EXPANET: A PATHWAY LEVEL ANALYSIS TOOL THROUGH GRAPH EXPANSION USING MARKOV CHAINS AND RANDOM WALKS

MASTER'S THESIS

FIKAS NIKOLAOS

AUGUST 2018

Academic Supervisors:
Michalis Aivaliotis
Christoforos Nikolaou
Pavlos Pavlidis

Dedicated to Pyr & Tainia

Abstract

Moving from protein deregulation-level statistical analysis to the ones that take into account the deregulation levels of functional protein groups and pathways, the statistical power of the results increases and a systemic approach towards understanding the biological question is offered. This approach, 20 years after its birth, resulted in the creation of a variety of statistical approaches like GSEA, PAGE, GAGE etc. These approaches belong to the gene-set analysis category, which use at their basis, the lists of biological processes and pathways offered by online databases like KEGG, MSigDB, Reactome, BioCyc, etc, and analyze data from micro-arrays, next generation sequencing and recently proteomics methods. One major drawback of all these approaches is that they do not take into account the interactions between proteins of different pathways because neither topological, nor dynamic information of the analyzed networks is fed into their algorithms. In order to surpass the above disadvantage, this work aimed to develop a new package in R, based on the work of Dupont et al. [18], where by modeling limited random walks in graph using Markov Chain properties a relevant sub-network extraction achieved. These extracted relevant sub-networks represent expanded forms of the known biological pathways that when compared between different conditions obtain a pathway-level deregulation score. In the current work, several gene-expression lymphoma data-sets were used for the validation and evaluation of the new tool. In addition, a small scale proteomic data-set from a currently running project in the lab in *Plasmodium* was analyzed by ExpaNET in order to evaluate its applicability in proteomic data.

Keywords: protein interactions network, graph, gene-expression, proteomics

Acknowledgements

I would like to thank my supervisor

Michalis Aivaliotis

and my two co-supervisors

Christoforos Nikolaou and Pavlos Pavlidis

Contents

List of Tables

List of Figures

Chapter 1

Introduction

> Science is not only a disciple of reason but, also, one of romance and passion.
>
> *Stephen Hawking*

1.1 Motivation

Following the technological advances of the last decades, on big complex data statistical and graph analysis for an accurate study and insightful interpretation of the massive relational networks that occur from the increasingly use of the social media, one can realize that many new mathematical models and theories have been developed in order to shed light and decrypt the inner structures and functions of such objects.

In contrary with the computer science, which has developed the previously mentioned theory and tools, biology, a completely different discipline, who have been in the era of omics and big data during the few last decades, seems that is facing difficulties and moves slowly to the adaptation of these new procedures and approaches, while at the same time seems to be anchored to outdated - old protocols which many of them have been questioned.

This gap is starting to decrease slowly the last years with the emergence of a multidisciplinary filed called bioinformatics. The main objective of this research is to contribute towards this direction.

1.2 Research Topic

The main subject of the current work is the actual development of a tool based on the

the graph theory and Markov Chains property, belonging to the pathway analysis tools category. This tool will have as its main purpose, a score calculation describing the degree of change between two conditions of each specified biological network given as input to.

1.3 Research Problem

Need for the development of more sophisticated bioinformatics tools based on the latest advances of computer science in order to replace the ones that face many limitations.

1.4 Research Overview

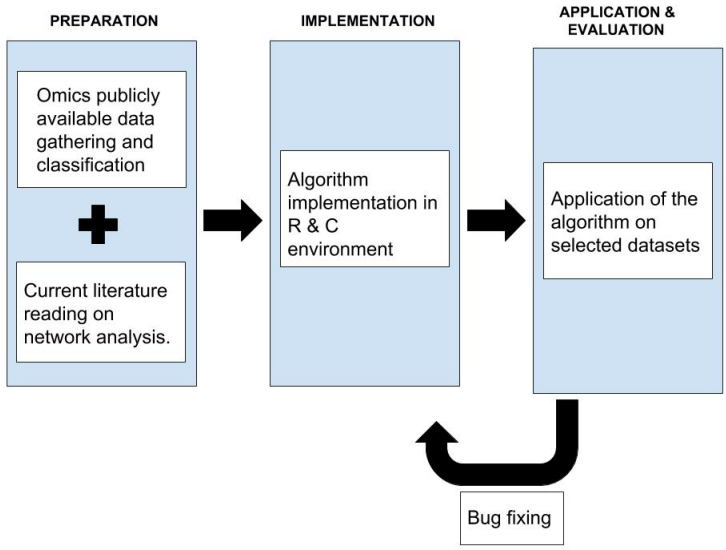

Figure 1.1: Research process of the Master's thesis

1.5 Common Abbreviations

- MS - Mass spectrometer
- NGS - Next Generation Sequencing
- GO - Gene Ontoology
- ORA - Over Representation Analysis
- FCS - Functional Class Scoring
- PT - Pathway Topology

Chapter 2

Literature Review

2.1 Omics era and data struggling in biology

The recently and quickly developed technological advances in molecular biology field, have dramatically increased our ability to perform biologically relevant researches that produce immense amount of data. The potential of those advances made clear the gap and still broadens it, between the production of such data and their imminent analysis. Sure enough, the storage, the management and the analysis of those data is one of the primary concern at the time, in molecular's biology field. The challenge now is to be able to read and study properly the whole data volume and infer through them the basic causes that lie underneath, the clinical observation one studies. From the other hand, integration of different -omic platforms will meliorate the diagnosis, the monitoring and the therapy of specific disease, while permitting at the same time the discovery of potentially new biomarkers who will bring the precision medicine reality, one step closer.

One, by following the development of the technological advances in molecular's biology field throughout a historical basis, perceives the reading of the human genome [10, 57] as a historical point which allowed the further development and organization of a plethora of -omics sciences, like genomics, transcriptomics, epigenomics, proteomics, metabolomics and more in scale that few years earlier, such power and scale would be unthinkable [36, 57]. These new technologies made true the production of plenty of GigaBytes of high quality and accurate data in daily basis. That feature is responsible for the liberation of the molecular biology from the fetters of the more traditional analytic approaches imposed to it. The combination of the high throughput ability and the high quality of results with the increasingly reduced cost, is the main reason why in about just 10 years the cost for a whole human genome sequencing reduced from 10 millions into 1000 dollars [26]. By using such powerful technologies like NGS (Next Generation Sequencing) for studying whole genomes [13, 28], exosomes [53], transcriptomes [44, 51], small RNAs [2], epigenomes [45] microbiome [12, 14] implementing even single cell analysis [42], can lead one to shed

light to the molecular mechanisms that are responsible for a specific disease into all the previously mentioned levels. The understanding of those mechanisms does not only help in the understanding of the disease by itself, but is also a tool that helps in the identification of new biomarkers for the more accurate detection of the disease, the stratification of patients, disease monitoring and development of new more targeted medicines and therapies. There is also a 'side effect' of using such technologies and that became clear the last decade of using them. The huge amount of the data that these technologies produce in daily basis, make mandatory the development of new approaches and infrastructures for the management, storage and analysis of them through bioinformatic protocols [31, 46].

The main widely used NGS approaches that have been turn into standard in molecular biology due to the high sequencing coverage, the unbiased view of the complex systems and the low cost is the RNA-seq, which have replace the previous generation microarrays, since it has theoretically the ability to read all the RNA molecules, known and unknown, to detect alternative spliced isoforms, long non-coding RNAs and small RNAs. Besides the RNA-seq, Chip-seq technology is shown to be very informative in respect of studying the regulatory networks of the specified transcription factors. The NGS technology does not stop here. As already mentioned, epigenomic methodologies, such as micrococcal nuclease sensitive sites sequencing (MNase-seq), DNase I hypersensitive sites sequencing (DNase-seq), formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq), and assay for transposase-accessible chromatin using sequencing (ATAC-seq), have been developed and have shown their reliability for the study of chromatin accessibility at the genome-wide level to identify the epigenetic changes responsible for differential gene expression, cell proliferation, functional diversification and disease development.

It would be an oversight, if there was no mention to an also relatively new high-throughput method that has been developed and used in molecular's biology field. High-thoughput mass spectrometers (MS), have been developed in order to explore the whole proteome and/or the metabolome of the cells, biological fluids and tissues [30, 49]. Since proteins are the final product of the gene expression and the actual effector in all biological processes, the direct study of them and of the metabolites in combination with the data of the other -omic technologies, can help and boost shedding more light on diseases' pathobiology and biomarker development. MS has the ability for simultaneously analysis of hundred of thousands of peptides and proteins, while untargeted approaches have been developed for the discovery of novel molecules [30, 49]. MS though, presents some limitations too. The main reason is that the detection of the peptides and the metabolites are based on already existed databases that lack redundancy and still present limitations and especially for metabolites analysis. But as further technological progress will be achieved, and more and more data will be collected on those databases, many of these limitations will eventually disappear. Furthermore,de novo protein sequencing algorithms were developed for the identification of proteins from organisms with non-sequenced genomes (ref).

2.2 An overview of biological data analysis

Based on all the above, it is easy to predict that a plethora of tools and methods have been developed the last decades while the next few years will feature further development of statistical, mathematical and information technology (IT) instruments in the -omic context. This will completely change many central concepts of medicine and biology, and will also require careful regulation to avoid the risks related to a data-driven view. Bellow follows an overview of the different statistical/mathematical methods developed and used over time for the interpretation of biological data.

2.2.1 Gene level statistical analysis

One of the first steps after obtaining the results from the high throughput procedures, is to spot those genes that behave differently between the sample groups. For messenger RNA and miRNA expression studies, where the expression level data are continuous, genes are tested for differential expression between groups by employing a t–test. In the case that more than two groups (conditions) are available, more sophisticated linear models, like ANOVA can be used for the same purpose. From the other hand, in studies implementing SNP or other sequences analyses where the results are categorical, the x^2 test is used. In both cases though, the yield statistic is the used for a comparison with the appropriate distributions (either t or x^2) in order to obtain a p-value for the significance inference. For such analysis, dedicated packages like limma [48] have been developed and imported to BioConductor's [24] library and can be used through the R environment.

