1-1-2016

Systems Biology Approaches For The Analysis Of High-Throughput Biological Data

Michele Donato
Wayne State University,

SYSTEMS BIOLOGY APPROACHES FOR THE ANALYSIS OF HIGH-THROUGHPUT BIOLOGICAL DATA

by

MICHELE DONATO

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2015

MAJOR: COMPUTER SCIENCE
(Bioinformatics)

Approved By:

_____
Advisor Date

_____

_____

_____

_____

DEDICATION

To

My parents, Armando and Annamaria, for their endless love and support, and to my entire family, who taught me how to be curious.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Sorin Drăghici, who taught me everything I know about research, from finding problems worth investigating, to finding solutions to those problems, and all the determination and attention to details that stays in between those two ends. Also, I am grateful to the unending patience he showed during my studies.

Many thanks to my committee members for their reviews and suggestions. I would like to thanks Dr. Michael Tainsky for the support he showed when I spent a period in his laboratory, and for the invaluable lessons he taught me about working in such an environment.

I would like to thank all the current and past members of the Intelligent Systems and Bioinformatics Laboratory for their encouragement and support. In particular, I am thankful to Dr. Călin Voichiţa, for the support he gave me through our collaboration and friendship, and who helped me being a better researcher and a better person. I would like to thanks the many collaborators I had the honor to work with during my graduate studies.

I would like to thank my family, without whom I would have not been able to achieve so much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

The final goal of a high-throughput biology experiment is the effective translation of large amounts of data into knowledge of biological phenomena. In the past two decades, since the introduction of the first full genome screening techniques, there has been a steady effort to bridge the gap between the constantly growing volume of *experimental data* and the ability of researchers to derive precise and accurate information from it. In many cases, the data to be analyzed comes from the comparison between two phenotypes. In such cases, the expression level of each gene is compared between these two phenotypes, and the data takes the form of a list of genes along with their measured *differential expression value*, i.e. the ratio between the expression values of each gene in the two phenotypes, and, in most cases, a *p-value* expressing the likelihood of obtaining such differential expression value, or a more extreme one, just by chance.

Recently, a new type of information has gained popularity: *signaling pathways* describe the complex signal transduction mechanisms in which genes and gene products are involved, and that carry out specific cell functions. Pathways are represented as *systems* composed by biological entities, and these systems are parts of a larger, more complex system (i.e. the organism). The information about signaling pathways is provided by a number of repositories that gather information from many experiments whose aim is to discover how genes interact with each other to carry out biological processes.

When this information became available, pathway analysis methods were developed to use this new knowledge to interpret the deluge of biological data and to obtain insights into biological phenomena of interest. These methods are used for several purposes, from phenotype detection, e.g. when a pathway describes a particular condition, to mechanism discovery, when the biological processes underlying the condition in analysis are not known. This last aspect is particularly useful in the field of drug development, when accurate knowledge of the mechanisms of action of a certain disease increase the chances of finding the appropriate treatment, and pathway-specific therapy is emerging as a more effective alternative to

single-gene therapy.

Complex diseases like cancer, for example, are the result of a multitude of different biological processes happening at system level, and only a system level analysis can give a complete overview of the whole disease. A notable example is the PTEN-PI3K-AKT pathway, known to be a mediator of signaling phenomena in a number of cancer types, and different activity in various parts of the pathway are linked to different prognoses [56]. In particular, *anomalies in the activity of this pathway* can lead to metastasis and poor prognosis in several types of cancer [95], and testing for such anomalies can help to identify those patients that might need a more aggressive line of treatment. There are, however, some issues with the current approaches to pathway analysis. The base concept of systems biology is that genes are not independent entities, but interact with each other to carry out specific biological processes. These biological processes are not independent, but they interact with each other through signaling and through sharing of sub-processes. In the case of the PTEN-PI3K-AKT pathway this is a well known issue [19], i.e. the crosstalk of the PI3K pathway, due to the overlap among this pathway and pathways related to other diseases. For example, the PI3K gene itself, central to the pathway, belongs to 70 pathways, many of which are unrelated to cancer processes, such as *Non-alcoholic fatty liver disease* or *Amoebiasis*. The PTEN gene, which negatively regulates the PI3K cascade, belongs to 16 pathways, among which there is *Hepatitis B*. Finally, the AKT gene, just a step downstream the PI3K gene, is present in 65 pathways, including the *Insulin signaling pathway*.

Unfortunately, existing pathway analysis methods do not take into account such phenomenon, and they consider all the genes as acting independently in each pathway they belong to, analyzing pathways as independent entities separated from each other. These issues lead to two important limitations of pathway analysis approaches that this thesis aims to solve.

First, existing pathway analysis methods assume that all genes are equally important in the biological processes they are involved in. However if, for example, many genes of the

same family perform the same biological function, it is intuitive to think that if only one of such genes is not active, this inactivity would not have a dramatic effect on the biological function performed by the gene family, i.e. the *single gene* activity level is somewhat less important than the one of a gene that, for example, is the only gene upstream of a cascade of signaling events, without whose activity the given biological function is not carried out. In order to overcome this limitation of existing pathway analysis methods, we developed an evolutionary computation approach for **determining the different contribution that individual genes make** to the phenomenon of interest, and including this information in the analysis. This is the first available approach for the systematic estimation of gene weights to be used in the analysis of signaling pathways. In addition, this method solves another issue related to pathway analysis methods. Most methods need the user to set a number of parameters in order to perform the analysis. These parameters are often set based on trial and error on a small number of case studies, or, in some cases, simulated data. Instead, the framework developed here can be used to estimate any parameter of pathway analysis methods.

