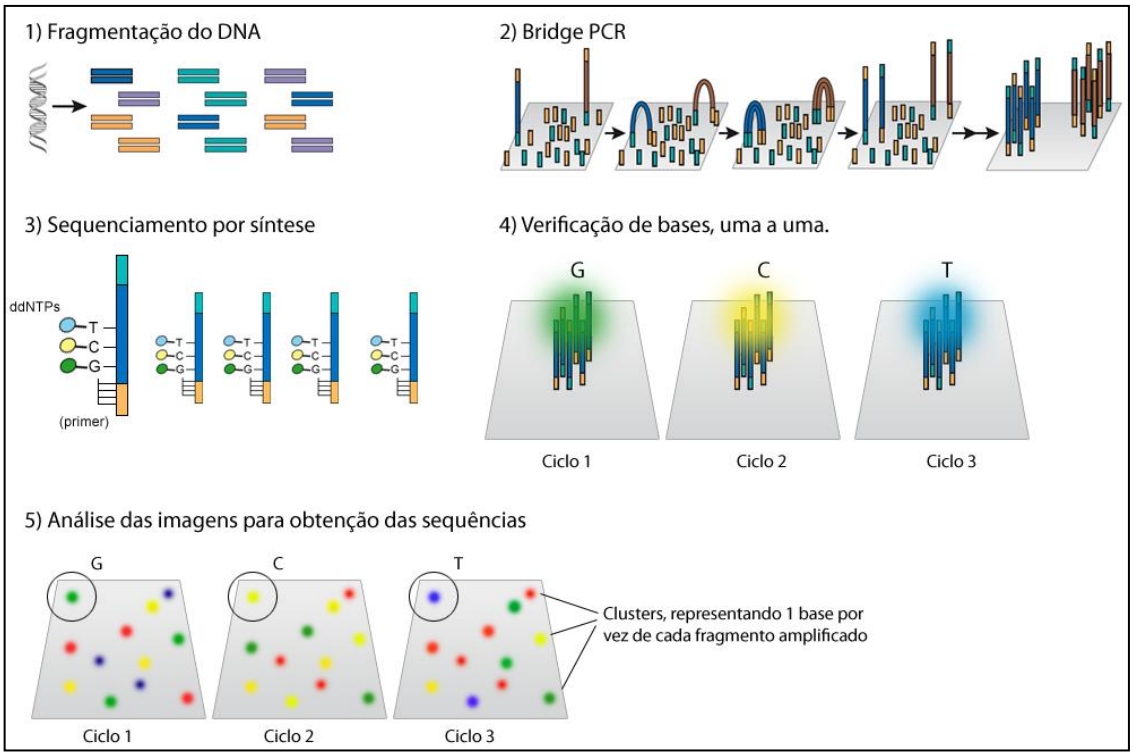

FIGURA 1.3 - SEQUENCIAMENTO PELO MÉTODO DE VETOR CÍCLICO
FONTE: Adaptação (SHENDURE *et al*., 2008)

Todos os passos mostrados acima são realizados por um único equipamento denominado Illumina Genome Analyzer. Essa plataforma teve origem no trabalho de Turcatti (FEDURCO *et al*., 2006; TURCATTI *et al*., 2008) e é resultado da junção de quatro companhias: (i) Solexa, (ii) Lynx Therapeutics, (iii) Manteia Predictive Medicine e (iv) Illumina (SHENDURE *et al*., 2008).

1.2.2 Montagem das leituras em uma sequência contígua

Independentemente do método de sequenciamento utilizado, o resultado é uma coleção de leituras de sequência de DNA cujo número pode ser de centenas, milhares, ou milhões, representando pequenas partes do genoma alvo do sequenciamento. A correta ordenação destas leituras e a criação de uma sequência única que represente o genoma inicial é um dos processos mais complexos dentre todas as etapas de sequenciamento e montagem genômica e envolve desde a sobreposição das leituras para confirmação e junção de informação, até o retorno para etapas anteriores do processo a fim de obter novos dados que eventualmente sejam necessários ou confirmar as informações (SHENDURE *et al*., 2008).

Outro desafio da montagem são sequências que contêm regiões exatamente iguais, conhecidas como repetições. A identificação destas regiões e posicionamento na sequência final é uma tarefa extremamente complexa já que, se tal região exceder o tamanho das leituras, nenhum fragmento vai conter a região inteira, não sendo possível orientá-la em relação ao resto das sequências (IDURY *et al*., 1995; LEMOS *et al*., 2003; HAVLAK *et al*., 2004). A FIGURA 1.4 exemplifica a criação de grafos de sobreposição quando a sequência inicial possui regiões repetitivas. Em (A), é mostrada uma sequência inicial formada por regiões únicas e três regiões repetitivas, e as leituras do sequenciamento; em (B), um grafo de sobreposição por consenso ilustra a complexidade da resolução das regiões repetitivas: cada círculo representa uma leitura e as setas ligações entre as leituras; todas as leituras que possuem partes da região repetitiva apontam para outras leituras que também possuem tal região, não permitindo a identificação do caminho a seguir.

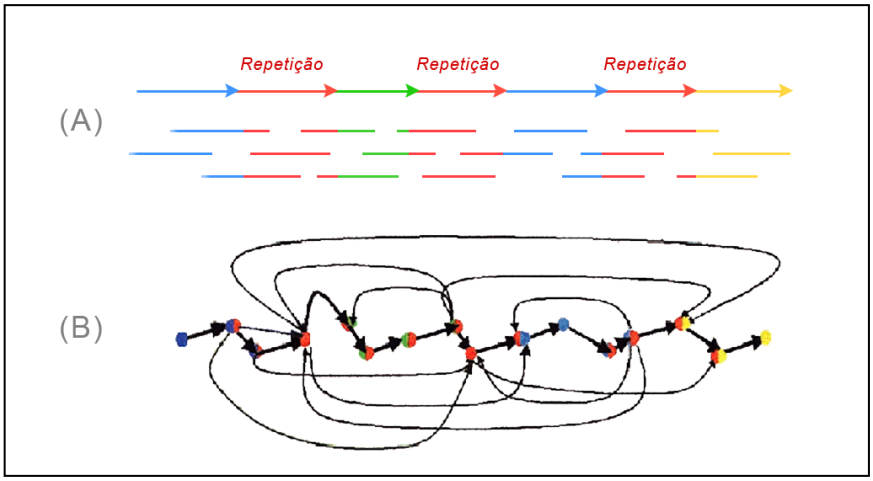

FIGURA 1.4 - EXEMPLO DE CRIAÇÃO DE UM GRAFO DE SOBREPOSIÇÃO COM REGIÕES REPETITIVAS
FONTE: Adaptação (LEMOS *et al*., 2003)

1.2.2.1 Etapas de um processo de montagem

A montagem da sequência genômica consiste na sobreposição das leituras oriundas do processo de sequenciamento e criação de sequências contíguas ou *contigs*; caso existam indícios de ligação física entre dois ou mais *contigs*, mas não

exista a sequência para completar as regiões indicadas pelos indícios de ligação, é possível ainda a criação de *scaffolds* (BATZOGLOU *et al*., 2002).

A FIGURA 1.5 e a FIGURA 1.6 ilustram os passos gerais do processo de montagem. Na FIGURA 1.5, em (A), está representado o conjunto de leituras oriundo do método de sequenciamento, os alinhamentos e sobreposições entre tais leituras e, por fim, a criação de *contigs* resultantes das leituras sobrepostas; em (B), um indício de ligação entre os dois *contigs* gerados em (A) é observado, pois as leituras em destaque são pareadas (como mostrado na lista dos conjuntos de leituras em (A)); em (C), uma falha (*gap*) tem seu tamanho estimado e é criada para representar a ligação entre os dois *contigs* (BATZOGLOU *et al*., 2002).

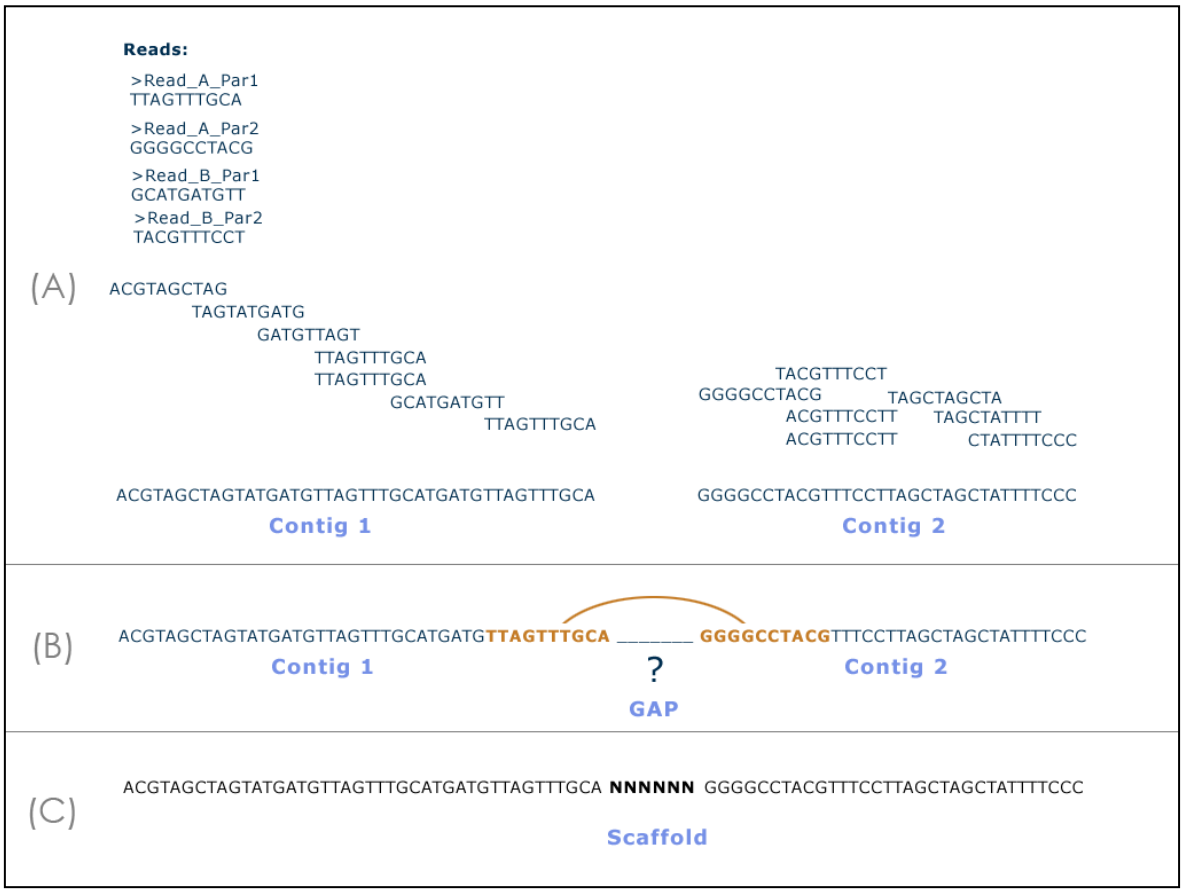

FIGURA 1.5 - PASSOS GERAIS DE UM PROCESSO DE MONTAGEM: SOBREPOSIÇÃO DE LEITURAS, *CONTIGS* E *SCAFFOLDS*

A FIGURA 1.6 completa a descrição do processo de montagem. Em (A), o *scaffold* criado na FIGURA 1.5 tem suas pontas estendidas por novas leituras que não entraram na montagem e/ou por outro conjunto de leituras para verificar eventuais sobreposições; em (B), as bases circuladas simulam uma sobreposição

entre duas das leituras agregadas à montagem, indicando que as extensões das pontas da falha se encontraram, possibilitando a união das sequências em um único *contig*, representado pelo *contig* final.

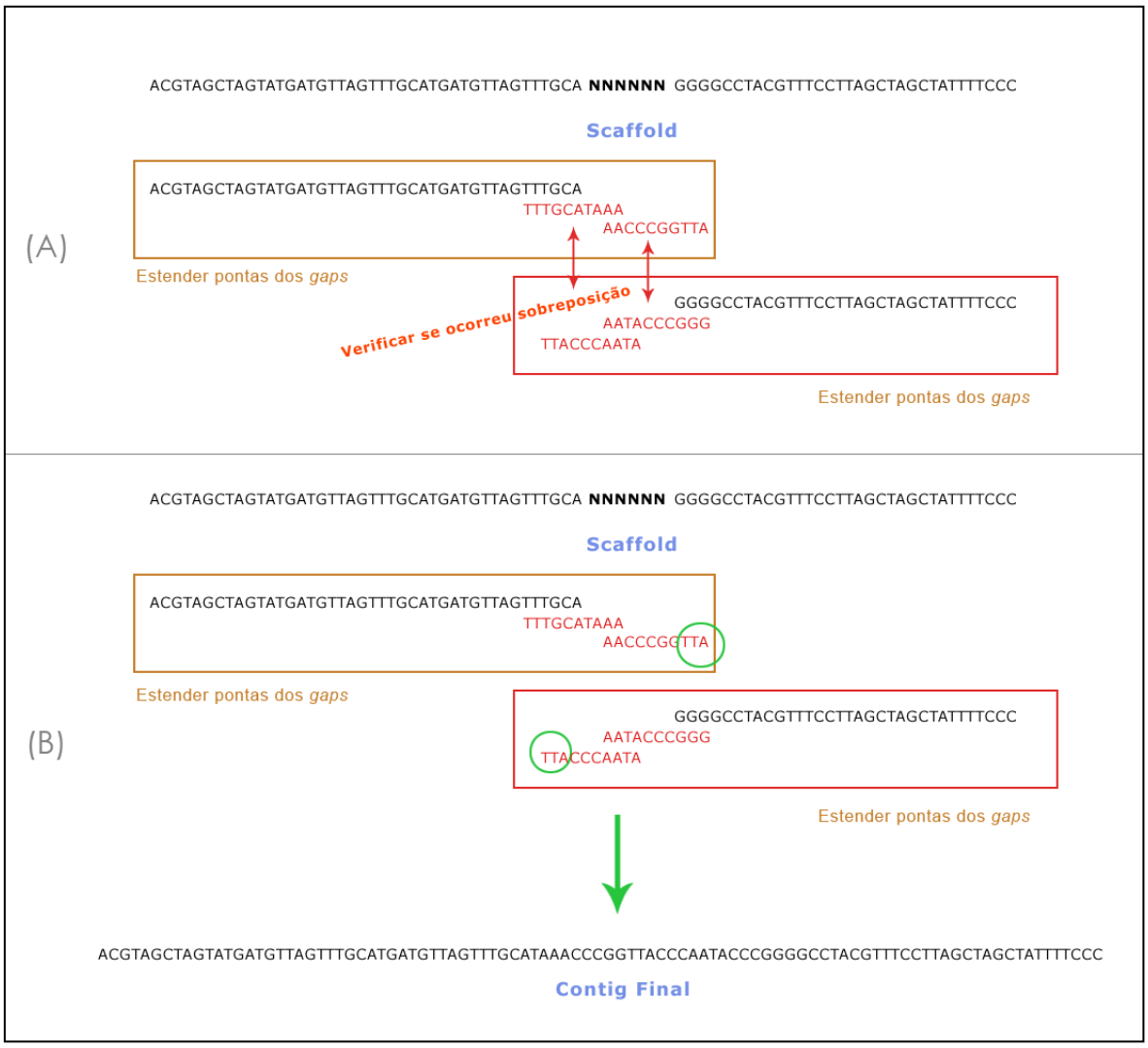

FIGURA 1.6 - PASSOS GERAIS DE UM PROCESSO DE MONTAGEM: FECHAMENTO DE FALHAS E *CONTIG* FINAL

1.2.2.2 Montadores

Muitos algoritmos foram desenvolvidos com o intuito de realizar as tarefas de montagem de maneira automática e propuseram estratégias de correção de erros para solucionar as falhas presentes na metodologia de sequenciamento e montagem genômica. Entre os montadores direcionados ao método WGS, destacam-se: Atlas (HAVLAK *et al*., 2004), Arachne (BATZOGLOU *et al*., 2002), Celera (MYERS *et al*.,

2000), Pcap (HUANG *et al*., 2003), Phrap (GREEN, 1999), Phusion (MULLIKIN *et al*., 2003), entre outros. Todos estes são baseados em um conceito de sobreposição utilizando um grafo de sobreposição por consenso (BATZOGLOU, 2005) onde cada leitura é tratada como um nó do grafo e cada sobreposição um arco entre os nós (ZERBINO *et al*., 2008).

