A COMPUTATIONAL PLATFORM FOR GENE EXPRESSION ANALYSIS

Diogo André Rocha Teixeira

Dissertation conducted under the supervision of *Prof. Rui Camacho* and co-supervision of *Nuno Fonseca*
at *Faculdade de Engenharia da Universidade do Porto*

1. Context

Molecular biology is a branch of biology that studies biological activities of living beings, at a molecular level. The grounds for this field of study were set in the early 1930s, although it only emerged in its modern form in the 1960s, 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 which 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 or molecular biology, but has crucial applications in fields such as medicine. One example of such an application is the usage of gene expression analysis in the diagnosis and treatment of cancer patients [1].

With the advent of *Next Generation Sequencing* (NGS) 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, presenting a more challenging problem, from a computational standpoint [2].

2. Domain Problem

Despite its great advancements in the past decades, molecular biology is still a relatively new subject and, as such, there are still some unknowns and partial knowledge in this area. In respect to gene expression, some mechanisms of this intricate process are yet to be fully understood. One such mechanism is the one that regulates the transcription speed of RNA. One of the objectives of this thesis is to understand how the final sequences of a gene's exons are responsible for the speed at which the exons themselves are transcribed. The other objective is to understand how RNA-binding protein (RBP) manipulation can be used to better understand an organism's gene expression. These are, however, complex tasks that can be further decomposed in the three main problems that will be addressed in the thesis, namely:

Sequencing read alignment against a reference genome and differential expression analysis between samples of different individuals (of the same species). This is effectively one of the most complex problems addressed in the thesis. We will use data obtained through a sequencing method called RNA Sequencing[2].

Gene enrichment and RBP analysis. This part of the work aims to collect as much relevant information as possible about the particular genes being studied at the time, to help biologists to better understand their function. RBP knowledge is particularly important for gene manipulation and a very useful tool for better understanding gene expression.

Further analysis of the produced data, using machine learning techniques for data mining, specifically for clustering analysis. These techniques will be employed in an effort to give biologists more relevant information about gene expression, uncovering possible relationships in the retrieved information.

Solving these problems requires the use of computational tools. As such, the development of a computer system (or multiple systems) to tackle these problems emerges as a secondary objective of the thesis.

3. Motivation and Objectives

Gene expression analysis is essential for modern day molecular biology. Among many of the possible applications of this information, we can highlight: better classification and diagnosis of diseases, assessing how cells react to a specific treatment, and others.

While nowadays powerful computational tools exist to target almost any biology problem, many of those tools require a very specific set of technical skills and have a steep learning curve. Possibly the most important motivation behind this thesis, and ultimately its main objective, is to provide researchers with powerful yet simple and user friendly tools. This means developing a system simple enough that any user can learn to operate it in a short period of time with minimal effort, but sufficiently advanced to suit the user's research needs.

Another typical problem that biology researchers face nowadays is information dispersion and the repetitive and lengthy task of compiling that information. Researchers frequently have to manually join infor-

[1] A *read* is a single fragment of a genome/transcriptome, obtained through sequencing techniques.
[2] RNA Sequenceing (RNA-Seq) is also referred to as *Whole Transcriptome Shotgun Sequencing*, or WTSS.

mation originating from a multitude of different platforms, which use inconsistent formats and notations. Our second objective is therefore to provide a system that is able to take this burden off the user, making the process faster and simpler.

4. Project

The project itself revolves around the development of a prototype computer system, capable of solving the aforementioned problems. Due to the complexity of the complete system, its development followed a modular organization. The envisioned system architecture is divided into three major components.

The differential expression analysis pipeline is responsible for aligning reads against a reference genome and compare contrasts between different samples. The pipeline is based on the preexisting iRAP pipeline. The pipeline's capabilities are further enhanced with both job configuration automation and differential expression results consolidation (combining results from multiple differential expression tools).

The RNA-binding protein analysis workflow aggregates information about RBPs from multiple biologic web databases (Ensembl, NCBI, UniProt, etc.) and organizes it in ways that are useful to biology researchers. Moreover, this information is clustered using data mining techniques, in order to reveal groups of genes and RBPs that may hold biologic relevance.

The web platform is responsible for storing and managing genetic data, coordinating interaction between the other components of the system and providing a web interface for user interaction. This component is based mainly on typical web technologies, that is, a document based database for data storage (MongoDB), a web framework for business logic implementation (Padrino) and web markup and styling languages for interface implementation (HTML, CSS).

5. Case Study

Two case studies were conducted in order to assess the quality of the developed system.

The first case study focused on the enhancements that were developed for the gene expression analysis pipeline. It was based on a past experience (ArrayExpress experiment *E-GEOD-48829*) that studied the *E. coli* bacteria. The objective of this case study was to ascertain if the developed pipeline enhancement could help improve the researchers' confidence in differential expression results by combining the best results from multiple tools. The results were compared with both the raw results and the results filtered by *p-value*[3]. We conclude that the result combining tool significantly reduced the number of genes in the results, increasing confidence in those results and creating an easier to analyse data set for future experiments.

The second case study was conducted in collaboration with IBMC (*Instituto de Biologia Molecular e Celular*) experts. The studied data set was composed by twenty three genes from the *RhoGTPase* family, from *Rattus norvegicus* (commonly known as *norway rat*). This case study had three main goals: assess the general usefulness of the developed RBP analysis tool; comparing the developed solution with the existing ones; and assess the impact of different hardware in the overall performance of the tool. The obtained results were verified both in terms of their completeness and their biological correction and relevance. The biological validation was also performed by IBMC experts. We concluded that the developed tool could mimic the same results an expert would obtain, in a fraction of the time and providing much more useful information.

6. Conclusions

Our objectives, in terms of studying the problem at hand and developing a solution to the problem, were completely fulfilled. The proposed solution corresponds to all of our expectations. However the implementation of the RNA-Seq data analysis pipeline system was not completed, due to time constraints. As such, our objective of prototyping and testing the complete system could not be completely achieved.

7. Future Work

The obvious continuation of the proposed work would be to finish the implementation and integration of the RNA-Seq data analysis pipeline. This would allow our solution to work as designed, integrating the complete analysis pipeline, from sequencing data to gene clustering and result visualization. Furthermore, it would be interesting to study the developed tools in terms of performance, under large volumes of information and requests. Whilst the tools were developed taking in consideration their performance, making them available in a large scale would take another kind of infrastructure.

References

[1] 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.

[2] 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.

[3] Steven Goodman. A dirty dozen: Twelve p-value misconceptions. *Semin Hematol*, 45:135–140.

[3] A *p-value* is used to assert the statistical significance of results. It represents the probability of obtaining the same results as before in a new sample, given that the null hypothesis is true [3].