Introduction

Currently omics generate large amounts of information about the different types of molecules that exist from the simplest to the most complex organisms. There is five main omics that exist currently and each one of these, produce information from specific types of molecules. Genomics, the field of science that is focused on the study of DNA, more precisely in the study of genomes and genes from different organisms. Transcriptomics, focused on the study of RNA. Proteomics is the area that focus his efforts on the study of proteins, using technologies as mass spectrometry and immunoassays to protein profiling and protein detection respectively. Metabolomics, the study of the compounds produced by the cells metabolism. It uses identification and quantification techniques as different types of chromatography and mass spectrometry to separate and identify the different metabolites that are produced. Fluxomics, the study of the rates of the reactions that produce different metabolites, this area of research is fundamental for biological engineering and other areas. Generally, it uses labelled carbon atoms that are quantified using techniques as mass spectrometry.

The transcriptome is the group of all transcripts in a cell in a specific condition. The study of these transcripts is fundamental to uncover the constitution of cells, tissues and characterize what is present in the cells in specific diseases or conditions. The main objective of transcriptomics is to identify the transcripts that are produced, including mRNA, small RNA (siRNA,miRNA…) and non-coding RNA. It also focusses on the transcriptional structure of genes, splicing patterns and post-transcriptional modifications. And one of the most interesting objectives and capabilities of transcriptomics is the measurement of differential expression levels of different genes in different sceneries, like different physiological states, during development and pathologies. [RNA-Seq: a revolutionary tool for transcriptomics]

There are two types of methods used in transcriptomics, the sequence-based methods and the hybridization-based methods. The hybridization-based methods generally use labelled molecules of cDNA to quantify expression on microarrays. While microarrays are inexpensive, easy to use and robust, they have some technical limitations. The limitations go from the necessity of previous knowledge of the genome, high levels of background, high difficulty in comparison among different experiments to limited range of detection due to signal saturation and background expression. [RNA-Seq: a revolutionary tool for transcriptomics]

Opposed to hybridization methods, sequence-based methods as RNA-seq directly identifies the sequence of the cDNA. As it generates high amounts of data, it requires higher computational power to analyse the data generated and is generally more expensive than microarrays. But it doesn’t require previous knowledge of the genome to work, it has a dynamic range to quantify gene expression (over 8000 fold), it is able to distinguish isoforms, it requires low amounts of sample to be executed and the cost can be considered low when it’s used to map transcriptomes of large genomes. [RNA-Seq: a revolutionary tool for transcriptomics]

There are several databases for transcriptomics where different data can be consulted. One of the best know is Gene Expression Omnibus (GEO). This database introduced MIAME and MINSEQE and is hosted by NCBI. There is ArrayExpress that will be further discussed later. The Expression Atlas is a database tissue-specific for different types of organisms and it’s hosted by EBI. Genevestigator is a privately curated database which is peculiar because individual experiments submitted to the database are normalized to permit comparison among different experiments. The RefEx database is a general database that also compiles the transcriptomes of human, rat and mouse in over 40 different organs. Another peculiar database is NONCODE which is a database of non-coding RNAs (NcRNAs) sequenced by RNA-seq. [Wikipedia]

ArrayExpress is a repository where it’s archived functional genomics data, more precisely microarrays and sequence platforms. This database supports MIAME and MINSEQE, making data that follows these guidelines easier to be submitted. The experiments data present in ArrayExpress are directly submitted to the platform or are imported from NCBI GEO. The raw high-throughput sequencing data are stored in European Nucleotide Archive while data from microarrays are archived in ArrayExpress. The team behind this platform is part of Functional Genomics group at EMBL-EBI. [https://www.ebi.ac.uk/arrayexpress/about.html]

The study of gene expression allows us to study which are the genes responsible for a specific phenotype or compound in a determined disease or environmental condition. It is studied among all fields of research on biomedical research, from degenerative diseases to immunologic diseases. The study of the transcriptome is used to understand the human development and compare cells among tissues. It’s also used to comparing the genetic expression among species capable of regenerating limbs and medulla. The comparison between healthy and cancer cells is majorly done using transcriptomics techniques microarrays and RNA-seq are used to discover which are the genes that promote the development and activity of cancer cells.

The problem with using databases like ArrayExpress is that to assess and evaluate the available data, it’s a time spending job that isn’t efficient. It’s inefficient because you need to search among experiments, files and protocols manually, opening each of these data and analyzing it manually. There’s the need to optimize this type of processes to decrease the time that is spent analyzing these data.

Methods

The approach made was using an Application Programming Interface (API) from Array Express. API is a group of steps and protocols that allow the communication and utilization of an application with third party applications. It allows entities to share their information with users more efficiently. Software developers can use the APIs to create third-party applications capable of obtaining complex data in a more efficient way. Users can benefit from these third-party applications that can obtain data from different sources. Complex applications can be created because software developers can use different APIs to integrate data from multiple sources, producing more complete data in an easier way to obtain to the user. So, an API is powerful tool that allows the access to data from different sources in a quicker and easier way. [A Review on Genomics APIs]

The specific type of APIs in these case is Genomic APIs. Typically, a genomic API deals with complex information generated from different methods. One major problem is that there are a high diversity of data and a variety of file formats. Because of this, there isn’t a unique API capable of accessing all the information available from different sources. [A Review on Genomics APIs]

The API from Array Express works with REST-style queries. Representational State Transfer (REST) is a hybrid architectural style derived from some network-based architectural styles combined with a group of architectural constraints: null style, client-server, stateless, cache, uniform interface, layered system and code-on-demand. [ http://www.ics.uci.edu/~fielding/pubs/dissertation/rest\_arch\_style.htm] The API available from Array Express is an REST API. That means that the API follows the REST architecture. It also means we use a uniform interface to do the requests (e.g.: “HTTP”) and the output is in JSON or XML format. [ <http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm>]

The Array Express website is really to use to access information. It’s only needed the keywords or criteria that it’s needed, and search for the experiments. Inside the experiment you can access to all the information from the experiment but also separate files as the protocols used and files containing data as investigation description, raw data, processed data and arrays. When using the Array Express API you can access three types of data: experiments, protocols and files. Either type of data you request, as the API follows REST architecture, the output will be in XML or JSON format. The format of the output can is chosen by the user. [https://www.ebi.ac.uk/arrayexpress/help/programmatic\_access.html]

To use the API from Array Express it was created a class in Python 3. The programming language used is Python 3 because it is a powerful high-level language. It is of friendly usage and easy to learn but also versatile as it has packages almost anything.

This class uses the package *requests* to obtain the information from the database. This package allows to send a request using an URL that follows the HTTP protocol. It’s useful because it’s simple to create a base URL and add the criteria or keywords needed to base URL. Then the request is done using the function get() from the package to send the generated URL and obtain the XML or JSON with the query results. This process is fully automated, the user only needs to choose the criteria and/or keywords and the format of the output.