1
Introdução

Diversas espécies e organismos já foram sequenciados e montados usando tecnologias de sequenciamento tradicionais. Destes a mais conhecida é o método Sanger[1]. Nos últimos anos, uma nova geração de sequenciadores (NGS- Next-Generation Sequencing) trouxe grandes avanços em máquinas de sequenciamento, como Roche 454 System[1], Illumina[1] e AB SOLiD System[1]. Com a chegada desta nova geração, houve redução de custo e aumento da eficiência, através do emprego de variadas técnicas de sequenciamento altamente paralelas[2]. Esse avanço vem tornando possível o mapeamento de genomas cada vez maiores a um custo cada vez mais baixo.

A nova geração de sequenciadores tem como característica a geração de fragmentos com tamanho pequeno (*short reads*), quando comparados aos métodos mais tradicionais. Além de gerar uma quantidade de erros de sequenciamento maior também se comparado aos métodos tradicionais.

Estas duas características somadas aumentam significativamente o volume de dados a ser processado, fazendo com que o processo de montagem de fragmentos seja um grande desafio computacional[3]. Além deste desafio, os programas de montagem enfrentam outra dificuldade, o alto consumo de memória principal.

Os programas de montagem de fragmentos com eficácia conhecida, tem a sua execução recomendada em máquinas com alta disponibilidade de memória principal. Quando executados em máquinas convencionais, a sua eficiência é reduzida ou até mesmo inviável.

Como motivação adicional para este trabalho, houve uma demanda de pesquisadores do Departamento de Bioquímica da UFRJ quanto à dificuldade de processar a montagem de fragmentos em seus próprios laboratórios. Onde são realizados importantes estudos sobre a cana-de-açúcar, o qual não possui genoma de referência e sim uma estrutura complexa, com diversos genes homólogos. Perante a isso, é necessário sequenciar uma grande quantidade de material genético,

tornando o problema da montagem dos fragmentos ainda maior em relação a outros genomas mais estudados como o do ser humano, vírus e bactérias.

Um dos programas de montagem de fragmentos mais utilizados, inclusive pelo grupo da UFRJ, é o Velvet [4] com alta qualidade de resultados de montagem. O Velvet tem um módulo inicial, chamado VelvetH, que permite retirar boa quantidade de erros de pré-processamento, antes da montagem propriamente dita. Entretanto, este módulo inicial também enfrenta problemas de alto consumo de memória principal, dificultando o uso do programa como um todo por parte dos especialistas em bioinformática.

Estudos vêm sendo realizados visando reduzir o consumo de memória principal em alguns destes programas, como é o caso do Velvet[3, 5-8].

Cabe observar que não se considera aqui como adequadas as soluções baseadas em tecnologia como, por exemplo, aqueles programas que assumem memória RAM virtualmente infinita, exigindo computadores e processadores com baixa relação custo-benefício. Buscamos entender o problema no contexto de computadores com componentes de prateleira (*off-the-shelf*), presentes na grande maioria dos laboratórios de bioinformática no Brasil e no mundo.

Vale ressaltar que o estudo descrito neste trabalho, é parte da pesquisa realizada no laboratório BioBD (PUC-Rio) sobre o gerenciamento de memória principal para montadores de fragmentos curtos.

1.1. Objetivos da Dissertação

Esta dissertação de mestrado tem como objetivo:

1. Entender o funcionamento do programa Velvet e as causas do alto consumo de memória principal do módulo VelvetH.

2. Descrever em detalhes o funcionamento do VelvetH e propor melhorias na sua implementação.

3. Propor e implementar o VelvetH-DB. Uma solução robusta, baseada em banco de dados, para o módulo VelvetH.

1.2. Estrutura da Dissertação

O capítulo 2 apresenta as definições e conceitos necessários para o entendimento desta dissertação e o do contexto em que está inserida.

O capítulo 3 apresenta o processo de montagem de fragmentos e o problema computacional no processamento dos fragmentos.

O capítulo 4 descreve o programa Velvet e sua implementação para montagem de fragmentos, foco de estudo desta dissertação de mestrado. Detalha o funcionamento do módulo VelvetH, além de uma análise do consumo de memória RAM. Neste capítulo também são apresentadas duas alterações realizadas no código com o intuito de demonstrar onde está o gargalo no consumo de memória principal.

O capítulo 5 é apresentada a implementação do VelvetH-BD, uma nova implementação do módulo VelvetH baseada em banco de dados.

Por fim, o capítulo 6 apresenta as conclusões, objetivos alcançados e as possíveis extensões e continuação de pesquisa desta dissertação.