The second limitation of existing pathway analysis methods is related to how they deal with *the different roles that each gene has in the pathways it belongs to.* These pathways describe very different biological processes, often mutually exclusive, and in some cases processes that happen in different tissues. From the biological perspective it is intuitive to think that a gene, in a specific condition, will be involved with some biological process (pathway) more than with others. In a specific condition, for example, the PTEN gene in the above example will be involved more with some of the 16 pathways it belongs to, and less with others. Existing methods completely ignore this fact, and assume that all genes make the same contribution in all the pathways they belong to, and the consequence is that when such *crosstalk* effect is predominant, the results of pathway analysis methods are riddled with both false positives and false negatives. In the second part of this thesis we developed the first method able to overcome this limitation of pathway analysis methods, by i) **identifying**

such crosstalk effects, and ii) **correct for it**. This work is the first work that objectively analyzes the effects of pathway crosstalk on the results of pathway analysis methods. We show that the three major categories of pathway analysis methods are severely influenced by these effects, and that this phenomenon is related to the structure of the pathways. The correction of crosstalk effects leads to a more meaningful ranking among pathways in a specific condition, removing both false positives and false negatives due to crosstalk from the results. Lastly, the method is able to identify *novel functional modules* that can play an independent role, and have different functions, than the pathways they are located on, allowing a better understanding of individual experiment results, as well as allowing for a more refined definition of existing signaling pathways that is bound to a given phenotype.

CHAPTER 2 PATHWAY ANALYSIS

The purpose of pathway analysis methods is to translate the lists of genes and differential expression values coming from high-throughput experiments into knowledge of the biological phenomena underlying the phenotypes in analysis, in the context of the systems described by signaling pathways. Many methods achieve this result by identifying the signaling pathways that are mostly impacted in a given experimental condition, calculating a p-value that aims to quantify the statistical significance of the impact of each pathway in the experimental condition.

From this description of pathway analysis, it is easy to define the three aspects that constitute it: the input data, the analysis method, and the pathways.

The data is probably the simplest aspect: a common input for pathway analysis methods is a list of genes together with their measured activity, or *expression level* in two phenotypes, e.g. disease and normal. Some methods require genes to be labeled as *differentially expressed* (DE) if the expression levels are different between the two phenotypes. A *p-value* is assigned to each gene, assessing the statistical significance of the difference in expression values. The input to pathway analysis methods differs between methods. In its simplest form is just the list of genes and the DE/Non-DE label for each gene. Other methods accept as an input the measured difference, while other methods accept both measured difference and statistical significance.

From the description of pathway analysis provided above, it is clear that the pathways themselves represent a fundamental aspect contributing to the results of the analysis. For this reason, this chapter is divided in two parts. The first part presents the fundamental concepts that are at the base of the description and representation of signaling pathways, and the consequences that different representations have on existing analysis methods. The second part presents an overview of the available approaches for pathway analysis.

2.1 Curated databases of biological pathways

After the introduction of high-throughput technologies researchers started facing the key challenge in functional genomics: once data for tens of thousands of genes was retrieved, the focus shifts to the interpretation of such data. One of the first resources whose goal was to facilitate the interpretation of high-throughput data was the Gene Ontology (GO) [5]. The GO represents a collection of three ontologies that aim to describe how various terms, representing *biological processes, molecular functions*, and *cellular components* are related with each other. Terms in GO are organized in a hierarchical tree structure, and they are ordered from the most generic term to the most specific. Each ontology in GO takes its name from the most generic term in the tree. Genes and proteins are *annotated* with one or more terms, indicating that they are involved with that particular entity (e.g. involved in a particular biological process or molecular function, or they are found to be related to a certain cellular component). In the context of signaling pathway, the gene ontology (and more specifically, the Biological Process ontology) contains a particular term, the *signal transduction* term. Children of this term are terms describing signaling pathways, and genes associated with these terms are genes that are known to be involved in a specific signaling pathway. One first negative aspect of such basic description is that a signaling pathway is represented as a *set* of genes that, in some non-specified way, interact with each other to carry out the process described by the specific signaling pathway. For example, the genes associated with the signaling pathway *Intrinsic apoptotic signaling pathway* are known to be involved in apoptosis (programmed cell death) but no information is given about *how these genes interact in the process.*

More recently there have been several initiatives aimed to fill this lack of advanced information on signaling pathways. Pathway databases such as KEGG [60], BioCarta [12], DC-Atlas [20], and Reactome [26] describe signaling pathways not only as mere gene sets, but they include the knowledge about the interactions among the genes, with the type of information provided depends on the specific pathway database. Another difference among the

database is how pathways are described. For example, the KEGG database uses the KEGG Markup Language (KGML), a proprietary language based on XML, Reactome is based on the BioPAX language and the DC-Atlas database uses BCML [8], while other databases describe their pathways in the Simple Interaction Format (SIF), a language that is only able to describe if an interaction exists among two genes, without providing any other information such as direction or type of interaction. Although the language used in the description of pathways seems to be a trivial factor in the context of the analysis of signaling pathways, it plays an important role in determining the type of analysis that can be performed. For example, a pathway database describing pathways in SIF might allow only elementary types of analysis, while adding other types of information such as type of interaction, or direction, or weight of an interaction allows for more refined analyses. This is why the next sections, which describes three widely used pathway databases, places particular focus on the versatility of the description language towards its use in computational methods.

