Universidade Federal de São Paulo
Instituto de Ciência e Tecnologia
Departamento de Ciência e Tecnologia

Projeto de Pesquisa

Grafos de Bruijn para identificação de variações genéticas usando genotipagem-por-sequenciamento

Mestrado em Ciência da Computação
Programa de Pós-Graduação em Ciência da Computação

Proponente: Marcos Castro de Souza
Orientador: Prof. Dr. Reginaldo Massanobu Kuroshu

Setembro de 2015

RESUMO

Com o surgimento de tecnologias NGS (*Next-Generation Sequencing*), teve-se uma revolução no processo de sequenciamento devido ao custo reduzido, eficiência e versatilidade. O surgimento dessas novas tecnologias trouxe desafios para a Bioinformática devido a um grande volume de dados produzidos pelos sequenciadores. Para lidar com essa grande quantidade de informação, tem-se uma demanda em relação a construção de algoritmos eficientes e eficazes para a manipulação dessa imensa quantidade de dados com o objetivo de extrair informações relevantes visando a um melhor entendimento do genoma de uma espécie.

A cana-de-açúcar é uma planta que possui uma grande importância econômica devido à sua múltipla utilidade como é o caso da produção de biocombustível. Apesar dos grandes avanços tecnológicos, essa espécie de planta possui uma complexa organização genômica e, portanto, ainda não possui seu genoma totalmente sequenciado. Ferramentas que buscam entender melhor o genoma da cana-de-açúcar identificando variações genéticas visam contribuir para o melhoramento genético de uma das espécies economicamente mais importantes do país.

Várias ferramentas de alinhamento e montagem surgiram com as novas tecnologias de sequenciamento. Porém, as soluções que existem atualmente não permitem a descoberta de SNPs (*single nucleotide polymorphism)* de todas as regiões sequenciadas por GBS (*Genotyping By Sequencing*). Este projeto tem como objetivo construir um método baseado em grafos *de Bruijn* para identificação de variações genéticas a partir de dados obtidos por GBS. Após a construção do método, serão feitos testes com dados simulados e dados reais de uma população de mapeamento de cana-de-açúcar.

ABSTRACT

With the emergence of technologies NGS (Next-Generation Sequencing, there was a revolution in the sequencing process due to the reduced cost, efficiency and versatility. The emergence of these new technologies has brought challenges for Bioinformatics because of a large amount of data produced by the sequencers. For to deal with this amount of information, has been a demand for the construction of efficient and effective algorithms for handling this large amount of data in order to extract relevant information aimed at better understanding the genome of a species.

The sugarcane is a plant that it has a great importance economic because of its multiple uses such as the production of biofuels. Despite major technological advances, this plant species has a complex genomic organization and, therefore, does not have its genome fully sequenced. Tools that seek to better understand the genome of sugarcane identifying genetic variations aims to contribute to the genetic improvement of one of the most economically important species of the country.

Several alignment and assembly tools have emerged with the new sequencing technologies. However, the solutions that there currently do not allow the discovery of SNPs (single nucleotide polymorphism) from all regions sequenced by GBS (Genotyping By Sequencing). This Project aims to build a method based on Bruijn graphs to identify genetic variations from data obtained by GBS. After the construction of the method, will be made tests with simulated and actual data from a population of sugarcane mapping.

1. Introdução

Este documento tem como objetivo apresentar o plano de trabalho do projeto de mestrado em Ciência da Computação. A subseção 1.1 apresenta um breve histórico sobre sequenciamento, definições e tecnologias de sequenciamento. A subseção 1.2 apresenta a Bioinformática. A subseção 1.3 descreve a técnica de genotipagem por sequenciamento (GBS). A subseção 1.4 tem como objetivo versar sobre a identificação de variações genéticas em cana-de-açúcar, complexidade e desafios. Na subseção 1.5 é apresentado o problema da montagem de *reads*. Na subseção 1.6 são descritos alguns paradigmas para a montagem *de novo*. A subseção 1.7 descreve os grafos *de Bruijn*. Na seção 2 é apresentada a justificativa desse projeto. A seção 3 descreve os objetivos a serem alcançados. A seção 4 apresenta a metodologia que será seguida para atingir os objetivos. Na seção 5 é apresentado o cronograma de atividades.

1.1 Sequenciamento

Em biologia molecular, o sequenciamento é o processo de determinar a ordem dos nucleotídeos de um DNA, RNA ou proteína. O sequenciamento do genoma de uma determinada espécie oferece muitas informações a respeito dela. Pode-se citar como exemplo o Projeto Genoma Humano (PGH) que consistiu em um esforço conjunto de vários laboratórios com o intuito de sequenciar 3,1 bilhões de bases nitrogenadas do genoma humano [1]. O PGH tinha como objetivo decifrar o genoma humano através de um processo chamado mapeamento genético humano. São vários os benefícios provenientes do PGH tais como a prevenção ou até mesmo a cura de doenças genéticas contribuindo para uma melhoria na qualidade de vida das pessoas.

