Knowledge graph Final Project

Luca Farinola, i6326662

1 Significance

Bioinformatics is an interdisciplinary field that combines biology, computer science, mathematics, and statistics to analyze and interpret biological data. It is a rapidly growing field that plays a critical role in understanding the vast amounts of data generated in biotechnology and genomics labs. Due to technology advancements, in fact, an enormous increase in data is faced, necessitating the creation of massive databases, such as USC NCBI and Uniprot[i]. Knowledge graph can be very useful for storage and mining of such data. At the same time the field of bioinformatics is constantly evolving, and new computational and experimental approaches are being developed to handle the increasing complexity of biological data. In bioinformatics, data analysis helps in understanding biological systems, including tissues functions, protein interactions, and gene expression patterns. In the latter example, understanding patterns of gene expression in different biological conditions can lead to the identification of potential targets for disease diagnosis and treatment. A very common way to extract such data is by using RNA sequencing[ii].Dealing with RNA means performing transcriptomic studies which requires comparison of two situations and the extraction of genes whose expression profile is different in the two cases[iii]. In a standard analysis, genes that are significantly differentially expressed would be selected. Significance is established through t-test statistic corrected for multiple testing, generally using False Discovery Rate (FDR). One limitation of this approach is that this kind of analysis do not capture the relationship between genes. In fact, a possible follow up analysis can be related to building and mining networks, for example using gene expression levels to calculate correlation among genes[iv]. The major objective of this project is to build such networks and convert them into an RDF knowledge graph that can be expanded to enable answers to increasingly complex questions through queries.

2 Related Work

Network medicine is an emerging area of research that can allow a more holistic understanding of complex biological systems. Different kind of interactions among genes, and therefore different kind of biological networks have been described and analyzed. Protein-protein interaction (PPI), gene regulation or signaling are just few example[v]. In this project I want to use correlation to build a gene co-expression network where genes are linked together and edges are weighted according to their correlation[vi]. Weighted Gene Correlation Network (WGCNA) is a very good example of an algorithm developed to study those kind of graphs[vii]. Differentially expressed genes can be used also to perform enrichment analysis or over representation analysis (ORA)[viii]. This is where semantic web and ontologies comes into play as those analysis will use information stored in different biomedical datasets, with appropriate statistical analysis to link genes to specific biological pathways or molecular functions. To make an example Gene Ontology is probably the most used system of classification for describing genes and their products across all organisms[ix]. In fact, genes can be linked to a disease or be targeted by a chemical compound as well as located in a specific chromosome. The field of Drug Discovery for example heavily rely on huge databases that can give detailed information on drugs and chemical compounds[x] or link genes with other diseases to investigate on possible comorbidities[xi]. The growing number of databases made it difficult to retrieve this different information all at once. Bio2RDF is an open-source project that provide a unique URL in the form of http://bio2rdf.org/namespace:id in order to provide a common interface for accessing and integrating data from various biomedical databases[xii]. In this small project I want to use different dataset and combine it with information from gene co-expression network.

3 Goal

In this project the main goal is generating an RDF knowledge graph that contains both general as well as case-specific information. With general I refer to information that can be extracted from huge database while gene correlation network it's case specific since it's built from a standard differential expression analysis from an RNA sequencing experiment. The main Idea of this project is to expand an already exiting knowledge graph with experimental data in order to be able to answer questions through SPARQL queries. The advantage of

using a knowledge graph is that it allows you to integrate and reason over diverse sources of data in a more flexible and extensible way. The expected outcome is therefore a knowledge graph containing additional information on differentially expressed genes such as fold change and correlation. This would be an interesting approach to analyze gene expression data as it enables scientists to dive deeper into biological processes linked to their collected data. If successful this approach can be further developed not only with more advanced and specific queries but also by applying algorithms to further investigate and mine this network performing tasks such as link predictions.

4 Methodology

I will start by choosing the Dataset and performing a differential expression analysis. Differentially expressed genes will be picked for the construction of a correlation network. a threshold for the correlation will be used to filter only relevant edges. The following is the construction and expansion of a Knowledge graph. The aim is to use already existing ontology such as Bio2RDF or DisGeNET. As a final step I'm willing to mine such graph performing some queries that can be of relevant interest to answer domain specific questions according to the gene expression dataset.