Porém, outra categoria de montadores, mais atual e voltada para as novas tecnologias de sequenciamento, tem trabalhado com leituras curtas utilizando diferentes estratégias não baseadas em grafos de sobreposição por consenso, já que tais grafos tornam-se muito custosos computacionalmente se considerado o volume de dados produzido pelos novos métodos de sequenciamento (ZERBINO *et al*., 2008). Os novos montadores que se destacam são SSAKE (WARREN *et al*., 2007) e VCAKE (JECK *et al*., 2007), SHARCGS (DOHM *et al*., 2007), Velvet (ZERBINO *et al*., 2008) e ABySS (SIMPSON *et al*., 2009).

1.3 Anotação

A sequência obtida do genoma de um organismo através dos processos de sequenciamento e montagem genômica é uma importante fonte de informação para estudos biológicos servindo como elemento necessário para um processo de anotação. O objetivo da anotação é identificar as principais características do genoma, seus genes e produtos, sendo responsável pela ligação entre a sequência de DNA e a biologia do organismo. Os recursos ligados à anotação estão cada vez mais aprimorados e fundamentam diversas pesquisas biológicas (STEIN, 2001).

1.3.1 Etapas de um processo de anotação

O processo de anotação genômica consiste no reconhecimento de padrões nas sequências de DNA e dos significados destes padrões.

A FIGURA 1.7 mostra um fluxo geral dos processos envolvidos na anotação genômica. Na etapa inicial de anotação (A) são identificados padrões nas sequências que levam a identificação de códons de início e final de tradução, regiões codificadoras de proteínas, RBSs (sítios de ligação ribossomal), regiões reguladoras e promotoras, entre outros; além disso, regiões de sequências conservadas ou previamente conhecidas são mapeadas – como tRNAs (ácido

ribonucleico transportador), rRNAs (ácido ribonucleico ribossomal), elementos repetitivos, etc. (STEIN, 2001).

Após identificar as regiões de interesse na sequência, é necessário compreender qual a informação contida em tais sequências (FIGURA 1.7 (B)) através de análises realizadas a fim de obter uma relação completa das proteínas contidas no organismo, com seus respectivos nomes e funções; tais análises são feitas através de comparações com genes conhecidos de organismos normalmente próximos taxonomicamente do genoma estudado. Na prática, a tarefa de identificação das proteínas não é simples, já que apenas parte do conjunto de genes codifica proteínas com funções conhecidas e bem caracterizadas, o resto dos possíveis genes codifica proteínas hipotéticas ou conservadas sem função claramente definida (STEIN, 2001).

A identificação das proteínas possibilita a inferência de suas funções e suas relações com processos biológicos (FIGURA 1.7 (C)). Conjuntos de genes podem ser associados a funções permitindo o agrupamento das proteínas em categorias, tais como em processamento e armazenamento de informação, processos celulares e sinalização, metabolismo, entre outras. Além disso, é possível reconstruir vias metabólicas, caracterizar sistemas de transporte e secreção, etc. (STEIN, 2001).

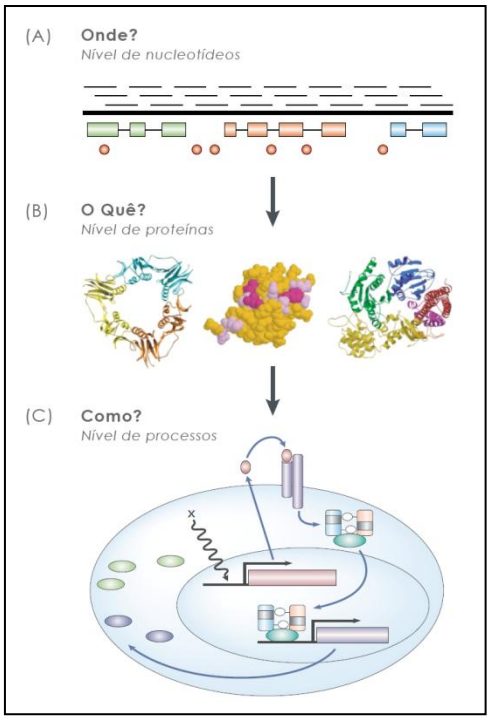

FIGURA 1.7 - PASSOS GERAIS DE UM PROCESSO DE ANOTAÇÃO
FONTE: Adaptação (STEIN, 2001).

1.3.2 Ferramentas para anotação

Muitos algoritmos e conjuntos de processos foram desenvolvidos a fim de automatizar o processo de anotação ou partes dele. São comumente utilizados em pesquisas biológicas, e mais usados a nível local, programas de computadores e ferramentas direcionadas à: (i) detecção de ORFs (fase ou quadro aberto de leitura) e de motivos de sequências conservados, por exemplo os programas Glimmer (SALZBERG *et al*., 1998) e RBS Finder (SUZEK *et al*., 2001); (ii) busca de similaridade de sequências com genes relatados, por exemplo o programa BLAST (ALTSCHUL *et al*., 1990) e COG (TATUSOV *et al*., 2000); (iii) visualização e edição das sequências anotadas, como Artemis (RUTHERFORD *et al*., 2000; CARVER *et al*., 2008); e (iv) anotação funcional, através de comparação com bases de dados específicas como KAAS (MORIYA *et al*., 2007; WEISS, 2010).

1.4 Objetivos

1.4.1 Objetivo geral

Montagem e análise preliminar da sequência genômica da bactéria diazotrófica *Bradyrhizobium elkanii* 587.

1.4.2 Objetivos específicos

- Montagem de *contigs* do genoma de *B. elkanii* 587 utilizando leituras obtidas com sequenciador Illumina e pelo método tradicional de Sanger;
- Ordenação dos *contigs* obtidos;
- Estimativa do tamanho de falhas (*gaps*) entre *contigs* e criação de *scaffolds*;
- Fechamento de falhas nos *scaffolds* utilizando leituras de origem Sanger e alinhamento com genoma de referência;
- Identificação de características estruturais e pré-anotação do genoma de *B. elkanii* 587.

2 MATERIAL E MÉTODOS

A metodologia principal consistiu na combinação entre tratamento dos dados iniciais, estratégias automáticas, ajustes manuais e anotação dos resultados.

A FIGURA 2.1 ilustra o processo geral utilizado para a montagem do genoma de *Bradyrhizobium elkanii* 587. O primeiro passo consistiu na obtenção e tratamento dos dados. Em seguida, os dados foram submetidos aos montadores automáticos. Os parâmetros foram alterados empiricamente a fim de obter a montagem com as melhores características. Uma vez obtida a montagem, foram feitos ajustes manuais para fechamento de falhas ou validação de ligação de *contigs*.

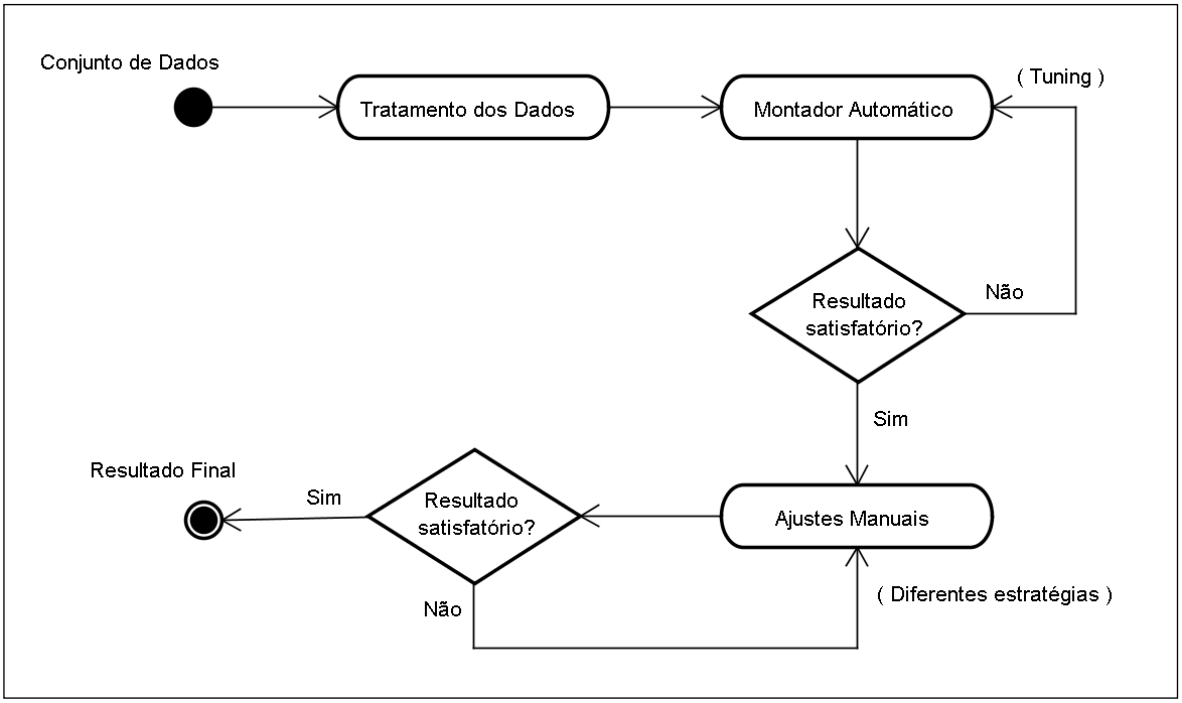

FIGURA 2.1 - FLUXOGRAMA DO PROCESSO DE MONTAGEM DO GENOMA DE *Bradyrhizobium elkanii* 587

A FIGURA 2.2 exibe a metodologia específica utilizada neste desenvolvimento para a montagem e análise preliminar do genoma de *Bradyrhizobium elkanii* 587. Primeiramente foi realizada uma montagem automática utilizando os dados iniciais sem tratamento; após, foi selecionado um genoma de referência para auxiliar e validar o processo de montagem. Em um terceiro momento, duas montagens automáticas foram realizadas, desta vez com tratamento dos dados de entrada. Com a melhor montagem obtida pelos montadores

automáticos, foi realizado um processo de fechamento de falhas e obtido o conjunto final de *scaffolds*. Por fim, com tal conjunto obtido, a etapa de pré-anotação pôde ser executada.

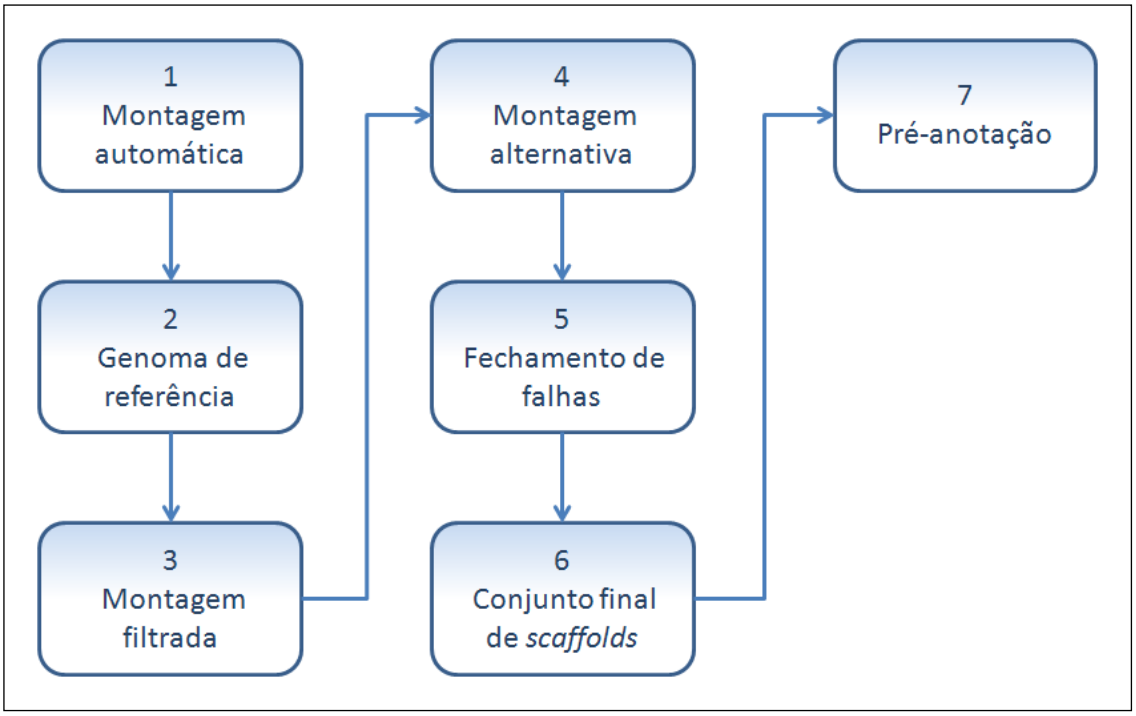

FIGURA 2.2 - METODOLOGIA ESPECÍFICA PARA MONTAGEM E ANÁLISE PRELIMINAR DO GENOMA DE *Bradyrhizobium elkanii* 587

2.1 Origem dos dados

Foi obtido um conjunto final de 42.785 leituras através do método Sanger (leituras Sanger), e de 6.020.858 leituras através do sequenciamento por vetor cíclico (leituras Illumina), totalizando aproximadamente 266 Mpb sequenciadas. As leituras Sanger foram gentilmente cedidas pela Professora Doutora Eliana Lemos (UNESP) e pelo Professor Doutor Jackson Marcondes (UNESP) para este trabalho. As leituras Illumina foram cedidas pelo INCT-FBN coordenado pelo Professor Doutor Fábio de Oliveira Pedrosa (UFPR).

2.1.1 Leituras de sequências de DNA

O conjunto de leituras oriundo do sequenciamento pelo método Sanger totalizou 38.064.206 bases, distribuídas em 42.785 leituras pareadas, ou seja,

leituras das duas extremidades de clones aleatórios com tamanho aproximado de inserto de 2 Kpb. O tamanho aproximado destas leituras foi de 900 bases.

O conjunto de leituras oriundo do sequenciamento por vetor cíclico totalizou 228.792.604 bases, distribuídas em 6.020.858 leituras pareadas com tamanho de 38 pb e tamanho aproximado de inserto de 300 pb. A TABELA 2.1 mostra os detalhes destes conjuntos de dados.

TABELA 2.1 - CARACTERÍSTICAS DOS DADOS ORIUNDOS DO SEQUENCIAMENTO PELOS MÉTODOS SANGER E ILLUMINA

Característica \ Tipo de leitura	**Leituras Sanger**	**Leituras Illumina**
Tipo de dado	Longo, final pareado	Curto, final pareado
Tamanho médio da leitura	890 pb	38 pb
Distância entre extremidades pareadas	Variável	~300 pb
Total de leituras	42.785	6.020.858
Total de bases	38.064.206 pb	228.792.604 pb

2.2 Montagem automática

Com o conjunto de leituras selecionado e tratado, foram realizadas montagens automáticas a fim de obter *contigs* do genoma de *B. elkanii* 587, estimativas do tamanho de falhas (*gaps*) entre os *contigs*, e criação de *scaffolds*.

2.2.1 Montadores

Para a etapa de montagem automática o montador Phrap e o montador de leituras curtas Velvet foram escolhidos para a montagem das leituras de origem Sanger e das leituras obtidas com o sequenciador Illumina, respectivamente.

2.2.1.1 Phrap

O montador Phrap é um programa pertencente ao pacote de programas Phred/Phrap/Consed utilizado para montagem genômica de conjuntos de sequências de DNA obtidos através de sequenciamento pelo método WGS (GREEN, 1999).