2.1.1 The Kyoto Encyclopedia of Genes and Genomes - KEGG

Since it was introduced in 1995, the Kyoto Encyclopedia of Genes and Genomes (KEGG) provides information on a variety of biological entities. KEGG is composed by 17 databases related to various biological entities such as genes, genomes, compounds, diseases, drug, pathways, etc. These databases are divided in four categories: Systems Information, Genomic Information, Chemical Information, and Health Information. The KEGG Pathway database contains information about molecular networks, representing systems present in cells and organisms. The knowledge on interaction comes from both manual and automated curation of experimental knowledge. The pathway database contains information about two types of pathways: metabolic pathways and signaling pathways. The difference between these two types of pathways is that in metabolic pathways nodes represent compound and enzymes, and edges represent biochemical reactions, while in signaling pathways nodes represent genes or gene products, and edges represent signals passed from one node to another. Signaling pathways in KEGG describe how genes and gene products interact together to carry out a

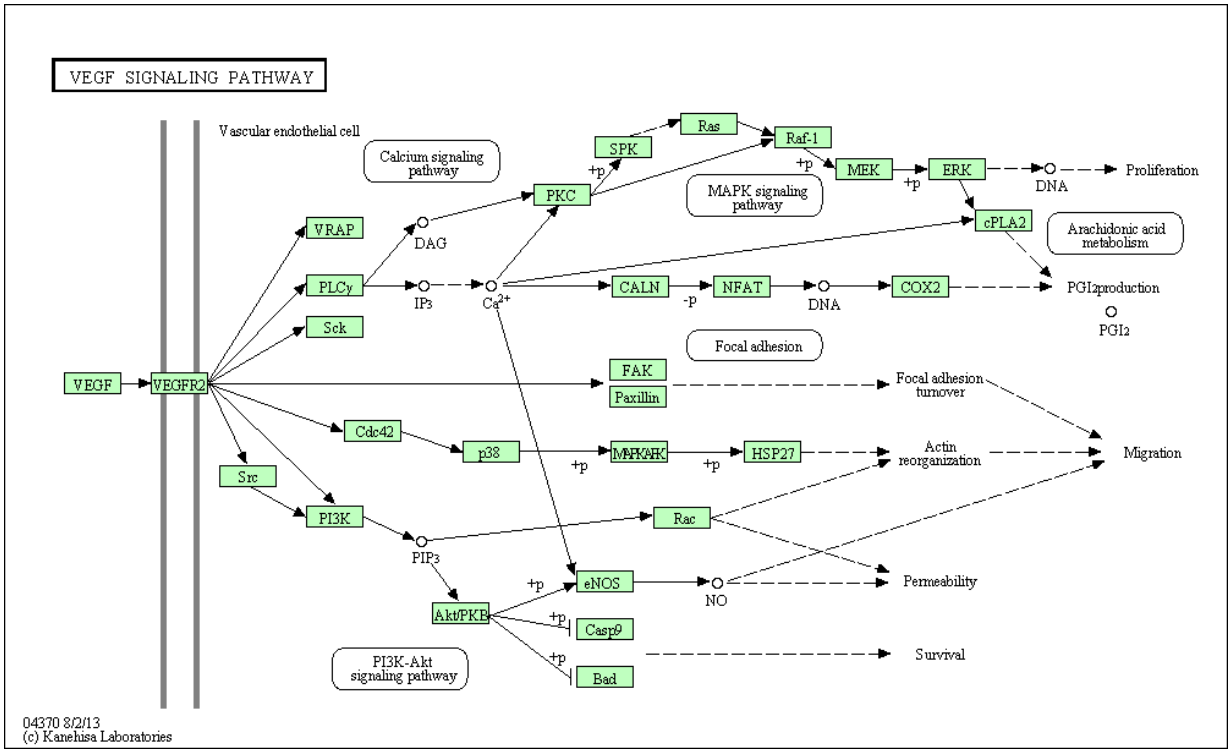


Figure 2.1: The VEGF signaling pathway in KEGG. This pathway describes the process through which vascular endothelial growth factor triggers events leading to proliferation and migration of endothelial cells. Genes (or gene families) and gene products are represented by the green nodes, while the interactions among them are represented by the arrows connecting the nodes. Different types of edges represent different types of reactions. For example, the two leftmost nodes in the pathway indicate that the VEGF protein (node with label VEGF) *activates* the VEGF receptor (node with label VEGFR2). Figure 2.2 describes the meaning of the various types of edges.

particular biological process. Figure 2.1 shows an example of a KEGG signaling pathway, the Vascular Endothelial Growth Factor (VEGF) signaling pathway. This pathway describes the process through which vascular endothelial growth factor triggers events leading to proliferation and migration of endothelial cells. Genes (or gene families) and gene products are represented by the green nodes, while the interactions among them are represented by the arrows connecting the nodes. Different types of edges represent different types of reactions. For example, the two leftmost nodes in the pathway indicate that the VEGF protein (node with label VEGF) *activates* the VEGF receptor (node with label VEGFR2). Figure 2.2 describes the meaning of the various types of edges.

