



Marcos Vinicius Marques da Silva

**VelvetH-DB: Uma abordagem robusta de
banco de dados no processo de
montagem de fragmentos de seqüências
biológicas**

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós-graduação em Informática da PUC-Rio como requisito parcial para obtenção do título de Mestre em Informática.

Orientador: Prof. Sérgio Lifschitz

Rio de Janeiro
Março de 2016



Marcos Vinicius Marques da Silva

**VelvetH-DB: Uma abordagem robusta de
banco de dados no processo de montagem
de fragmentos de sequências biológicas**

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Sérgio Lifschitz

Orientador

Departamento de Informática – PUC-Rio

Prof. Edward Hermann Haeusler

Departamento de Informática – PUC-Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão

Departamento de Informática – PUC-Rio

Prof. Luiz Fernando Bessa Seibel

Departamento de Informática – PUC-Rio

Prof. Márcio da Silveira Carvalho

Coordenador Setorial do Centro

Técnico Científico – PUC-Rio

Rio de Janeiro, 30 de março de 2016

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Marcos Vinicius Marques da Silva

Formou-se em Análise de Sistemas e Gestão da Informação na Universidade Candido Mendes em 2006. Kursou Análise e Desenvolvimento de Sistemas na Universidade Candido Mendes em 2010. Atuou em empresas como Analista de Sistemas. Atua como Coordenador de Projetos no laboratório Tecgraf(PUC-Rio).

Ficha Catalográfica

Silva, Marcos Vinicius Marques da

VelvetH-DB : uma abordagem robusta de banco de dados no processo de montagem de fragmentos de sequências biológicas / Marcos Vinicius Marques da Silva ; orientador: Sérgio Lifschitz. – 2016.

66 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2016.

Inclui bibliografia

Incluí referências bibliográficas.

1. Informática – Teses. 2. Banco de dados. 3. Bioinformática. 4. Fragmentos. 5. Montagem. 6. Genoma. I. Lifschitz, Sérgio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

Agradecimentos

A Deus por sempre guiar os meus caminhos e desta forma permitiu que tudo isso acontecesse, não somente nestes anos que passei no desenvolvimento deste trabalho, mas ao longo de toda a minha vida. Agradeço a Ele que em todos os momentos é a minha maior força e fonte de conhecimento.

Aos meus pais, que sempre me mostraram a importância dos estudos e que fizeram o possível e muitas vezes o impossível para que eu chegasse até aqui.

Aos meus amigos, que durante toda a jornada estiveram ao meu lado.

Ao professor Sérgio Lifschitz, pela orientação, apoio e confiança.

A Tecgraf, que acreditou no meu potencial e me apoiou nos estudos.

Aos meus amigos cubanos do LaBBio, em especial a Ema, que em muito ajudaram na construção deste trabalho.

E finalmente um agradecimento mais que especial a minha esposa Eleinne, sem a qual não conseguiria ter chegado até aqui. Uma companheira formidável de vida e que esteve ao meu lado em todos os momentos difíceis sempre me motivando e dando muita alegria.

Resumo

Silva, Marcos Vinicius Marques; Lifschitz, Sérgio. **VelvetH-DB: Uma abordagem robusta de banco de dados no processo de montagem de fragmentos de sequências biológicas**. Rio de Janeiro, 2016. 66p. Dissertação de Mestrado. — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Avanços tecnológicos recentes, tanto nos métodos de sequenciamento quanto nos algoritmos de montagem de fragmentos, têm facilitado a reconstrução de todo o DNA de espécies sem a necessidade de um genoma de referência. A montagem da cadeia completa envolve a leitura um grande volume de fragmentos do genoma (*short reads*), um desafio significativo em termos computacionais. Todos os principais algoritmos de montagem de fragmentos existentes têm como gargalo principal o alto consumo de memória principal. Consonante a isso, essa dissertação de mestrado visa estudar a implementação de um destes algoritmos, Velvet, que é amplamente usado e recomendado. A mesma possuiu um módulo, VelvetH que realiza um pré-processamento dos dados com o intuito de reduzir o consumo de memória principal. Após um estudo minucioso do código e alternativas de melhorias, foram feitas alterações pontuais e proposta uma solução com persistência de dados em memória secundária visando obter eficácia e robustez.

Palavras-chave

Banco de dados; bioinformática; fragmentos; montagem; genoma.

Abstract

Silva, Marcos Vinicius Marques; Lifschitz, Sérgio (Advisor). **VelvetH-DB: a robust database approach for the assembly process of biological sequences**. Rio de Janeiro, 2016. 66p. MSc.Dissertation. — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Recent technological advances, both in assembly algorithms and in sequencing methods, have enabled the reconstruction of whole DNA even without a reference genome available. The assembly of the complete chain involves reading a large volume of genome fragments, called short-reads, which makes the problem a significant computational challenge. A major bottleneck for all existing fragment-assembly algorithms is the high consumption of RAM. This dissertation intends to study the implementation of one of these algorithms, called Velvet, which is widely used and recommended. The same possessed a module, VelvetH that performs a pre-processing data with the aim of reducing the consumption of main memory. After a thorough study of code improvements and alternatives, specific changes have been made and proposed a solution with data persistence in secondary memory in order to obtain effectiveness and robustness.