O módulo Phred é responsável pela interpretação de eletroforetogramas (os eletroforetogramas contêm os dados brutos gerados pelo método de

sequenciamento de didesoxinucleotídeos marcados com fluoróforos), informando ao programa Phrap sequências de bases resultantes do sequenciamento e a qualidade relativa a cada uma delas.

A FIGURA 2.3 ilustra um eletroforetograma de uma leitura Sanger. A distinção de bases é feita de acordo com a cor de cada pico, atribuída de acordo com o sinal emitido durante o processo de sequenciamento. A atribuição de valores da qualidade para cada base é realizada através de métricas envolvendo o sinal componente do eletroforetograma, amplitude e distinção dos picos, permitindo avaliar a incerteza na identificação da base (EWING *et al*., 1998).

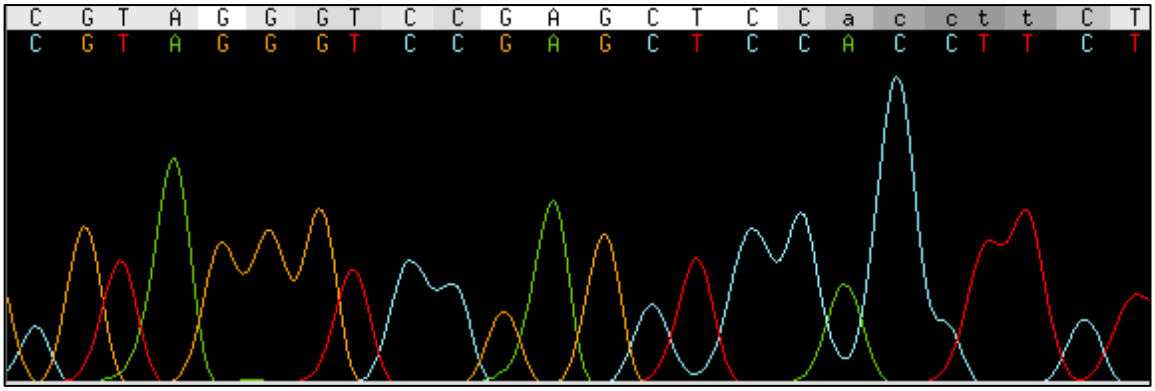

FIGURA 2.3 - EXEMPLO DE ELETROFORETOGRAMA DE UMA LEITURA SANGER

O programa Phrap realiza a montagem das leituras oriundas da chamada (identificação) de bases realizada pelo módulo Phred. O Phrap é um montador de leituras longas que utiliza grafos de sobreposição para junção das leituras. A FIGURA 2.4 simula um grafo de sobreposição simples; bases com alinhamento perfeito (indicados pelas linhas de ligação) podem ser mescladas em um *contig*. A figura também exemplifica um eventual problema de montagem que deve ser resolvido utilizando informações adicionais: a leitura 1 tem sua extremidade exatamente alinhada com a extremidade da leitura 3, porém, a junção de um único *contig* não deve ser realizada já que o restante da leitura é conflitante (GREEN, 1999).

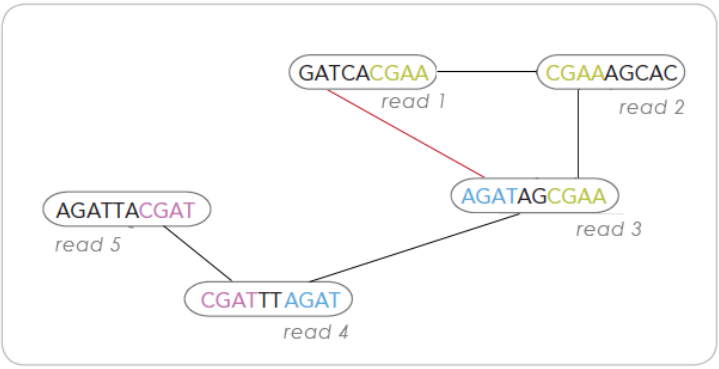

FIGURA 2.4 - EXEMPLO DE GRAFO DE SOBREPOSIÇÃO UTILIZADO PELO PROGRAMA PHRAP
FONTE: Adaptada de *Technical Note*: Illumina® *Sequencing - De Novo Assembly Using Illumina Reads* (2009)

A qualidade das bases também é utilizada no processo de construção dos grafos de sobreposição pelo montador Phrap, permitindo o uso de toda a leitura e não apenas da parte de ótima qualidade. Além disso, parâmetros informados pelo usuário podem aumentar a performance dos algoritmos executados pelo montador no tratamento de erros ou repetições durante a montagem. A TABELA 2.2 mostra os parâmetros utilizados durante o processo de alinhamento de sequências e criação do grafo de sobreposição para a montagem do genoma de *B. elkanii*. Os parâmetros referem-se aos parâmetros padrões de alinhamento de sequências do montador Phrap.

TABELA 2.2 - DESCRIÇÃO DOS PRINCIPAIS PARÂMETROS E SEUS RESPECTIVOS VALORES UTILIZADOS NO MONTADOR PHRAP

Parâmetro	Valor
Base alinhada	+1
Base não alinhada	-1
Abertura de falha (*gap*)	-4
Extensão de falha (*gap*)	-3
Início de leitura aparada	Não

2.2.1.2. Velvet

Velvet é um pacote de algoritmos projetado para realizar a montagem genômica de conjuntos de dados contendo leituras curtas. (ZERBINO *et al*., 2008; ZERBINO *et al*., 2009)

A principal característica do montador de leituras curtas Velvet que o distingue de montadores tradicionais é o uso do grafo *de Bruijn* (PEVZNER *et al*., 2001) para o processo de comparação e montagem de leituras. Montadores tradicionais tratam cada leitura como um nó em um grafo de sobreposição, o que, considerando a quantidade de informação gerada pelos sequenciadores de leituras curtas, torna o processamento do grafo extremamente custoso computacionalmente. Já o grafo *de Bruijn* é composto por uma representação das leituras em pequenas palavras com tamanho pré-definido *k*, chamadas *k-mers*. Os *k-mers* são gerados através da informação contida nas leituras e podem tratar naturalmente a alta quantidade de dados gerada pelos sequenciadores de leituras curtas e a eventual redundância de informação, já que são gravados apenas uma vez tendo a quantidade de vezes que eles aparecem e os respectivos caminhos para outros *k-mers* apenas referenciados no grafo e não explícitos (ZERBINO *et al*., 2008).

A FIGURA 2.5 ilustra a criação de um grafo *de Bruijn*. Neste caso, vários *k-mers*, com um tamanho de três pares de bases, são criados a partir da leitura original; o grafo é construído através da sobreposição dos *k-mers* utilizando o alinhamento perfeito de *k - 1* bases (neste caso, dois pares de bases). O tratamento de redundância é exemplificado pelo *k-mer* “GAT”, que é representado apenas uma vez no grafo, mas referenciado duas vezes; permitindo que o algoritmo utilize o *k-mer* para obtenção da sequência original, sem desconsiderar dados existentes ou reutilizá-los.

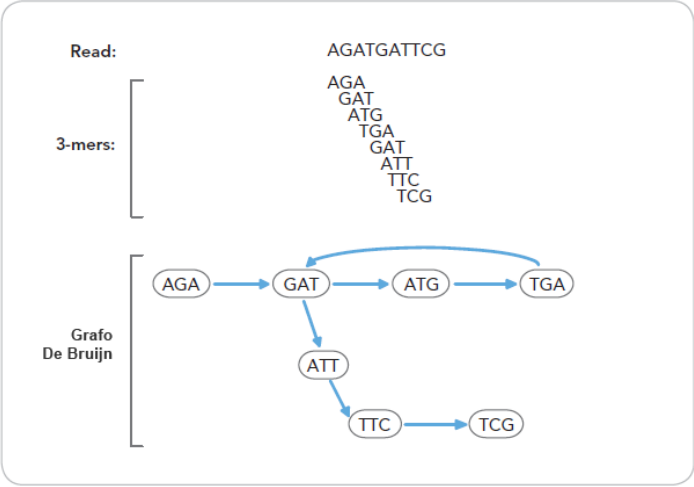

FIGURA 2.5 - EXEMPLO DE GRAFO DE BRUIJN

FONTE: Adaptada de *Technical Note*: Illumina® *Sequencing - De Novo Assembly Using Illumina Reads* (2009)

Outro exemplo de tratamento de redundância, desta vez especificamente no tratamento de repetições, é mostrado na FIGURA 2.6. O grafo *de Bruijn* simplifica a resolução das regiões repetitivas identificando as regiões e referenciando o número de ocorrências da mesma. Além disso, exemplifica um problema de montagem, pois mostra que o problema, neste caso, não tem solução, já que dois caminhos podem ser corretos: (i) A-B-C-B-D-B-E ou (ii) A-B-D-B-C-B-E (LEMOS *et al*., 2003).

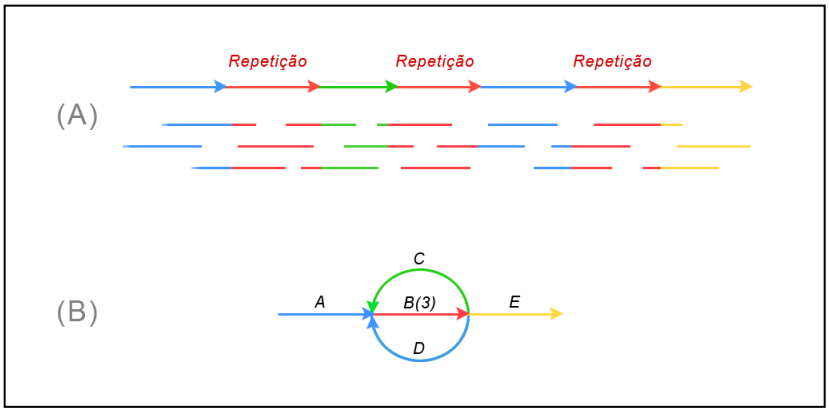

FIGURA 2.6 - EXEMPLO DE CRIAÇÃO DE UM GRAFO *DE BRUIJN* COM REGIÕES REPETITIVAS

FONTE: Adaptação (LEMOS *et al*., 2003)

Outra importante característica do montador de leituras curtas Velvet é a possibilidade de entrada de diferentes conjuntos de dados para a realização da montagem, mas sem uso efetivo destes dados. Por exemplo, leituras longas são usadas apenas para tratamento de repetições e erros de montagem. Além desta característica, o montador Velvet não faz uso da qualidade de bases (ZERBINO *et al*., 2008).

O montador Velvet é composto por dois módulos: *velveth* e *velvetg*. O módulo *velveth* é responsável pela criação do conjunto de dados a ser usado pelo módulo *velvetg*; para isso, interpreta o arquivo de sequência passado como entrada e cria uma tabela de *k-mers*. O módulo *velvetg* cria o grafo *de Bruijn* e o manipula, percorrendo o gráfico de diferentes maneiras, de acordo com os parâmetros informados pelo usuário (ZERBINO *et al*., 2008).

As TABELAS 2.3 e 2.4 mostram os parâmetros e valores usados nos módulos *velveth* e *velvetg*, respectivamente, do montador Velvet.

TABELA 2.3 - PARÂMETROS E SEUS RESPECTIVOS VALORES UTILIZADOS NO MÓDULO *VELVETH* DO MONTADOR VELVET

Parâmetro	**Valor**
Tipo de dado	*shortPaired*
Tamanho de palavra (*k-mer*)	19

TABELA 2.4 - PARÂMETROS E SEUS RESPECTIVOS VALORES UTILIZADOS NO MÓDULO VELVETG DO MONTADOR VELVET

Parâmetro	**Valor**
Tamanho de fragmento (*ins_length*)	336
Profundidade de cobertura esperada (*exp_cov*)	16
Cobertura mínima (*cov_cutoff*)	3
Tamanho mínimo de *contig* exportado (*min_contig_lgth*)	5000
Criação de *scaffolds* (*scaffolding*)	*yes*

Os valores mostrados nas TABELAS 2.3 e 2.4 foram determinados experimentalmente testando diferentes parâmetros e comparando os resultados obtidos para montagem do genoma de *B. elkanii*.

No módulo *velveth*, o tamanho de palavra foi definido com valores variando de 15 a 33. Tamanhos menores de *k-mer* são testados com a intenção de aumentar a sensibilidade do grafo através de uma maior profundidade de cobertura; já os *k-mers* maiores trazem especificidade ao grafo evitando falsas sobreposições, mas com profundidades de cobertura menores (ZERBINO *et al*., 2008).

No módulo *velvetg*, parâmetros adicionais aos mostrados na TABELA 2.4 não alteraram os resultados. Dentre os parâmetros expostos, os seguintes testes foram realizados:

- Tamanho de fragmento (*ins_length*): definido com valores entre 80% e 120% da distância aproximada entre as leituras pareadas.
- Profundidade de cobertura esperada *(exp_cov)*: definido com valores entre 1/3 e 5/3 da profundidade de cobertura esperada.
- Cobertura mínima *(cov_cutoff)*: definido com valores representando 1/4 e 1/5 da profundidade de cobertura esperada.
- Tamanho mínimo de *contig* exportado *(min_contig_lgth)*: valores entre 100 e 5.000, a fim de obter mais informações com valores menores e menos redundância com valores maiores.
- Criação de *scaffolds* (*scaffolding*): definido com valores “*no*” e “*yes*”, a fim de verificar os conjuntos de *contigs* e *scaffolds*.

2.2.2 Estimativa de tamanho de genoma

O genoma de *B. elkanii* 587 teve seu tamanho estimado entre 7 e 8 Mpb através de uma eletroforese de campo pulsado, de acordo com estudo realizado pelo Laboratório de Bioquímica de Microrganismos e Plantas (LBMP) da Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista (UNESP) (KISHI *et al.*, 2005; KISHI, 2007).

Com o tamanho estimado do genoma, tornou-se possível verificar a profundidade de cobertura de sequenciamento esperada para ambos os conjuntos de leituras. A profundidade de cobertura foi calculada dividindo a quantidade total de bases sequenciadas pelo tamanho do genoma em pares de bases. A EQUAÇÃO 2.1 detalha o cálculo da profundidade de cobertura.

$$C_{esperada} = \frac{N_{reads} t_{reads}}{T_{genoma}} \quad (2.1)$$

onde:

$C_{esperada}$ = profundidade de cobertura esperada

N_{reads} = número de leituras disponíveis;

t_{reads} = tamanho médio, em pares de bases, das leituras; e,

T_{genoma} = tamanho, em pares de bases, do genoma alvo de montagem.

A profundidade de cobertura obtida para o conjunto de leituras Illumina variou entre 32x e 28x (de acordo com os valores estimados de 7 e 8 Mpb) e para o conjunto de leituras Sanger entre 5x e 4x.

2.3 *Alinhamento a genoma de referência*

O número de genomas completamente sequenciados aumenta a cada ano e vários projetos de sequenciamento de genomas de organismos muito próximos taxonomicamente já foram concluídos ou estão em andamento. A existência da

sequência genômica similar a do organismo de interesse pode ser aproveitada em processos de montagem genômica utilizando, por exemplo, um genoma como referência para ordenação do conjunto de dados ou fechamentos de falhas (DELCHER *et al*., 1999).

A seguir são descritas as ferramentas utilizadas para realizar o alinhamento de sequências utilizando um genoma de referência.

2.3.1 Ferramentas para alinhamento de sequências

Os programas MUMmer e Consed, foram utilizados para alinhamento e comparação de sequências. O primeiro foi utilizado para alinhamentos contra genomas completos, e o segundo para alinhamentos parciais e/ou ancoragem de leituras em sequências de *scaffolds* para fechamento de falhas, por exemplo.

2.3.1.1 MUMmer

MUMmer é um pacote de programas utilizado para alinhamento de sequências (DELCHER *et al*., 1999; DELCHER *et al*., 2002; KURTZ *et al*., 2004).