O primeiro método de sequenciamento foi o método de Sanger (primeira geração) que mostrou que era possível compreender genomas através das bases do DNA: adenina (A), citosina (C), timina (T) e guanina (G). Após Sanger, veio a segunda geração de sequenciadores com o surgimento do sequenciamento de nova geração (*next generation sequencing* - NGS) que trouxe como benefícios um melhor custo por pares de bases, rapidez, eficiência e versatilidade [2]. *Roche 454*, *Solid* e *Illumina* são exemplos de tecnologias NGS [3].

As tecnologias NGS têm contribuído para avanços científicos significativos nos estudos de evolução das espécies, doenças humanas e agricultura. Essas tecnologias não necessitam de um grande volume de DNA (poucas amostras) e geram milhares de fragmentos de DNA (*reads*) por amostra ao contrário do método de *Sanger* que gera apenas um *read* por amostra. Com a geração de um grande volume de dados, as tecnologias NGS permitem uma melhor visão sobre os genomas facilitando o entendimento dos mesmos.

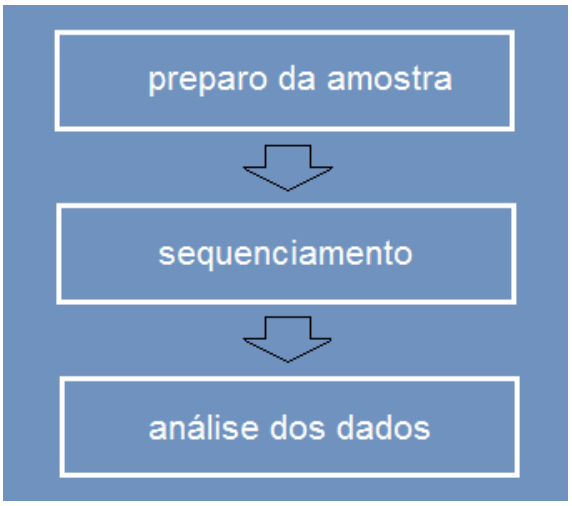

Figura 1: *Pipeline* do processo de sequenciamento NGS

- Preparo da amostra: o DNA é fragmentado através de algum processo químico ou enzimático. Cada fragmento desses é um modelo (*template*). Após a fragmentação, adaptadores (sequências de DNA artificiais conhecidas) são incorporados nas extremidades dos *templates* e, assim, realizam a sua extensão.
- Sequenciamento: a máquina sequenciadora executa uma série de reações químicas produzindo os *reads*.
- Análise dos dados: são gerados arquivos com os *reads* onde cada *read* contém as bases de um fragmento sequenciado. Esses fragmentos precisam ser montados para obter a sequência completa.

1.2 Genotipagem por sequenciamento

A técnica de genotipagem por sequenciamento (*genotyping by sequencing* – GBS) é um método que utiliza a plataforma de segunda geração *Illumina*. A técnica GBS se baseia no uso de enzimas de restrição para reduzir a complexidade do genoma do organismo em estudo através da fragmentação do DNA em diversos sítios reconhecidos pela enzima obtendo pequenos fragmentos de DNA [4].

Estes fragmentos são ligados a adaptadores *barcodes* para a produção de bibliotecas multiplexadas de amostras a serem sequenciadas, sendo que cada *barcode* identifica unicamente um indivíduo da população que está sendo estudada. O método GBS consegue evitar, na prática, a inclusão de regiões repetitivas criando uma representação reduzida do genoma, facilitando, portanto, o processo de análise computacional.

GBS possibilita, a um custo bastante reduzido, a descoberta de um grande número de marcadores moleculares de polimorfismo de base única (*single nucleotide polymorphism* – SNP) que são fundamentais para análises de genomas. SNP é uma variação genética que ocorre quando um único nucleotídeo (A, T, C ou G) é alterado e essa alteração é mantida ao logo das gerações.

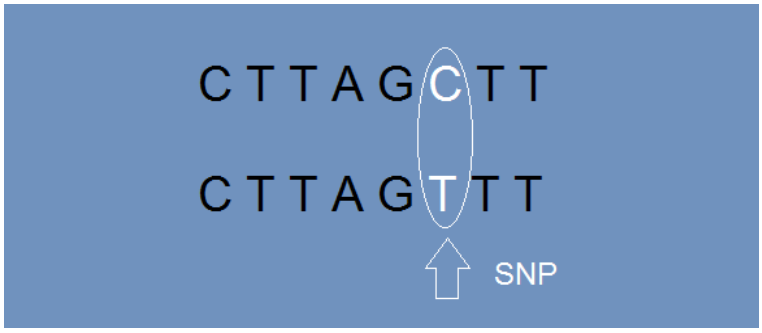

Figura 2: Marcador molecular SNP

O uso de marcadores SNPs se baseia no dogma central da biologia molecular: a informação genética do nosso organismo é unidirecional.