For each process described by a pathway, KEGG defines a *reference pathway*, and each different organism in KEGG has its own specific pathway object derived from the reference

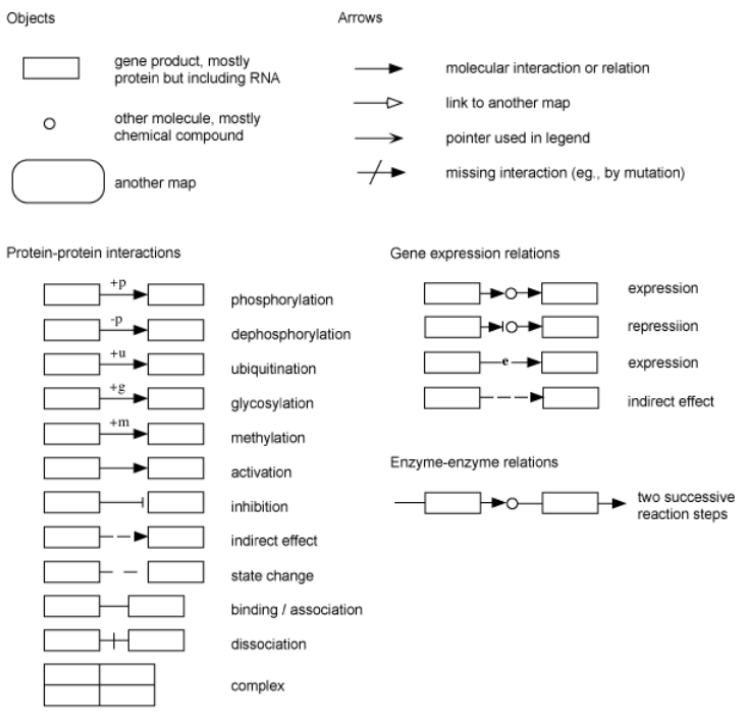


Figure 2.2: Various types of nodes and edges representing different entities and types of interactions in a KEGG pathway.

pathway. For example, from the reference *Apoptosis* pathway, a specific pathway for *Homo Sapiens* (human) is generated, then another for *Mus Musculus* (mouse), and so on. The main issue with this strategy is that, when a reference pathway is modified due to new knowledge, another manual process has to be performed to adapt all the organism-specific pathways. Retrieving pathway information from KEGG is possible through a number of options. The most convenient is through their REST server via the KEGGREST package from the Bioconductor repository for R [38, 94].

2.1.2 Reactome

Reactome is a free, open-source curated pathway database resulting from the collaboration from the Ontario Institute for Cancer Research, Cold Spring Harbor Laboratory, New York University Medical Center and the European Bioinformatics Institute.

The Reactome database collects information for more than 1,000 peer-reviewed, expert-curated pathways that include metabolic and signaling pathways. Pathways are represented, when appropriate, in a hierarchical fashion not dissimilar to GO. For example, the *Apoptosis*

group of pathways includes four interconnected pathways. One of them, *Intrinsic Pathway for Apoptosis*, is formed by other sub-pathways, and so on until we reach the specific reactions that act together to carry out that particular process.

Although the basic concept is the same as KEGG, i.e. making pathway maps available through a centralized system, the Reactome database presents several unique advantages over KEGG. While KEGG restricts some uses to paid customers only, the Reactome database is completely free. Furthermore, Reactome is dynamically connected to many external sources of information such as BioGRID [103], ChEMBL [122], or MINT [21], and the very same KEGG, de facto extending enormously the amount of information available. This flexibility allows Reactome to rely, with little effort, on the most advanced knowledge bases of biological information. Another strong point of Reactome is the visualization aspect. Reactome pathways graphs are automatically generated, and users can download them and visualize them with free visualization softwares (e.g. Cytoscape [98]).

Analysis of Reactome pathways is possible directly from the website. Users can upload their list of genes, proteins, or small molecules and perform over-representation analysis on the specified pathways. The last important aspect of the Reactome database is the fact that pathways can be exported in the BioPAX format. This is important since BioPAX is the most widely used format for the exchange of information on biochemical interactions (see Sec. 2.2 for details).

An example of a Reactome pathway can be seen in the left panel of Figure 2.3, while the right panel of Figure 2.3 shows the meaning of the graphical elements in a Reactome pathway.

2.1.3 DC-Atlas

DC-Atlas is a publicly available database that integrates the efforts of 32 European research groups providing information on pathways involved in dendritic cells functions. This approach departs from Reactome and KEGG approach. Whereas these two database provide a generic representation of a signaling pathway, disregarding differences in tissue type,

