

# A Computational Platform for Gene Expression Analysis

Diogo Teixeira

Supervisors: Rui Camacho<sup>1</sup>, Nuno Fonseca<sup>2</sup>

<sup>1</sup>LIAAD INESC & DEI FEUP, Universidade do Porto

<sup>2</sup>EMBL-EBI, Cambridge, UK

July 2014

# Outline

## 1 Introduction

- Domain Problem
- Motivation and Objectives

## 2 Developed Solution

- Overview
- RNA-Seq Analysis Pipeline
- RBP Analysis Pipeline (PBS Finder)
- Integration

## 3 Case Studies

- RNA-Seq Analysis Pipeline
- RBP Analysis Pipeline (PBS Finder)

## 4 Conclusions

- Objectives Fulfilment
- Future Work

# Domain Problem

## Introduction

- Molecular biology is a young field of study, with a lot of unknowns and partial knowledge.
- Studying gene expression is crucial to understand the mechanisms that control living organisms.
- Two problems sparked the interest of biologists:
  - Under different contexts which genes became active and which became inactive?
  - What is the mechanism that regulates gene transcription speed?

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

### **Information is scattered**

Information is easy to acquire, but is often scattered through multiple platforms, services and institutions.

# Motivation and Objectives

## Introduction

### **Tools are complex**

Tools for biological data analysis often require a very technical set of skills.

### **Create simpler tools**

Any user should be able to use the tools, with little to no training.

### **Tasks are repetitive**

Analysing high quantities of data can be repetitive, especially if executed manually.

### **Automate tasks**

Automated systems should perform repetitive tasks, so that users can focus on their work.

### **Information is scattered**

Information is easy to acquire, but is often scattered through multiple platforms, services and institutions.

### **Gather information**

Information should be contextually aggregated, allowing for quick access of relevant information.

# Overview (Primary Objectives)

## Developed Solution

- Take sequencing reads, align them against a reference genome and perform differential expression analysis.
- Take a list of gene identifiers, cross reference them between multiple online platforms and discover potential RBPs (as well as additional relevant information).
- Uncover implicit relationships present in the collected data that may be useful to biologists, using data mining techniques.



# Overview (Secondary Objectives)

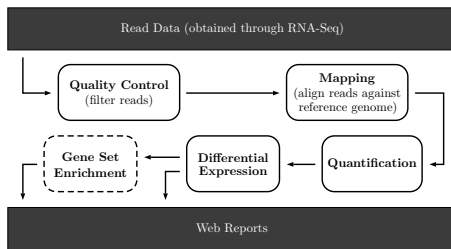
## Developed Solution

- Develop two distinct pipelines: one for differential expression analysis and another for RBP discovery and analysis.
- The two pipelines may also be chained for a complete study.
- Both pipelines should be available through web platforms that can manage user accounts, user jobs (analysis tasks), and view and export results.
- Both the platforms and pipelines should be as modular as possible, to easily allow future development.

# RNA-Seq Analysis Pipeline

## Developed Solution

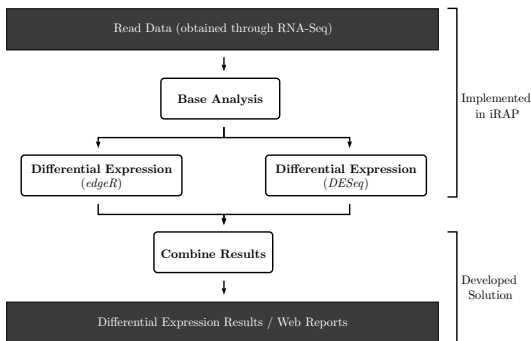
- Uses iRAP (an RNA-Seq analysis pipeline) as its base.
- iRAP is flexible and allows users to choose what tool should be used in each step of the process.
- But, if users have little experience with these tools, how can they choose the best one for the job?



# RNA-Seq Analysis Pipeline

## Developed Solution

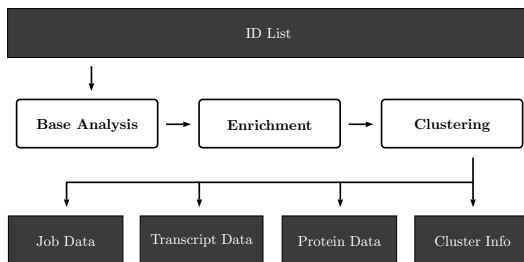
- Uses multiple differential expression analysis tools (edgeR e DESeq).
- Combines the results of those tools.



# RBP Analysis Pipeline (PBS Finder)

## Developed Solution

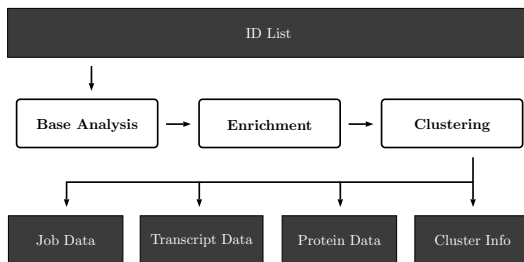
- Uses DB2DB, Ensembl and NCBI to convert identifiers, identify gene species, obtain basic information and extract genetic sequences (5' UTR, 3' UTR, 3' UTR downstream).