However, in all cases, the huge number of hypothesis tests that taken under consideration requires a multiple test correction [17] of p-values. That is, at a threshold of $p \leq 0.05$, we expect a gene to be mistakenly called a significant 5% of the time, leading to thousands of such false positives when the number of genes tested is of the order of 10^5. While the simple Bonferroni correction (in which the significance threshold is set to 0.05 divided by the number of genes tested) can be used, it is considered too conservative. Also, in the Bonferroni adjustment it is assumed that each gene is strictly independent of every other, a case known to be wrong for genomic data. Instead, false positives should be checked using the False Discovery Rate (FDR) [7] , which has been shown to exert robust error rate control even when the multiple hypothesis testing have dependencies [9]. Alternatively, assumption–free but computationally–intensive permutation procedures [27] may be used.

2.2.2 Functional modules identification

The lists of major genes obtained from the assays described above provide limited mechanistic knowledge without additional biological context. In order to understand the biology of the systems, it is necessary to collect genes' information to determine gene sets and interactions between them that meet specific biological functions. Typically, this is done by either finding clusters of genes behaving in the same way in the experiment or by incorporating expertise from pathway databases for the analysis.

2.2.2.1 Cluster analysis

Nowadays it is widely known that the genes interact with each other in transcriptional groups and that these groups in turn interact with other groups [11]. Due to these relationships, genes that work together often show direct or inversely related expression. The simplest method for identifying these groups and linkages is the clustering of genes into groups whose expression is similar across the set of samples [15].

Two are the most commonly used clustering methods. Hierarchical and k-means clustering. Their great popularity is due to their computational and intuitional simplicity. However, because both of them ask from the user to determine the number of clusters, they are prone to artificially separate genes that should belong to the same cluster (e.g when the user determines more clusters than they really exist). Also they present limitations while trying to detect clusters with complex shapes. In order to address these constraints, improvements have been proposed for both systems and described below.

Hierarchical clustering

Commonly used hierarchical clustering algorithm [20] sorts genes based on the similarity of their expression, producing a tree that can be separated into clusters. For each pair of genes i and j, hierarchical clustering calculates a D_{ij} distance measurement by using most often the Euclidean distance measurement ($D_{ij} = \sqrt{\sum_s (g_{i,s}^2 + g_{j,s}^2)}$ where $g_{i,s}$ denotes the gene expression of gene i in sample s). Then, starting with each gene as its own cluster, it repeatedly merge clusters together by accounting the smallest D_{ij} (also called linkage) among them. Except from the Euclidean distance measurement there are also different alternatives for the distance metric like Manhattan or Mahalanobis and correlation–based distances (e.g $D_{ij} = 1 - Cor(g_i, g_j)$). One important note to be added here is that the hierarchical clustering method belongs to the agglomerative or bottom-up subcategory of clustering. That means that the algorithm begins by defining each separate initial gene expression as a cluster by itself and throughout the repetition the merging of the most similar genes is happening until one vast cluster to be created, including all genes. However, while hierarchical clustering has a long history in microarray analysis, it is extremely sensitive to the choice of distance measurement, and therefore this technique should be considered as an exploratory tool rather than an analytical one.

K-means clustering

As mentioned earlier, the other commonly used clustering algorithm is the k-means [29]. In contrast with the hierarchical clustering, k-means belongs to the divisive subcategory of clustering also known as top down. By defining an initial big cluster, the algorithm repeatedly calculates the points that determine the centers of the groups: starting with a user specified number of clusters k, selects the first k genes randomly as the initial centroids and merges all the genes based on their closest to them centroid. For each of

the resulting clusters k, new centroids are calculated based on the mean expression of the genes corresponding to each cluster. The genes then are re-clustered in relation to the new centroids and the procedure is repeated until full converge succeeded. It turns out that k-means is less error prone and faster than the hierarchical clustering but since user must specify the number of clusters same drawbacks and limitations occur.

To remedy these drawbacks, there are several improvements that proposed. Based on spectral graph theory techniques (e.g spectral clustering) [35, 40] are able to articulate clusters with nonlinear boundaries, allowing the discrimination of complex relationships between genes. Various solutions have also been proposed for estimating the number of clusters from the data themselves rather than relying on user specifications [22, 55]. An interesting and scalable approach, called consensus clustering [38], is a method that can be wrapped around any clustering algorithm (hierarchical, k-means, spectral clustering, etc.) to provide an estimate of the number of clusters for the data as well as and a measure of the robustness of clustering. In that approach, data is randomly splitted, so only a portion of the genes and samples are used each time. Then the clustering algorithm that was chosen to be wrapped, repeatedly clusters the samples or genes in $k = 2, 3, 4, 5...$ groups for each different random subset of the initial data. For each k, a consensus matrix is obtained where the i, j-th value represents the percentage of the times that the i and the j gene are assigned to the same cluster in the multiple random subsets. For a really robust clustering of them, it is expected that all records will be close to 1 or 0, meaning that i and j are placed in the same cluster or are placed in different clusters constantly. By taking into account all these different consesus matrices (for each k) the algorithm then suggests the optimum k of which the corresponding consesus matrix comes closest to the ideal of pure 1's and 0's. In a 2005 work of Monti S. et.al [39], consensus clustering used to identify molecular subtypes of diffuse large B-cell lymphoma, ending up to highly-reproducible signatures.

2.2.2.2 Dimensionality reduction

Since the number of analyzed genes is enormous, it is often interesting to find a small number of representative patterns describing the majority of the variance observed in the data and on which the whole gene expression can be modeled (profiled) instead of dealing with the entire set of data. This problem is closely related to clustering: by locating dominant patterns of gene expression , one can then find clusters of genes that match specific patterns.

Principal Component Analysis (PCA)

One of the most-known and simplest dimensionality reduction technique is the Principal Component Analysis (PCA) [1], which has the ability to convert a set of observations of possibly correlated variables like gene expression measurements, into a new set of coordinates called the principal components (PCs). This transformation is defined in such way that the first principal component is drawn along the direction of the greatest variance of

the data, representing most of the overall variance in gene expression between samples. Each following principal component (PC2, PC3, .etc) is drawn, in turn, along the direction of the next highest variance with the restriction that it will always be perpendicular to the previous component. Given that, most of the statistical variance of the data is explained within only the first Principal Components and thus only these first components can be used instead of the initial 10^5 dimensions when analyzing further the data. Often this dimensionality reduction makes clusters to emerge that can then be examined for any common regulatory elements.

Non-linear Dimensionality Reduction

One of the main drawbacks of PCA analysis is the assumption of orthogonality. Assumption that expects linearity from the data. Any pattern that will emerge from a PCA analysis will imply a linear combination of the gene expression measurements. However, it is well know that biological patterns are not always linear. A non-linear behaviour is likely to arise from regulatory networks with feedback loops. Such behaviour neither the standard PCA nor the SVD (Singular Value Decomposition) can discover these patterns. Instead, non-linear dimensional reduction techniques (NLDR) should be used. The NLDR can be considered as a non-linear version of the PCA where the coordinates are "threaded" along the direction of greatest variability. An efficient and optimal detection of these coordinates is a mathematical and computational challenge and several methods have been proposed, such as the kernel PCA, Laplacian eigenmaps, IsoMaps and spectral integration [4, 5]. From these approaches, the neural network based self-organizing map (SOM) [6] is often represented in omics analysis literature.

However, while the NLDR provides a more accurate and probably more biologically compatible dimensionality reduction than PCA or SVD, it should also be noted that transformation from the newly created dimensionaly reduced space to the initial one of the genes is not such a straight-forward task. This is obviously a consequence of their non-linearity and its a feature that clearly places those methods in the opposite site of the PCA and SVD, from which it is easy to retrieve the original (gene) co-ordinates. That characteristic feature causes problems when one tries to identify the genes that play the main role for the emerged pattern. In other words by using one of the previously mentioned techniques is a trade-off between accuracy and interpretability and one should choose, depending on the end goal of the analysis.

2.2.3 Gene-set / Pathway analysis

Many investigators, realized that the common list with the most significant deregulated genes does not always provide mechanistic insights of the underlying biology of the condition being studied. For that reason new methods had to be developed in order to approach that challenge. Such a method is to to simplify the analysis by grouping long lists of individual genes into smaller groups of related genes or proteins. The so-called gene-sets. Gene-sets definitions may be extracted from a growing number of databases, including the

Pathway Interaction Database, KEGG, Reactome, and InnateDB, amongst others. This method has two major advantages. First, it allows the grouping of hundreds of thousands of genes in several hundred pathways, reducing the complexity of the following analysis and secondly, identifying active pathways that differ between two conditions can provide more explanatory power and mechanistic knowledge than a simple list of genes.

These advantages have given rise to many different pathway analysis approaches over the last years as well as plenty of tools have been developed based on these approaches. Such tools are available, either as free, open–source R software from the BioConductor project or as commercial tools such as Ariadne Genomics Pathway Studio and Ingenuity Pathway Analysis. Although the number of the tools is big enough, their underlying methodologies can easily be grouped into three categories/generations. This grouping and methodology discussion is following.

Over-Representation Analysis (ORA) Approaches

The need for a functional analysis of microarray gene expression data and the use of GO during this period resulted in the emergence of the over representation analyses (ORA) [23] that statistically assess the fraction of genes in a particular pathway found among the set of genes showing changes in expression. More specifically it tries to address statistically the question : given a set of genes of a known pathway and given the list of differentially expressed genes (e.g., with FDR ≤ 0.05), is there greater overlap than would be expected by chance alone? That is, do the significant genes seem to aggregate in certain pathways? So, as input the ORA tools are gaining a list of gene-sets with their corresponding genes and a list of over- or under-expressed genes. Then for each pathway an overlap between the genes of the pathway and the deferentially expressed genes is calculated. Finally the actual overlap is statistically tested against a hypergeometric distribution that gives the probability of having an overlap of m or more genes when there are M significant genes out of N genes assayed, and n genes in total on the pathway by using the following equation:

$$Pr(X \geq m|N, M, n) = \sum_{r=m}^{n} \frac{\binom{M}{r}\binom{N-M}{n-r}}{\binom{N}{n}}$$

ORA based tools present some drawbacks. Since they use only the most significant genes (chosen for their <2 fold change and a p-value < 0.05) and discards the marginally less significant (e.g fold change = 1.999 and p-value = 0.051), result in information loss. Second, the statistical analysis they provide, does not take into account the gene expression values associated with each gene, but only the number of genes. That means that ORA algorithms treat all genes as equal and in turn they drive to the assumption that each gene is independent from each other, ignoring the correlation between genes. Lastly ORA, treats each pathway as independent by not taking into account any molecular functions from the GO or signaling pathways from KEGG. Both of the last two ORA assumptions is known to be incorrect in biological context and that makes ORA prone to biased and incorrect results.

