FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Information and Data Analysis System for Gene Expression

Diogo André Rocha Teixeira

DISSERTATION PLANNING

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rui Camacho (PhD)

Second Supervisor: Nuno Fonseca (PhD)

January 26, 2014

Information and Data Analysis System for Gene Expression

Diogo André Rocha Teixeira

Mestrado Integrado em Engenharia Informática e Computação

January 26, 2014

Abstract

Resumo

Contents

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

New structure idea:
- Context: same as context and motivation right now;
- Motivation and objectives: talk about exon analysis, and other high level objectives for the project; just reference the technical aspects of the problem, a let the bulk prototype be described in project;
- Project: describe the project itself, the prototype and development phases; refer to planning;
- Structure: it's already good;

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].

1.2 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 2.2.1.
- 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

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

be used for tackling the biology aspects of the problem and, as such, is likely to suffer significant alterations.

1.3 Structure of the Report

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 data mining algorithms and how they will be applied to this work.

Chapter 3 outlines the main steps in the development of this thesis (and the respective software prototype) and attempts to provide a feasible schedule for the work's execution. It also presents the datasets that will be studied and used in this work, their origins and features, as well as the validation methods that will be used to ascertain the quality of our results.

Chapter 4 sums up the what has been defined in the report, emphasizing the problem that the thesis addresses and the work that will be executed towards solving that problem. It will also give a brief idea of what are the expected results at the end of the project.

Chapter 2

State-of-the-Art

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

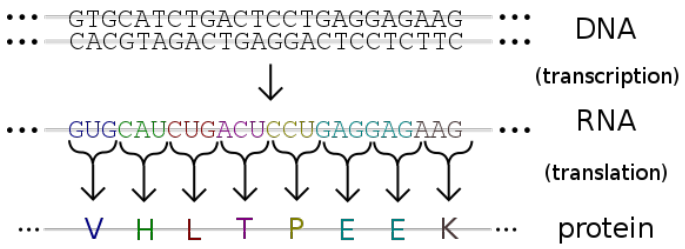

Figure 2.1: Representation of the gene expression process[1]

[1] Image taken from http://en.wikipedia.org/wiki/File:Genetic_code.svg.

2.2 Genome Assembly and RNA Sequencing

2.2.1 RNA Sequencing Tools

2.2.2 Relevant Standard File Formats

2.3 Data Mining

2.3.1 Data Analysis Algorithms

2.3.2 Data Analysis Tools

2.4 Chapter Conclusions

Chapter 3

Work Plan

This chapter describes the general work plan for the thesis, in terms of activities and their respective timings. Furthermore, we will discuss the datasets that will be used in the work, their characteristics and provenience. Lastly, we will address the subject of work evaluation and validation, explaining how it will be conducted, both during and at the end of the project.

3.1 Planning

Aside the preparation phase (already completed), the time available for the thesis will span from February to July, 2014, roughly totalling twenty weeks. It is essential to define a top level schedule beforehand, to ensure that sufficient time will be allotted for every phase of the project and that the timings of those phases are feasible. As such, Figure 3.1 represents the division of the six main phases of the project, during the available period of twenty weeks. Although each phase comprises several smaller tasks, we believe that such a small granularity planning is not needed in this phase and will be defined during the work's execution, as needed. Each main phase of the project is composed as follows:

Information system development comprises the design and development of the data management component of the project and will take roughly six weeks. Despite not being the most critical component, making it the first in the development timeline facilitates later data intensive phases like the transcriptome assembly and, at the same time, allows extensive testing and performance evaluation through usage. A substantial time allotted for this phase will be spent tackling the performance aspects of implementing a system for such large quantities of data, both in terms of database size and response times.

Assembly pipeline development consists of the construction of the tool pipeline responsible for assembling the transcriptomes. At first, several of the already mentioned tools will be studied and tested against small datasets, in an effort to ascertain which are best suited to our

