FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Information and Data Analysis System for Gene Expression

Diogo André Rocha Teixeira

TECHNICAL REPORT

U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

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

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

cDNA	Complementary DNA
DNA	Deoxyribonucleic Acid
FEUP	Faculty of Engineering of the University of Porto *(Faculdade de Engenharia da Universidade do Porto)*
IBMC	Institute for Molecular and Cell Biology *(Instituto de Biologia Molecular e Celular)*
mRNA	Messenger RNA
NGS	Next Generation Sequencing
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
tRNA	Transfer RNA
WTSS	Whole Transcriptome Shotgun Sequencing

Chapter 1

Introduction

This chapter aims at giving a general overview about the themes address by this thesis. We will address the context in which the thesis is inserted, as well as the motivation that led to its proposal. Furthermore there will be brief description of this thesis' main objectives and the methods that will be used to achieve those objectives.

1.1 Context and Motivation

Molecular biology is a branch of biology that studies biological activities of living being, at a molecular level. The early grounds for this field of study were set in the early 1930's, although only emerging in its modern form in the 1960's, with the discovery of the structure of DNA. Among the processes studied by this branch of biology is gene expression. Gene expression is the process by each DNA molecules are transformed into useful genetic products, typically proteins, which are essential for living organisms. This knowledge is not only important in fields like evolutionary biology or molecular biology, but may have crucial applications in fields such as medicine. One example of such an application is the usage of gene expression analysis in the treatment of cancer patients [PASH03].

With the advent of NGS (Next Generation Sequencing) techniques, researchers have at their disposal huge amounts of sequencing data, that is not only cheaper and faster to produce, but also more commonly available. This data can then be used to obtain relevant information about organisms' gene expression. But, as the cost of sequencing genomes was reduced, the cost of processing such information was increased. NGS techniques tend to produce much smaller reads[1] than previously used techniques, which present a much harder problem, from a computational standpoint [Wol13].

[1] A *read* is a single fragment of a genome/transcriptome, obtained through sequencing techniques.

1.2 Objectives

While defining the concrete objectives of this thesis it becomes relevant to separate them in two groups: strictly biology research related objectives and more general, software solution development objectives. Despite this division, both objectives are tightly interconnected, and each complements the other.

From a molecular biology standpoint, the main objective of this thesis will be to try to understand the mechanisms that regulate the speed of transcription for coding regions of the DNA, in other words, to understand the mechanisms that regulate gene expression. This information will be obtained using the RNA Sequencing method, that will be further discussed in Chapter 2. There are several intermediate objectives for this particular problems, as follows:

- Alignment of the given sequencing reads into a known reference genome. This is one of the first steps in the RNA Sequencing process and is effectively one of the most complex problems addressed by this thesis. Some of the tools used in this particular step of the process will be referenced in Section 2.2.2.

- Further analysis of the RNA Sequencing results using machine learning algorithms, applied to data mining. These techniques will be used in an effort to try to understand the already mentioned transcription mechanisms. This topic will be developed in Section 2.3.

The last objective of this thesis is the development of a software platform prototype. This prototype comes as a materialization of the work done along the previous objectives, combining the developed genetic data processing pipeline, with a web information system and with data mining tools. When completed, the prototype should allow for users to store, search and manipulate their genome sequencing data. This data can them be assembled using the tool pipeline developed for the analysis of our own experimental dataset. Lastly, the prototype should integrate data mining tools, that would allow users to reproduce the types of data analysis that were done in this thesis, on their own results.

This document, however, will not dwell in the details of the implementation of such a platform, but rather in the molecular biology section of the overall problem. This is largely due to the fact that the development of the web platform is highly dependent on the tools and methods that will be used for tackling the biology aspects of the problem and, as such, is likely to suffer significant alterations.

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/transcriptome 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

- explain gene expression
    - explain importance and applications of gene expression profilling
    - explain that nowadays sequencing data is easier and cheaper to obtain, but harder to process
    - explain that there are several techniques to obtain gene expression information
    - explain that in the thesis only RNA-Seq will be analysed

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

2.4 Conclusions

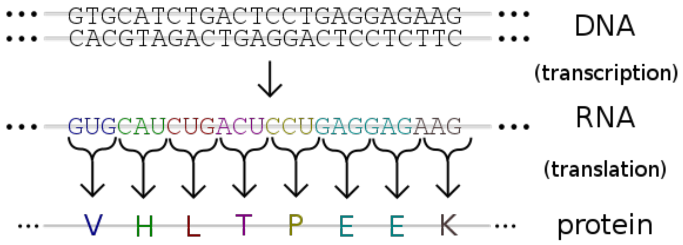

Figure 2.1: Representation of the gene expression process [Bal06]

Chapter 3

Datasets and Validation

3.1 Datasets

3.2 Result Validation

3.3 Conclusions

Chapter 4

Work Plan

4.1 Development Phases

4.2 Schedule Planning

4.3 Conclusions

Work Plan

References

[Bal06] Madeleine Price Ball. Genetic code.svg. Available at http://en.wikipedia.org/wiki/File:Genetic_code.svg, last access on January 2014, May 2006.

[PASH03] Lajos Pusztai, Mark Ayers, James Stec, and Gabriel N Hortobágyi. Clinical Application of cDNA Microarrays in Oncology. *The Oncologist*, 8(3):252–258, January 2003.

[Wol13] Jochen B W Wolf. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular ecology resources*, 13(4):559–72, July 2013.