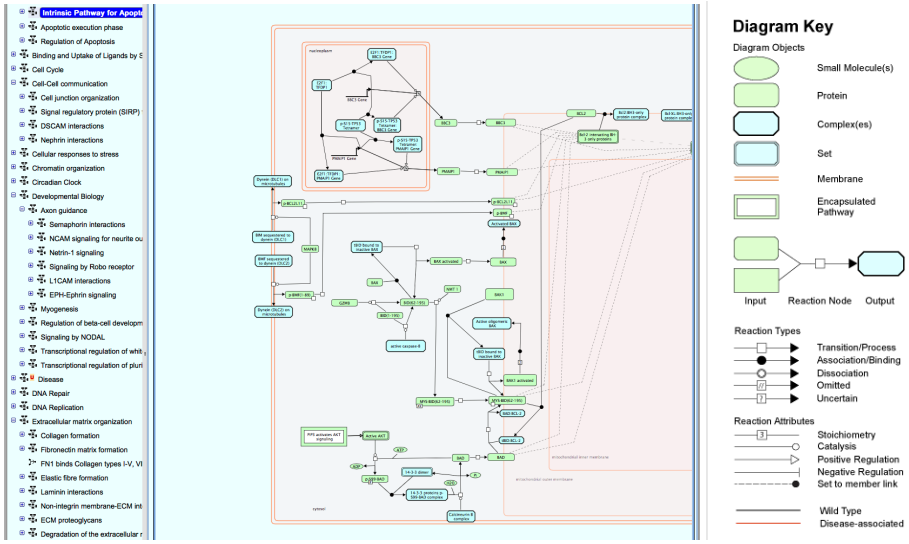


Figure 2.3: Left: detail of the *Intrinsic Pathway for Apoptosis* in the Reactome database. This panel shows an example of the visualization of a pathway in the Reactome database. Similarly to the visualization in KEGG, nodes represent gene products or other biochemical entities, and edges represent interactions among them. The Reactome on-line visualization tool is much more advanced than the one found on the KEGG website, as it allows for analysis and visualization of publications related to the pathway in analysis. Right: Reactome legend. Various types of nodes and edges representing different entities and types of interactions in a Reactome pathway.

DC-Atlas focuses on a specific type of cell, resulting in a very accurate description of each network. The rationale behind this approach is that biological processes in cell types can be dramatically different, and therefore analysis performed on general pathways will not be able to capture tissue specific effects, yielding sub-optimal results.

Dendritic cells are specialized cells of the immune system that are involved in many aspects of the immune response of an organism. Their involvement in a number of diseases such as HIV, many types of cancer and autoimmune diseases makes dendritic cells a key focus for the understanding of how such diseases work.

Pathways in DC-Atlas are annotated using the BCML format, one of the most advanced format for the description of biochemical pathways. It is important to note that pathways in the BCML atlas also follow the SBGN specification (described in Sec. 2.2).

One interesting aspect of the DC-Atlas database is the modular structure with which the pathways are described. In each pathway, cascades downstream a receptor are divided in three types of modules, each module interconnected sequentially with the others. These types

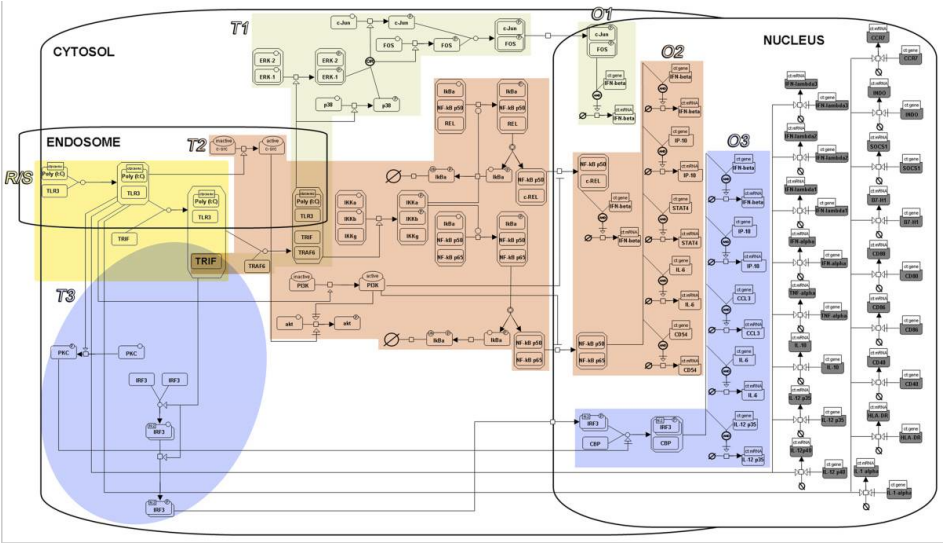


Figure 2.4: DC-Atlas: representation of the TLR3 pathway. The different color represent different signal cascades. This pathway presents one Receptor/Sensing module (R/S, yellow) spans the endosome and the cytosol. Three transduction modules (T1 - transduction 1 - light yellow, T2 - transduction 2 - orange, T3 - transduction 3 - light blue) connect to three outcome modules (O1- light yellow, O2 - orange, O3 - light blue). The T2 module is partly in the endosome. All the outcome modules are located in the nucleus.

of modules are the *receptor and sensing*, containing components of the pathway that interact with the stimulus and pass the appropriate signal to the second module, the *transduction module*. This module contains the components that carry the signal to the nucleus. The third module is the *outcome module*, beginning with a transcription factor and representing the final effect that the initiating signal has on the pathway. DC-Atlas pathways also specify the cellular region in which a reaction happens.

Figure 2.4 shows the representation of the *Toll Like Receptor 3* (TLR3) pathway in DC-Atlas. The different color represent different signal cascades. This pathway presents one Receptor/Sensing module (R/S, yellow) spans the endosome and the cytosol. Three transduction modules (T1- light yellow, T2 - orange, T3 - light blue) connect to three outcome modules (O1- light yellow, O2 - orange, O3 - light blue). The T2 module is partly in the endosome. All the outcome modules are located in the nucleus.

The retrieval of pathway information from DC-Atlas is possible through a suite of tools provided by the maintainer of the project. Due to the compatibility with SBGN, pathways retrieved from DC-Atlas benefit from numerous applications available for the visualization,

manipulation, and analysis of SBGN compliant networks (see `http://www.sbgn.org/SBGN_Software` for an up-to-date list of software).