5 Milestone and Deliverables

Week1 -> I will choose the dataset and perform differential expression analysis

Week2 -> construction of correlation network and selection of Metadata Schema

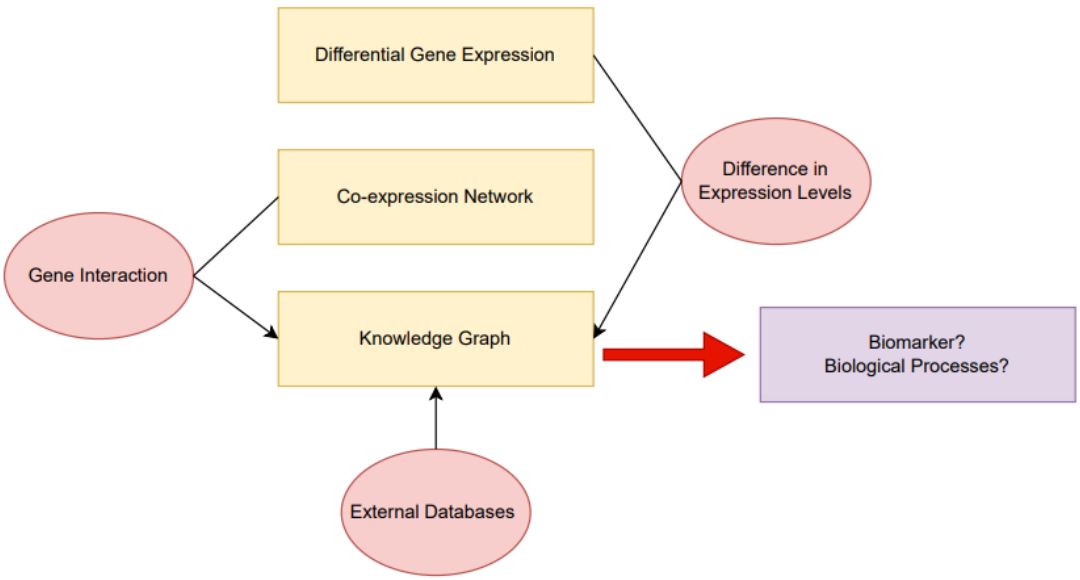
Week3 -> construction and integration of KG

Week4 -> query to answer relevant research questions

6 Anticipated Result

The main output will be a pipeline that can suggest possible follow-up analysis in the field o biomedical research using Knowledge graph. Hopefully this will enables biologist and bioinformatician to retrieve relevant information and speed up the research in the field of molecular biology

General Workflow



ABSTRACT

For this project I'm using a transcriptomics dataset from the MAGNet consortium. To generate this dataset, left ventricular free-wall tissue was harvested at the time of cardiac surgery from subjects with heart failure undergoing transplantation and from unused donor hearts with normal function. In this way I was able to compare Dilated Cardiomyopathic (DCM) and normal patients or non-failure (NF). As a first step I have performed a differential expression analysis. I have then used the retrieved differentially expressed genes to build co-expression network weighted according to patients expression profiles. In parallel a knowledge graph was constructed using the information present in the bio2rdf open-source project[xiii], biomart and gene ontologies, this was then enlarged with calculations made on those experimental data such as p-values and fold change as well as co-expression among genes and gene ontologies. This was later used to retrieve relevant information using SPARQL queries.

RESULTS:

1.1 Extraction of differentially expressed genes

The R programming language was used to carry out the initial stages. R is in fact extensively used in the field of bioinformatics and is thus ideal for working with those types of data. Some exploratory plots were created prior to the differential expression analysis in order to make some observations and determine the quality of the data. Heatmap displaying the most variable genes across the various etiologies is also produced to better study and visualize gene expression. In order to compare genes, the CPM count is converted into FPKM. This is done because FPKM normalization accounts for gene length witch is needed to compare expression levels across genes. PCA is also shown, a clear separation between non failure and other etiologies is observed.