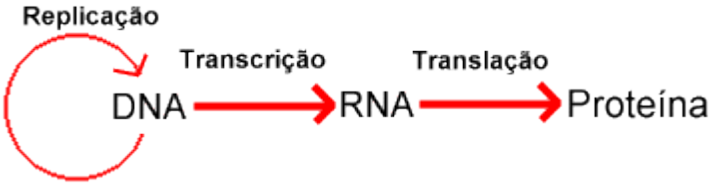

Figura 3: Dogma central da biologia molecular

A detecção de SNPs ajuda a compreender as variações genéticas entre indivíduos [5]. Esses marcadores são utilizados no estudo de diversas doenças humanas e vem sendo utilizados também para interpretar a variação de cada gene em espécies, entre elas plantas poliploides, como a cana-de-açúcar. Marcadores SNPs são muito utilizados em programas de melhoramento genético das características de interesse econômico. No caso da cana-de-açúcar, o melhoramento genético é realizado com o objetivo de desenvolver variedades mais produtivas, com maior resistência às pragas, maior tolerância ao estresse hídrico dentre outras melhorias.

1.3 Bioinformática

A bioinformática é uma área que busca analisar informações provenientes de estudos biológicos. Trata-se de uma área multidisciplinar que combina conhecimentos de várias áreas com o objetivo de resolver problemas da biologia através de ferramentas computacionais. Os estudos de genomas produzem um grande volume de dados, por isso existe uma necessidade cada vez maior do desenvolvimento de novos programas de computador para o armazenamento, relacionamento e análise computacional desses dados. A geração de dados deve ser acompanhada por uma extração de conhecimento que possa servir como ponto de partida para a produção de novos conhecimentos científicos. Tão importante quanto extrair algum conhecimento, é também realizar essa tarefa de forma eficiente por causa da imensa quantidade de dados a serem processados.

1.4 Identificação de variações genéticas utilizando genotipagem por sequenciamento de uma população de cana-de-açúcar

A cana-de-açúcar (*Saccharum* ssp.) é uma espécie de planta que possui várias utilidades. Existe um grande interesse em seus derivados tais como o açúcar e etanol fazendo com que essa espécie possua um elevado interesse econômico. O Brasil lidera a lista dos países produtores de cana-de-açúcar e também é o primeiro do mundo na produção de açúcar e etanol [6]. O uso do biocombustível como alternativa energética impulsiona os investimentos em programas de melhoramento com o objetivo de entender o genoma da cana-de-açúcar que, até o momento, ainda não foi totalmente sequenciado, dado sua complexidade genética.

Esta espécie é um organismo poliploide (múltiplos cromossomos homólogos), alógama (grande variabilidade genética logo na primeira geração de cruzamento), possui uma frequente aneuploidia (alteração genética) e um genoma de 10Gbp [7]. Por conta desses fatores, a cana-de-açúcar possui um genoma altamente complexo que ainda não é conhecido completamente. A técnica de GBS busca otimizar o processo de melhoramento da cana-de-açúcar fazendo com que ela possa se beneficiar de importantes recursos genômicos que contribuirão para o seu melhoramento e conservação, mesmo não tendo a sua sequência genômica total decifrada.

A tarefa de identificar variações genéticas a partir de sequências obtidas por GBS é desafiadora. Isso se deve ao grande volume de dados gerados pelas novas tecnologias de sequenciamento. Cada tecnologia gera dados com características particulares como a quantidade de fragmentos lidos, comprimentos desses fragmentos e padrão de erros na leitura. A maior dificuldade a ser enfrentada com o uso de fragmentos curtos é a baixa especificidade em solucionar sequências repetitivas, que é agravada na presença de erros de sequenciamento e polimorfismos [8]. Por conta disso, a dificuldade em analisar repetições é um problema computacional que merece uma atenção especial.

Ferramentas de alinhamento e montagem (*assembly*) de sequências vêm sendo desenvolvidas utilizando novas técnicas para satisfazer as necessidades surgidas com os sequenciadores de segunda geração. Identificar variações genéticas requer precauções para que variações reais possam ser diferenciadas de erros de leitura. Para isso, métodos que considerem a complexidade do genoma bem como a profundidade de sequenciamento devem ser desenvolvidos para que essas variações possam ser identificadas com precisão.