2.2 Languages for the description of signaling pathways

2.2.1 KGML

In addition to the graphical representation, pathways in KEGG are described in the KEGG Markup Language (KGML). KGML is an XML-based format developed by KEGG. The images of pathways that are found on the KEGG website are not, however, obtained directly from the KGML file describing each pathway. This means that an inefficient and tedious manual process of translation from KGML to image is necessary, with potential discrepancies between them. In addition, this level of decoupling makes it necessary to manually update the images every time the KGML file is updated. Despite the versatility of the XML format from which KGML is derived, this usefulness of this format is limited to the simple description of the relationship among gene products, and any additional information has to be added out of the standard. For example, should a researcher desire to show how the genes in the pathway react to a certain experiment, by mapping gene expression levels on the pathway, she would have to manually add that capability in the schema.

2.2.2 BioPAX

The Biological Pathway Exchange language (BioPAX) [28] is the result of an international collaborative effort aimed to provide a standard for integration, description and visualization of biological pathway data. BioPAX relies on RDF/OWL, therefore it benefits from all the characteristics that make OWL an excellent choice for the representation of knowledge.

BioPAX versioning is described by "levels". The current version is BioPAX Level 3. Each level adds functionalities to the previous level. In addition, extensions are available for each particular level, adding features that are often in the process of being included in the successive level. Level 3 was the first level where signaling pathways were officially introduced.

BioPAX level 4, currently in development, will include support for Semantic Web tech-

nologies as an effort to facilitate seamless integration among knowledge bases.

The main advantage of BioPAX is the wide availability of tools for manipulation and visualization of pathways, as well as the wide adoption of the format. Pathway databases offering BioPAX export include Reactome, BioCyc [1], WikiPathways [61], Pathway Commons [22], and Panther [111, 79].

Many tools are available for visualization, manipulation, and analysis of BioPAX pathways, including Paxtools. Paxtools is a Java library developed by the same group behind BioPAX, ensuring high compatibility with the language, and guaranteeing efficient, complete, and consistent access to all the resources made available by BioPAX.

Although it is the most widely used format for describing biological pathways, BioPAX has an important limitation that stands in the fact that it is mainly oriented towards consultation of the information contained in pathways. The types of analysis that are available for BioPAX pathways are, to this date, only basic, since most of the groups involved in its development are focused on the biological aspects of the description of a pathway, and mostly oblivious to advanced methods for the analysis of signaling pathways. The other limitation of BioPAX is linked to the size of its governance group, size that results in a sort of *inertia* towards changes. A clear example is the planned introduction of compatibility with semantic web technologies: although the usefulness of linked data capabilities is clear, and although BioPAX is already described in RDF, talks regarding this change have started as far back as 2008, with little progress so far.

2.2.3 Graphical notation of pathways: the Systems Biology Graphical Notation

The Systems Biology Graphical Notation is a standard for the graphical representation of biochemical pathways. Similarly to BioPAX, it has been crafted by a community of researchers (often overlapping with the BioPAX community) over the course of several years.

SBGN is not formally a language for the *description* of biochemical pathways, as much

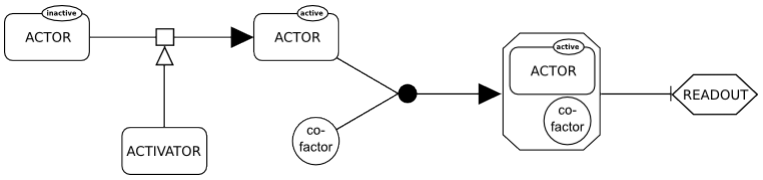


Figure 2.5: SBGN Process Description diagram for the example network. This language describes a biological network in terms of the temporal steps that reactions and interactions follow to carry out the desired function.

as a format for the *graphical representation* as the name states. However, it is possible to exploit the clarity of SBGN networks for pathway analysis purposes, and this is why it is included here.

SBGN is divided in three *languages* for describing different aspects of biological network: the Process Description language (PD) describes a biological network in terms of the temporal steps that reactions and interactions follow to carry out the desired function. This type of description can be used if the purpose of the analysis is to capture time dependent phenomena, such as in the case of time series gene expression experiments. Figure 2.5 shows an example of an interaction described with PD. This interaction is of type *activation*, in which the activator entity (e.g. a gene product) activates the "actor" entity. *After* the activation, the actor entity is presented as active. Then, the active actor binds with a co-factor, resulting in a combined entity. The Readout state represents a point in time where the result of the interaction can be screened.

The Entity Relationship (ER) language shows, as the name states, relationships among entities, regardless by the temporal aspects of such relationships. Figure 2.6 shows the same interaction shown in Figure 2.5, but this time described in ER. The arrows now represent the fact that the entities have a relationship among them.

Finally, the Activity Flow (AF) diagram represent the transduction of activity (signals) throughout the network. Figure 2.7 shows again the same reaction described above. In AF the arrow between the activator and the actor defines the type of interaction (in this example

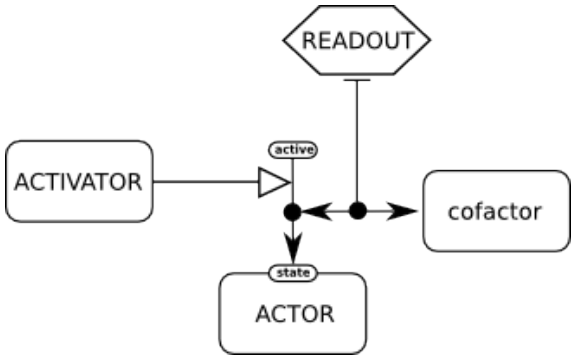


Figure 2.6: SBGN Entity Relationship diagram for the example network. The arrows now represent the fact that the entities have a relationship among them.