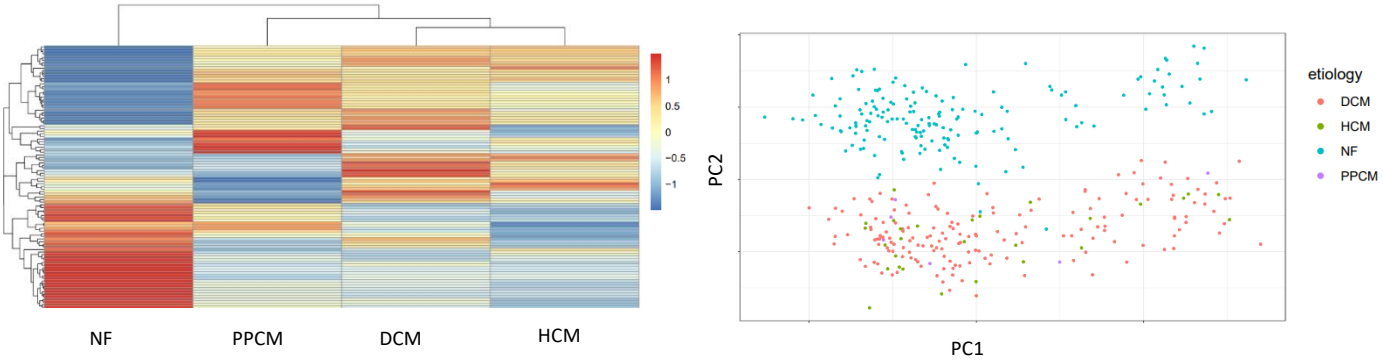


Fig 1) Top 20 variable genes (left), PCA (right)

I chose to focus on DCM simply for an higher number of samples. Differential gene expression analysis was then performed using the limma package[xiv]. Important covariates (gender, age and diabetes) have been added to the linear model. In this project I decided to focus on DCM patients so only DCM vs NF contrast is taken into consideration. 1061 genes resulted differentially expressed in this comparison, taking as threshold a logarithmic fold change of 1 and adjusted p-value, corrected for multiple testing using false discovery rate (FDR), of 0.05. In fig 2 a volcano plot is shown to give a graphical representation of such selection.

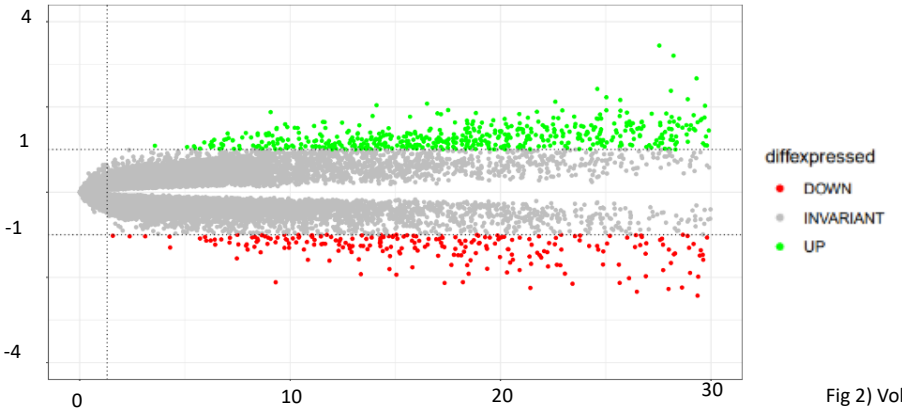


Fig 2) Volcano Plot

1.2 Co-expression network

Those selected genes were used to construct a gene co-expression network. In order to do so spearman correlation was used to form correlation matrix. Spearman correlation is chosen over Pearson because it can capture also non monotonic relationship. Edges are formed taking as a threshold the correlation absolute value of 0.7. Note that by using the absolute value both positive and negative correlation are taken into consideration. Using Network-x, a correlation matrix is therefore transformed into a weighted graph. In this graph 613 genes have at least one edge. Those genes will be used to construct the knowledge graph that will capture protein co-expression information.