O pacote MUMmer é formado principalmente pelos programas, NUCmer e PROmer. Os dois programas diferem quanto aos dados utilizados nos alinhamentos: o programa NUCmer trabalha com sequências de nucleotídeos em suas comparações e o programa PROmer trabalha as sequências traduzidas nas seis fases de leituras (DELCHER *et al*., 1999; DELCHER *et al*., 2002; KURTZ *et al*., 2004).

A principal característica do programa MUMmer é encontrar e registrar similaridades entre duas sequências e utiliza tais registros para exibir os alinhamentos em forma de gráficos de pontos.

Para os alinhamentos e consequentes gráficos de comparação de sequências deste trabalho foi utilizado o programa PROmer com seus parâmetros e valores padrões.

2.3.1.2 Consed

O programa Consed, módulo do pacote de programas Phred/Phrap/Consed descrito na seção 2.2.1.1, é uma ferramenta para visualização de montagens genômicas, permitindo a edição de sequências e finalização das montagens. O programa apresenta vantagens como o suporte a diferentes tipos de dados, atenção para áreas problemáticas através de ergonomia de interface, edição guiada por qualidade das sequências oriunda do processo de chamada de bases do módulo Phred, visualização geral da montagem, entre outras (GORDON *et al*., 1998).

Neste estudo, a ferramenta Consed foi utilizada para alinhamentos gerais, incluindo ancoragem de leituras, e para fechamento de falhas e finalização de montagens utilizando diferentes conjuntos de leituras e comparação entre os conjuntos de dados.

2.3.2 Genoma de referência

A escolha de um genoma de referência é necessária para verificação e/ou validação das montagens obtidas de maneira automática contra um genoma fechado, além de este ser utilizado para eventuais ajustes manuais que precisem ser executados (por exemplo, através de alinhamento dos *contigs*/*scaffolds* com a sequência de referência).

Para escolha do genoma de referência foram pré-selecionados genomas próximos taxonomicamente a *B. elkanii* 587. A TABELA 2.5 detalha os genomas pré-selecionados. Os possíveis genomas de referência selecionados correspondem a todos os genomas fechados do gênero *Bradyrhizobium*.

TABELA 2.5 - POSSÍVEIS GENOMAS DE REFERÊNCIA BASEADOS NOS GENOMAS FECHADOS DO GÊNERO *Bradyrhizobium*

Referência no GenBank®	**Identificação do genoma**	**Tamanho do genoma (pb)**
NC_009485	*Bradyrhizobium* sp. BTAi1; circular	8.264.687
NC_004463	*Bradyrhizobium japonicum* USDA 110; circular	9.105.828
NC_009445	*Bradyrhizobium* sp. ORS278; circular	7.456.587

Os genomas de referência foram pré-selecionados com base em dois critérios: identidade da sequência 16S rRNA e alinhamento com *scaffolds*.

2.3.2.1 Alinhamento do gene 16S rRNA

Os genes 16S rRNA dos possíveis genomas de referência mostrados na TABELA 2.5 foram comparados com os *scaffolds* obtidos em montagem automática das leituras de origem Illumina.

O alinhamento das sequências foi realizado com a ferramenta Consed e a pontuação de alinhamento foi atribuída seguindo os valores padrões utilizados no algoritmo de alinhamento do montador Phrap, descritos na TABELA 2.2. A TABELA 2.6 exibe a identidade entre a sequência contida nos *scaffolds* e a dos possíveis genomas de referência.

TABELA 2.6 - IDENTIDADE DO GENE 16S rRNA DE *Bradyrhizobium elkanii* 587 COM O GENE HOMÓLOG DOS POSSÍVEIS GENOMAS DE REFERÊNCIA

Genoma	**Base alinhada (pt)**	**Base não alinhada (pt)**	**Falha (*gap*) (pt)**	**Extensão de falha (*gap*) (pt)**	**Identidade (%)**
Bradyrhizobium sp. BTAi1	1446	-30	-20	-3	96,46
Bradyrhizobium japonicum USDA 110	1449	-28	-16	-6	96,66
Bradyrhizobium sp. ORS278	1438	-36	-16	-6	96,12

A pontuação de alinhamento do gene 16S rRNA obtida pelos três possíveis genomas de referência coincide com a homologia de 97% apresentada em uma comparação entre tal gene de *B. elkanii* 587 e *B. japonicum* USDA 110 (KISHI, 2007).

Considerando a semelhança entre os valores de identidade obtidos para o gene 16S rRNA dos organismos de referência pré-selecionados, o alinhamento do gene 16S rRNA não foi utilizado como parâmetro exclusivo para escolha do genoma de referência.

2.3.2.2. Alinhamento aos *scaffolds*

Utilizando o programa MUMmer, os *scaffolds* disponíveis da montagem automática realizada foram plotados contra os possíveis genomas de referência.

A TABELA 2.7 exibe o porcentual de alinhamento dos *scaffolds* em relação a cada genoma de referência.

TABELA 2.7 - PORCENTUAL DE ALINHAMENTO DOS *SCAFFOLD*S EM RELAÇÃO AOS POSSÍVEIS GENOMAS DE REFERÊNCIA.

Genoma de referência	***Scaffolds* alinhados (%)**
Bradyrhizobium sp. BTAi1	70,6
Bradyrhizobium japonicum USDA 110	85,3
Bradyrhizobium sp. ORS278	70,6

A FIGURA 2.7 mostra a comparação gráfica entre o conjunto de *scaffolds* e os genomas de *B.* sp. BTAi1, *B. japonicum* USDA 110, *B.* sp. ORS278 respectivamente. O alinhamento dos *scaffolds* com os genomas de *B.* sp. BTAi1 e *B.* sp. ORS278 exibiram padrões com falhas. Já o alinhamento do conjunto de *scaffold*s contra o genoma de *B. japonicum* USDA 110 é mais consistente e apresentou regiões sintênicas em toda a extensão do genoma.

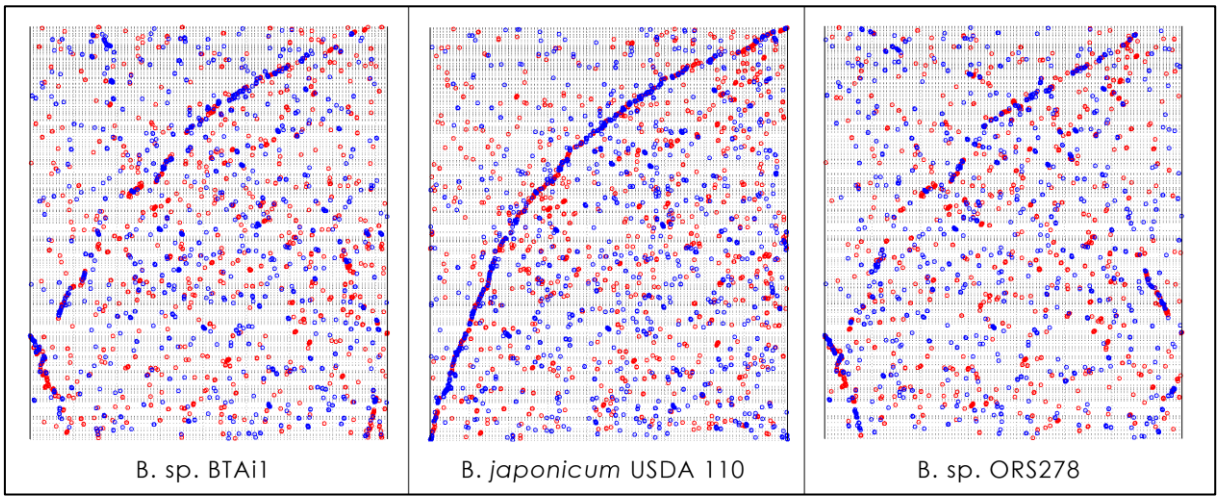

FIGURA 2.7 - ALINHAMENTO DOS *SCAFFOLDS* CONTRA OS POSSÍVEIS GENOMAS DE REFERÊNCIA.

2.4 Criação de conjunto alternativo de leituras

Como discutido na seção 2.2.1, os montadores genômicos requerem dados de tipo específico. Por exemplo, o montador Phrap não pode processar leituras curtas produzidas pelo sequenciador Illumina devido a capacidade computacional necessária para processar estes dados. No caso do montador Velvet, leituras longas obtidas pelo método de Sanger não representam ganhos na montagem já que a informação é usada apenas como guia para os algoritmos de correção de erros e não como dados efetivos.

Com a finalidade de melhorar o desempenho dos montadores automáticos, foi desenvolvida uma estratégia alternativa para a entrada de dados de diferentes plataformas. As extremidades das leituras Sanger foram aparadas para eliminar bases de baixa qualidade no início e final da leitura e então fragmentadas em leituras curtas.

Esta fragmentação das leituras Sanger criou um novo conjunto de dados simulando leituras Illumina, a fim de criar um conjunto de dados de entrada único para o montador Velvet. Em contrapartida, a informação das leituras longas não foi perdida, já que a mesma pôde ser utilizada em processos futuros para eventuais fechamentos de falhas através do alinhamento das leituras longas aos *contigs*/*scaffolds* criados com o conjunto de dados alternativo.

Os dois processos da criação do conjunto alternativo de dados consistiram na (i) eliminação de bases de baixa qualidade no início e final das leituras Sanger e (ii) fragmentação destas leituras em leituras curtas.

2.4.1 Remoção de regiões de baixa qualidade das leituras Sanger

A remoção de regiões de baixa qualidade das leituras é um processo onde as bases de baixa qualidade localizadas nas extremidades destas leituras são descartadas para reduzir sobreposições discrepantes.

Neste estudo, o processo de remoção de regiões de baixa qualidade foi realizado pelo programa Phred (EWING *et al*., 1998).

A remoção de regiões de baixa qualidade foi necessária já que o montador Velvet ignora a confiabilidade da identificação (qualidade) das bases. Assim o

montador consideraria bases identificadas com alta confiabilidade com o mesmo peso que bases identificadas com baixa confiabilidade.

2.4.2 Fragmentação das leituras Sanger

O conjunto de leituras Sanger aparadas foi submetido a um algoritmo simples de fragmentação para criar um novo conjunto de leituras para entrada no montador Velvet.

O algoritmo utilizado para a fragmentação das leituras Sanger realiza os seguintes processos:

- Abertura de arquivo contendo as leituras longas pareadas;
- Criação de uma sequência da base *n* até a base *n + 38* pares de bases;
- Criação das leituras curtas pareadas;
- Início de *leitura* (*n*) definido como *n + 19*;

A FIGURA 2.8 ilustra os processos realizados para a criação do conjunto alternativo de dados. Em (A), a leitura Sanger, longa pareada, original; em (B) a mesma leitura após o processo de remoção das bases de baixa qualidade das extremidades. Por fim, em (C), a criação de leituras curtas pareadas baseadas na leitura longa pareada Sanger.

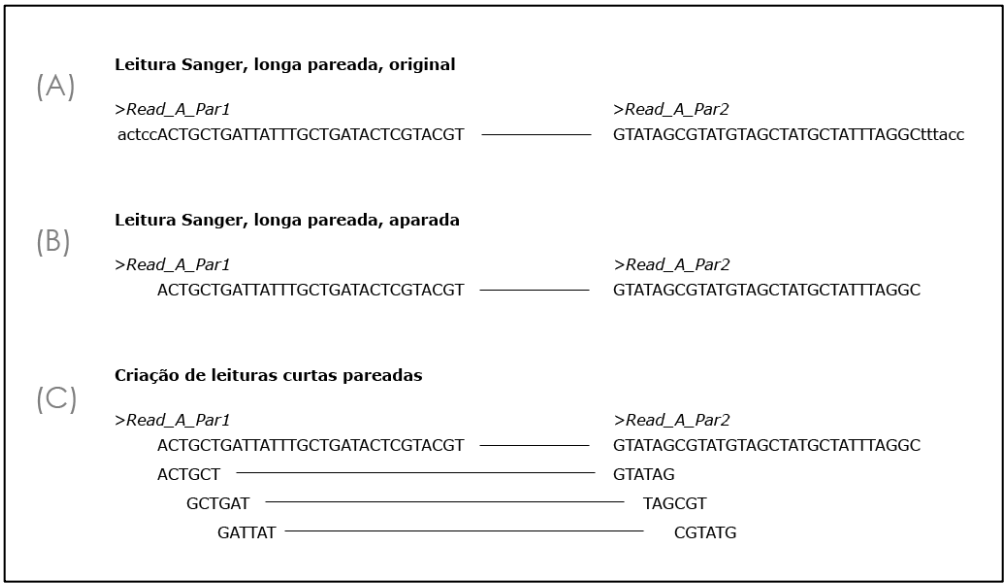

FIGURA 2.8 - PROCESSOS PARA A CRIAÇÃO DO CONJUNTO ALTERNATIVO DE DADOS

O processo de fragmentação resultou em um conjunto de leituras curtas pareadas, com tamanho de 38 pares de bases a intervalos de 19 pares de bases entre seus inícios. A TABELA 2.8 estrutura detalhes das leituras.

TABELA 2.8 - CARACTERÍSTICAS DO CONJUNTO ALTERNATIVO DE LEITURAS

Conjunto alternativo de leituras	
Tipo de dado	Curto, final pareado
Tamanho de leitura	38 pb
Distância entre leituras pareadas	Variável
Total de leituras	748.002
Total de Bases	28.424.076 pb

2.5 Fechamento de falhas

A montagem resultante dos processos automáticos apresenta falhas (*gaps*) internas aos *scaffolds*, com tamanhos estimados e com indícios de ligação entre ambas as extremidades; e falhas externas sem indício ligação e que provavelmente ligam dois *scaffolds*.

O processo de fechamento de falhas foi realizado com a verificação manual de cada falha através da ancoragem de leituras aos *scaffolds* e através da junção de *scaffolds* considerando a sobreposição de suas extremidades.

2.5.1 Fechamento manual de falhas

As falhas internas presentes nos *scaffolds* foram submetidas à análise manual utilizando a ferramenta Consed.

Para tal, os *scaffolds* resultantes do processo de montagem automática foram congelados em uma montagem referência, e as leituras Illumina e Sanger foram ancoradas nesta montagem para preencher as falhas com informações não utilizadas pelos montadores automáticos.

2.5.1.1 Fechamento de falhas utilizando a ferramenta Consed

Para o fechamento manual de falhas utilizando a ferramenta Consed foram verificadas eventuais sobreposições de leituras ancoradas ao conjunto de *scaffolds*.

A FIGURA 2.9 ilustra a nomenclatura de um caso padrão simulado de sobreposição de leituras e fechamento de uma falha na ferramenta Consed. A sequência consenso representa uma sequência de referência, *contigs* ou *scaffolds* fixados, servindo de base para ancoragem de leituras. Abaixo da sequência consenso estão as quatro leituras sobrepostas. A leitura 1 possui qualidade alta (sombreamento branco e bases em caixa alta); as leituras 2 e 3 são de qualidade média (sombreamento cinza e mistura de caixa alta e baixa). A leitura 4 (sem sombreamento e com caixa baixa) representa uma leitura de qualidade baixa.