Figure 2.7: SBGN Activity Flow diagram for the example network. In this language the arrow between the activator and the actor defines the type of interaction (in this example activation).

activation). This diagram is more similar to the description provided by KEGG.

2.3 The Biological Connection Markup Language

The languages described so far share an important limitation: none of them was designed to be used with advanced methods for the analysis of signaling pathways. All these languages have the purpose of providing pathways that can be use as a reference, not unlike a static map that biologists can consult when in need to confirm results of some wet lab experiment, or to draft preliminary experimental design. However, the recent development of methods for the analysis of signaling pathways clashes against this shortcoming. This is why we developed the **Biological Connection Markup Language** (BCML) [8]. BCML is defined as a complete *framework* that allows all the aspects of signaling pathways, from the design, to the manipulation, the integration with external data sources, to the analysis.

The analysis aspect, in particular, is where BCML distinguishes itself from existing solutions. Rather than incorporating basic type of analyses, BCML allows the exporting of pathways in a number of format compatible with the most advanced methods for the analysis of signaling pathways, such as the SBGN Activity Flow diagram, or the Gene Matrix Text (GMT).

BCML is XML based, and in order to be SBGN compatible, it follows the complete SBGN specification. In addition to the basic features, many additional features can be added to BCML entities. In contrast, adding feature to any KEGG pathway results in a non-KGML compatible object.

In addition to graphical components and elements of SBGN notation, BCML users can embed additional information on the entities in the network, such as a set of identifiers for the entity (e.g. Entrez Gene IDs or Uniprot Accession Numbers). Different species can be linked to specific identifiers in the set. Probably the most innovative feature is the capability of associating *Findings*, i.e. *condition-specific information*, to each entity. *Findings* represent of biological information that has been found to be relevant to a specific entity, in a real world experiment. The first version of BCML supports the following types of information: organism, tissue, type of cell, biological environment in which the evidence was proven and the type of the specific type of experiment from which the evidence was gathered. Users must use a controlled vocabulary that has been developed from state-of-the-art medical ontologies. This guarantees consistency between findings that are submitted by different users.

Another function provided by the schema is the possibility of splitting pathways into independent units carrying out specific functions in a pathway. These units are called *macro modules*.

Lastly, BCML goes beyond its main focus, i.e. biological data and analysis, by providing support for advanced visualization of pathways. Users can personalize elements such as background, border, etc. The software suite that we developed is able to process these graphical elements during parsing, with the consequence that users can assign specific meanings to them.

It is important to note that the additions to SBGN are not mandatory: users can choose to take advantage of the added features, or to just follow the plain SBGN specification. Also, since BCML follows a layer structure, users can, at any time, obtain a fully SBGN compliant description of a pathway without the additional features.

2.3.1 BCML software suite

A software suite is available for BCML provides a series of tools that allow users to properly describe, manipulate, and visualize pathways written in BCML. BCML files can be validated through a validation tool, to make sure that the structure of the file follows the BCML and SBGN criteria. First, the file is checked against XML specifications, and after that for SBGN specifications, and elements that break any specification are reported. Additionally, elements with duplicated identifiers are reported, as well as elements that are disconnected from any other element in the network.

Graphical representation.

The BCML software suite allows for conversion of BCML files into a number of formats that allow direct visualization. Since BCML files contain all the necessary information to produce SBGN-compliant graphs, graphical output of BCML files is straightforward. The default conversion is from BCML into GraphML, one of the widely used standards for the representation of graphs. BCML files converted to GraphML can be opened by several free software tools such as the yEd or Cytoscape and then successively exported in many other formats.

Pathway analysis.

One of the main goals of BCML is to allow to use pathways with advanced methods for pathway analysis. The tools provided with the software suite allow the extraction of identifiers (genes) lists from a BCML file, allowing for their use with analysis methods such as functional class scoring methods and over-representation methods. In addition to this basic conversion, BCML overcomes one of the main limitations of existing formats by allowing straightforward conversion to formats that are suitable for topological analysis of pathways. This conversion takes into account of basic elements such as type of interaction, entities involves, and advanced elements such as previous knowledge (through the experimental data information stored along with each entity), allowing for a fine tuning of the parameters of the analysis.

BCML and the software suite are freely available and open source, available upon request from `http://dc-research.eu`. An example of the graphical representation of a BCML pathway can be seen in Figure 2.4.

2.4 Pathway analysis methods

The state of the art of methods for pathway analysis can be best understood from an historical perspective. Before pathways were created to describe in detail complex biological processes, the only available data was provided by the Gene Ontology (GO). As explained in Sec. 2.1, in GO genes are associated with terms to known biological processes and cellular component. When presented with a list of genes of interest, researchers would consult GO through a tedious manual process to identify which terms were most likely to be associated with the genes of interest. The first approach to automate this process was *functional profiling* [31, 63]. Since most people are usually interested in the GO categories that are enriched in DE genes, this approach has become known as *over-representation analysis* (ORA). An alternative approach is the one followed by methods in the *Functional Class Scoring* category. Methods belonging to this category do not rely on a selection of DE genes needed by over-representation approaches. In addition, these methods take into account the correlations among expression profiles of genes. The Gene Set Enrichment Analysis (GSEA) [83, 105, 112] is the most widely used FCS approach. GSEA ranks all genes based on the correlation between their expression and the given phenotype, and calculates a score that reflects the degree to which a given pathway is represented at the extremes of the ranked list.

