Pontifícia Universidade Católica
do Rio de Janeiro

**Marcos Vinicius Marques da Silva**

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

**Marcos Vinicius Marques da Silva**

Formou-se em Análise de Sistemas e Gestão da Informação na Universidade Candido Mendes em 2006. Cursou 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.

CDD: 004

## 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

## Lista de figuras