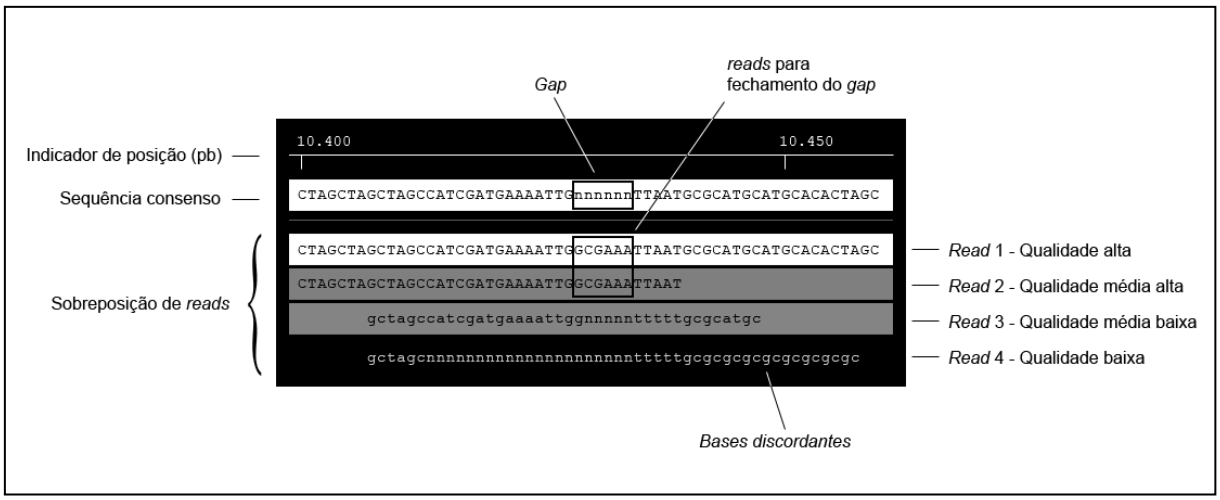

FIGURA 2.9 - TERMINOLOGIA UTILIZADA DURANTE O PROCESSO DE FECHAMENTO MANUAL DE *GAPS* ATRAVÉS DA FERRAMENTA CONSED

Além de um caso padrão de sobreposição, a FIGURA 2.9 ilustra o método comum para fechamento de uma falha (*gap*). A falha, representada pelas bases indeterminadas na sequência consenso, pode ser preenchida com as bases correspondentes das leituras 1 e 2, já que ambas as leituras possuem boa qualidade e se sobrepõem à sequência consenso . As leituras 3 e 4, discrepantes das duas primeiras, podem ser desconsideradas nesse caso, pois não possuem uma boa qualidade.

2.5.2 Junção de scaffolds através de alinhamento das extremidades

Durante a análise manual para fechamento de falhas internas utilizando a ferramenta Consed, ambas as extremidades de cada *scaffold* analisado foram

utilizadas para identificar sequências sobrepostas que indicassem a junção de dois ou mais *scaffolds* em uma única sequência.

O processo de junção de *scaffolds* consistiu em duas etapas: identificação de possíveis sobreposições das extremidades dos *scaffolds* e validação da ligação. Na primeira etapa, sequências de 10 a 15 pares de bases de alta qualidade das extremidades de cada *scaffold* foram utilizadas para procura de identidade de sequência contra todo o conjunto de *scaffolds*; na segunda etapa, as sobreposições identificadas foram selecionadas e alinhadas, desta vez estendendo a região de sobreposição para 25 a 50 pares de bases.

Além destas duas etapas, um processo adicional foi realizado verificando as mesmas condições de alinhamento, mas utilizando sequências distantes de 50 a 100 pares de bases das extremidades dos *scaffolds*; em casos positivos de alinhamentos, os *scaffolds* foram mesclados em uma única sequência e as bases excedentes das extremidades dos *scaffolds* foram descartadas assumindo que as mesmas foram inseridas ao *scaffold* de forma errônea pelo montador automático.

2.6 Pré-anotação

Para obtenção da informação contida nas sequências de DNA montadas em todos os processos descritos, foi realizada uma pré-anotação do conjunto final de *scaffolds.*

A pré-anotação foi realizada através do sistema de anotação automática GAAT (PISA *et al.*, 2010).

2.6.1 Sistema GAAT

O sistema GAAT (*Genome Assembly and Analysis Tool*) é uma plataforma de anotação automática desenvolvida no Laboratório de Bioinformática do Núcleo da Fixação Biológica de Nitrogênio da UFPR que integra módulos responsáveis por gerir projetos de anotação e revisão de genomas.

Neste trabalho foi utilizado o módulo GAnM (*Genome Annotation Module*). Este módulo de anotação automática realiza a carga dos dados (sequência de nucleotídeos) no formato FASTA ou GBK, detecta possíveis ORFs com as ferramentas Glimmer e RBS Finder, e compara as possíveis ORFs encontradas com

bancos de dados públicos, como o NCBI GenBank® e o COG (*Clusters of Orthologous Groups*), utilizando a ferramenta BLAST. Os resultados obtidos são integrados a um banco de dados e disponibilizados ao usuário via interface (PISA, 2010)

2.6.1.1 Glimmer

O programa Glimmer (*Gene Locator and Interpolated Markov ModelER*) tem como principal característica a busca por ORFs e identificação de genes especialmente em genomas de bactérias. O sistema utiliza Cadeias de Markov para distinção entre regiões codificadoras e não codificadoras e tem a possibilidade de ser treinado com o próprio conjunto de dados resultante do seu processo antes de apresentar um resultado final (SALZBERG *et al*., 1998).

2.6.1.2 RBS Finder

O programa RBS Finder (Ribosomal Binding Sites Finder) tem como característica principal a busca por sítios de ligação ribossomal e é utilizado normalmente como um processo posterior à busca por ORFs realizada pelo programa Glimmer.

O algoritmo analisa as extremidades 5’ do gene para identificar prováveis sítios de ligação (SUZEK *et al*., 2001).

2.6.1.3 BLAST

A ferramenta BLAST (*Basic Local Alignment Search Tool*) consiste de um pacote de algoritmos desenhados para realizar comparações entre sequências de DNA. O programa realiza a comparação de uma sequência com bancos de dados de sequências de DNA e retorna todas as comparações que satisfaçam um determinado nível de semelhança através de significância estatística dos alinhamentos (ALTSCHUL *et al*., 1990).

Neste trabalho, a ferramenta BLAST foi utilizada para a comparação das possíveis ORFs, encontradas com os programas Glimmer e RBS Finder, com informações extraídas dos bancos de dados públicos NCBI GenBank® e COG. A

execução ocorreu de maneira local e utilizou um banco de genomas contendo apenas os organismos próximos taxonomicamente à *B. elkanii* 587.

2.6.1.3.1 Banco de sequências para a ferramenta BLAST

Para execução da ferramenta BLAST foi construído um banco de sequências para comparações locais contendo genomas próximos taxonomicamente à *B. elkanii* 587. O banco de sequências foi construído através da própria ferramenta através do uso dos comandos de criação de bancos locais.

As sequências utilizadas na criação do banco consistem de todos os genomas completos de bactérias da Família *Bradyrhizobiaceae*, listadas abaixo:

- *Bradyrhizobium*
  - *Bradyrhizobium japonicum* USDA 110 *chromosome*
  - *Bradyrhizobium* sp. BTAi1
  - *Bradyrhizobium* sp. ORS278
- *Nitrobacter*
  - *Nitrobacter hamburgensis* X14
  - *Nitrobacter winogradskyi* Nb-255
- *Oligotropha*
  - *Oligotropha carboxidovorans* OM5
- *Rhodopseudomonas*
  - *Rhodopseudomonas palustris* BisA53
  - *Rhodopseudomonas palustris* BisB18
  - *Rhodopseudomonas palustris* BisB5
  - *Rhodopseudomonas palustris* CGA009 *chromosome*
  - *Rhodopseudomonas palustris* HaA2
  - *Rhodopseudomonas palustris* TIE-1

2.6.1.4 COG

O COG (*Cluster of Orthologous Groups*) é um banco de dados contendo sequências de proteínas ortólogas identificadas em genomas bacterianos. O banco

foi criado através da comparação entre sequências de proteínas de genomas completos e classificados em grupos funcionais que apresentam as mesmas características. O banco possui cerca de 140 mil proteínas de 66 genomas procarióticos (TATUSOV *et al*., 2000).

Neste trabalho, os genes candidatos, encontrados pelos programas Glimmer e RBS Finder e comparados pela ferramenta BLAST, foram classificados de acordo com suas funções, preditas pelo banco de dados COG.

A TABELA 2.9 mostra as categorias funcionais propostas pelo COG.

TABELA 2.9 - CATEGORIAS FUNCIONAIS COG

Código	**Descrição**
Processamento e armazenamento de informação	
J	Tradução, estrutura ribossomal e biogênese
A	Transformação e modificação de RNA
L	Transcrição
L	Replicação, recombinação e reparação
B	Estrutura e dinâmica da cromatina
Processos celulares e sinalização	
D	Controle do ciclo celular, divisão celular, e particionamento do cromossoma
Y	Estrutura nuclear
V	Mecanismos de defesa
T	Mecanismos de transdução de sinal
M	Biogênese da parede/membrana celular
N	Motilidade celular
Z	Citoesqueleto
W	Estruturas Extracelular
U	Tráfico intracelular, secreção e transporte vesicular
O	Modificação pós-traducional, renovação de proteínas, e chaperonas
Metabolismo	
C	Produção e conversão de energia
G	Transporte e metabolismo de carboidratos
E	Transporte e metabolismo de aminoácidos
F	Transporte e metabolismo de nucleotídeos
H	Transporte e metabolismo de Coenzima
I	Transporte e metabolismo de lipídios
P	Transporte e metabolismo de íons inorgânicos
Q	Biossíntese, transporte e catabolismo de metabólitos secundários
Não Caracterizadas	
R	Função geral predita
S	Função não conhecida

FONTE: Adaptada de COG Database. Disponível em: <ncbi.nlm.nih.gov/COG/grace/fiew.cgi>. Acesso em: 07/12/2010.

3 RESULTADOS E DISCUSSÃO

3.1 Obtenção das sequências

O DNA cromossomal de *B. Elkani* 587 foi gentilmente cedido pelo laboratório da Professora Doutora Eliana Lemos (UNESP, Jaboticabal) e enviado para a empresa Fasteris (Suíça) que realizou o sequenciamento utilizando o sequenciador Illumina. No total foram obtidas 6.020.858 leituras de 38 pb pareadas, com uma separação média de 300 pb. Foram também utilizados cerca de 40.000 leituras de sequência obtidas pelo método de Sanger com tamanho médio de 900 bp, que foram gentilmente cedidas para este trabalho pela Professora Eliana Lemos (UNESP, Jaboticabal).

3.2 Montagem automática

O processo de montagem automática descrito na seção 2.2 resultou em dois conjuntos de dados representando as montagens obtidas pelo montador Velvet com leituras Illumina e pelo montador Phrap utilizando leituras obtidas pelo método Sanger.

3.2.1 Estatísticas da montagem utilizando Velvet

A TABELA 3.1 mostra a estatística da montagem obtida com o montador de leituras curtas Velvet utilizando leituras Illumina.

TABELA 3.1 - ESTATÍSTICAS DA MONTAGEM AUTOMÁTICA UTILIZANDO VELVET

Característica \ Tipo de conjunto de dados	***Contigs***	***Scaffolds***
Tamanho total de genoma obtido (pb)	9.581.185	9.862.276
Número de *contigs*/*scaffolds*	7.673	1.249
Bases indeterminadas (pb)	0	296.901
Contig/*scaffold* N50	2.149	16.958
Contig/*scaffold* máximo (pb)	16.958	137.259
Leituras usadas (%)	90,4	90,5

O tamanho do genoma predito a partir de ambos os conjuntos, *contigs* e *scaffolds*, foi de aproximadamente 9,6 Mpb e 9,9 Mpb, respectivamente. Embora o

maior *scaffold* seja maior que 100 Kpb, o *scaffold* N50 é curto, não chegando a 13% do valor do *scaffold* máximo. Isso ocorreu porque o número total de *scaffolds* é alto.

O genoma deve ser representado no menor número possível de *contigs*/*scaffolds.* Idealmente, apenas um *contig* deve representar todo o genoma, sendo ele, consequentemente, o *contig* máximo e também o *contig* N50 (50% do tamanho total do genoma representado em *contigs*/*scaffolds* com pelo menos o tamanho do *contig*/*scaffold* N50, em pares de bases). Em outras palavras, os resultados de montagens devem procurar diminuir o número de *contigs*/*scaffolds*, e aumentar o tamanho dos *contigs*/*scaffolds* máximo e N50.

3.2.2 Estatísticas da montagem utilizando Phrap

A TABELA 3.2 mostra as características da montagem obtida com o montador Phrap utilizando leituras Sanger. Diferentemente dos resultados com o montador Velvet, são exibidas na tabela apenas as informações relativas ao conjunto de *contigs*, em acordo com a saída padrão do montador Phrap, que não permite a criação de conjuntos distintos de *contigs* e *scaffolds*. As sequências retornadas pelo montador podem conter bases indeterminadas que representam falhas ocasionadas por erros de sequenciamento; neste caso, as falhas não configuram a formação de um *scaffold*, já que as bases indeterminadas não são entendidas como indícios de ligação entre dois *contigs*, mas como bases reais de baixa qualidade.

TABELA 3.2 - ESTATÍSTICAS DA MONTAGEM AUTOMÁTICA UTILIZANDO PHRAP

Característica	***Contigs***
Tamanho total de genoma obtido (pb)	6.837.868
Número de *contigs*	4.183
Bases indeterminadas (pb)	33.674
Contig N50	1.939
Contig máximo (pb)	112.233
Leituras usadas (%)	100

O tamanho total dos *contigs* foi de aproximadamente 6,8 Mpb. Apesar de possuir um *contig* máximo de tamanho semelhante ao da montagem Velvet e menos bases indeterminadas, o *contig* N50 tem apenas 2% do tamanho do *contig* máximo.

3.2.3 Adição de leituras Sanger à montagem utilizando Velvet

As montagens obtidas com ambos os montadores automáticos não apresentaram bons resultados, visto que o número de *contigs* e/ou *scaffolds* permaneceu sempre alto e não houve diferença significativa da *contig*/*scaffold* N50.

A fim de melhorar o desempenho do montador Velvet, o conjunto de leituras Sanger foi adicionado à montagem com a intenção de informar um conjunto de dados de entrada único para o montador e usar as leituras longas para a criação de *scaffolds*.

No entanto, os resultados não foram satisfatórios, com a melhor montagem obtida apenas igualando os resultados alcançados sem o uso das leituras Sanger. As possíveis causas para isso podem ser a falta de sobreposição necessária das leituras Sanger para montar *scaffolds* e a presença de bases discordantes nas leituras e/ou nos *contigs*/*scaffolds* que não permitiram o alinhamento das sequências.

3.2.4 Alinhamento ao genoma de referência

Os conjuntos de *contigs* retornados pelos montadores Velvet e Phrap foram alinhados contra o genoma de referência *B. japonicum* USDA 110 a fim de validar os conjuntos de sequências retornados por ambos os montadores.

A FIGURA 3.1 mostra os alinhamentos dos *contigs* Velvet e Phrap, respectivamente, obtidos com o programa MUMmer. Nos *contigs* obtidos com o montador Velvet, 3.778.987 bases alinharam com a referência (39,44%), contra 2.655.517 bases (38,83%) nos 4.183 *contigs* obtidos com o montador Phrap.

