FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Information and Data Analysis System for Gene Expression

Diogo André Rocha Teixeira

TECHNICAL REPORT

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rui Camacho (PhD)

Second Supervisor: Nuno Fonseca

January 5, 2014

Information and Data Analysis System for Gene Expression

Diogo André Rocha Teixeira

Mestrado Integrado em Engenharia Informática e Computação

January 5, 2014

Abstract

Resumo

Acknowledgements

Diogo André Rocha Teixeira

Contents

CONTENTS

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

cDNA	Complementary DNA
DNA	Deoxyribonucleic Acid
IBMC	Institute for Molecular and Cell Biology *(Instituto de Biologia Molecular e Celular)*
mRNA	Messenger RNA
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
tRNA	Transfer RNA
WTSS	Whole Transcriptome Shotgun Sequencing

Chapter 1

Introduction

1.1 Context and Motivation

1.2 Objectives

1.3 Document Outline

Besides the introduction chapter, this document is composed three additional chapters. These chapters have the following structure:

Chapter 2 introduces some basic Biology and RNA Sequencing concepts, that are essential to understand the problems with which this document deals. Furthermore, we describe the main techniques used for genome sequencing and assembly, their differences and applications and the tools and data formats typically used on those areas. Lastly, we give some insight about machine learning algorithms and how they will be applied to this work.

Chapter 3 presents the datasets that will be studied and used in this work, their origins and features. We will also refer the validation methods that will be used to access the quality of our results.

Chapter 4 outlines the main steps in the development of this thesis (and the respective software prototype). In the last part of this chapter we will attempt to provide a feasible schedule for this work's execution.

Chapter 2

Background

2.1 Introduction

2.2 Genome Assembly and RNA Sequencing

2.2.1 Assembly Methods

2.2.2 RNA Sequencing Tools

2.2.3 Common Data Formats

2.3 Machine Learning

Background

Chapter 3

Datasets and Validation

3.1 Datasets

3.2 Result Validation

Chapter 4

Work Plan

4.1 Development Phases

4.2 Schedule Planning

References