Existem várias ferramentas para identificação de variantes como substituições e pequenas inserções e deleções (*indels*), assim como existem ferramentas específicas para a análise de dados obtidos por GBS. Porém, a maioria dessas ferramentas requer a existência de uma sequência referência, preferencialmente o genoma. Contudo, a sequência do genoma total da cana-de-açúcar ainda não está disponível, o que existem são sequências transcritas obtidas por RNA-Seq (forma mais acurada de determinar os níveis de expressão gênica) [9] e sequências de clones BACs (Cromossomo Artificial de Bactéria) para serem usadas como referências.

As soluções que existem atualmente não permitem a descoberta de SNPs de todas as regiões sequenciadas por GBS. Tassel [10] e GATK [11] são exemplos de *softwares* que implementam

métodos para identificação e análises de variações genéticas. Neste projeto, propõe-se o uso de grafos *de Bruijn* para analisar os dados obtidos através de GBS e identificar as variações genéticas como alternativa aos métodos existentes.

1.5 O problema da montagem de reads

Os *reads* produzidos pelas tecnologias NGS são bem menores que os produzidos pela tecnologia *Sanger*. O tamanho e a quantidade de fragmentos produzidos pelas tecnologias NGS representam um grande desafio em Bioinformática para a montagem de *reads* curtos (*short reads*). É mais complexo trabalhar com *short reads* por causa do aumento da possibilidade de ocorrerem sobreposições entre as regiões diminuindo a probabilidade de unir corretamente estas sobreposições.

A necessidade de novos métodos para manipular um volume maior de dados bem como *reads* menores é fundamental para acompanhar as mudanças provenientes da evolução das tecnologias de sequenciamento. O objetivo da montagem é a obtenção da sequência original a partir dos *reads*. A montagem é o ponto de partida dos trabalhos de pós-sequenciamento onde se busca conhecer a sequência de DNA de um indivíduo ou de um gene.

O problema da montagem de *reads* é dependente de tecnologias de sequenciamento. Uma boa solução computacional além de ser fundamental para o processo de montagem desses *reads*, é muito importante para o sucesso das tecnologias de sequenciamento. O genoma da cana-de-açúcar, assim como o genoma de outras espécies, possui regiões repetitivas que dificultam muito o processo de montagem. É necessário o estudo e desenvolvimento de algoritmos eficientes e eficazes que possam tirar proveito das informações contidas nos *reads* e que solucionem satisfatoriamente o problema de repetição.

O resultado de uma montagem é um conjunto de contigs. Um contig (*contiguous assembly*) é uma sequência montada a partir da sobreposição dos *reads.* A partir de cada contig é gerada uma sequência consenso (*consensus sequence*) que deve representar a sequência de bases original do DNA fragmentado. A sequência consenso é uma sequência de nucleotídeos ou aminoácidos similares ou idênticos entre regiões de homologia em diferentes sequências de DNA, RNA ou proteína.

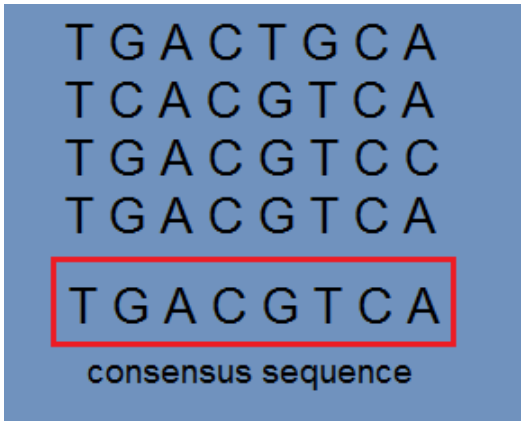

Figura 4: Sequência consenso

Quando não há um genoma de referência, a montagem é chamada *de novo*. Existem algumas medidas que avaliam uma montagem *de novo* tais como o tamanho médio dos contigs, número de contigs maiores que 1000 pb (pares por base) e N50 [12]. Para o cálculo do N50 por exemplo, suponha que um genoma de 300 pb produziu 8 contigs de tamanhos 3, 3, 15, 24, 39, 45, 54 e 117. Para obter o N50, basta ordenar esses contigs em ordem decrescente de tamanho e somar um a um. Quando a soma ultrapassar 150 (metade de 300), o tamanho do contig da vez é o N50. No caso do exemplo em questão, o N50 seria o contig de tamanho 54, ou seja, o tamanho do segundo maior contig.

Conjuntos de contigs que podem ser colocados numa mesma região são chamados de supercontigs ou *scaffolds*. Mesmo que não se tenha a sequência entre dois contigs, a informação de que eles são vizinhos é muito importante. A construção de *scaffolds* pode ser modelada em grafos através do caminho de custo mínimo.

