Pedro Filipe
Carneiro Venâncio

Aplicação de novos algoritmos bioinformáticos na análise de dados de Next Generation Sequencing (NGS)

Application of new bioinformatic algorithms in Next Generation Sequencing (NGS) data analysis.

DOCUMENTO PROVISÓRIO

Pedro Filipe
Carneiro Venâncio

Aplicação de novos algoritmos bioinformáticos na análise de dados de Next Generation Sequencing (NGS)

Application of new bioinformatic algorithms in Next Generation Sequencing (NGS) data analysis.

DOCUMENTO PROVISÓRIO

"*The greatest challenge to any thinker is stating the problem in a way that will allow a solution*"

— Bertrand Russell

Universidade de Aveiro
2024

Pedro Filipe Carneiro Venâncio

Aplicação de novos algoritmos bioinformáticos na análise de dados de Next Generation Sequencing (NGS)

Application of new bioinformatic algorithms in Next Generation Sequencing (NGS) data analysis.

Relatório de estágio curricular apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Bioinformática Clínica, especialização em Bioinformática do Genoma , realizado sob a orientação científica da Doutora Gabriela Maria Ferreira Ribeiro de Moura, Professora auxiliar do Departamento de Ciências Médicas da Universidade de Aveiro, e supervisão da Doutora Alexandra Filipa Lopes, membro da entidade de acolhimento UNILABS.

Dedico este trabalho à minha esposa e filho pelo incansável apoio.

o júri / the jury

presidente / president

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

vogais / examiners committee

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

Prof. Doutor João Antunes da Silva
professor associado da Faculdade de Engenharia da Universidade do Porto

agradecimentos / acknowledgements

Agradeço toda a ajuda a todos os meus colegas e companheiros.

Palavras Chave NGS, arquitetura, história, construção, materiais de construção, saber tradicional.

Resumo

Um resumo é um pequeno apanhado de um trabalho mais longo (como uma tese, dissertação ou trabalho de pesquisa). O resumo relata de forma concisa os objetivos e resultados da sua pesquisa, para que os leitores saibam exatamente o que se aborda no seu documento.
Embora a estrutura possa variar um pouco dependendo da sua área de estudo, o seu resumo deve descrever o propósito do seu trabalho, os métodos que você usou e as conclusões a que chegou.
Uma maneira comum de estruturar um resumo é usar a estrutura IMRaD. Isso significa:

- Introdução
- Métodos
- Resultados
- Discussão

Veja mais pormenores aqui:
`https://www.scribbr.com/dissertation/abstract/`

Keywords textbook, architecture, history, construction, construction materials, traditional knowledge.

Abstract

An abstract is a short summary of a longer work (such as a thesis, dissertation or research paper).
The abstract concisely reports the aims and outcomes of your research, so that readers know exactly what your paper is about.
Although the structure may vary slightly depending on your discipline, your abstract should describe the purpose of your work, the methods you've used, and the conclusions you've drawn.
One common way to structure your abstract is to use the IMRaD structure. This stands for:

- Introduction
- Methods
- Results
- Discussion

Check for more details here:
`https://www.scribbr.com/dissertation/abstract/`

Contents

List of Figures

List of Tables

Lista de Excertos de Código

Glossário

CHAPTER 1

Introduction

"The only source of knowledge is experience." - Albert Einstein

1.1 Internship Context and Framework

This document represents the final report of the internship carried out as an integral part of the Internship Course (49991) of the second year of studies of the Master's Degree in Clinical Bioinformatics, specializing in Genome Bioinformatics, at the University of Aveiro. The internship lasted for nine months, starting on November 21, 2023, and concluding on July 19, 2024, totaling 1296 hours of work.

During this period, the intern had the opportunity to apply the knowledge acquired throughout the course and to engage in practical projects related to bioinformatics and genomics. Unilabs, a renowned company in the healthcare sector, provided a professional environment where the intern could collaborate with experienced professionals and actively participate in projects relevant to clinical bioinformatics. This report addresses the activities developed during the internship and the contributions to the projects in which the intern was involved.

This introductory section aims to provide an overview of the context in which the internship was carried out, establishing the groundwork for understanding the activities and results presented throughout the report.

1.2 Motivation and Objectives of the Work

Currently, Unilabs employs a genomic intelligence platform that utilizes natural language processing to analyze new genetic publications and incorporate them into an always-updated knowledge base. This platform is particularly useful in identifying pathogenic variants, interpreting them, and generating clinical reports, thus enabling increasingly personalized care.

However, at the time of this internship, this platform had some limitations regarding the presentation of essential coverage metrics for compliance with guidelines and recommended practices for Next Generation Sequencing (NGS). These metrics are important for assessing data quality, indicating how well the target regions were covered by sequencing. Additionally, coverage depth directly influences the ability to detect genetic variants: regions with low coverage may result in undetected or underestimated variants. Furthermore, coverage metrics are also useful for optimizing sequencing protocols by adjusting experimental parameters to ensure adequate coverage of target regions and minimize unnecessary costs.

Therefore, a software was developed, described in this report, aimed at addressing the gap of not presenting Average Read Depth and Coverage Percentage at 1x, 10x, 15x, 20x, 30x, 50x, 100x, and 500x per gene and per panel in gene panel analysis. Additionally, besides presenting metrics per panel, analysis was also implemented for single genes and exomes.

1.3 Document Structure

This document is divided into five chapters, each with several sections.

The first chapter, Introduction, begins with contextualization and framing of the internship, followed by a brief presentation of the motivation and objectives of the work. The hosting entity is also described, along with a brief theoretical foundation that supports the solution presented.

The second chapter, Analysis Tool, addresses several sections. The first is Development, which explains the bioinformatics process involved in creating the software, including setting up the work environment, the programming languages and software used, and the calculations performed. The following section, Operation, details step-by-step how the software operates to obtain the desired metrics. The Validation section compares the metrics obtained by the software with other tools. Then, the Documentation section describes the process of documenting the software to ensure its reproducibility. Finally, the Distribution section explains how the software can be distributed and implemented.

The third chapter, Discussion, is intended for the discussion of ideas associated with the work, analyzing and interpreting the results in the context of existing knowledge in the field. It emphasizes potential practical applications, limitations, and implications of the results.

The last two chapters are, respectively, Final Considerations, where possible improvements and new features for future versions of the software are presented, and Bibliography, which lists all the sources consulted and cited in the work.

1.4 Characterization of the Host Entity and Work Plan

1.4.1 Unilabs

1.4.2 Unilabs Genetics

1.4.3 Schedule

1.5 Theoretical Framework

CHAPTER 2

Analysis tool

"The only source of knowledge is experience." - Albert Einstein

2.1 Internship Context and Framework

This document represents the final report of the internship carried out as an integral part of the Internship Course (49991) of the second year of studies of the Master's Degree in Clinical Bioinformatics, specializing in Genome Bioinformatics, at the University of Aveiro. The internship lasted for nine months, starting on November 21, 2023, and concluding on July 19, 2024, totaling 1296 hours of work.

During this period, the intern had the opportunity to apply the knowledge acquired throughout the course and to engage in practical projects related to bioinformatics and genomics. UNILABS, a renowned company in the healthcare sector, provided a professional environment where the intern could collaborate with experienced professionals and actively participate in projects relevant to clinical bioinformatics. This report addresses the activities developed during the internship and the contributions to the projects in which the intern was involved.

This introductory section aims to provide an overview of the context in which the internship was carried out, establishing the groundwork for understanding the activities and results presented throughout the report.

APPENDIX A

Additional content