1.3 Knowledge graph Construction and Analysis

Different Datasets and bioinformatics tools are used to construct the knowledge graph. In order to be consistent and comprehensible I have decided to take advantage of Biolink[xv], that provides a standard set of semantic types and relationships for the biomedical domain. The weighted gene expression network is therefore converted into a knowledge graph by using rdflib objects in python. In this graph Both positive and negative correlations are taken into consideration using the different predicates to distinguish the two situations. Logarithmic Fold change and FDR corrected p values are also added as litterals. Biomart[xvi] was used to extract gene ontologies. Gene Ontologies can fall into three main categories: Molecular functions, Cellular Components and Biological Processes. Each term is linked to a stable identifier and contains a set of genes that are involved in that specific biological process, have that specific molecular function or are found in a particular cell compartment (Cellular Components). Gene ontologies can be very general or extremely specific and the number of genes that are found in a specific category is variable. For this reason simply linking genes to ontologies does not give relevant information. It is instead important to identify significant over-representation of specific biological terms through statistical tests. Through EnrichR[xvii] significance is established using T-test and p values that are again corrected for multiple testing. The constructed Knowledge graph contains triples linking genes to significant ontologies. Adjusted p-values are also added as literals both for genes significant change in expression values (Fold change), calculated in the differential gene expression, and for significant Gene Ontologies enrichments. Biotordf is also used to retrieve genes symbols, chromosome position and protein family as well as PubMed articles.