Functional Class Scoring (FCS) Approaches

In order to address the limitations of the ORA methods, a new class of approaches have been developed. The so called FCS [3]. In contrast with the ORA, FCS methods get as input the whole list of the genes and their respective expression level values. At first, they compute a gene-level statistic and rank the genes according to their significance for each gene-set. Commonly on those methods this gene-level statistic is calculated using t-test, ANOVA and fold of change. Then, for each specific gene-set, the aggregation of its gene-levels statistics is converted into a pathway-level statistic. This pathway level statistic is obtained by answering the question: what is the probability that the genes in this pathway lie as near the top of the ranked list as we observe them to be? To answer that question Kolmogorov-Smirnov or Wilcoxon rank sum statistics are most commonly used. The power of a pathway-level statistic can depend on the size of the pathway, the proportion of differentially expressed genes in each pathway and the amount of correlation between the genes of each pathway. Finally in order to assess the statistical significance of the pathway level statistic a null hypothesis must be formulated. In the case of FCS, there are two different kind of null hypothesis that can be used. Competitive and Self-contained null hypothesis. The self-contained null hypothesis creates permutations of the class labels (i.e phenotype) for each sample and compares the set of genes in a given pathway with itself, while ignoring the genes outside that pathway. On the other hand a competitive null hypothesis create permutations on the gene labels level for each pathway, and compares the set of genes in the pathway with a set of genes that are not in the pathway.

Functional Class Scoring algorithms achieved to address some of the basic drawbacks of the previously used ORA approaches. Within these achievements is the usage of the whole available information without the need of an arbitrary cut-off threshold of differentially expressed genes. As a consequence they can detect differences between pathways that are barely passing the differentially expressed thresholds and the ones that are passing them with significance levels. Also while ORA completely ignores molecular measurements (i.e gene expressions) for the pathway level inference, FCS methods use this information in order to detect coordinated changes in the expression of genes in the same pathway. Finally by calculating the coordinated changes between the genes, FCS methods have the ability to identify the most relevant genes that might play a key role in the whole pathway.

Nonetheless, FCS approaches are bound to their own limitations. One of the most significant disadvantages is that similar to ORA, FCS analyzes each pathway independently. As discussed earlier this is something being know that is not true in a biological context because a gene might be a part and function in more than one pathway. So any kind of cross-talk/overlap information between the pathways escapes under that kind of investigation.

Pathway Topology Approaches

Previously discussed methodologies take as input a gene-level information and a set of pathways with their corresponding genes. Nowadays though, there are plenty of online

databases that gather, annotate and make publicly available information for the topology of the underlying networks of pathways. Those databases provide information about gene that interact with each other in a given pathway, how they interact (e.g., activation, inhibition, etc.), and where they interact (e.g., cytoplasm, nucleus, etc.). Some of these knowledge databases are KEGG , MetaCyc, Reactome, RegulonDB and BioCarta. Similar to the FCS, PT-based methods they perform along the same general steps, but they add pathway topology for assessing statistical relevance of the pathways and this is the key difference between them. More precisely one of the most used and established tool in that category called SPIA [16, 54], it calculates the gene-level statistic as a sum of of its measured change in expression and a linear function of the perturbation factors of all genes in a pathway based on the topological information of that pathway.

Although still evolving, the algorithms of that categories carries their own limitations as well. One such obvious limitation is that the actual pathway topology is dependent on the conditions of the study as well as on the type of cell due to cell-specific gene expression profiles. However, this information is in a way available but is fragmented in different kind of databases. The lack or inability of the PT-based methods to model dynamic states of a system and the inability to consider interactions between pathways is another worth mention drawback.

Studying and having all these different kind of approaches aside with their limitations in mind, we decided to develop a new methodology that literally lies between the FCS and Pathway Topology approaches and its main concern to be the addressing of the limitation of all previously presented methods of not taking into account the interaction between internal and external genes of the pathway and between pathways.

Chapter 3

Research Methodology

As mentioned in the previous chapter, the main aim of this research is the development of an R compatible package that takes as input high-throughput expression data (microarrays , RNA-seq, proteomics) along with a list of gene-sets and returns a list of these gene-set with a score that defines the how much each specific pathway differ between two conditions. The core of this package is based on the work of P. Dupont, et.al [18] on how to extract relevant subgraphs using random walks in a graph.

3.1 Theory introduction

The problem that the current methodology is trying to resolve is the extraction of a subgraph that best explains the relationships between any $k \geq 2$ given nodes of interest in a graph. For example in a large metabolic network which can be represented as a directed/non-directed connected graph of reactions and bio-molecules one could set at will 2 or more nodes of interest on that particular network and try to to extract a relevant subgraph that explains the relationships between these nodes. Considering a theoretical graph depicted in Figure 3.1, with edges of either the same or different weight and for the sake of simplicity two nodes of interest 1 and 9 colored in green.

In this particular example graph, the best path between nodes 1 and 9, corresponds here to the shortest distance 1-3-6-9 or, equivalently 9-6-3-1 (inversely) between them. Based on that, one could consider using known algorithms which calculate the k-shortest paths [21] for figuring out the second best path. In such a case, the second best route wouldn't be just one but many (e.g 1-5-3-6-9, 1-3-6-8-9, 1-3-11-6-9 and so forth), because all of them have the same length equal to 4. From that it is obvious that using such algorithms all these paths are treated as equally important. According now the the research paper of Dupont, et.al [18], the second best path in this particular problem is the one that pass through the nodes 1-3-11-6-9. That is because given a random walker starting from node

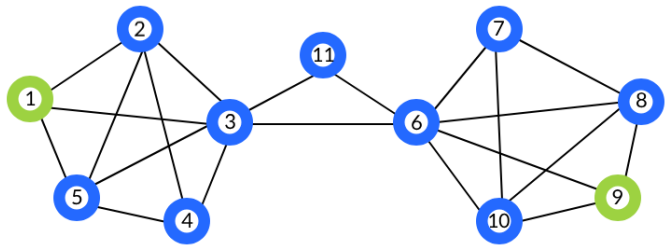

Figure 3.1: A graph describing all the possible routes between a set of nodes

1 and eventually reaching node 9 (or even the reverse path) is more likely to pass through node 11 than any other node as an alternative to the firs best route. In contrast, if a random walker will choose the 1-2 edge instead of 1-3 is less likely to choose 2-3 as the next edge, because there are many other options to leave from the node 2. The approach that Dupont and colleagues proposed is based on the expected number of times a given node or a given edge is visited along any random walk connecting the nodes of interest.

The resulting graph that is illustrated in Figure 3.2a the edge's width corresponds to the relative frequency of visit along the random walks. This relative frequency is translated in to the relevance of each edge to explain relations between nodes of interest. The rejection of any edge whose relevance falls below a threshold defines a relevant subgraph. A stricter threshold would lead to smaller subgraph as shown in Figure 3.2b.

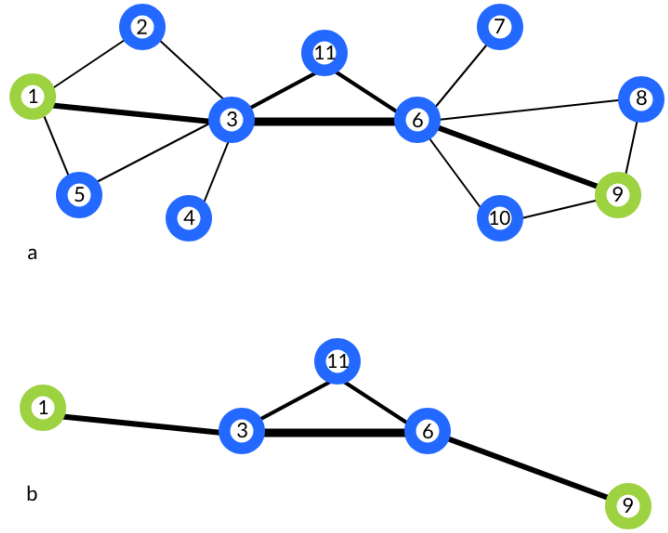

Figure 3.2: a. Relative edge relevance based on random walks between nodes 1 and 9. and b. same as a. but with a more strict relevance threshold.

The approach proposed previously is based on the interpretation of the graph as a Markov chain. The nodes of the original graph represents the states of such a model and the probabilities of transition are defined according to the weights of the edges. The theory of the Markov chain [33, 41] allows one to calculate the nodes and edges most frequently visited when performing random walking between any pair of separate nodes of interest. The walks are random, but the probability distribution over all possible walks is far from uniform. Therefore, the likelihood of any given walk is really important in calculating the relevance. These ideas are formally explained in the next section.

A relevant subgraph should in many cases be as small as possible, since it gathers the most information to explain the relationships between the nodes of interest. More clearly, one would need to make distinctions between important and less important edges or nodes based on their relevance. The limited k-walks approach presented here is not necessarily extreme discriminant, since relevance values are based on a large, possibly infinite, set of walks in the graph. Whether or not the extracted subgraph is small relative to the initial graph also depends on the graph topology, the initial edge weights and the chosen nodes of interest. In any case, one can impose more discrimination between the edges by following the inflating method described on [18].

We have talked so far on how to define a measure of relevance in the edges and nodes of a graph to better explain the relationships between k nodes that interest us. Keeping only the edges whose relevance score is above a defined threshold is a direct way to extract a subgraph. In the following section the mathematical formulation of these ideas are extensively analyzed.

3.2 Limited k-walks

The limited k-walks method is a general algorithm for constructing a more relevant subgraph linking the nodes of interest to a large graph which in our case it describes a metabolic network. The relevance of each edge is calculated as the expected number of visits of a random walker connecting nodes nodes of interest. These expected passage times reflect both the topology of the network and the weights of the edges. They occur from the interpretation of the graph as a Markov chain characterized by a transition probability matrix T.