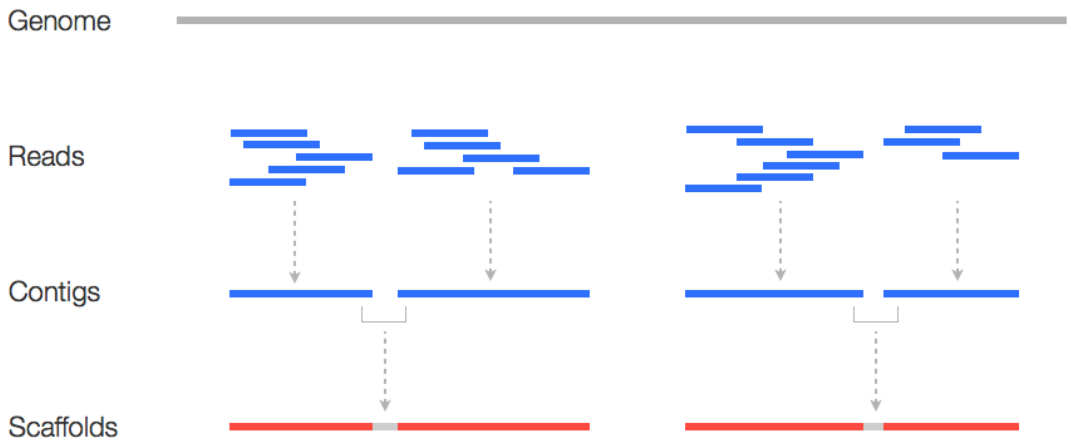

Figura 5: Genoma, reads, contigs e scaffolds

Os *reads* são fragmentos sequenciados que através de um processo de montagem, são identificadas as sequências que se sobrepõem formando os contigs que por sua vez podem formar *scaffolds*.

1.6 Paradigmas de montagem de novo

A montagem *de novo* não utiliza nenhum tipo de referência. Existem três paradigmas muito utilizados na montagem *de novo:* guloso (*greedy*), *overlap-layout-consensus* (OLC) e grafos *de Bruijn* [13].

Ao usar o paradigma guloso, o *assembler* sempre faz escolhas locais, ou seja, sempre junta os *reads* que se sobrepõem melhor. Dado um conjunto de fragmentos, o problema se resume em encontrar a menor supersequência comum (*shortest common supersequence*).

Os passos do paradigma guloso podem ser descritos da seguinte forma:

1) Realizam-se os alinhamentos *pairwise* de todos os fragmentos;
2) Escolhe-se dois fragmentos com a maior sobreposição;

3) Junta-se os fragmentos escolhidos;
4) Repete-se os passos 2 e 3 até restar somente um fragmento.

O resultado da abordagem gulosa é uma solução subótima. Por conta de ser um paradigma guloso (faz escolhas locais), dificulta o uso de informação global prejudicando na montagem de genomas repetitivos como é o caso do genoma da cana-de-açúcar. Essa abordagem não funciona bem com *reads* curtos (Illumina/Solid).

A abordagem *overlap-layout-consensus* (OLC) usa a estrutura de grafo sobreposto (*overlap graph*). Cada *read* representa um nó e as arestas são as sobreposições entre pares de *reads*. É realizado um alinhamento par a par entre todos os *reads* sequenciados para detectar sobreposições. *Layout* é a ordenação (orientação) dos *reads* de acordo com as sobreposições. *Consensus* refere-se à reconstrução da sequência do genoma através do alinhamento múltiplo dos *reads* de acordo com o *layout.* OLC é uma abordagem robusta para *reads* longos, mas é desvantajosa para o sequenciamento de nova geração por causa do custo de tempo que chega a ser quadrático. Esse paradigma foi bastante utilizado até o surgimento de tecnologias de sequenciamento de segunda geração.

1.7 Grafos de Bruijn

Grafos de sobreposição (*overlap graphs*) podem ser muito custosos para manipular dados provenientes de tecnologias NGS, pois mesmo para organismos simples, é preciso milhões de *reads* tornando o grafo de sobreposição muito grande. Por conta disso, grafos de sobreposição não escalam bem com o aumento do número de *reads*, por isso a maioria dos *assemblers* para NGS utilizam grafos *de Bruijn* com o objetivo de reduzir o esforço computacional através da quebra de *reads* em sequências ainda menores chamadas *k-mers* onde a letra "k" corresponde ao tamanho em bases dessas sequências.

A escolha do "k" é um *trade-off* entre especificidade e sensibilidade. Um "k" grande proporciona mais especificidade ao passo que um "k" pequeno proporciona mais sensibilidade. Quanto menor o "k", aumenta-se a chance de obter um número maior de ocorrências de um dado *k-mer* em uma sequência alvo (menor especificidade). A sensibilidade indica a chance de encontrar sobreposição que de fato exista entre os fragmentos, portanto, quanto menor a especificidade, maior a chance de que a sobreposição de fato exista com um "k" menor. Quando se aumenta o "k", existe uma chance maior de não encontrar sobreposição real caso existam bases com erro de sequenciamento. Pode-se dizer que quanto menor o "k", maior a chance de se encontrar a sobreposição entre *reads* de fato próximos (sensibilidade), mas isso reduz a especificidade, pois a chance de se encontrar sobreposição entre *reads* não relacionados de fato aumenta.