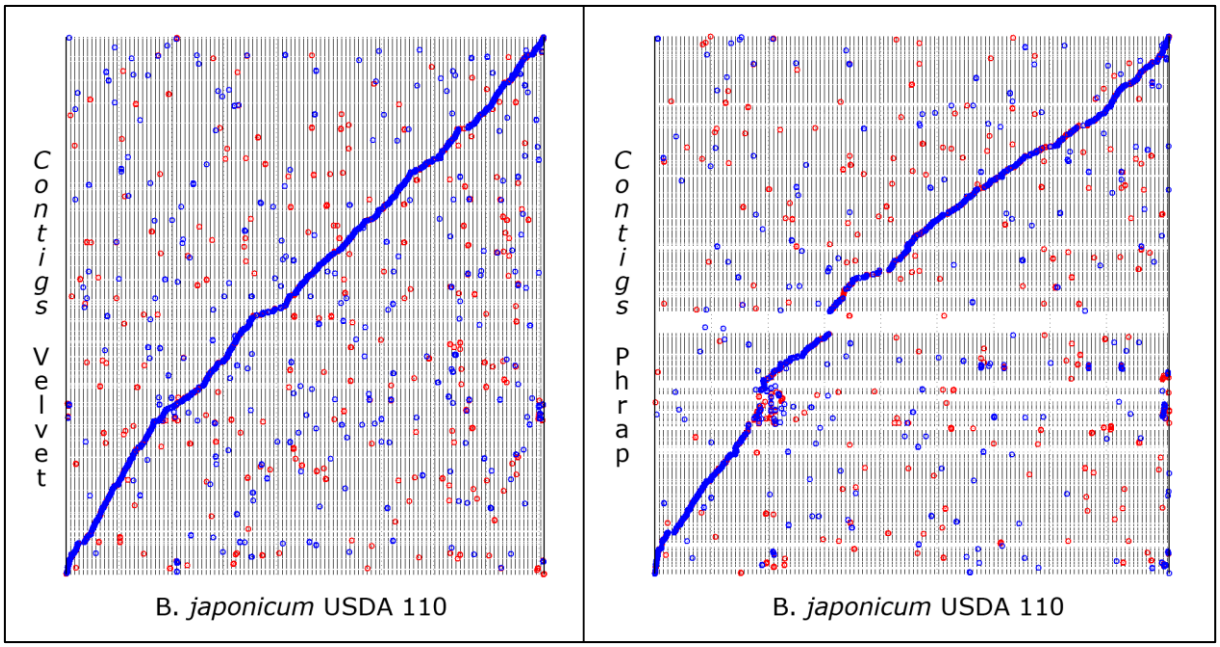

FIGURA 3.1 - ALINHAMENTO DOS *CONTIGS* VELVET E DOS *CONTIGS* PHRAP INICIAIS AO GENOMA DE REFERÊNCIA

3.3 Montagem automática alternativa

Considerando as montagens automáticas obtidas utilizando dados iniciais, estratégias alternativas de montagens automáticas precisaram ser desenvolvidas para alcançar um maior desempenho dos montadores automáticos, já que o número de *contigs*/*scaffolds* sempre se mostrou elevado, e os *contigs*/*scaffolds* máximos e/ou *contigs*/*scaffolds* N50 foram muito baixos.

As seções 3.3.1, 3.3.2, e 3.3.3 descrevem montagens alternativas realizadas e exibem suas respectivas características.

3.3.1 Montagem Phrap utilizando contigs Velvet como leituras simuladas

A fim de integrar de maneira automática o resultado retornado por ambos os montadores e melhorar o desempenho da montagem automática, foi realizada uma montagem utilizando os 7.673 *contigs* Velvet, descritos na seção 3.2.1, simulando leituras longas adicionais à montagem Phrap, descrita na seção 3.2.2. Tal estratégia realiza um processo contrário ao descrito na seção 3.2.3, onde ocorreu a adição de leituras Sanger à montagem utilizando Velvet. Ambas as estratégias tiveram como

objetivo integrar os conjuntos de entradas de dados e/ou os resultados de ambos os montadores.

A TABELA 3.3 mostra as características da montagem obtida com o montador Phrap utilizando leituras Sanger, acrescidas de leituras simuladas pelos *contigs* Velvet.

TABELA 3.3 - ESTATÍSTICAS DA MONTAGEM ALTERNATIVA UTILIZANDO LEITURAS SANGER MAIS *CONTIG*S VELVET

Característica	***Contigs***
Tamanho total de genoma obtido (pb)	9.834.150
Número de *contigs*	2.635
Bases indeterminadas (pb)	9.864
Contig N50	6.829
Contig máximo (pb)	66.233
Leituras usadas (%)	100

A montagem Phrap obtida utilizando *contigs* Velvet como leituras simuladas apresentou vantagens e desvantagens em relação às montagens automáticas iniciais. O tamanho total do genoma apresentou-se estável em relação ao montador Velvet; o número de *contigs* apresentou uma melhora significativa, já que se aproximou do menor número obtido até então (1.249 *scaffolds* na montagem inicial utilizando o montador Velvet), e com um número reduzido de bases indeterminadas. Por outro lado, os tamanhos dos *scaffold* máximo e N50 foram menores do que os obtidos na montagem inicial utilizando o montador Velvet.

A FIGURA 3.2 apresenta o gráfico de alinhamento do conjunto de *contigs* retornado pelo montador Phrap ao genoma de referência *B. japonicum* USDA 110, onde 1.136 *contigs* alinharam à referencia, representando 6.325.627 pb (64,32%).

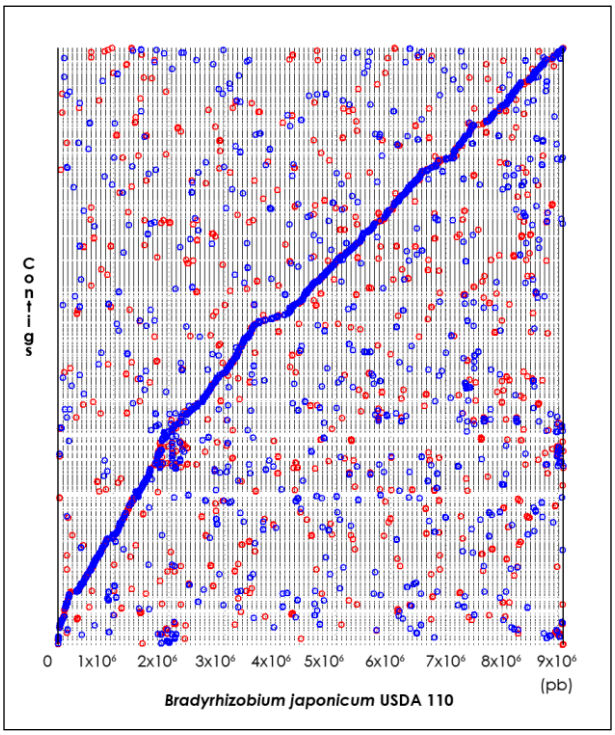

FIGURA 3.2 - ALINHAMENTO DOS *CONTIGS* DA MONTAGEM ALTERNATIVA UTILIZANDO LEITURAS SANGER MAIS *CONTIGS* VELVET AO GENOMA DE REFERÊNCIA

3.3.2 Montagem com linha de corte de 5 Kpb

Considerando o tamanho dos *contigs*/*scaffolds* N50 de todas as montagens obtidas de maneira automática, procurou-se filtrar a informação retornada pelos montadores a fim descartar informações redundantes ou pouco informativas.

Para tanto, estipulou-se uma linha de corte para exportação de *contigs*/*scaffolds* no montador Velvet de 5.000 pb, com o intuito de obter fontes de informação mais consistentes e excluir possíveis pequenas regiões repetidas, por exemplo.

A TABELA 3.4 mostra as características da montagem obtida utilizando o montador de leituras curtas Velvet com leituras Illumina e exportando apenas *contigs*/*scaffolds* com tamanho maior ou igual a 5.000 pb. Apesar do conjunto de *scaffolds* apresentar bons resultados, o conjunto de *contig*s é muito pouco representativo do genoma.

TABELA 3.4 - ESTATÍSTICAS DA MONTAGEM ALTERNATIVA UTILIZANDO LINHA DE CORTE PARA EXPORTAÇÃO DE *CONTIGS*/*SCAFFOLDS*

Característica \ Tipo de conjunto de dados	***Contigs***	***Scaffolds***
Tamanho total de genoma obtido (pb)	1.301.280	8.960.901
Número de *contigs*/*scaffolds*	200	367
Bases indeterminadas (pb)	0	264.980
Contig/*scaffold* N50	2.136	31.471
Contig/*scaffold* máximo (pb)	16.958	152.738
Leituras usadas (%)	13	82

O conjunto de *scaffolds* utilizando a estratégia com linha de corte de 5 Kpb mostrou melhores resultados que as montagens iniciais e também que as montagens com integração dos conjuntos de dados. O número de *scaffolds* foi o menor até então (367 *scaffolds*), o *scaffold* N50 foi de aproximadamente 31 Kpb, e o *scaffold* máximo foi substancialmente maior do que o da montagem utilizando o montador Velvet.

O conjunto de *scaffolds* não exportado pelo montador Velvet foi recuperado e armazenado para eventuais adições de dados ou realização de novas estratégias. Em sua composição, 560.405 pb formam 882 *scaffolds*, com conteúdo GC de aproximadamente 61,5%, e 3,6% de bases indeterminadas. A FIGURA 3.3 apresenta o gráfico de alinhamento do conjunto de *scaffolds* não exportado pelo montador Velvet ao genoma de referência *B. japonicum* USDA 110; dos 882 *scaffolds*, 129 alinharam com a referência, representando 185.581 pb (33,1%).

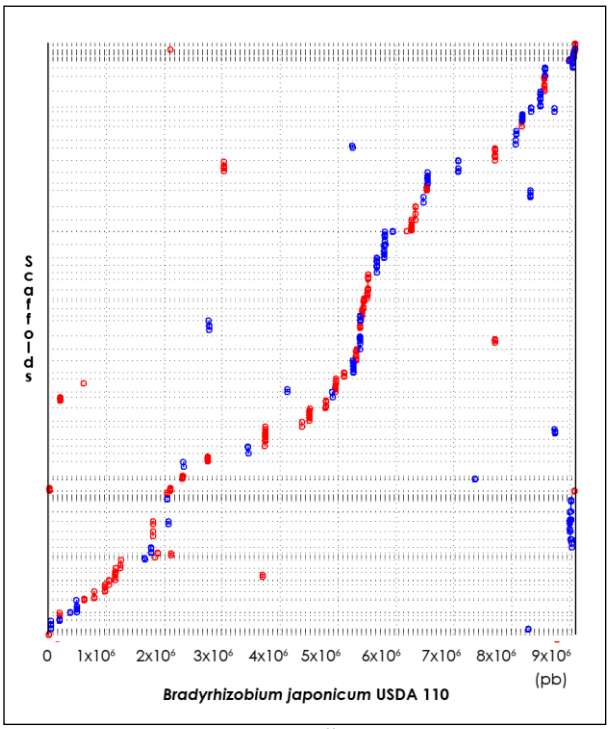

FIGURA 3.3 - ALINHAMENTO DOS *SCAFFOLDS* NÃO EXPORTADOS PELO MONTADOR VELVET AO GENOMA DE REFERÊNCIA

A estratégia de armazenamento dos *contigs*/*scaffolds* com tamanho inferior a 5 Kpb mostrou-se útil filtrando informações redundantes e facilitando assim o tratamento manual do conjunto de dados através de visualização; e não representou perdas de informação, já que o conjunto de dados não exportado foi armazenado para uso futuro.

3.3.3 Montagem com conjunto de dados alternativo

O conjunto de dados alternativo, produzido pela eliminação de bases de baixa qualidade e fragmentação das leituras Sanger em leituras curtas (descrito na seção 2.4), foi submetido ao montador Velvet juntamente com o conjunto de dados iniciais contendo as leituras Illumina.

A TABELA 3.5 mostra as características da montagem obtida utilizando tal conjunto de dados alternativos.

TABELA 3.5 - ESTATÍSTICAS DA MONTAGEM ALTERNATIVA UTILIZANDO O CONJUNTO DE DADOS ALTERNATIVOS

Característica	***Scaffolds***
Tamanho total de genoma obtido (pb)	8.778.872
Número de *scaffolds*	274
Bases indeterminadas (pb)	173.716
Scaffold N50	51.344
Scaffold máximo (pb)	259.954
Leituras usadas (%)	77

O conjunto de dados alternativo oriundo das leituras Sanger contribuiu para melhorar a montagem trazendo informações diferentes das disponíveis no conjunto de leituras Illumina. O ganho mais visível é no *scaffold* máximo, de 259.953 pb. É possível que este *scaffold* tenha sido obtido pela junção das extremidades de dois ou mais *scaffolds* através da sobreposição de leituras componentes do conjunto de dados alternativo. Além disto, o número de *scaffolds* foi reduzido para 274, o *scaffold* N50 foi 20 Kpb maior em relação à montagem inicial, e o número de bases representa menos de 2% do tamanho total do genoma.

3.4 Comparação entre montagens automáticas

A TABELA 3.6 estrutura uma comparação entre as características das montagens obtidas de maneira automática. A montagem inicial (1) refere-se à montagem automática utilizando os dados iniciais e o montador automático Velvet; a filtrada (2), representa a montagem obtida utilizando o filtro de 5 Kpb para exportação de *contigs*/*scaffolds*; por fim, a montagem alternativa (3), refere-se à montagem utilizando o filtro de 5 Kpb para exportação de *contigs*/*scaffolds* mais o conjunto de dados alternativos criado.

TABELA 3.6 - COMPARAÇÃO ENTRE AS MONTAGEM AUTOMÁTICAS

Característica \ Montagem	**(1) Inicial**	**(2) Filtrada**	**(3) Alternativa**
Tamanho total de genoma obtido (pb)	9.862.276	8.960.901	8.778.872
Número de *contigs*/*scaffolds*	1.249	367	274
Bases indeterminadas (pb)	296.901	264.980	173.716
Contig/*scaffold* N50	16.958	31.471	51.344
Contig/*scaffold* máximo (pb)	137.259	152.738	259.954
Leituras usadas (%)	90,5	82	77

O número de *contigs*/*scaffolds* obtido de maneira automática teve uma redução significativa de 1.249 na montagem inicial para 274 na alternativa. As bases indeterminadas foram reduzidas em aproximadamente 120 Kpb. O *contig*/*scaffold* passou a ter aproximadamente 51 Kpb e o tamanho máximo de *contig*/*scaffold* quase 260 Kpb.

3.5 Fechamento de falhas

Após os processos de montagem utilizando os montadores automáticos Velvet e Phrap, foi realizado o fechamento manual de falhas restantes na montagem através do uso da ferramenta Consed, como descrito na seção 2.5.

O processo de fechamento de falhas resultou na diminuição das bases indeterminadas contidas no conjunto de *scaffolds*, e, consequentemente, no aumento do número de *contigs* componentes do conjunto de *scaffolds*. As mudanças obtidas no conjunto de dados estão na TABELA 3.7.

TABELA 3.7 - ALTERAÇÕES NAS CARACTERÍSTICAS DO CONJUNTO DE DADOS APÓS FECHAMENTO DE FALHAS

Característica \ Conjunto de dados	**Antes do fechamento de falhas (*gaps*)**	**Após o fechamento de falhas (*gaps*)**
Tamanho total de genoma obtido (pb)	8.778.872	8.832.066
Número de *contigs*/*scaffolds*	274	260
Contigs	5/274	20/260
Scaffolds	269/274	240/260
Bases indeterminadas (pb)	173.716	108.132
Bases indeterminadas (%)	2	1,2
Contig/*scaffold* máximo (pb)	259.954	259.411

No total o processo de fechamento de falhas resultou em diminuição de 0,8% no número de bases indeterminadas e no aumento de 5 para 20 *contigs* na composição do conjunto de *contigs*/*scaffolds*. O aumento no tamanho total do genoma é resultado do fechamento de falhas subestimadas, onde o número de bases indeterminadas não refletia a quantidade real de bases componentes da lacuna.

3.6 Conjunto final de *scaffolds*

3.6.1 Estatísticas da montagem

A TABELA 3.8 mostra as características da montagem final obtida. Os valores exibidos entre parênteses nas linhas de *scaffolds* N50 e N90 são relativos às posições dos *scaffolds* na lista ordenada decrescente.

TABELA 3.8 - ESTATÍSTICAS DA MONTAGEM FINAL OBTIDA

Característica \ Tipo de conjunto de dados	***Scaffolds***
Tamanho total de genoma obtido (pb)	8.832.066
Número de *contigs*/*scaffolds*	260
Contigs	20/260
Scaffolds	240/260
Scaffold N50 (pb)	54.941 (46)
Scaffold N90 (pb)	15.072 (168)
Contig/*scaffold* máximo (pb)	259.411
Bases indeterminadas (pb)	108.132
Bases indeterminadas (%)	1,2
Leituras Illumina usados (%)	82
Leituras Sanger usados (%)	100

O tamanho total de genoma obtido (aproximadamente 8,8 Mpb) é coerente com o tamanho estimado entre 7 e 8 Mpb, já que deve sofrer alterações em seu valor com o fechamento das falhas restantes e verificação de possíveis regiões repetidas. Dentre o conjunto de 260 *contigs*/*scaffolds*, 20 são *contigs* e 240 são *scaffolds*. O *scaffold* N50 indica que 50% do genoma está representado em 46 *scaffolds* com pelo menos 54.941 pb.