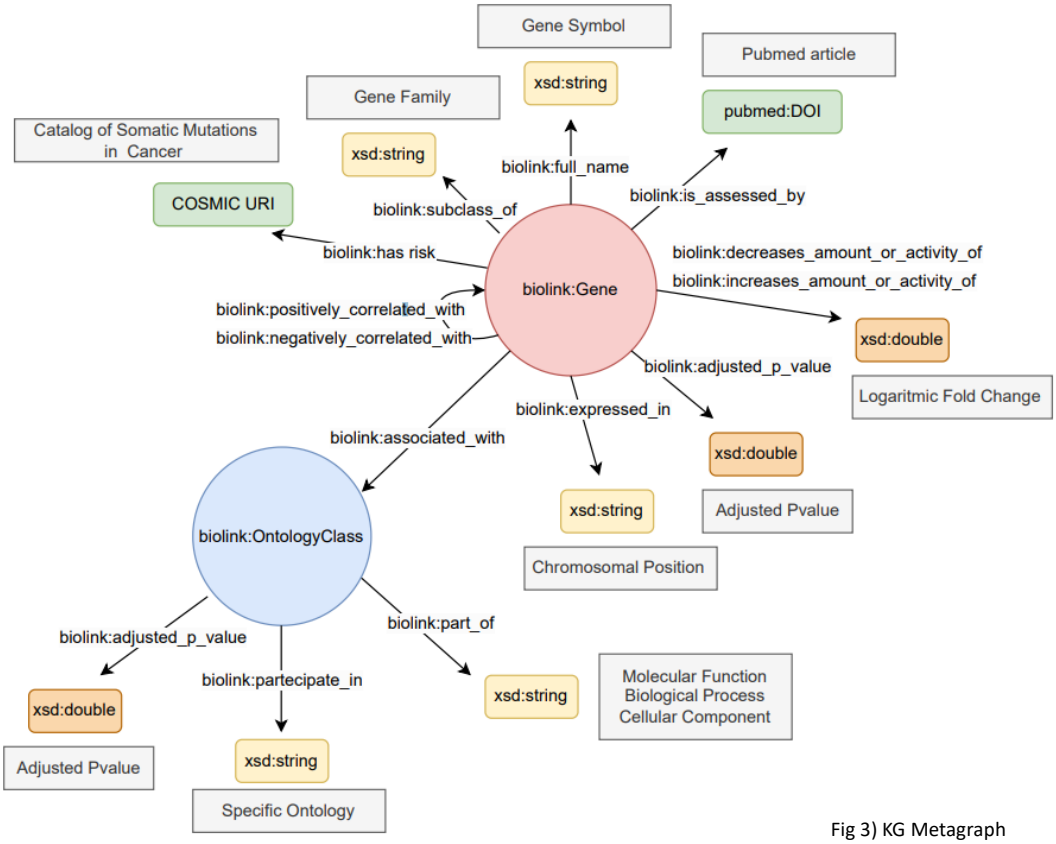


Fig 3) KG Metagraph

This graph contains 4254 ontologies and 595 genes with an average degree of 4.01. By looking at the fundamental network analysis parameter, I found that the gene ISLR has both high degree and centrality. This gene is connected to the immune response and is involved in adhesion or binding to other proteins. Importantly, this conclusion is congruent with earlier literature. In fact, ISLR, LUM, and ASPN were named as the HF hub genes and suggested as possible biomarkers in this study[xviii]. Similarly interesting observations can be done by looking at PubMed articles associated to co-expressed genes. Most of the articles are associated to a single gene, but interestingly some of them are observed in more than one triple in the knowledge graph. The following article captured my attention[xix]. Interestingly, Serpins are implicated in regulation of the cardiovascular system and have been found to regulate blood pressure.

1.4SPARQL

SPARQL queries are found in txt files called SPARQL_queries.txt . SPARQL was used to extract meaningful information out of the data obtained. The following tables shows top 6 up regulated genes and most enriched Biological processes ordered according to their significance.It is interesting to notice here that Ontologies and Up regulated genes descriptions are mostly related to extracellular matrix formation and collagen secretion which is coherent with what we know about dilated cardiomyopathies. This is in fact probably related to left ventricular (LV) remodelling[xx].

Genes	Description	Chromosomal Position	Log FoldChange	Adj.P.val
FNDC1	Fibronectin type III domain containing	6q25	3.006045	6.263961e-90
FRZB	Secreted frizzled-related proteins	2q32.1	2.142725	3.981418e-78
SFRP4	Secreted frizzled-related proteins	7p14.1	3.468365	2.980358e-77
MFAP4	Fibrinogen C domain containing	17p11.2	1.66582	1.925156e-66
COL22A1	Collagens	8q24.3	3.489646	2.790672e-66

Table 1) Up regulated genes

Gene Ontologies	Adj.P.value
collagen-containing extracellular matrix	7.833393e-35
collagen trimer	2.05994e-13
extracellular matrix organization	3.739119e-12
blood microparticle	4.813465e-11
immunoglobulin complex	5.660651e-11
glycosaminoglycan binding	4.032397e-10
extracellular matrix structural constituent	8.41023e-10

Table 2) Enriched Ontologies

DISCUSSION:

The following project provide a possible workflow to deal with RNA-sequencing data. The main Idea is to be able to combine experimental data with information that can be retrieved from other sources and databases. Nowadays being able to integrate and organize different kind of data is of crucial importance especially in this field. In this project I have managed to parse a knowledge graph (biotordf) and map specific genes of interest to create a different network by updating the schema and adding new triples. The main limitation of the BiotoRDF parsed data on their own is that information are mainly describing single nodes, it is not very interconnected which makes mining very difficult. On the other hand having information on protein interactions can improve network connectivity. Also Ontologies in this case are meant to connect this nodes further. More than one gene can be involved in the same biological process or even have the same function Moreover the schema was updated taking advantage of classes and objects present in bioink. Stable and reusable Identifiers are adopted using HGNC symbol for genes, the GO for Ontologies, DOI for articles …

The main problem here is that some information is lost in the process of selecting differentially expressed genes and only connected nodes as well as converting genes from ensemble identifiers to HGNC, needed to parse info from biotordf. As it is very case specific the graph is probably very small compared to a more general knowledge graph. In fact, Amie rule mining algorithm has been applied but with no good results. Rules extracted have very low confidence, probably this is due to the dimension of the KG. the tsv file containing this rules is found in the folder.

CONCLUSION:

As a final remark I would say that the overall results are not striking but in line with expectations. Often in biology we want to study changes in biological activity that can be observed in different conditions. Networks can be applied to add different layers of information. As a System Biology student, I probably still see such types of analyses from the perspective of interpretation and less from the technical side. I still believe I was able to develop a method that can be used on similar datasets to expose intriguing insights.