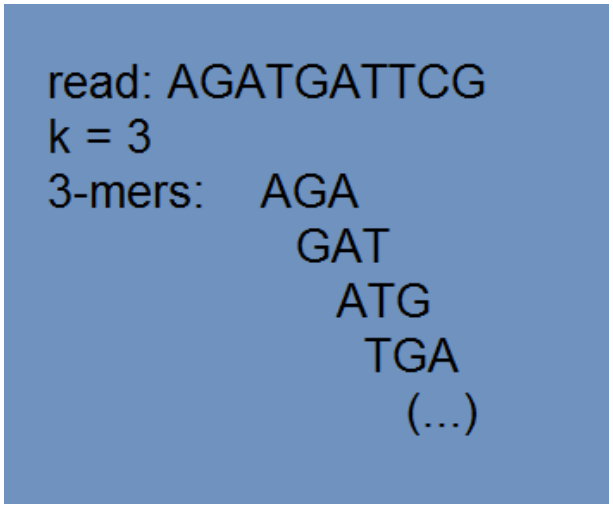

Figura 6: Quebra do read em sequências menores chamadas k-mers

Grafos *de Bruijn* são grafos direcionados que capturam sobreposições de tamanho k – 1 entre os *k-mers* e não entre os *reads* originais. Os nós são sequências de k – 1 caracteres. As arestas são inseridas entre os pares de vértices (u, v) onde o sufixo k – 2 de "u" é igual ao prefixo de tamanho k – 2 de "v". Os *reads* são implicitamente representados como caminhos do grafo *de Bruijn*. A sequência pode ser recomposta através de um caminho euleriano (percorre cada aresta apenas uma vez) [14]. Se o grafo *de Bruijn* não for euleriano, pode-se simplificá-lo ao máximo para encontrar subgrafos eulerianos.

Ao reduzir o conjunto de dados até *k-mer* sobreposições, o grafo *de Bruijn* reduz a alta redundância nos conjuntos de dados (*short reads*). O valor do parâmetro "k" tem uma influência significativa na qualidade do *assembly* fazendo com que seja necessário estimar, antes da montagem, um bom valor para "k".

Erros de sequenciamento geram um grafo maior e, portanto, um maior consumo de memória. Além disso, esses erros geram topologias comuns em grafos *de Bruijn* tais como: pontas (*tips*), bolhas (*bubbles*) e repetições (*repeats*).

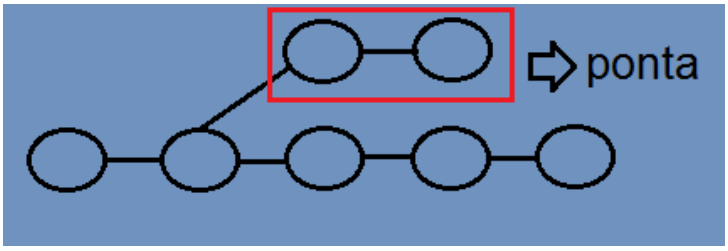

Figura 7: Ponta (tip)

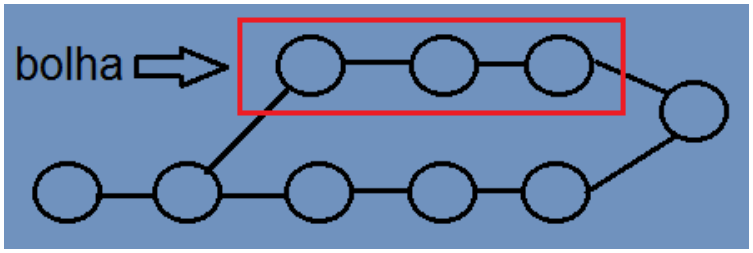

Figura 8: Bolha (bubble)

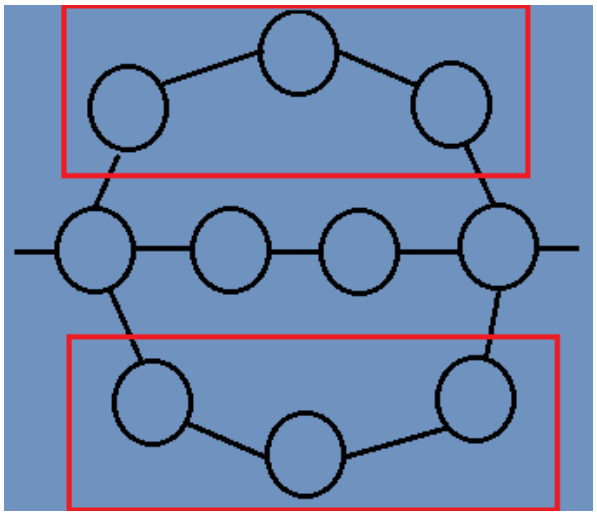
Figura 9: Repetição (repeat)