Given a graph $G = (V, E)$ consisted by a set of V vertices (nodes) and a set of E edges with $E \subseteq V \times V$ the $n = |V|$ denotes the order of the graph. In particular since such a graph is considered weighted, it represented by its weighted adjacency $n \times n$ matrix A. Each a_{ij} entry on A denotes the weight of the edge that connect the node i to node j. Edges that does not belong to the graph are assigned with 0. In any other case the weight between any pair of connected nodes is assumed strictly as a positive number and should be defined in a way such that the larger its value the easier the connection/communication from i to j. Undirected graphs are expected to have a symmetric adjacency matrix A. The

diagonal degree matrix is defined as $D = diag(d_1, d_2, ..., d_n)$ where $d_i = \sum_{j=1}^{n} a_{ij}$ and is interpreted as the weighted degree of node i.

Given a weighted adjacency matrix A describing a graph $G = (V, E)$ and a subset $K \subseteq V$ of nodes of interest with $|K| \geq 2$ with $k = |K|$ number of nodes of interest much smaller that n; an edge relevance function is defined $er(A, K)$ which maps any edge to its relevance index $E \to \mathbb{R}^+$. As noted in the previous section relevance indices should based on all the possible ways that connect the k nodes of interest by accounting the different likelihoods of each way and not on shortest distance methods. Technically, this is achieved by defining the relevance of the edge to be proportional to the expected passage times that a random walker used this edge along a certain walk starting from one node of interest and reaching a distinct node of interest. In order this function to be able to calculate these quantities, the theory of absorbing Markov chains is being used [33].

A way to model a random walk in a graph is by using a Markov chain describing the sequence of nodes visited during the walk. A state of the Markov chain is associated with each node of the graph. In that way, each state of of the Markov chain at time t is described with the random variable $X(t)$ while the probability of transiting to state j at time $t + 1$, given that the current state is i at time t, is defined as:

$$P[X(t+1) = j | X(t) = i] = p_{ij} = \frac{a_{ij}}{d_i} \qquad (1)$$

This definition makes clear, that the probability of any jump from the state i to state j is proportional to the weight a_{ij} of the edge from i to j. Thus, the whole transtion matrix $P = [p_{ij}]$ of the Markov chain is related to the degree D and adjacency A matrices as $P = D^{-1}A$. By its construction P is a row-stochastic matrix with $0 \leq p_{ij} \leq 1$ and $\sum_{j=1}^{n} p_{ij} = 1$. Knowing that the random walks are start from one node of interest x and end up in any other node of interest $K \backslash \{x\}$ is getting easy to define the absorbing states as described in equation (2) below. A state in a Markov chain, is called absorbing if and only if any walk reaching that state will stay forever on this, with probability 1.

$${}^{x}P_{ij} = \begin{cases} 1 & \text{if } i \in K \backslash \{x\} \text{ and } i = j, \\ 0 & \text{if } i \in K \backslash \{x\} \text{ and } i \neq j, \\ P_{ij} & \text{otherwise} \end{cases} \qquad (2)$$

Where the ${}^{x}P$ denotes a modified transition matrix, for which all nodes of interest except x (the starting one) have been turned into absorbing states. The rest of the states including x are kept as transient states, from which there is a positive probability to leave these nodes. In the newly transformed ${}^{x}P$ and without loosing any information for the whole graph, its states can be reordered such that ${}^{x}P$ receive the following block structure described in equation (3).

$${}^{x}P = \begin{bmatrix} {}^{x}Q & {}^{x}R \\ 0 & I \end{bmatrix} \qquad (3)$$

Where ${}^{x}Q$ denotes the $(n - k + 1) \times (n - k + 1)$ sub-matrix between transient states, I

denotes the so called identity matrix of $(k-1)\times(k-1)$ order containing the absorbing states and finally the ${}^{x}R$ is the $(n-k+1)\times(k-1)$ matrix which express the probabilities of the walks being in transient state to be absorbed. More specific the notation ${}^{x}R_{ir}$ denotes the probability of a walk to be absorbed in state r in the next step given that is currently in the transient state i. The ${}^{x}Q_{xi}$ express the probability of transiting from one transition state to another transition state in one step. According now to the Markov chain theory, the $({}^{x}Q)^{l}$ is the l^{th} power of the ${}^{x}Q$ matrix and thus the notation $[({}^{x}Q)^{l}]_{xi}$ represents the xi entry of the $({}^{x}Q)^{l}$ matrix and represents the probability of transiting from one transition state to another transition state in l steps.

As clarified in the title of the current section, core of the current research is the limited k-walks. That means that the relationships between the nodes of interest are not explained based on any length of walks on the graph but instead can be explained using walks of any length up to a pre-specified length L_{max}. Having set all of the the above, the main interest is the calculation of the conditional expectation $E[e(x,i,j)|L]$ that expresses the expected number of times the edge i,j is visited while starting the random walk from the x , given that the walk length is L. The way to calculate the this conditional probability can be expressed as shown in equation (4) by the fraction of two different conditional probabilities.

$$E[e(x,i,j)|L] = \sum_{l=0}^{L-1} \frac{P[X_l = i, X_{l+1} = j, L|X_0 = x]}{P[L|X_0 = x]} \qquad (4)$$

In this equation, the random variable X_l denotes the state visited at step l of a walk with total length L. The numerator $P[X_l = i, X_{l+1} = j, L|X_0 = x]$ denotes the joint probability of visiting the edge i,j between step l and $l+1$ of a walk with total length L given that the walk started from the x state. In the same way, the denominator $P[L|X_0 = x]$ expresses the probability of walk of length L given that the walk started from the x state.

In order to calculate the previously defined conditional probabilities, described in equation (4), the sub-matrices of the ${}^{x}P$, ${}^{x}Q$ and ${}^{x}R$ are going to be used. In particular the $P[L|X_0 = x]$ is given by

$$P[L|X_0 = x] = \sum_{r\in K\setminus\{x\}} [({}^{x}Q)^{L-1}({}^{x}R)]_{xr} \qquad (5)$$

because a walk of length L transits $L-1$ times through transient states of ${}^{x}Q$ and then it gets absorbed in any state $r \in K\setminus\{x\}$. In the same manner, the probability of visiting the edge i,j in a walk of length L can computed by the equation (6), in the case of j being a transient and not absorbing state.

$$P[X_l = i, X_{l+1} = j, L|X_0 = x] = \sum_{r\in K\setminus\{x\}} [({}^{x}Q)^{l}]_{xi}[{}^{x}Q]_{ij}[({}^{x}Q)^{L-l-2}({}^{x}R)]_{jr} \qquad (6)$$

This equation is explained, since such a walk reaches the transient state i in the l^{th} step

- $[({}^{x}Q)^{l}]_{xi}$ - , then transits in one step from state i to transient state j - $[{}^{x}Q]_{ij}$ -, after the state j trasits again $L-l-2$ times through transient states - $({}^{x}Q)^{L-l-2}$ - until its finally got absorbed in any state $r \in K \backslash \{x\}$ - ${}^{x}R$ -. Now taking into consideration the possibility of j state, be an absorbing state the equation (6) is transformed into

$$P[X_l = i, X_{l+1} = j, L | X_0 = x] = [({}^{x}Q)^{L-1}]_{xi}[({}^{x}R)]_{ij}, \forall j \in K \backslash \{x\} \tag{7}$$

because for such a walk to have a length equal to L before getting absorbed, the one last step should describe the absorption stage from i to j.

The equations defined so far, didn't seem to take into account a specified maximum walk length L_{max}. For that reason we define a new one that computes the limited edge passage times for a maximal walk length equal to L_{max} shown in equation (8).

$$E[e(x,i,j)|L \leq L_{max}] = \sum_{L=1}^{L_{max}} E[e(x,i,j)|L] \tag{8}$$

The last step is to calculate the relevance of each edge based on the limited edge passage times (equation (8)). Edge relevances also optionally depend on the definition of an initial probability distribution ι that contains weights describing the relative importance of each node of interest. Given that, the definition of a function $er(i,j)$ that calculates the edge relevance based on the limited k-walk algorithm described previously, is

$$er(i,j) = \begin{cases} \sum_{x \in K} \iota_x E[e(x,i,j)|L \leq L_{max}] & \text{if G is directed} \\ \sum_{x \in K} \iota_x |E[e(x,i,j)|L \leq L_{max}] - E[e(x,j,i)|L \leq L_{max}]| & \text{if G is undirected} \end{cases} \tag{9}$$

The algorithm of the limited k-walks on a graph, presented in this chapter, constitutes the core of the current research and consequently is the core of the expanet package that developed in order to propose a different approach for a pathway level analysis on biological data. The following chapter describes in details the idea behind the tool, how it is implemented and contains a brief documentation on its functions.

Chapter 4

Tool design

In this chapter we analyze in detail the workflow of the expanet package. Expanet, is a pathway-level analysis tool written mostly in R and partially in C languages, and can be used as a package in R environment. The Figure 4.1 bellow shows in generic the whole procedure. Form the data fetching and pre-processing to the exported list with the scores of each biological pathway, that represent the deregulation the pathway has been imposed due to the different experimental conditions. It is clear from the Figure 4.1 that expanet is consisted from two modules. One that is responsible for the execution of the limited k-walks algorithm and one that takes care of the meta-analysis of the relevance scores that the first module created in order to obtain the pathway-level score.