A TABELA 3.9 mostra a distribuição dos tamanhos estimados das falhas: 50% das falhas são de no máximo 300 pb, e apenas 11% das falhas devem ser maiores que 1.000 pb.

TABELA 3.9 - DISTRIBUIÇÃO DE BASES INDETERMINADAS PELOS *SCAFFOLDS*

Intervalo de bases (pb)	**Porcentual de *scaffolds* (%)**
N <= 100	17%
100 < N <= 200	21%
200 < N <= 300	18%
300 < N <= 400	07%
400 < N <= 500	06%
500 < N <= 1000	20%
N > 1000	11%

3.6.2 DotPlot

A FIGURA 3.4 apresenta o gráfico de alinhamento do conjunto final de *scaffolds* de *B. elkanii* 587 ao genoma de referência *B. japonicum* USDA 110.

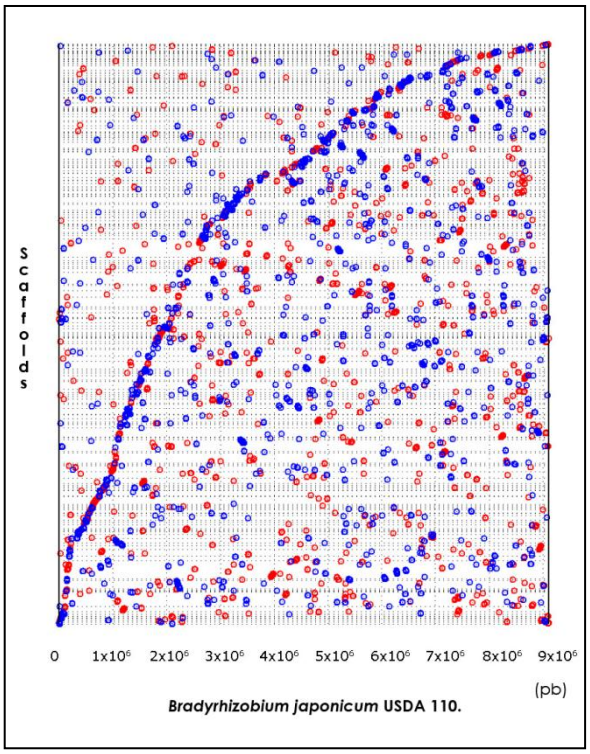

FIGURA 3.4 - ALINHAMENTO DO CONJUNTO FINAL DE *SCAFFOLDS* DE *Bradyrhizobium elkanii* 587 AO GENOMA DE REFERÊNCIA *Bradyrhizobium japonicum* USDA 110

Dos 260 *scaffolds* presentes no conjunto de dados, 226 alinharam com o genoma de *B. Japonicum* USDA 110, representando aproximadamente 87% da sequência de referência. Este resultando é concordante com a homologia de 77% do DNA total entre as duas espécies (SOARES, 2009).

3.6.3 GCSkew

O GCSkew, ou desvio do conteúdo GC, de um genoma é calculado através da razão da diferença entre o conteúdo de citosinas e guaninas (C-G) pela soma do conteúdo de guaninas e citosinas (G+C) presentes no genoma multiplicado por 100, resultando na porcentagem de excesso de citosinas sobre guaninas (LOBRY, 1996; GRIGORIEV, 1998; LOBRY, 1999).

A EQUAÇÃO 3.1 detalha o cálculo do GCSkew.

$$GCSkew_{(\%)} = \frac{C-G}{C+G} \quad (3.1)$$

onde:

$GCSkew_{(\%)}$ = desvio de conteúdo GC, em %;
C = número de citosinas presentes no genoma; e,
G = número de guaninas presentes no genoma.

A FIGURA 3.5 apresenta o gráfico GCSkew cumulativo do conjunto final de *scaffolds* de *B. elkanii* 587. O GCSkew de genomas completos apresenta um aumento contínuo e relativamente regular até um ponto máximo e então queda do valor de excesso de citosinas. O ponto de inflexão em geral está associado com a origem de replicação do genoma. A irregularidade apresentada no gráfico pode ser justificada pelo ordenamento incorreto de *contigs*/*scaffolds* sugerindo que o ordenamento baseado no alinhamento com o genoma de *B. Japonicum* USDA 110 deve ser corrigido futuramente.

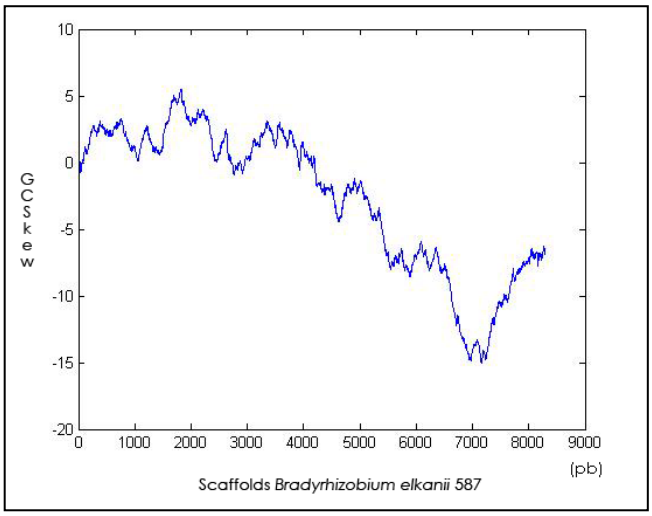

FIGURA 3.5 - GRÁFICO GCSKEW DO CONJUNTO FINAL DE *SCAFFOLDS* DE *Bradyrhizobium elkanii* 587

Em comparação, a FIGURA 3.6 apresenta o gráfico GCSkew cumulativo do genoma de referência *B. Japonicum* USDA 110.

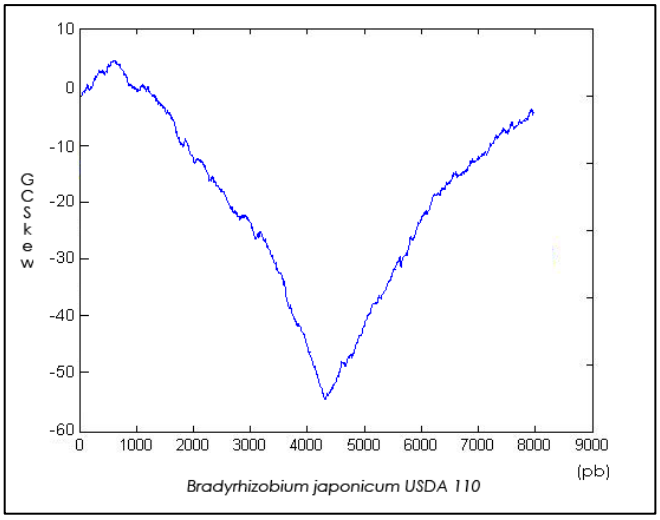

FIGURA 3.6 - GRÁFICO GCSKEW DO GENOMA DE REFERÊNCIA *Bradyrhizobium Japonicum* USDA 110

3.7 Pré-Anotação

O conjunto final de *scaffolds* obtido foi utilizado para obter informação sobre o conteúdo gênico de *B. elkanii* 587.

3.7.1 Características gerais do genoma de B. elkanii 587

A TABELA 3.10 mostra as características gerais do genoma de *B. elkanii* 587, obtidas pelo programa de anotação automática GAAT.

TABELA 3.10 - CARACTERÍSTICAS GERAIS DO GENOMA DE *Bradyrhizobium elkanii* 587

Características	**Valores**
Tamanho total de genoma obtido (pb)	8.832.066
Bases (pb)	A: 1.585.543 C: 2.779.086 G: 2.773.792 T: 1.585.513 N: 108.132
Conteúdo GC (%)	62,87
Total de regiões codificadoras (%)	81,75
ORFs codificadoras para proteínas	9.859
ORFs com função conhecida	5.028 (51%)
Tamanho médio das ORFs (pb)	732
Menor ORF (pb)	105
Maior ORF (pb)	7.548

O conteúdo GC (G+C) de 62,9% é próximo ao conteúdo de *B. japonicum* USDA 110, de 64,1%. Foram identificadas 9.859 ORFS potencialmente codificadoras de proteínas, dentre as quais 51% possuem função conhecida. Em *B. japonicum* USDA 110 foi identificado um número semelhante de 8317 ORFs sendo 52% de função conhecida (KANEKO *et al*., 2002). Uma vez que o genoma não está fechado e a anotação não foi verificada manualmente, estes números deverão ser revistos.

A FIGURA 3.7 exibe o alinhamento das ORFs encontradas pelo sistema GAAT ao genoma de *B. Japonicum* USDA 110. Dentre as 9.859 ORFs, 1947 (19,74%) alinham ao genoma de referência, representando aproximadamente 30% do total de regiões codificadoras.

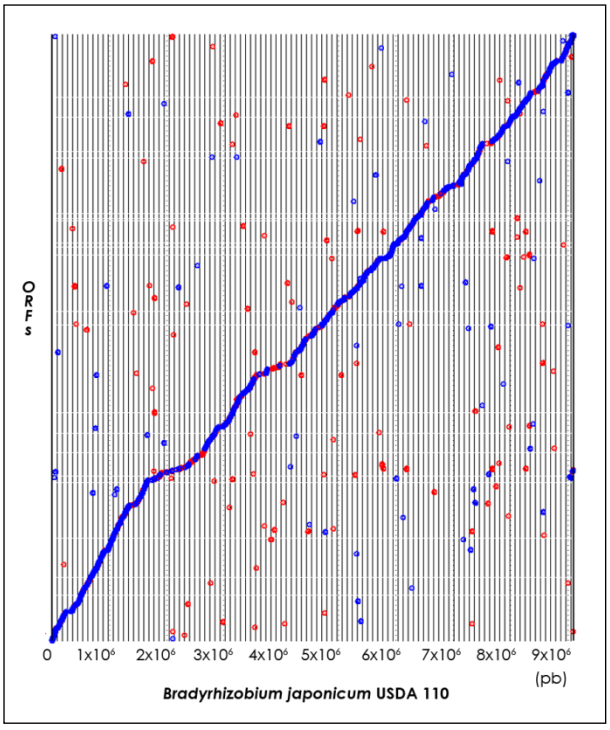

FIGURA 3.7 - ALINHAMENTO DAS ORFS DE *Bradyrhizobium elkanii* 587 AO GENOMA DE REFERÊNCIA *Bradyrhizobium japonicum* USDA 110

3.7.2 Grupos funcionais COG

A TABELA 3.11 apresenta a classificação dos genes identificados no genoma de *B. elkanii* 587 de acordo com os grupos funcionais COG.

TABELA 3.11 - GENES DE *Bradyrhizobium elkanii* 587 CLASSIFICADOS DE ACORDO COM OS GRUPOS FUNCIONAIS COG

Código	Descrição	Genes	Genes (%)
Processamento e armazenamento de informação			
J	Tradução, estrutura ribossomal e biogênese	151	1,53
A	Transformação e modificação de RNA	0	0,00
L	Transcrição	261	2,65
L	Replicação, recombinação e reparação	261	2,65
B	Estrutura e dinâmica da cromatina	4	0,04
Processos celulares e sinalização			
D	Controle do ciclo celular, divisão celular, e particionamento do cromossoma	31	0,31
Y	Estrutura nuclear	0	0,00
V	Mecanismos de defesa	110	1,12
T	Mecanismos de transdução de sinal	341	3,46
M	Biogênese da parede/membrana celular	327	3,32
N	Motilidade celular	102	1,03
Z	Citoesqueleto	0	0,00
W	Estruturas Extracelulares	2	0,02
U	Tráfico intracelular, secreção e transporte vesicular	131	1,33
O	Modificação pós-traducional, renovação de proteínas, e chaperonas	237	2,40
Metabolismo			
C	Produção e conversão de energia	499	5,06
G	Transporte e metabolismo de carboidratos	556	5,64
E	Transporte e metabolismo de aminoácidos	959	9,73
F	Transporte e metabolismo de nucleotídeos	92	0,93
H	Transporte e metabolismo de Coenzima	242	2,45
I	Transporte e metabolismo de lipídios	512	5,19
P	Transporte e metabolismo de íons inorgânicos	454	4,60
Q	Biossíntese, transporte e catabolismo de metabólitos secundários	343	3,48
Não Caracterizadas			
R	Função geral predita	864	8,76
S	Função não conhecida	414	4,20
Não anotadas			
-	-	4417	44,80

4 CONCLUSÃO

A montagem da sequência genômica parcial da bactéria diazotrófica *Bradyrhizobium elkanii* 587 foi obtida através da montagem de *contigs*/*scaffolds* utilizando as leituras de origem Illumina e Sanger e resultou em um conjunto de 260 *scaffolds*, com tamanho total de aproximadamente 8,8 Mpb e conteúdo GC de 62,87%. Para completar a sequência genômica, deve ser necessário obter novos dados de sequenciamento uma vez que os dados utilizados apresentaram uma cobertura relativamente baixa (entre 28x e 32x e entre 4x e 5x para leituras curtas e longas, respectivamente).

Para obter a sequência genômica parcial de *B. elkanii* 587 foram desenvolvidas estratégias utilizando montadores automáticos com o objetivo de se ter um conjunto de *contigs*/*scaffolds* a partir de um conjunto de dados relativamente restrito de leituras de duas plataformas de sequenciamento.

Foram desenvolvidos processos alternativos para a integração dos dados de diferentes plataformas em um único montador, resultando em uma grande melhoria na montagem obtida de maneira automática e trazendo uma nova visão do tratamento dos dados para os montadores em relação ao tipo de dado de entrada, sugerindo a necessidade de esforços para a integração automática de todo o conjunto de dados em um processo de montagem genômica.

O fechamento de falhas nos *scaffolds* utilizando leituras de origem Sanger e alinhamento com genoma de referência através da ferramenta Consed foi executado manualmente, resultando na diminuição de 14 *scaffolds* e de 0,8% no número de bases indeterminadas no conjunto final de *scaffolds* do genoma de *B. elkanii* 587.

A anotação preliminar da sequência obtida utilizando o programa GAAT permitiu uma primeira análise das características estruturais do genoma de *B. elkanii* 587. O conteúdo GC de *B. elkanii* 587 (62,9%) é semelhante com o conteúdo GC do genoma de referência *B. Japonicum* USDA 110 (64,1%). A anotação preliminar permitiu a identificação de conjunto de proteínas codificado pelo genoma de *B. elkanii* 587 sendo que aproximadamente 51% destas têm função conhecida.

4.1 Conclusões

- Uma sequência genômica parcial do genoma da bactéria diazotrófica *Bradyrhizobium elkanii* 587 foi obtida e está distribuída em 260 *scaffolds*, com tamanho total de aproximadamente 8,8 Mpb e conteúdo GC de 62,9%.
- Uma metodologia alternativa para a utilização de dados de sequência obtidos de plataformas diferentes (Illumina e Sanger) em montadores automáticos foi desenvolvida.

4.2 Desenvolvimentos Futuros

- Obtenção de novas sequências para resolução de repetições e aumentar a profundidade de cobertura de sequenciamento, com objetivo de preencher as falhas e completar a determinação da sequência genômica de *B. elkanii* 587.
- Desenvolvimento de metodologias e/ou ferramentas para a automatização do processo de fechamento de falhas.
- Anotar completamente a sequência parcial do genoma de *B. elkanii* 587, validando e verificando funções e produtos das ORFs encontradas.

REFERÊNCIAS

ADESSI, C., MATTON, G., AYALA, G., TURCATTI, G., MERMOD, J.-J., MAYER, P. *et al*. (2000). **Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms**. *Nucleic Acids Research, 28(20)*, 1-8.

ALBERTON, O., KASCHUK, G., & HUNGRIA, M. (2006). **Sampling effects on the assessment of genetic diversity of rhizobia associated with soybean and common bean**. *Soil Biology & Biochemistry, 38*, 1298-1307.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., & LIPMAN, D. J. (1990). **Basic Local Alignment Search Tool**. *Journal of Molecular Biology, 215*, 403-410.