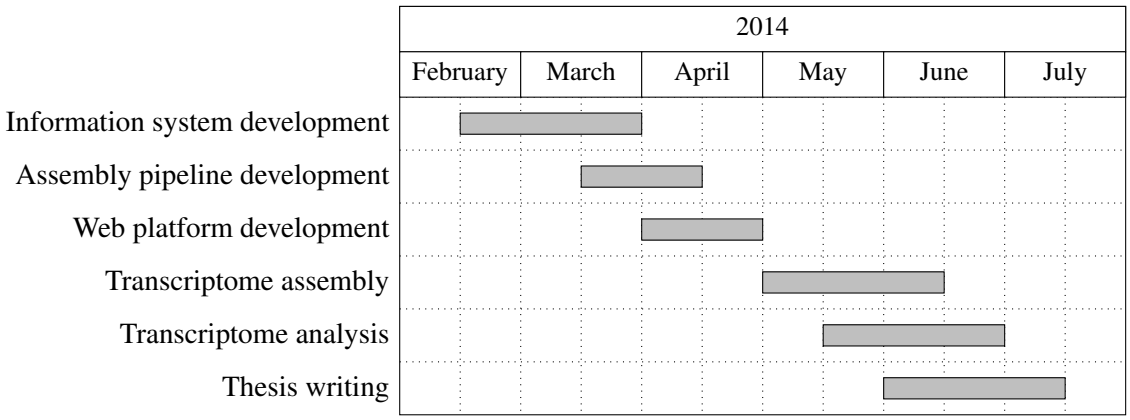

Figure 3.1: Work distribution planning

particular problem. As the tools are selected, the pipeline itself will take shape, integrating the tools in sequence. Any actual development effort in this phase will be in the form of simple data format conversion scripts, since we will use existing assembly tools. Because of this, the estimated duration of this phase is only one month or four weeks, despite its critical importance to the project.

Web platform development will take four weeks and comprises the design and implementation of the system's web front-end. The web platform will integrate the information and assembly systems, providing a user friendly interface for genetic data storage, management and assembly. From a technical standpoint it's a fairly trivial system, which explains why only four weeks were reserved to this phase.

Transcriptome assembly is the first phase after concluding the development of the main components of the system. In this phase the developed system will be used to produce the assembled transcriptome, employing the given production dataset. This phase will take about six weeks, despite no implementation work taking place (saving some small system tweaks). This is because genome and, in this case, transcriptome assembly are resource and time intensive processes that can take several days, making the extra time necessary for both new and repeat experiments.

Transcriptome analysis will consist in the usage of several data mining tools in order to try to explain the already mentioned RNA transcription mechanisms. This phase is expected to last about six weeks. Although not as resource demanding as the transcriptome assembly phase, this will require choosing and testing a new set of tools and possibly integrate them with the developed system.

Thesis writing is the last phase of the project, with an expected six weeks allocated time. These last six weeks refer to a period to collect and report the obtained results and to make the final reviews to the produced content. However, it is expected that the thesis report will

be worked on continuously from the start of the project, in parallel with the other project phases.

3.2 Experimental Data

During this project there will be essentially three types of datasets used: read data, genome data and test data. Each type of dataset has its own nature, origin and purpose.

The read data will be made available during the project through IBMC. This data consists of several short sequencing reads of the *Drosophila melanogaster* genome. It is this dataset that will be ultimately used for assembly and posterior data mining analysis. It should be noted that this will be real data, experimentally obtained in a laboratory for this project.

The genome data will consist of already assembled *Drosophila melanogaster* genome(s), that will be used as a reference in our own assembly process. This data will be obtained through FlyBase (`www.flybase.org`). FlyBase is an online and publicly accessible database of *Drosophila* genes and genomes. This database allows its data to be downloaded in several formats, that can be either directly used in our assembly pipeline, or be automatically converted by one of the created conversion tools.

Lastly, we will use some small scale datasets for the test and calibration of the assembly pipeline. Such datasets are usually shipped with the assembly tools themselves. If needed, a combination of the two previous datasets can be used to produce small scale test data for this purpose.

3.3 Thesis Work Evaluation

In the second part of the project, that is the transcriptome assembly and analysis phases, results evaluation and validation is essential. Even more so when such results are typically evaluated from a molecular biology standpoint and therefore are out of the scope of knowledge of the thesis itself. In such cases we will have two evaluation methods at our disposal.

The first method is based on relevant metrics for the problems at hand, from both the transcriptome assembly and data mining parts. Such metrics are usually produced by the tools themselves. As for these metrics there is usually a well defined range of expected results, which makes them a very important method of early result evaluation, in the sense that they can be interpreted without a profound knowledge about molecular biology.

The second method available is the evaluation by IMBC's technicians, that will assist us whenever expert biology knowledge is required. This will ultimately be the method that will provide a real measure of the success of the project.

Furthermore, IMBC's technicians will be essential during the entirety of the project. They will help steer the project into its intended direction, giving some insight about their expectations towards the system. Project phases like the implementation of the information system or the transcriptome analysis will be driven by their feedback, giving us a sense about what should be

done. Lastly, they will also be present throughout the project to help with any biology related questions that arise.

Chapter 4

Conclusions

Conclusions

References

[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.