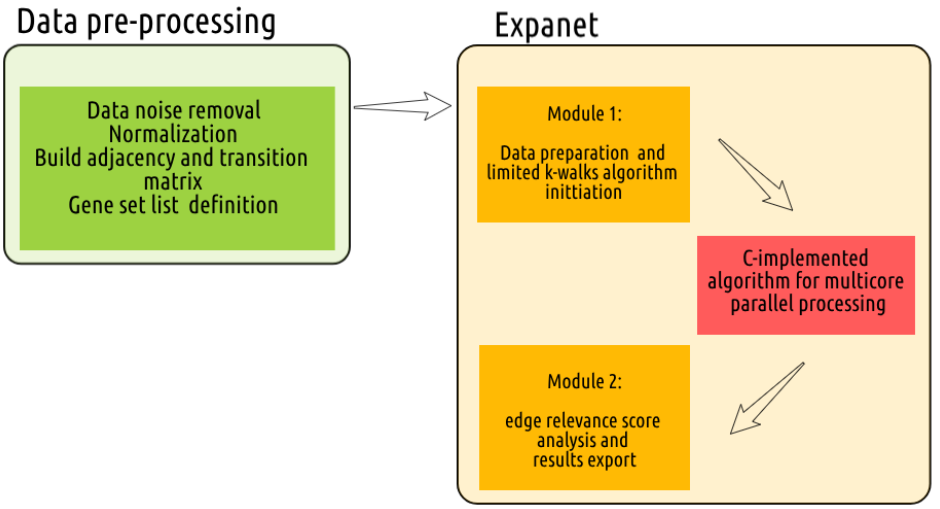

Figure 4.1: Expanet package general workflow

4.1 Data pre-processing

Expanet can take as input any kind of omic data derived either from microarrays, RNAseq and proteomics. The raw data of these platforms, before they get used to the following procedures of any analysis, a series of pre-processing methods should be applied. Such methods is the proper annotation using globally accepted gene names and IDs. Filter out genes/proteins that their expression levels are near the noise level, remove background noise and finally normalize data. These methods are the result of very important and constantly evolving statistical and mathematical models. A plethora of books, articles and free tools have been published and keep publishing [8, 19, 32, 56, 58], dealing with this first-step crucial subject of data pre-processing. For the rest of the discussion we assume that this first step has been completed.

4.1.1 Weighted gene-gene interaction network construction

As expected, input to the program should be a transition matrix which derived from an adjacency matrix describing a biological network. In the case of expanet this initial adjacency matrix is built using two different information sources. The first one is the known and predicted Protein-Protein interaction networks and the second is the gene-coexpression network (GCN). The procedure is shown in Figure 4.2 and explained below.

The PPI network is a useful framework for demystifying the functional organization of the proteome. However, the information to build these networks is obtained under different experimental designs and conditions and often different kind of algorithms are used to extract/predict it [25]. In addition, the interaction between proteins varies in different cells or tissues. Thus, PPIs can not accurately describe interactions between proteins in specific conditions just by themselves and its a good tactic one to use PPI databases that are based on similar cell types and pathology.

The gene co-expression network describes an undirected graph where nodes correspond to genes and nodes are connected with each other through an edge if there is a significant co-expression relationship between them [50]. High-throughput gene expression analysis methods are the best candidates to gather enough data for a number of genes for several samples or experimental conditions and use them to construct a gene co-expression network for those genes that show a similar expression pattern across the samples. In this study, the weight of each pair of genes is calculated using the function *bicorr()* of the R package **WGCNA** [34] that calculates a robust Pearson's correlation coefficient. These coefficients correspond to a possible interaction between the genes while the absolute values indicates the intensity of one gene being related to its co-expressed gene. However, as noted these gene-gene interactions are possible interactions and such a gene co-expression network does not guarantee the existence of them. Instead, it only suggests that there may be an interaction between the proteins.

In order to accurately extract information for the changes between gene-gene interactions

for several samples or experimental conditions, we constructed weighted gene-gene interaction networks for control and condition by combining the information from PPIs and GCN (see Figure 4.2).

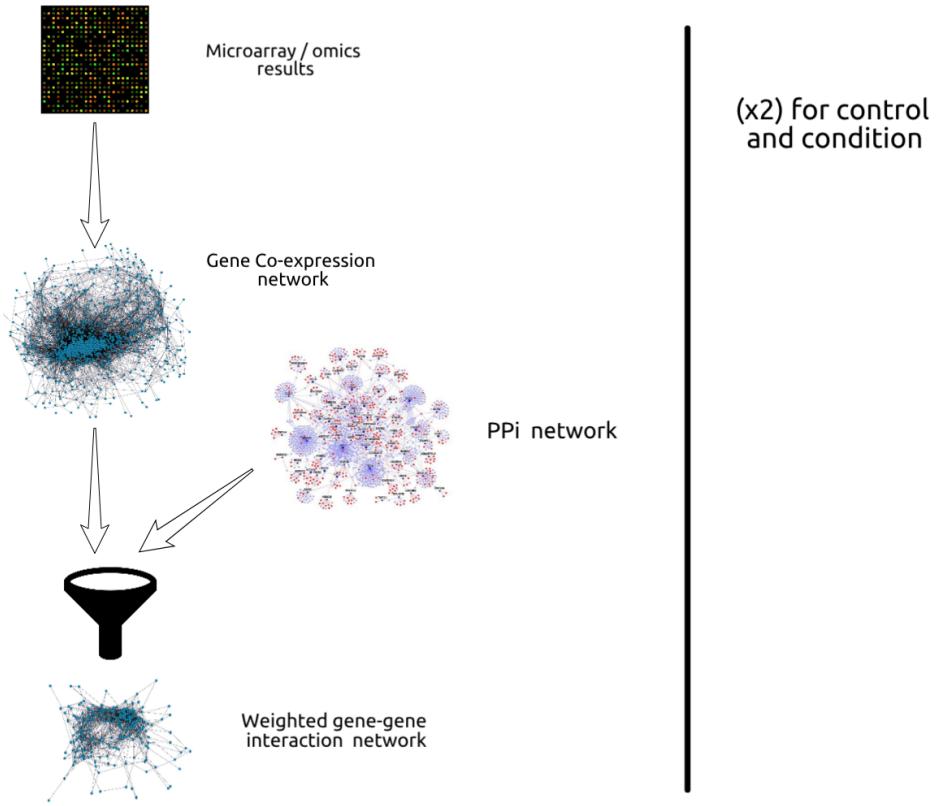

Figure 4.2: Weighted gene-gene interaction network construction workflow

The construction of the GCN as mentioned earlier, is based on the Pearson's correlation coefficients between gene expressions. This Gene Co-expression Network, can be interpreted mathematically as a graph $G = (V, E)$ with a set of vertices/nodes/genes V and a set of edges E as the weighted gene-gene interactions where $E \in V \times V$ and $n = |V|$ the number of genes. That G is derived from an $n \times n$ adjacency matrix A in which each a_{ij} denotes the weight of the connection between gene i and gene j. The a_{ij} values are computed according to the equation (10) that folows.

$$a_{ij} = \begin{cases} |cor(x_i, x_j)|^{\beta} & \text{Where } x_i \text{ and } x_j \text{ are the expressions of genes } i \text{ and } j \\ 0 & Otherwise \end{cases} \tag{10}$$

4.2 Expanet execution

4.2.1 Sub-network expansion

In the next step the chosen PPi network is mapped in the two weighted gene co-expression networks (control & condition) and only the a_{ij} values of gene interactions that explained from the PPi network are kept inside the A. All the rest are set to 0. After the adjacency matrix that describe the **weighted gene-gene interaction network** being ready, should be transformed into a *row-stochastic* transition matrix P. By defining the degree of each node i the $d_i = \sum_{j=1}^{n} a_{ij}$ where $i \in (1, 2, ..., n)$, each element of the transition matrix is calculated as $P_{ij} = \frac{a_{ij}}{di}$ which denotes the probability of transiting from node i to node j. That transition matrix describes the whole biological network of a sample/condition. In order to use this transition matrix according to theory with the algorithm of expanet, a list of different sets S_p where p is an index of each gene-set to be expanded and $|S_p| > 2$ should be given. Then for each such set of genes the transition matrix is transformed as discussed in the previous chapter such as

$$ {}^{x}P_{ij} = \begin{cases} 1 & \text{if } i \in S\backslash\{x\} \text{ and } i = j, \\ 0 & \text{if } i \in S\backslash\{x\} \text{ and } i \neq j, \\ P_{ij} & \text{otherwise} \end{cases} \qquad (11) $$

and it's fed into the pathway expansion algorithm. The algorithm is extensively described in chapter 3 and what it comes up with is a relevance score for each visited node calculated with the equation's (9) part for undirected graphs.

$$ er(i, j) = \sum_{x \in K} \iota_x |E[e(x, i, j)|L \leq L_{max}] - E[e(x, j, i)|L \leq L_{max}]| $$

Finally a threshold θ is applied to those scores. This threshold is chosen as the maximal relevance score that can lead to a connected subgraph. The edges that satisfy the applied threshold are kept and considered parts of the expanded network. The whole procedure is repeated for every experimental condition and is shown in the Figure 4.3.

Expanet package handles this procedure by calling the function *runExpanet (condition.label, condition.tm, results.dir, gsc, l.of.walk, verbose, n.cores)* where ***condition.label*** is the label of the current condition of class *character* and ***condition.tm*** is its transition matrix of class *matrix*, ***results.dir*** is the argument that accepts the output directory to store the results of the expansion of class *character*, ***gsc*** is a gene set collection of class *list*, ***l.of.walk*** is the maximal allowed walk representing the L_{max} and is an object of class *numeric*. Finally there is a boolean argument called ***verbose*** that enables or disables the verbose output of the program for debugging and a *numeric* argument where user defines the number of cores to be used called ***n.cores***.

Depending on the number of cores a user has specified, the above function spawns an external program written in C that implements the limited k-walks algorithm. Each process,

spawned in a different CPU thread/core allowing a more efficient way to calculate the expansion for more than one gene-set at the same time. At the end of that step, each pathway has been expanded twice. Once for the control and once for the condition (see Figure 4.3).

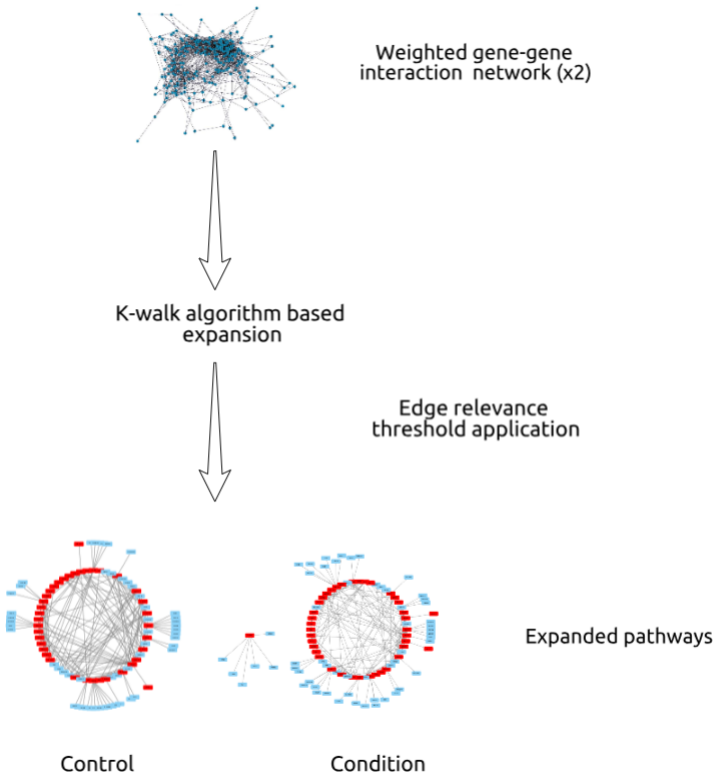

Figure 4.3: K-walk algorithm application for pathway expansion. Red nodes are the initial ones while the blue are the expanded.