ASMANN, Y. W., WALLACE, M. B., & THOMPSON, E. A. (2008). **Transcriptome profiling using next-generation sequencing**. *Gastroenterology, 135*, 1466-8.

BATZOGLOU, S., JAFFE, D. B., STANLEY, K., BUTLER, J., GNERRE, S., MAUCELI, E. *et al*. (2002). **ARACHNE: A Whole-Genome Shotgun Assembler**. *Genome Research, 12*, 177-189.

BATZOGLOU, S. (2005). **Algorithmic challenges in mammalian whole genome assembly**. *Encyclopedia of Genomics, Proteomics and Bioinformatics, .*, 1-16.

BENTLEY, D. R. (2006). **Whole-genome re-sequencing**. *Curr. Opin. Genet. Dev., 16*, 545-552.

CARVER, T., BERRIMAN, M., TIVEY, A., PATEL, C., BÖHME, U., BARRELL, B. G. *et al*. (2008). **Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database**. *Bioinformatics, 24(23)*, 2672-2676.

CHAN, E. Y. (2005). **Advances in sequencing technology**. *Mutation Research, 573*, 13-40.

CHEN, G. (2008). **DNA sequencing and short reads assembly**. *Roskilde University*, 1-60.

DELCHER, A. L., KASIF, S., FLEISCHMANN, R. D., PETERSON, J., WHITE, O., & SALZBERG, S. L. (1999). **Alignment of whole genomes**. *Nucleic Acids Research, 27(11)*, 2369-2376.

DELCHER, A. L., PHILLIPPY, A., CARLTON, J., & SALZBERG, S. L. (2002**). Fast algorithms for large-scale genome alignment and comparison**. *Nucleic Acids Research, 30(11)*, 2478-2483.

DHIMAN, N., SMITH, D. I., & POLAND, G. A. (2009). **Next-generation sequencing: a transformative tool for vaccinology**. *Expert Rev Vaccines., 8*, 963-7.

DOHM, J. C., LOTTAZ, C., BORODINA, T., & HIMMELBAUER, H. (2007). **SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing**. *Genome Research, 17*, 1697-1706.

EWING, B., HILLIER, L., WENDL, M. C., & GREEN, P. (1998). **Basecalling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Research, 8*, 175-185.

FEDURCO, M., ROMIEU, A., WILLIAMS, S., LAWRENCE, I., & TURCATTI, G. (2006). **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies**. *Nucleic Acids Research, 34(3)*, 1-13.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R. *et al*. (1995). **Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae Rd***. *Science, 269*, 1-15.

GARRITY, G. M., BELL, J. A., & LILBURN, T. (2005). **Family VII. *Bradyrhizobiaceae* fam. nov. In Bergey's Manual of Systematic Bacteriology - Second edition, Volume two. The Proteobacteria - part C (The Alpha-, Beta-, Delta-, and Epsilonproteobacteria)**. (G. M. Garrity, Ed.) New York: Springer.

GIONGO, A. (2007). **Diversidade de *Bradyrhizobium elkanii* e *B. japonicum* que nodulam soja em solos do Rio Grande do Sul**. Tese (Genética e Biologia Molecular). *Universidade Federal do Rio Grande do Sul*, 1-168.

GORDON, D., ABAJIAN, C., & GREEN, P. (1998). **Consed: a graphical tool for sequence finishing**. *Genome Research, 8*, 195-202.

GREEN, P. (1999). **Phrap**. Disponível em: *http://phrap.org*. Último acesso em: 24/01/2011.

GRIGORIEV, A. (1998). **Analyzing genomes with cumulative skew diagrams**. *Nucleic Acids Research, 26*, 2286-2290.

HARRIS, T. D., BUZBY, P. R., BABCOCK, H., BEER, E., BOWERS, J., BRASLAVSKY, I. *et al*. (2008). **Single-Molecule DNA Sequencing of a Viral Genome**. *Science, 320*, 1-5.

HAVLAK, P., CHEN, R., & DURBIN, K. J. (2004). **The Atlas genome assembly system**. *Genome Research, 14*, 721-732.

HUANG, X., WANG, J., ALURU, S., YANG, S.-P., & HILLIER, L. (2003). **PCAP: A Whole-Genome Assembly Program**. *Genome Research, 13*, 2164-2170.

HUNGRIA, M., DE, L. M., COCA, R. G., & MEGÍAS, M. (2001). **Preliminary characterization of fast growing rhizobial strains isolated from soyabean nodules in Brazil**. *Soil Biology and Biochemistry, 33*, 1349-1361.

HUNGRIA, M., FRANCHINI, J. C., CAMPO, R. J., & GRAHAM, P. H. (2005). **The Importance of Nitrogen Fixation to Soybean Cropping in South America**. *Nitrogen Fixation in Agriculture, Forestry, Ecology, and the Environment, 3*, 25-42.

HUNKAPILLER, T., KAISER, R., KOOP, B., & HOOD, L. (1991). **Large-scale and automated DNA sequence determination**. *Science, 254*, 59-67.

IDURY, R. M., & WATERMAN, M. S. (1995). **A New Algorithm for DNA Sequence Assembly**. *Journal of Computacional Biology, 2(2)*, 291-306.

IMELFORT, M., DURAN, C., & ANDDAVID, J. B. (2009). **Discovering genetic polymorphisms in next-generation sequencing data**. *Plant Biotechnology J., 7*, 312.

JECK, W. R., REINHARDT, J. A., BALTRUS, D. A., HICKENBOTHAM, M. T., MAGRINI, V., MARDIS, E. R. *et al*. (2007). **Extending assembly of short DNA sequences to handle error**. *Bioinformatics, .23*, 2942-2944.

JORDAN, D. C. (1982). **Transfer of *Rhizobium japonicum Buchanan* 1980 to *Bradyrhizobium* gen. nov., a Genus of Slow-Growing, Root Nodule Bacteria from Leguminous Plants**. *International Journal of Systematic Bacteriology, 32*, 136-139.

KANEKO, T., NAKAMURA, Y., SATO, S., MINAMISAWA, K., UCHIUMI, T., SASAMOTO, S. *et al*. (2002). **Complete Genomic Sequence of Nitrogen-fixing Symbiotic Bacterium *Bradyrhizobium japonicum* USDA 110**. *DNA Research, 9*, 189-197.

KISHI, L. T., *et al*. (2005). **Avaliação do tamanho do genoma de *Bradyrhizobium elkanii* por Eletroforese em Campo Pulsado (PFGE)**. In: *XXIII Congresso Brasileiro de Microbiologia*, Santos. Anais do Congresso Brasileiro de Microbiologia, 23, 277.

KISHI, L. T., *et al*. (2007). **Análise Comparativa das sequências de genes da ilha simbiótica entre *Bradyrhizobium elkanii* e *Bradyrhizobium japonicum***. Tese (Microbiologia Agropecuária). *Faculdade de Ciências Agrárias e Veterinárias - Universidade Estadual Paulista*, 1-75.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. *et al*. (2004). **Versatile and open software for comparing large genomes**. *Genome Biology, 5(2)*, 1-9.

KUYKENDALL, D. L., SAXENA, B., DEVINE, T. E., & UDELL, S. E. (1992). **Genetic diversity in *Bradyrhizobium japonicum* Jordan 1982 and a proposal for *Bradyrhizobium elkanii* sp.nov**. *Canadian Journal of Microbiology, 38(6)*, 501-505.

LEMOS, M., BASÍLIO, A., & CASANOVA, M. A. (2003). **Um Estudo dos Algoritmos de Montagem de Fragmentos de DNA**. Monografia (Ciência da Computação). *Pontifícia Universidade Católica do Rio de Janeiro*, 1-42.

LOBRY, J. R. (1996). **A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria**. *Biochimie, 78*, 323-326.

LOBRY, J. R. (1999). **Genomic landscapes**. *Microbiology Today, 26*, 164-165.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A. *et al*. (2005). **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature, 437*, 376-80.

MAXAM, A. M., & GILBERT, W. (1977). **A new method for sequencing DNA**. *Vol. 74, No. 2, pp. 560-564, February 1977 Biochemistry, 74(2)*, 560-564.

MENNA, P., HUNGRIA, M., BARCELLOS, F. G., BANGEL, E. V., & PABLO, E. M.-R. (2006). **Molecular phylogeny based on the 16S rRNA gene of elite rhizobial strains used in Brazilian commercial inoculants**. *Systematic and Applied Microbiology, 29*, 315-332.

MORIYA, Y., ITOH, M., OKUDA, S., YOSHIZAWA, A. C., & KANEHISA, M. (2007). **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Research, 35*, 1-4.

MOROZOVA, O., & MARRA, M. A. (2008). **From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors**. *Biochemistry and Cell Biology, 86*, 81-91.

MULLIKIN, J. C., & NING, Z. (2003). **The Phusion Assembler**. *Genome Research, 13*, 81-90.

MYERS, E. W., SUTTON, G. G., DELCHER, A. L., DEW, I. M., FASULO, D. P., FLANIGAN, M. J. *ET AL*. (2000). **A Whole-Genome Assembly of *Drosophila***. *Science, 2196-2204*, 1-10.

PEVZNER, P. A., TANG, H., & WATERMAN, M. S. (2001). **An Eulerian Path Approach to DNA fragment assembly**. *Proceedings of the National Academy of Sciences of the United States of America, 98(7)*, 9748-9753.

PISA, F. R. O., *et al*. (2010). **Desenvolvimento de programa integrado de montagem e anotação de genomas**. In: *18º Evento de Iniciação Científica (EVINCI - UFPR)*, 18, 148.

RHIJN, P. V., & VANDERLEYDEN, J. (1995). **The Rhizobium-Plant Symbiosis**. *Microbiological Reviews, 59(1)*, 124-142.

RUMJANEK, N. G., DOBERT, R. C., BERKUM, P. V., TRIPLETT, E. W. (1993). **Common Soybean Inoculant Strains in Brazil Are Members of *Bradyrhizobium elkanii***. *Applied and Environmental Microbiology, 59(12)*, 4371-4373.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M.-A. *et al*. (2000). **Artemis: sequence visualization and annotation**. *Bioinformatics, 16(10)*, 944-945.

SADOWSKY, M. J., KOSSLAK, R. M., MADRZAK, C. J., GOLINSKA, B., & CREGAN, P. B. (1995). **Restriction of Nodulation by *Bradyrhizobium japonicum* Is Mediated by Factors Present in the Roots of Glycine max**. *Applied and Environmental Microbiology, 61(2)*, 832-836.

SALZBERG, S. L., DELCHER, A. L., KASIF, S., & WHITE, O. (1998). **Microbial gene identification using interpolated Markov models**. *Nucleic Acids Research, 26(2)*, 544-548.

SANGER, F., AIR, G. M., BARRELL, B. G., BROWN, N. L., COULSON, A. R., FIDDES, J. C. *et al*. (1977). **Nucleotide sequence of bacteriophage phiX174 DNA**. *Nature, 265*, 687-695.

SHENDURE, J., MITRA, R. D., VARMA, C., & CHURCH, G. M. (2004). **Advanced sequencing technologies: methods and goals**. *Nat. Rev. Genet., 5*, 335-344.

SHENDURE, J., PORRECA, G. J., REPPAS, N. B., LIN, X., MCCUTCHEON, J. P., ROSENBAUM, A. M. *et al*. (2005). **Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome**. *Science, 309*, 1728-1732.

SHENDURE, J., & HANLEE, J. (2008). **Next-generation DNA sequencing**. *Nature Biotechnology, 26*, 1135-1145.

SIMPSON, J. T., WONG, K., JACKMAN, S. D., SCHEIN, J. E., JONES, S. J., & BIROL, I. (2009). **ABySS: A parallel assembler for short read sequence data**. *Genome Research, 19*, 1117-1123.

SOARES, R. A. (2009). **Diferenças genômicas entre a estirpe *Bradyrhizobium elkanii* SEMIA 587 e a estirpe de referência *B. japonicum* USDA 110***.*** Dissertação (Genética e Biologia Molecular). *Universidade Federal do Rio Grande do Sul*, 1-58.

SPRENT, J. I. (1995). **Legume Trees and Shrubs in the Tropics: N2 Fixation in Perspective**. *Soil Biology and Biochmestry, 27*, 401-407.

STEIN, L. (2001). **Genome annotation: from sequence to biology**. *Nature reviews. Genetics, 2(7)*, 493-503.

SUZEK, B. E., ERMOLAEVA, M. D., SCHREIBER, M., & SALZBERG, S. L. (2001). **A probabilistic method for identifying start codons in bacterial genomes**. *Bioinformatics, 17(12)*, 1123-1130.

SWERDLOW, H., WU, S.-L., HARKE, H., & DOVICHI, N. J. (1990). **Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette**. *J. Chromatogr, 516*, 61-67.

TATUSOV, R. L., GALPERIN, M. Y., NATALE, D. A., & KOONIN, E. V. (2000). **The COG database - a tool for genome-scale analysis of protein functions and evolution**. *Nucleic Acids Research, 28(1)*, 33-36.

TURCATTI, G., ROMIEU, A., FEDURCO, M., & TAIRI, A.-P. (2008). **A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis**. *Nucl, 36(4)*, 1-13.

VARSHNEY, R. K., NAYAK, S. N., MAY, G. D., & JACKSON, S. A. (2009). **Next-generation sequencing technologies and their implications for crop genetics and breeding**. *Trends Biotechnol., 27*, 522-30.

VOELKERDING, K. V., DAMES, S. A., & DURTSCHI, J. D. (2009). **Next-generation sequencing: from basic research to diagnostics**. *Clin Chem, 55*, 641-58.

WARREN, R. L., SUTTON, G. G., M., S. J., & HOLT, R. A. (2007). **Assembling millions of short DNA sequences using SSAKE**. *Bioinformatics, 4*, 500-501.

WEISS, V. A. (2010). **Estratégias De Finalização Da Montagem Do Genoma Da Bactéria Diazotrófica Endofítica Herbaspirillum Seropedicae SmR1**. Dissertação (Bioquímica). *Universidade Federal do Paraná*, 1-72.

WILLEMS, A. (2006). **The taxonomy of *rhizobia*: an overview**. *Plant And Soil, 287*, 3-14.

ZERBINO, D. R., MCEWEN, G. K., MARGULIES, E. H., & BIRNEY, E. (2009). **Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read de Novo Assembler**. *PLoS ONE, 4*, 1-9.

ZERBINO, D., & BIRNEY, E. (2008). **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Research, 18*, 821-829.

GLOSSÁRIO

***Contig* -** Sequência contígua de bases de DNA, sem a presença de falhas (*gaps*), representando todo, ou parte de, um genoma.

DNA - (*Deoxyribonucleic acid*) Ácido desoxirribonucleico.

***Draft* genômico -** Sequência de DNA antes de ser finalizada, contendo múltiplas falhas (*gaps*), áreas não representadas, e erros de montagem. Além disso, a taxa de erro de *drafts* é superior à taxa padrão de 1 em 10.000 para genomas fechados (Stein, 2001).

***Gap* -** Falha. Conjunto de bases indeterminadas (*N*), de tamanho estimado, representando uma ligação entre duas sequências comprovada por algum indício.

***k-mer* -** Palavra, ou semente, com tamanho pré-definido *k*, representante de uma ou mais leituras.

ORF - (*Open Reading Frame*) Fase ou quadro aberto de leitura.

pb - Pares de bases.

PCR - (*Polymerase Chain Reaction*) Reação em cadeia da polimerase.

***Primer* -** Oligonucleotídeos iniciadores do processo de PCR.

***read* -** Fragmento de leitura oriundo de um método de sequenciamento, composto por bases de DNA, representando parte de um genoma.

RBS - (*Ribosomal Binding Site*) Sítio de ligação ribossomal.

rRNA - (*Ribosomal ribonucleic acid*) Ácido ribonucleico ribossomal.

***Scaffold* -** Sequência de bases de DNA, com a presença de falhas (*gaps*), representando todo, ou parte de, um genoma, podendo ser composto por dois ou mais *contigs*.

tRNA - (*Transfer ribonucleic acid*) - Ácido ribonucleico transportador.