UMA PLATAFORMA COMPUTACIONAL PARA ANÁLISE DE EXPRESSÃO GÉNICA

Diogo André Rocha Teixeira

Dissertação realizada sob a orientação do *Prof. Rui Camacho* e co-orientação de *Nuno Fonseca*
na *Faculdade de Engenharia da Universidade do Porto*

1. Contexto

A biologia molecular é um ramo da biologia que estuda as atividades biológicas dos seres vivos, ao nível molecular. As bases para esta área de estudo foram criadas no início da década de 1930, embora apenas tenha emergido na sua forma mais moderna na década de 1960, com a descoberta da estrutura do DNA. Entre os processos estudados por este ramo da biologia está a expressão génica. A expressão génica é o processo através do qual moléculas de DNA são transformadas em produtos genéticos úteis, tipicamente proteínas, que são essenciais para os organismos vivos. Este conhecimento não é apenas importante em áreas como biologia molecular ou evolutiva, mas tem aplicações cruciais em áreas como medicina. Um exemplo de uma destas aplicações é a utilização de análise de expressão génica no diagnóstico e tratamento de pacientes com cancro [1].

Com o advento das técnicas de *Next Generation Sequencing* (NGS) os investigadores têm à sua disposição grandes quantidades de dados de sequenciação, cuja produção é não só mais barata e rápida, mas também vulgarmente mais disponível. Estes dados podem ser usados para obter informação relevante sobre a expressão génica de organismos. Mas, à medida que o custo da sequenciação de genomas é reduzido, o custo do processamento dessa informação aumenta. Técnicas NGS costumam produzir *reads*[1] mais curtas quando comparadas com aquelas produzidas por técnicas anteriores, apresentando um problema mais desafiante, de ponto de vista computacional [2].

2. Problema de Domínio

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 into 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. Motivação e Objetivos

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.

[1] Uma *read* é um fragmento de um genoma/transcriptoma, obtido através de técnicas de sequenciação.
[2] RNA Sequenceing (RNA-Seq) is also referred to as *Whole Transcriptome Shotgun Sequencing*, or WTSS.

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 information 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. Projeto

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. Caso de Estudo

A 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 *RhoGTPase* family, from *Rattus norvegicus* (commonly known *norway rat*).

The obtained results were validated both in terms of their ... 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 (see Tab. 1) and providing much more useful information.

Tab. 1 – Result time comparison between manual analysis (done by an expert) and both test machines.

	machine1	**machine2**	**Expert**
100 IDs	$9m\ 56s$	$11m\ 1s$	$\approx 50h$
500 IDs	$41m\ 47s$	$55m\ 51s$	$\approx 250h$
900 IDs	$1h\ 33m\ 32s$	$2h\ 7m\ 4s$	$\approx 450h$

6. Conclusões

Our objectives, in terms of studying the problem at hand and developing a solution to it, were completely fulfilled. The proposed solution corresponds to all of our expectations. However, as previously discussed, 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. Trabalho Futuro

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.

Referências

[1] Lajos Pusztai, Mark Ayers, James Stec, e Gabriel N Hortobágyi. Clinical Application of cDNA Microarrays in Oncology. *The Oncologist*, 8(3):252–258, Janeiro 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, Julho 2013.