4.2.2 Expanet-score calculation

Having calculated an expanded version of each gene set in S_p for the two different conditions, the identification of significantly altered pathways comes after. For this to happen, for each gene set, the two expanded networks (e.g Control_exp, Condition_exp) are merged into one by taking the union of their nodes (e.g Union_net). That network, then is mapped into the two adjacency matrices that describe the **weighted gene-gene interaction networks** for the two conditions (mentioned in 4.1.2). The result of this mapping is the re-

trieval of two edge weight vectors (e.g Control_w, Condition_w). In the next step the Pearson correlation coefficient between those vectors is calculated $cor(Control_w, Condition_w)$ and a final gene set score is obtained as $score_p = 1 - |cor(Control_w, Condition_w)|$ where p is the index of each gene-set. This score depicts the relevance between a particular gene set and the corresponding condition (disease, treatment, etc.).

For all gene sets' score calculation the expanet package, provides another user scope function called *analyzeExpanet(data.dir , control.label, treatment.label, control.adj, treatment.adj, gene.set, threshold, build.graphs, save.results)*. This function takes as input the following arguments. ***data.dir*** is the directory where the expanet saved the results from the first function (runExpanet()) of class *character*, ***control.label*** is the control's label and should be the same as the folder name in data.dir, of class *character*, ***treatment.label*** is the treatment's label and should be the same as the folder name in data.dir, of class *character*, ***control.adj*** is control's adjacency matrix of class *matrix*, ***treatment.adj*** is treatment's adjacency matrix of class *matrix* ***gene.set*** is a list with all gene sets, ***threshold*** is the edges' weight threshold of class *numeric* and if not provided algorithm calculates one by default as previously mentioned, ***build.graphs*** is a Boolean argument where if TRUE the function creates a directory with the expanded networks in .gml for further investigation with Cytoscape or other similar software. The last argument is the Boolean ***save.results*** which if is TRUE, the function stores the results into a csv file and not as an R environment variable.

Chapter 5

Validation

5.1 Analyzing biological data

In order to assess the expanet package on actual biological data, we conducted analyses on three different datasets. The first two were publicly available on GSE database studying Mantle Cell lymphoma (MCL), while the third one was derived from unpublished mass spectrometry data.

MCL is typically fast growing and generally need to be treated immediately. The disease's name arises from the fact that the tumor cells originally come from the "mantle zone" of the lymph node. This lymphoid neoplasia is derived from mature B cells. Their main genetic characteristic is the presence of t(11;14)(q13;q32) translocation that causes cyclin D1 overexpression. MCL is generally considered as aggressive lymphomas.

Using corresponding datasets of each of the mentioned studies, we applied two already widely used pathway analysis tools that belong respectively to the second and third generations. The GSEA[52] and SPIA [16]. Then we compared the results of those tools with the results of the expanet.The comparison with the two other tools was only possible for the first two datasets which derived from the GSE database. That because the third unpublished dataset described MS data, that are non compatible with the GSE and SPIA tools. Also as gene sets we used the KEGG's Databases pathways and for the Protein-Protein interaction filtering step we used the STRING Database.

5.1.1 Antitumoral activity of acadesine and rituximab in MCL (GSE47871)

Acadesine, a nucleoside analogue that has antitumoral effects, was used individually as well as in combination with Rituximab, an anti-CD20 monoclonal antibody, aiming the investigation of the possible synergistic action against MCL disease. After gene expres-

sion profiling analysis, researchers [37] concluded that acadesine had shown a crucial impact on the metabolic processes through the modulation of the immune response, actin cytoskeleton organization and metal binding. From the other hand rituximab shown also to have an effect on metal binding and immune responses.

5.1.2 Mantle Cell Lymphoma (GSE36000)

Recent studies have shown a subset of MCL with different behavior where following up molecular studies have identified genes such as SOX11 that belong to the SOX family, that can be used as discriminatory genes between these different lymphoma types. These genes encode for transcription factors that affect embryonic development and cell differentiation. Jara Paromero et.at [43] have investigated the gene and protein expression profiling of SOX11-positive and -knockdown MCL tumors, cell lines, and primary SOX11-positive and SOX11-negative MCL. The group identified PDGFA as a SOX11 direct target gene upregulated in MCL cells. The inhibition of the PDGFA pathway weakens angiogenic development both in vitro and in vivo but also MCL tumor growth in vivo, offering a promising novel therapeutic strategy for the treatment of aggressive MCL.

5.1.3 Mass Spectrometry dataset for Plasmodium using Nelfinavir and Ritonavir

Plasmodium is a well known parasite that can be transmitted by mosquitoes and causes Malaria disease. Although malaria is one of the most important diseases in the world, its role as an HIV coinfection has been poorly studied. Recent studies have shown that those two diseases have a number of interactions for example in HIV-1 progression and transmission. HIV is currently treated with a combination of drugs. One important class is HIV protease inhibitors (HIVPIs), that are protease inhibitors of an HIV protease. The protease plays an important role in creating mature virulent particles. Unpublished data from comparative proteomics analysis, using Nelfinavir and Ritonavir molecules, show that HIVPIS have differential effect parasites on protein level both quantitatively and qualitatively.

5.2 Analysis results

5.2.1 Score based

GSE47871 Dataset

Comparative results between the GSEA, SPIA and expanet tools for that specific dataset are shown in Figure 5.1.

What we conclude in thit case, is that expanet references much more pathways as differentiated between the conditions comparing to the other two platforms. This could also be explained as a result of the arbitrary way of setting a threshold to the expanet's final scores.

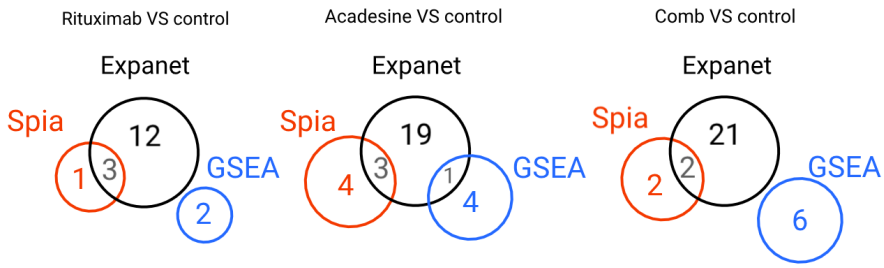

Figure 5.1: GSEA and SPIA comparison with Expanet GSE47871 Dataset

Besides that, by reviewing extensively the results, three unique, with a high expanet score and targeted enough pathways drew our attention. These pathways are the COMPLEMENT AND COAGULATION CASCADES with a score 0.9995, MAPK SIGNALING PATHWAY with a score 0.9743 and the NATURAL KILLER CELL MEDIATED CYTOTOXICITY with a score 0.9699. In contrast with the general enough results of the other tools, like PATHWAYS IN CANCER and PROTEOGLYCANS IN CANCER, these three pathways that expanet identified, are closely related with the mode of action mainly of the Rituximab and extensively described on the literature [47].

GSE36000 Dataset

Results of that dataset are still under investigation and thus, final conclusions cannot be made. Besides that, the overlap between expanet and the other tools is obviously bigger in this case according to the following Figure 5.2. In a first sight and without extensive study of the results, one unique and with high expanet score, pathway, called PRIMARY IMMUNODEFICIENCY (0.999568501) drew our attention due to its relevance with the underlying biological concept of the study.

MS Plasmodium Dataset

In this last dataset that we analyzed, we couldn't compare expanet results with the tools of GSEA and SPIA as previously mentioned and thus for the evaluation of the scores we turned into literature and co-leagues who are working on the Plasmodium for many years. After this kind of evaluation, it seems that two pathways which expanet shows with high scores, are highly related with the pathology of the malaria disease. Namely, these two pathways are the RIBOSOME BIOGENESIS IN EUKARYOTES (0.9978) and the PROTEASOME (0.9567).

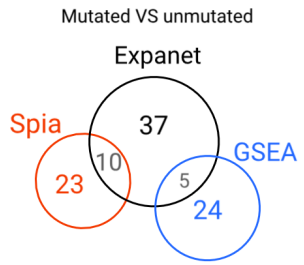

Figure 5.2: GSEA and SPIA comparison with Expanet for the GSE36000 Dataset

5.2.2 Graph based

As noted in the previous chapter, except from exporting the scores for each pathway, expanet, has the ability to store the expanded form of it into a Cytoscape file type. And maybe this is the most significant feature of that tool. Through these files, we can investigate the cross talk between the pathways, which was one of the basic reasons to develop that different kind of approach in pathway analysis. One such Cytoscape file when opened look like the following figure where the blue nodes represent the nodes of the initial pathway, the red nodes represent nodes that have been included through the expansion step and finally the edge thickness represents the edge relevance of each respective edge.

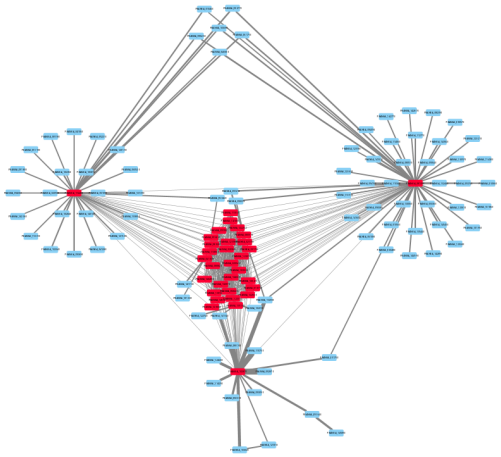

Figure 5.3: Example of Cytoscape extracted file from expanet

In order to gain insightful information from those graphs, we grouped the expanded nodes using a color code into the pathways that they actually belong to. After manual work on

the annotations, the same previously shown pathway looks like the Figure 5.3.