# RBP Analysis Pipeline (PBS Finder)

## Developed Solution

- Uses RBPDB to discover RNA binding proteins (based on the obtained genetic sequences).
- Uses UniProt and KEGG to enrich the obtained results and performs clustering analysis on those results.



# RBP Analysis Pipeline (PBS Finder)

## Developed Solution

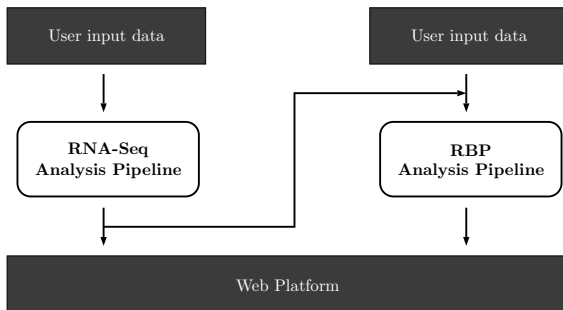
### Clustering analysis:

- uses  $k$ -medoids and hierarchical clustering, both with Jaccard and binary distance matrices;
- executes every possible combination of clustering setups (alternates algorithms, distance matrices, used features, etc.);
- results are filtered (acceptable solutions must have a minimum percentage of entries per cluster, clusters must have defining features, etc.);
- solution quality internally determined based on the average silhouette.

# Integration

## Developed Solution

A fully integrated solution: the analysis pipelines can be used separately or automatically executed in sequence; result visualization for both pipelines is isolated.



# RNA-Seq Analysis Pipeline

## Case Studies

### Objective

- Ascertain if combining the results of multiple tools has impact on the set of differentially expressed genes.

### Data set

- Reproduction of ArrayExpress experiment E-GEOD-48829 (*Escherichia coli*).
- Reference genome obtained from Ensembl Genomes and read data obtained from ENA Sequence Read Archive.



# RNA-Seq Analysis Pipeline

## Case Studies

Results (number of differentially expressed genes)

	Raw results	Filtered results	Combined results
edgeR	4494	386	191
DESeq	4494	204	

## Conclusions

- Combining results impacts the final differentially expressed gene list by reducing its size.
- The combined results will hopefully give researchers an higher confidence in the experimental results.

# RBP Analysis Pipeline (PBS Finder)

## Case Studies

### Objectives

- Assess the general usefulness of PBS Finder.
- Compare PBS Finder with the existing techniques of manual analysis.
- Assess the impact of differences in hardware performance in the overall performance of the platform.

### Data set

- 23 genes from the *RhoGTPase* family (*Rattus norvegicus*) provided by IBMC.

# RBP Analysis Pipeline (PBS Finder)

## Case Studies

Results (expert estimation of 30 minutes per gene analysed)

Number of IDs	Machine1	Machine2	Manual method
100	≈ 10m	≈ 11m	≈ 50h
500	≈ 42m	≈ 56m	≈ 250h
900	≈ 1h 34m	≈ 2h 7m	≈ 450h

## Conclusions

- PBS Finder can reproduce the results an expert would get.
- Months worth of an expert's manual work can be accomplished in a few hours.
- While hardware performance has a significant impact on analysis time, the platform achieves satisfactory performance on personal computer-level hardware.

# Objective Fulfilment

## Conclusions

- RBP analysis pipeline and web platform (PBS Finder) implemented and tested. PBS Finder has been in production for several months; during this time it was thoroughly tested by IBMC experts.
- RNA-Seq analysis pipeline implemented and tested (iRAP deployed and result consolidation tool implemented).
- Integration of both tools was not be accomplished.

# Future Work

## Conclusions

- Fully integrate the RNA-Seq analysis pipeline with the web platform (automatic job configuration, result visualization, etc.).
- Study the requirements for deploying the platform in large scale, and assess the feasibility of making it available internet-wide.

# Review

## 1 Introduction

- Domain Problem
- Motivation and Objectives

## 2 Developed Solution

- Overview
- RNA-Seq Analysis Pipeline
- RBP Analysis Pipeline (PBS Finder)
- Integration

## 3 Case Studies

- RNA-Seq Analysis Pipeline
- RBP Analysis Pipeline (PBS Finder)

## 4 Conclusions

- Objectives Fulfilment
- Future Work

# A Computational Platform for Gene Expression Analysis

Diogo Teixeira

Supervisors: Rui Camacho<sup>1</sup>, Nuno Fonseca<sup>2</sup>

<sup>1</sup>LIAAD INESC & DEI FEUP, Universidade do Porto

<sup>2</sup>EMBL-EBI, Cambridge, UK

July 2014