Keywords

Database; bioinformatics; assembly; genome.

Sumário

1	Introdução	12
1.1.	Objetivos da Dissertação	13
1.2.	Estrutura da Dissertação	14
2	Contexto Biológico	15
2.1.	Genômica	15
2.2.	Sequenciamento	16
2.3.	Cobertura	18
2.4.	Métodos de sequenciamento	18
2.4.1.	Método Tradicional (Sanger)	19
2.4.2.	Next-Generation Sequencing (NGS)	19
2.4.2.1.	Illumina (Solexa)	20
2.5.	Conclusão	21
3	Montagem de Fragmentos	22
3.1.	Abordagens <i>De Novo</i> para montagem de fragmentos	22
3.1.1.	De Novo gulosos	22
3.1.2.	De Novo por sobreposição	23
3.1.3.	De Novo baseados no grafo de Bruijn	24
3.2.	Problema da montagem de fragmentos NGS	27
3.2.1.	Erros de sequenciamento	27
3.2.2.	Consumo de memória principal	28
3.3.	Conclusão	29
4	Velvet	30
4.1.	Estrutura	30
4.2.	VelvetH e o Consumo de memória RAM	32
4.3.	Testes de execução	33
4.4.	Funcionamento do VelvetH	34
4.4.1.	Fase de padronização dos dados	34

4.4.2. Fase de geração do <i>Roadmap</i>	36
4.4.2.1. Leitura dos Dados	36
4.4.2.2. Geração das anotações	38
4.5. Modificações realizadas no VelvetH	40
4.5.1. Ajuste 1: Alocação de memória na leitura dos dados	40
4.5.2. Ajuste 2: Integração dos passos Leitura de Dados e Geração do <i>Roadmap</i>	42
4.5.3. Resultado experimentais	42
4.6. Conclusão	44
5 VelvetH-DB	46
5.1. Fluxo para geração do <i>Roadmap</i>	47
5.2. Implementação	48
5.2.1. Alteração do VelvetH	48
5.3. Modelagem em banco de dados relacional	49
5.4. Funções no SGBD	51
5.4.1. Maestra	52
5.4.2. Importa Arquivo	52
5.4.3. Processa Roadmap	52
5.4.4. GeraOutput	53
5.5. Testes de Execução	53
5.6. Conclusão	54
6 Conclusão	55
6.1. Resumo	55
6.2. Contribuições	56
6.3. Trabalhos Futuros	57
7 Referências	58
8 Anexo I – Código das funções PostgreSQL	60
8.1. Maestra	60
8.2. Importa Arquivo	61
8.3. Processa Roadmap	62

Lista de figuras

Figura 1 - Representação do DNA e RNA	16
Figura 2 - Exemplo de workflow do processo completo de sequenciamento de DNA e RNA[9]	17
Figura 3 - Custos por sequenciamento do genoma até meados de 2013[10]	20
Figura 4 - Erro na montagem de uma abordagem de sobreposição	24
Figura 5 - Extração dos k-mers e a construção do grafo de Bruijn	25
Figura 6 - <i>Workflow</i> dos algoritmos baseados nos grafos de Bruijn[18]	26
Figura 7 - Quantidade de k-mer distintos no genoma versus quantidade de k-mer distintos sequenciados. Genoma: <i>S. suis</i> P1/7, k=21 [12].	28
Figura 8 - Diferença de consumo de memória principal com uso do Roadmp [12].	32
Figura 9 - Consumo de memória Principal pelo VelvetH.	33
Figura 10 - Exemplo de um arquivo de sequências no formato fastq	35
Figura 11 - Exemplo de um arquivo <i>sequences</i> criado pelo VelvetH	35
Figura 12 - Esquemático da geração do Roadmap	36
Figura 13 - Estrutura de dados utilizada na fase de geração do <i>Roadmap</i>	37
Figura 14 - Exemplo de um arquivo <i>Roadmap</i>	40
Figura 15 - Erro na contagem de caracteres necessários para alocar o conteúdo da sequência	41
Figura 16 - Alteração realizada no código fonte do VelvetH	42
Figura 17 - Resultado das alterações no conjunto de dados com 50.000 sequências	43
Figura 18 - Resultado das alterações no conjunto de dados com 2.191.196 sequências	43
Figura 19 - Resultado das alterações no conjunto de dados com 12.671.416 sequências	44
Figura 20 - Fluxo de geração do <i>Roadmap</i> através do VelvetH sem modificações.	47
Figura 21 - Fluxo para a geração do <i>Roadmap</i> através do VelvetH-BD.	48
Figura 22 - Arquivo de saída do VelvetH alterado, colocando	

todas as informações da sequência em uma única linha	49
Figura 23 - Esquema relacional do VelvetH-DB	51
Figura 24 - Diagrama de sequência das funções do VelvetH-BD	52
Figura 25 - Resultados da execução do VelvetH e o VelvetH-BD	53