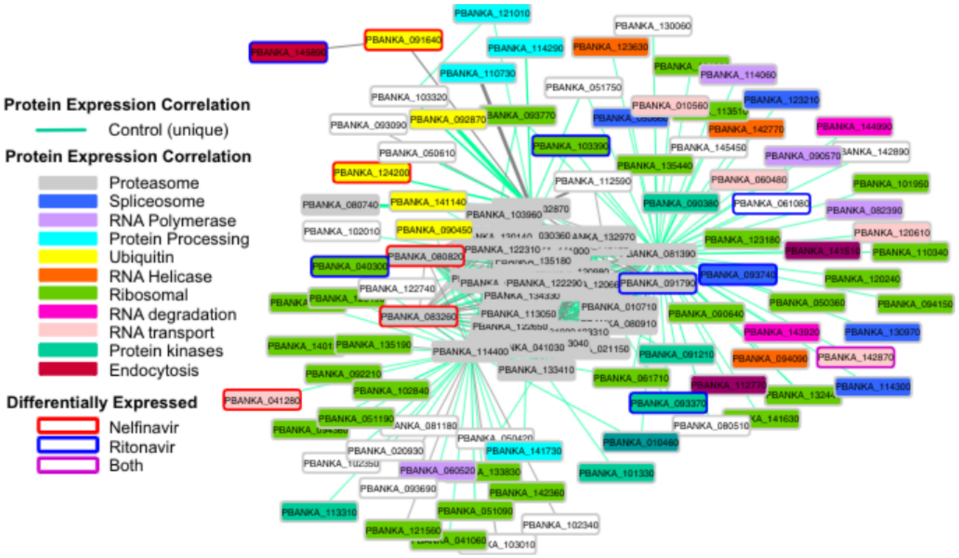

Figure 5.4: Example of Cytoscape extracted file after manual annotation

Since the information captured in those graphs is very condensed, we still investigate the underlying biological system through those graphs.

5.2.3 Limitations

In no case expanet can be considered as a panacea tool that solves all previous' generations issues and limitations. As a matter of human product, it also comes with its limitations. Some of these limitations and open questions have been pointed out and discussed in the context of this master thesis. By start naming them, one could argue that the expanet could be biased against the initial size of the pathway/gene-set and thus it could assign a bigger score to the bigger pathways. To answer this question we actually calculated the Pearson correlation between the initial gene-set size and its final expanet score for all the experiments that we conducted and the final average Pearson correlation was equal to -0.078. With such small value of the Pearson correlation coefficient, we couldn't support the idea that expanet is biased towards the initial size of the gene-set. By taking into account more than three datasets and calculating the same quantity we will be able to infer something different. Other open questions that need to be answered and can be possible limitations is the way that the edge relevance score threshold is calculated in order to conclude into the expanded network and if there is another possible way to calculate it. Also as previously mentioned in this chapter, until now, the threshold we applied on the final scores were chosen arbitrarily and not based on any statistical/mathematical method. One final limitation of the expanet's scores is that they don't give any insights on the

direction the studying pathway has changed e.g up/down-regulated. All these issue and more that will occur should be faced and deal with in order to make the expanet a more robust and rigid tool.

Chapter 6

Future Work

Aside from answering and providing better solutions for the limitations described in the previous section, there are two specific thoughts for future work related to further improvement of expanet.

The first one is to integrate topological information into the networks. This information will be translated into interdependence between the nodes and then as a final step the newly created network with the topological information integrated into it could be converted into a Bayesian network. The random walks then will occur on top of the Bayesian network by taking into account the interdependence between each pair of nodes. This will give the opportunity to treat each pathway not as a static network but as a dynamic one.

The second idea is to extract for each expanded pathway a signature, based on the expression levels of its nodes. This signature can be used against databases that store signatures from small molecules and/or medicines. By conducting this kind of mapping on those databases, one could possibly find out molecules that triggers the exact or the opposite of the given signature. Those results can then be used for a drug re-purposing/re-positioning studies.

Bibliography

[02] *Principal Component Analysis*. Springer-Verlag, 2002. DOI: 10.1007/b98835. URL: https://doi.org/10.1007/b98835.

[Ace+14] Serena Aceto et al. "The Analysis of the Inflorescence miRNome of the Orchid Orchis italica Reveals a DEF-Like MADS-Box Gene as a New miRNA Target". In: *PLoS ONE* 9.5 (May 2014). Ed. by M. Lucrecia Alvarez, e97839. DOI: 10.1371/journal.pone.0097839. URL: https://doi.org/10.1371/journal.pone.0097839.

[AS09] Marit Ackermann and Korbinian Strimmer. "A general modular framework for gene set enrichment analysis". In: *BMC Bioinformatics* 10.1 (2009), p. 47. DOI: 10.1186/1471-2105-10-47. URL: https://doi.org/10.1186/1471-2105-10-47.

[Ben+03] Yoshua Bengio et al. "Out-of-sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering". In: NIPS'03 (2003), pp. 177–184. URL: http://dl.acm.org/citation.cfm?id=2981345.2981368.

[Ben+04] Yoshua Bengio et al. "Learning Eigenfunctions Links Spectral Embedding and Kernel PCA". In: *Neural Computation* 16.10 (Oct. 2004), pp. 2197–2219. DOI: 10.1162/0899766041732396. URL: https://doi.org/10.1162/0899766041732396.

[Bey+18] samira Beyramysoltan et al. "Direct Analysis in Real Time-Mass Spectrometry & Kohonen Artificial Neural Networks for the Rapid Species Identification of Larvae, Pupae and Adult Life Stages of Carrion Insects". In: *Analytical Chemistry* (June 2018). DOI: 10.1021/acs.analchem.8b01704. URL: https://doi.org/10.1021/acs.analchem.8b01704.

[BH07] Author Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing". In: (2007).

[Bol] Benjamin M. Bolstad. “Pre-Processing DNA Microarray Data”. In: *Fundamentals of Data Mining in Genomics and Proteomics*. Springer US, pp. 51–78. DOI: 10.1007/978-0-387-47509-7_3. URL: https://doi.org/10.1007/978-0-387-47509-7_3.

[BY01] Yoav Benjamini and Daniel Yekutieli. “The Control of the False Discovery Rate in Multiple Testing under Dependency”. In: *The Annals of Statistics* 29.4 (2001), pp. 1165–1188. ISSN: 00905364. URL: http://www.jstor.org/stable/2674075.

[Con04] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome”. In: *Nature* 431.7011 (Oct. 2004), pp. 931–945. DOI: 10.1038/nature03001. URL: https://doi.org/10.1038/nature03001.

[Cse02] M. E. Csete. “Reverse Engineering of Biological Complexity”. In: *Science* 295.5560 (Mar. 2002), pp. 1664–1669. DOI: 10.1126/science.1069981. URL: https://doi.org/10.1126/science.1069981.

[DAr+16a] Valeria D’Argenio et al. “Metagenomics Reveals Dysbiosis and a Potentially Pathogenic N. flavescens Strain in Duodenum of Adult Celiac Patients”. In: *The American Journal of Gastroenterology* 111.6 (Apr. 2016), pp. 879–890. DOI: 10.1038/ajg.2016.95. URL: https://doi.org/10.1038/ajg.2016.95.

[DAr+16b] Valeria D’Argenio et al. “The complete 12 Mb genome and transcriptome of Nonomuraea gerenzanensis with new insights into its duplicated “magic” RNA polymerase”. In: *Scientific Reports* 6.1 (Dec. 2016). DOI: 10.1038/s41598-016-0025-0. URL: https://doi.org/10.1038/s41598-016-0025-0.

[DAr+17] Valeria D’Argenio et al. “The Cause of Death of a Child in the 18th Century Solved by Bone Microbiome Typing Using Laser Microdissection and Next Generation Sequencing”. In: *International Journal of Molecular Sciences* 18.1 (Jan. 2017), p. 109. DOI: 10.3390/ijms18010109. URL: https://doi.org/10.3390/ijms18010109.

[Dha05] Patrik D’haeseleer. “How does gene expression clustering work?” In: *Nature Biotechnology* 23.12 (Dec. 2005), pp. 1499–1501. DOI: 10.1038/nbt1205-1499. URL: https://doi.org/10.1038/nbt1205-1499.

[Dra+07] S. Draghici et al. “A systems biology approach for pathway level analysis”. In: *Genome Research* 17.10 (Sept. 2007), pp. 1537–1545. DOI: 10.1101/gr.6202607. URL: https://doi.org/10.1101/gr.6202607.

[DS07] Alain Dupuy and Richard M. Simon. “Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting”. In: *JNCI: Journal of the National Cancer Institute* 99.2 (Jan. 2007), pp. 147–157. DOI: 10.1093/jnci/djk018. URL: https://doi.org/10.1093/jnci/djk018.

[Dup+07] Pierre Dupont et al. “Relevant subgraph extraction from random walks in a graph”. In: (June 2007). URL: http://becool.info.ucl.ac.be/pub/papers/rr2006-07_walks.pdf.

[Dur08] Steffen Durinck. “Pre-Processing of Microarray Data and Analysis of Differential Expression”. In: (2008), pp. 89–110. DOI: 10.1007/978-1-60327-159-2_4. URL: https://doi.org/10.1007/978-1-60327-159-2_4.

[Eis+98] Michael B. Eisen et al. “Cluster analysis and display of genome-wide expression patterns”. In: *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14863–14868. ISSN: 0027-8424. eprint: http://www.pnas.org/content/95/25/14863.full.pdf. URL: http://www.pnas.org/content/95/25/14863.

[Epp] D. Eppstein. “Finding the k shortest paths”. In: (). DOI: 10.1109/sfcs.1994.365697. URL: https://doi.org/10.1109/sfcs.1994.365697.

[FR99] C. Fraley and A. E. Raftery. “MCLUST: Software for Model-based Cluster Analysis”. In: *Journal of Classification* (1999), pp. 297–306.