Repetições podem ser diminuídas com a utilização de grafos *de Bruijn*, pois esses grafos não levam a muitas sobreposições falsas, o que não significa que tais sobreposições possam ser facilmente resolvidas. Pode-se acumular sobreposições falso-positivos fazendo com que a resolução de repetições seja mais complexa, porém, essas sobreposições podem ser facilmente detectadas por alinhamento *pairwise* (alinhamento de duas sequências). O problema das repetições é um desafio para qualquer *software* independente do paradigma usado.

Na abordagem OLC, erros geralmente não afetam a topologia do grafo por causa do uso de alinhamentos globais entre *long reads*, porém, em grafos *de Bruijn*, a correção de erro é fundamental para utilizá-los em análise de dados de sequência. Cada SNP ou *indel* causa rompimento nos contigs do grafo, fragmentando o *assembly*.

Grafo *de Bruijn* é uma das estratégias mais utilizadas pelos *assemblers*. Velvet, SOAPdenovo e ABySS são exemplos de *assemblers* que utilizam grafos *de Bruijn* [15]. Velvet foi um dos primeiros *assemblers* para *short reads*, é atualmente muito utilizado e implementa várias correções de erro após a construção do grafo [16]. SOAPdenovo, ao contrário do Velvet, faz a correção de erro antes da construção do grafo [17]. Já o ABySS possui a vantagem de poder rodar em ambiente paralelo e, portanto, possui um potencial de montar genomas muito grandes [18].

Aplicações que utilizam grafos de *Bruijn* obtiveram resultados satisfatórios na detecção de variações genéticas (SNPs) como é o caso do *software Bubbleparse* que implementa um novo algoritmo para identificação de bolhas sem o uso de uma referência [19]. Outro exemplo é a implementação do *software Cortex* que utiliza grafos de *Bruijn* coloridos para detectar variações genéticas complexas em um indivíduo ou população [20].

2 Justificativa

Com o advento de novas tecnologias de sequenciamento, tornou-se fundamental o desenvolvimento de técnicas computacionais que possam tirar proveito da grande quantidade de dados (*big data*) gerados pelos sequenciadores. A exigência de métodos computacionais eficientes para análise desses dados promove um melhor entendimento biológico acerca dos processos evolutivos de uma determinada espécie.

A identificação de variações genéticas é muito importante para análise genômica. Neste trabalho, o desenvolvimento de um método baseado em grafos *de Bruijn* tem como objetivo identificar, com eficiência e precisão, variações genéticas que possam contribuir para uma melhor compreensão do genoma da cana-de-açúcar principalmente pelo fato dela ainda não ter seu genoma totalmente sequenciado.

Apesar dos grafos de *Bruijn* focarem na sobreposição de *reads*, não se faz a comparação par a par como na abordagem OLC tornando o uso de grafos de *Bruijn* competitivo em relação ao custo computacional. Para *short reads*, percebe-se um uso mais frequente de grafos de *Bruijn* principalmente nas ferramentas mais novas. Isso se deve principalmente ao fato de evitar a comparação par a par de sobreposição, já citado anteriormente, fazendo com que o uso de grafos de *Bruijn* represente um forte argumento para a sua utilização neste trabalho.

Além da utilização de grafos de *Bruijn*, esse trabalho se faz relevante pelo fato da importância econômica da cana-de-açúcar e pela alta demanda por métodos computacionais capazes de processar e analisar dados obtidos por GBS. O processamento e a análise desses dados buscam extrair informações relevantes em relação ao genoma da cana-de-açúcar que é um dos produtos agrícolas mais importantes do Brasil.

3 Objetivos

Este projeto tem como objetivo o desenvolvimento de pesquisa na área interdisciplinar de bioinformática com foco no tópico de identificação e análise de variação genética a partir de dados obtidos por GBS.

Objetivos específicos

- Investigar representações de grafos *de Bruijn* e algoritmos existentes para esses grafos e propor adaptações para dados de GBS;

- Propor método de busca para descobrir variações genéticas em grafos *de Bruijn* gerados com dados de GBS, implementar e testar esses algoritmos em dados simulados e reais de GBS de uma população de cana-de-açúcar.

4 Metodologia

1) Investigação de representações e algoritmos para grafos *de Bruijn* existentes.

Métodos existentes que se baseiam em grafos *de Bruijn* para a montagem de *reads* de sequenciamento de segunda geração serão estudados. Dentre esses métodos, destacam-se métodos voltados ao sequenciamento de genomas e outros específicos para dados de RNA-Seq e identificação de variações genéticas.