REFERENCES:

[i] The UniProt Consortium et al., «UniProt: The Universal Protein Knowledgebase in 2023», *Nucleic Acids Research* 51, fasc. D1 (6 gennaio 2023): D523–31, https://doi.org/10.1093/nar/gkac1052.
[ii] M.-A. Dillies et al., «A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis», *Briefings in Bioinformatics* 14, fasc. 6 (1 novembre 2013): 671–83, https://doi.org/10.1093/bib/bbs046.
[iii] W. Douglas Thompson, «Statistical Analysis of Case-Control Studies», *Epidemiologic Reviews* 16, fasc. 1 (1994): 33–50, https://doi.org/10.1093/oxfordjournals.epirev.a036143.
[iv] Shiyi Liu et al., «Three Differential Expression Analysis Methods for RNA Sequencing: Limma, EdgeR, DESeq2», *Journal of Visualized Experiments*, fasc. 175 (18 settembre 2021): 62528, https://doi.org/10.3791/62528.
[v] Albert-László Barabási, Natali Gulbahce, e Joseph Loscalzo, «Network Medicine: A Network-Based Approach to Human Disease», *Nature Reviews Genetics* 12, fasc. 1 (gennaio 2011): 56–68, https://doi.org/10.1038/nrg2918.
[vi] Joshua M. Stuart et al., «A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules», *Science* 302, fasc. 5643 (10 ottobre 2003): 249–55, https://doi.org/10.1126/science.1087447.
[vii] Peter Langfelder e Steve Horvath, «WGCNA: An R Package for Weighted Correlation Network Analysis», *BMC Bioinformatics* 9, fasc. 1 (dicembre 2008): 559, https://doi.org/10.1186/1471-2105-9-559.
[viii] Enrico Glaab et al., «EnrichNet: Network-Based Gene Set Enrichment Analysis», *Bioinformatics* 28, fasc. 18 (15 settembre 2012): i451–57, https://doi.org/10.1093/bioinformatics/bts389.
[ix] Gene Ontology Consortium, «The Gene Ontology (GO) Database and Informatics Resource», *Nucleic Acids Research* 32, fasc. 90001 (1 gennaio 2004): 258D – 261, https://doi.org/10.1093/nar/gkh036.
[x] David S. Wishart et al., «DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets», *Nucleic Acids Research* 36, fasc. suppl_1 (1 gennaio 2008): D901–6, https://doi.org/10.1093/nar/gkm958.
[xi] Janet Piñero et al., «The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update», *Nucleic Acids Research*, 4 novembre 2019, gkz1021, https://doi.org/10.1093/nar/gkz1021.
[xii] François Belleau et al., «Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems», *Journal of Biomedical Informatics* 41, fasc. 5 (ottobre 2008): 706–16, https://doi.org/10.1016/j.jbi.2008.03.004.
[xiii] Belleau et al.
xiv Matthew E. Ritchie et al., «Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies», Nucleic Acids Research 43, fasc. 7 (20 aprile 2015): e47–e47, https://doi.org/10.1093/nar/gkv007.
xv Deepak R. Unni et al., «Biolink Model: A Universal Schema for Knowledge Graphs in Clinical, Biomedical, and Translational Science», Clinical and Translational Science 15, fasc. 8 (agosto 2022): 1848–55, https://doi.org/10.1111/cts.13302.
xvi Damian Smedley et al., «BioMart – Biological Queries Made Easy», BMC Genomics 10, fasc. 1 (dicembre 2009): 22, https://doi.org/10.1186/1471-2164-10-22.
xvii Edward Y Chen et al., «Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool», BMC Bioinformatics 14, fasc. 1 (dicembre 2013): 128, https://doi.org/10.1186/1471-2105-14-128.

[xviii] Yang Guo et al., «Identification of Hub Diagnostic Biomarkers and Candidate Therapeutic Drugs in Heart Failure», *International Journal of General Medicine* Volume 15 (gennaio 2022): 623–35, https://doi.org/10.2147/IJGM.S349235.

xix Claire Heit et al., «Update of the Human and Mouse SERPINgene Superfamily», Human Genomics 7, fasc. 1 (dicembre 2013): 22, https://doi.org/10.1186/1479-7364-7-22.

xx Merry L. Lindsey et al., «A Novel Collagen Matricryptin Reduces Left Ventricular Dilation Post-Myocardial Infarction by Promoting Scar Formation and Angiogenesis», Journal of the American College of Cardiology 66, fasc. 12 (settembre 2015): 1364–74, https://doi.org/10.1016/j.jacc.2015.07.035.