[GB07] J. J. Goeman and P. Buhlmann. “Analyzing gene expression data in terms of gene sets: methodological issues”. In: *Bioinformatics* 23.8 (Feb. 2007), pp. 980–987. DOI: 10.1093/bioinformatics/btm051. URL: https://doi.org/10.1093/bioinformatics/btm051.

[Gen+04] Robert C Gentleman et al. In: *Genome Biology* 5.10 (2004), R80. DOI: 10.1186/gb-2004-5-10-r80. URL: https://doi.org/10.1186/gb-2004-5-10-r80.

[Hak+08] Luke Hakes et al. “Protein-protein interaction networks and biology—what’s the connection?” In: *Nature Biotechnology* 26.1 (Jan. 2008), pp. 69–72. DOI: 10.1038/nbt0108-69. URL: https://doi.org/10.1038/nbt0108-69.

[Hay14] Erika Check Hayden. “Technology: The $1,000 genome$”. In: *Nature* 507.7492 (Mar. 2014), pp. 294–295. DOI: 10.1038/507294a. URL: https://doi.org/10.1038/507294a.

[HKE09] Buhm Han, Hyun Min Kang, and Eleazar Eskin. “Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers”. In: *PLoS Genetics* 5.4 (Apr. 2009). Ed. by John D. Storey, e1000456. DOI: 10.1371/journal.pgen.1000456. URL: https://doi.org/10.1371/journal.pgen.1000456.

[Hor+17] Makiko Horai et al. “Detection of de novo single nucleotide variants in offspring of atomic-bomb survivors close to the hypocenter by whole-genome sequencing”. In: *Journal of Human Genetics* 63.3 (Dec. 2017), pp. 357–363. DOI: 10.1038/s10038-017-0392-9. URL: https://doi.org/10.1038/s10038-017-0392-9.

[HW79] J. A. Hartigan and M. A. Wong. “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Applied Statistics* 28.1 (1979), p. 100. DOI: 10.2307/2346830. URL: https://doi.org/10.2307/2346830.

[Jac+17] Minnie Jacob et al. “Metabolomics toward personalized medicine”. In: *Mass Spectrometry Reviews* (Oct. 2017). DOI: 10.1002/mas.21548. URL: https://doi.org/10.1002/mas.21548.

[KF17] Pranav Kulkarni and Peter Frommolt. “Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows”. In: *Computational and Structural Biotechnology Journal* 15 (2017), pp. 471–477. DOI: 10.1016/j.csbj.2017.10.001. URL: https://doi.org/10.1016/j.csbj.2017.10.001.

[Kos+11] A. Koschmieder et al. “Tools for managing and analyzing microarray data”. In: *Briefings in Bioinformatics* 13.1 (Mar. 2011), pp. 46–60. DOI: 10.1093/bib/bbr010. URL: https://doi.org/10.1093/bib/bbr010.

[KS60] J.G. Kemény and J.L. Snell. *Finite markov chains*. University series in undergraduate mathematics. Van Nostrand, 1960. URL: https://books.google.gr/books?id=WORLAAAAMAAJ.

[LH08] Peter Langfelder and Steve Horvath. “WGCNA: an R package for weighted correlation network analysis”. In: *BMC Bioinformatics* 9.1 (2008), p. 559. DOI: 10.1186/1471-2105-9-559. URL: https://doi.org/10.1186/1471-2105-9-559.

[Lux07] Ulrike von Luxburg. “A tutorial on spectral clustering”. In: *Statistics and Computing* 17.4 (Aug. 2007), pp. 395–416. DOI: 10.1007/s11222-007-9033-z. URL: https://doi.org/10.1007/s11222-007-9033-z.

[Mar08] Elaine R. Mardis. “The impact of next-generation sequencing technology on genetics”. In: *Trends in Genetics* 24.3 (Mar. 2008), pp. 133–141. DOI: 10.1016/j.tig.2007.12.007. URL: https://doi.org/10.1016/j.tig.2007.12.007.

[Mon+14] Arnau Montraveta et al. “Synergistic anti-tumor activity of acadesine (AICAR) in combination with the anti-CD20 monoclonal antibody rituximab in iin vivo/i and iin vitro/i models of mantle cell lymphoma”. In: *Oncotarget* 5.3 (Jan. 2014). DOI: 10.18632/oncotarget.1455. URL: https://doi.org/10.18632/oncotarget.1455.

[Mon03] Stefano Monti. In: *Machine Learning* 52.1/2 (2003), pp. 91–118. DOI: 10.1023/a:1023949509487. URL: https://doi.org/10.1023/a:1023949509487.

[Mon05] S. Monti. “Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response”. In: *Blood* 105.5 (Mar. 2005), pp. 1851–1861. DOI: 10.1182/blood-2004-07-2947. URL: https://doi.org/10.1182/blood-2004-07-2947.

[NJW01] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. “On Spectral Clustering: Analysis and an algorithm”. In: (2001), pp. 849–856.

[Nor98] J.R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN: 9780521633963. URL: https://books.google.gr/books?id=qM65VRmOJZAC.

[Nun+17] Marcella Nunziato et al. “Fast Detection of a BRCA2 Large Genomic Duplication by Next Generation Sequencing as a Single Procedure: A Case Report”. In: *International Journal of Molecular Sciences* 18.11 (Nov. 2017), p. 2487. DOI: 10.3390/ijms18112487. URL: https://doi.org/10.3390/ijms18112487.

[Pal+14] J. Palomero et al. “SOX11 promotes tumor angiogenesis through transcriptional regulation of PDGFA in mantle cell lymphoma”. In: *Blood* 124.14 (Aug. 2014), pp. 2235–2247. DOI: 10.1182/blood-2014-04-569566. URL: https://doi.org/10.1182/blood-2014-04-569566.

[Pan+17] Ioannis Panagopoulos et al. “Fusion of the genes ataxin 2 like, iATXN2L/i, and Janus kinase 2, iJAK2/i, in cutaneous CD4 positive T-cell lymphoma”. In: *Oncotarget* 8.61 (Oct. 2017). DOI: 10.18632/oncotarget.21790. URL: https://doi.org/10.18632/oncotarget.21790.

[Pu+17] Weilin Pu et al. “Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC)”. In: *Clinical Epigenetics* 9.1 (Dec. 2017). DOI: 10.1186/s13148-017-0430-7. URL: https://doi.org/10.1186/s13148-017-0430-7.

[Roy+18] Somak Roy et al. “Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines”. In: *The Journal of Molecular Diagnostics* 20.1 (Jan. 2018), pp. 4–27. DOI: 10.1016/j.jmoldx.2017.11.003. URL: https://doi.org/10.1016/j.jmoldx.2017.11.003.

[Sey+16] Narges Seyfizadeh et al. “A molecular perspective on rituximab: A monoclonal antibody for B cell non Hodgkin lymphoma and other affections”. In: *Critical Reviews in Oncology/Hematology* 97 (Jan. 2016), pp. 275–290. DOI: 10.1016/j.critrevonc.2015.09.001. URL: https://doi.org/10.1016/j.critrevonc.2015.09.001.

[Smy] G. K. Smyth. “limma: Linear Models for Microarray Data”. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer-Verlag, pp. 397–420. DOI: 10.1007/0-387-29362-0_23. URL: https://doi.org/10.1007/0-387-29362-0_23.

[SQE18] Charanjit Sandhu, Alia Qureshi, and Andrew Emili. “Panomics for Precision Medicine”. In: *Trends in Molecular Medicine* 24.1 (Jan. 2018), pp. 85–101. DOI: 10.1016/j.molmed.2017.11.001. URL: https://doi.org/10.1016/j.molmed.2017.11.001.

[Stu03] J. M. Stuart. “A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules”. In: *Science* 302.5643 (Oct. 2003), pp. 249–255. DOI: 10.1126/science.1087447. URL: https://doi.org/10.1126/science.1087447.

[Su+17] Yu-Ting Su et al. “Novel Targeting of Transcription and Metabolism in Glioblastoma”. In: *Clinical Cancer Research* 24.5 (Dec. 2017), pp. 1124–1137. DOI: 10.1158/1078-0432.ccr-17-2032. URL: https://doi.org/10.1158/1078-0432.ccr-17-2032.

[Sub+05] A. Subramanian et al. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proceedings of the National Academy of Sciences* 102.43 (Sept. 2005), pp. 15545–15550. DOI: 10.1073/pnas.0506580102. URL: https://doi.org/10.1073/pnas.0506580102.

[Tad+18] Hayato Tada et al. “Prominent Tendon Xanthomas and Abdominal Aortic Aneurysm Associated with Cerebrotendinous Xanthomatosis Identified Using Whole Exome Sequencing”. In: *Internal Medicine* 57.8 (2018), pp. 1119–1122. DOI: 10.2169/internalmedicine.9687-17. URL: https://doi.org/10.2169/internalmedicine.9687-17.

[Tar+08] Adi Laurentiu Tarca et al. “A novel signaling pathway impact analysis”. In: *Bioinformatics* 25.1 (Nov. 2008), pp. 75–82. DOI: 10.1093/bioinformatics/btn577. URL: https://doi.org/10.1093/bioinformatics/btn577.

[TWH01] Robert Tibshirani, Guenther Walther, and Trevor Hastie. “Estimating the number of clusters in a data set via the gap statistic”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (May 2001), pp. 411–423. DOI: 10.1111/1467-9868.00293. URL: https://doi.org/10.1111/1467-9868.00293.

[TWR16] Tsung-Heng Tsai, Minkun Wang, and Habtom W. Ressom. “Preprocessing and Analysis of LC-MS-Based Proteomic Data”. In: *Methods in Molecular Biology*. Springer New York, 2016, pp. 63–76. DOI: 10.1007/978-1-4939-3106-4_3. URL: https://doi.org/10.1007/978-1-4939-3106-4_3.

[Ven+01] J. Craig Venter et al. “The Sequence of the Human Genome.” In: *Science* (Feb. 2001).

[War+14] Charles D. Warden et al. “Detailed comparison of two popular variant calling packages for exome and targeted exon studies”. In: *PeerJ* 2 (Sept. 2014), e600. DOI: 10.7717/peerj.600. URL: https://doi.org/10.7717/peerj.600.