2) Desenvolvimento de um novo algoritmo de busca para descoberta de variações genéticas a partir de dados de GBS.

Métodos existentes que utilizam grafos *de Bruijn* serão considerados para a proposta de um novo algoritmo de busca que terá como objetivo a descoberta de variações genéticas a partir de dados de GBS. O objetivo do algoritmo será de agrupar sequências dos diferentes indivíduos que são provenientes do mesmo loco. Isso poderá ser feito através de um algoritmo de busca que, a partir de um dado *read*, procura outros *reads* que estão até a uma dada distância da sequência inicial em termos de diferenças entre as sequências. Após a identificação destas sequências, o trabalho de identificar as variações existentes em cada grupo se torna uma tarefa muito mais simples que pode ser resolvida alinhando-se as sequências de cada grupo. Para verificar a eficiência e eficácia do método a ser proposto, o algoritmo será implementado e deverá ser testado com um conjunto de dados simulados a partir do qual parâmetros iniciais poderão ser identificados antes do método ser aplicado a dados reais.

5 Cronograma

A Tabela 1 descreve a cronologia das atividades que serão realizadas durante este trabalho.

	2015				2016												2017	
	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F
1	X	X	X	X														
2	X																	
3		X	X	X	X	X	X											
4					X													
5			X	X	X	X	X	X	X	X	X	X						
6					X	X	X	X	X	X	X	X	X					

7						X	X	X	X	X	X							
8											X							
9										X	X	X						
10											X	X	X					
11											X	X	X	X	X	X		
12															X	X	X	
13																	X	
14																		X

Tabela 1: Cronograma de atividades

1. Disciplinas obrigatórias do mestrado.
2. Escrita do projeto de mestrado.
3. Estudos de representações e algoritmos existentes para grafos *de Bruijn*.
4. Estágio em laboratório de grupo de colaboradores (Unicamp).
5. Construção de um algoritmo de busca para detectar variações genéticas.
6. Implementação.
7. Criação de um conjunto de dados simulados e testes usando esse conjunto.
8. Exame de qualificação.
9. Comparação com ferramentas existentes.
10. Aplicação utilizando dados reais com GBS de cana-de-açúcar.
11. Escrita da dissertação.
12. Revisão final da dissertação.
13. Defesa da dissertação.
14. Correção da dissertação após a defesa.

6 Referências

1. Carla, A., Góes, D. S., Vinicius, B., & Oliveira, X. De. (2010). The Human Genome Project: a portrait of scientific knowledge construction by the Ciência Hoje magazine, 561–577.

2. Improvements on the previous technology. https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/improveme

3. Zhang, Jun et al. "The Impact of next-Generation Sequencing on Genomics." Journal of genetics and genomics = Yi chuan xue bao 38.3 (2011): 95–109. PMC. Web. 28 Sept. 2015.

4. Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. a., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.

5. Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *International Journal of Plant Genomics*, *2012*.

6. Kohlhepp, G. (2010). Análise da situação da produção de etanol e biodiesel no Brasil. *Estudos Avançados*, *24*(68), 223–253.

7. Dal-Bianco, M., Carneiro, M. S., Hotta, C. T., Chapola, R. G., Hoffmann, H. P., Garcia, A. A. F., & Souza, G. M. (2012). Sugarcane improvement: How far can we go? *Current Opinion in Biotechnology*, *23*(2), 265–270.

8. Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., … Frazer, K. a. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, *10*(3), R32.

9. CB Cardoso-Silva (2015). Análise do transcriptoma e de sequências genômicas de variedades comerciais de cana-de-açúcar.

10. Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS ONE*, *9*(2), e90346.

11. Schmidt, S. (2009). Measuring absorptive capacity. *The Genome Analysis Toolkit - A MapReduce framework for analyzing next-generation DNA sequencing data*, *20*, 254–260.

12. De Novo Assembly Using Illumina Reads. https://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf

13. Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly Algorithms for Next-Ganeration Sequencing Data. *Genomics*, *95*(6), 315–327.

14. Compeau, P. E. C., Pevzner, P. a, & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991.

15. Cláudio, A. (2012). Um protótipo de serviço de montagem de genomas a partir de dados de sequenciamento de próxima geração (NGS).

16. Zerbino, D. R. (2009). Genome assembly and comparison using de Bruijn graphs.

17. Luo et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012 1:18.

18. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., & Jones, S. J. M. (2009). ABySS : A parallel assembler for short read sequence data ABySS : A parallel assembler for short read sequence data, 1117–1123.

19. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, *44*(2), 226–232.

20. Leggett, R. M., Ramirez-Gonzalez, R. H., Verweij, W., Kawashima, C. G., Iqbal, Z., Jones, J. D. G., … Maclean, D. (2013). Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs. *PloS One*, *8*(3), e60058.