When pathway databases started to become available, describing in detail the interactions among genes, offered the potential for a more complex and useful analyses than the simple enrichment. However, at the beginning the methods originally developed for GO analysis were immediately used to analyze pathways. The extrapolation was very simple: consider a pathway as merely the set of the genes that are involved in it (discarding the interactions), and perform exactly the same analysis used for GO annotations. This was easy but, like many easy solutions, not optimal, as it has been discussed in the literature [62, 64, 82]. The

main limitations of these approaches when used to analyze pathways is that they treat the pathways as simple sets of genes, ignoring the very reason for their existence: the description of the complex interactions between their genes. More recently, an impact analysis [30, 108] has been proposed as approach that manages to take into consideration biologically important factors previously neglected by most existing pathway analysis tools. This approach has subsequently evolved into a category of *topology-based methods* (See [82] for a comprehensive list).

In this section we will briefly describe the two categories of pathway analysis methods, from the simplest analysis methods that consider pathways only as gene sets, to the most advanced methods that include other types of information such as the topology of each pathway and information on the differential expression of individual genes.

2.4.1 Gene set based analysis methods

Gene set based analysis methods belong to that category of methods that do not include topological information in the analysis. This approach does not assume the existence of any knowledge regarding the underlying interactions among genes, considering pathways as simple sets of genes. Most of the methods belonging to this category were originally developed for the Gene Ontology, where terms are indeed gene sets. Once pathway databases started providing pathway information, the same methods were *borrowed* for pathway analysis. The most widely used strategy for analyzing gene sets is *enrichment analysis*. This strategy involves determining if a gene set is *enriched* in genes of interest, where the definition of "gene of interest" varies based on the specific method. In most cases, a p-value is computed to assess the probability that the observed enrichment can be obtained by chance alone. An extensive review of enrichment methods [54] listed almost 70 existing available methods.

Over-representation analysis

Arguably the most common method for enrichment analysis is the over-representation analysis. This method was first introduced in 2002 for the functional analysis of GO [31, 63]. In this context, the input is a list of genes marked as differentially express (DE) or non

differentially expressed (NDE) based on their expression levels in a certain condition. In the case of phenotype comparisons, this is determined based on the difference in expression values for each gene in the phenotypes in analysis. A *p-value* is computed, expressing the probability of obtaining a difference equal or greater than what is observed. If this p-value is smaller than a certain arbitrary threshold, the gene is considered as DE, otherwise it is considered NDE. Methods for determining such p-value include t-test, fitting of linear models, ANOVA, etc. Each GO term is analyzed to determine if it is either over-represented (more DE genes than expected just by chance) or under-represented (less DE genes than expected just by chance) in the condition under study. This method was immediately used to analyze pathways with the same approach: if a pathway contained more DE genes than expected by chance, then it was considered involved in the condition in analysis. The hypergeometric model is one of the most commonly used methods for determining statistical significance of the observed over- or under-representation. This model computes a p-value that represents the probability of obtaining a number of DE genes in a pathway more extreme than the one observed, taking into account the total number of DE genes an the total number of genes screened. Assuming that N genes are screened, that K genes are found to be DE, that K_P genes are found to be DE in pathway P, and that pathway P has size N_P genes in total, the probability of obtaining exactly K_p DE genes can be computed as in Eq. 2.1.

$$P(X = K_P|N, N_P, K) = \frac{\binom{N_P}{K_P} \cdot \binom{N-N_P}{K-K_P}}{\binom{N}{K}} \tag{2.1}$$

The probability of obtaining a number of genes equal or higher than the observed value K_P can be obtained with Eq. 2.2.

$$P(X \geq K_P) = 1 - \sum_{i=0}^{K_P-1} \frac{\binom{N_p}{i} \cdot \binom{N-N_p}{K-i}}{\binom{N}{K}} \tag{2.2}$$

The hypergeometric p-value computed for each pathway is used to rank them, and it is interpreted as the *amount of involvement* of each pathway in the phenomenon that generated

the specific list of DE genes.

Currently, more than 40 tools using this or similar approaches are available, and an extensive survey can be found in [64].

Functional class scoring

The over-representation approach described in the previous section uses the simplest possible type of information available for pathways and for experimental data, i.e. pathway membership of each gene and the information regarding the differential expression of each gene (DE or not), respectively. This second type of information is based on an *a priori* selection of interesting genes: the list of genes is ranked based on some quantitative measure, e.g. fold change or p-value coming from the comparison of two phenotypes, and the list is *cut-off* at a certain point. For example, a commonly used such cut is performed choosing genes with p-value smaller than 0.05 (after correction for multiple comparisons) and the logarithm in base 2 of its fold change greater than 1. This *cut-off dependency* represents one of the biggest limitations of over-representation approaches, making them sensitive to the change of the cut-off parameters.

Functional class scoring approaches overcome this limitation by computing the association between pathways and phenotypes using all the genes measurements available.

Arguably the most widely used method in this category is the Gene Set Enrichment Analysis (GSEA) [83, 105]. GSEA first ranks all genes based on a measure expressing the correlation between the gene expression of each gene and the phenotypes in analysis, obtaining a ranked gene list L. The measure used can be as simple as the fold change between two phenotypes, or be more sophisticated, such as the moderated t-statistic [101].

Then, an *enrichment score* (ES) is computed, reflecting the degree to which a gene set S is represented at the top or at the bottom of the list L. The hypothesis is that if the set S is not associated with the phenomenon, then the genes that belong to it will not be concentrated either at the top or at the bottom of L. If the set is associated with the phenomenon, then the genes belonging to S will be mostly at either extreme of the list.

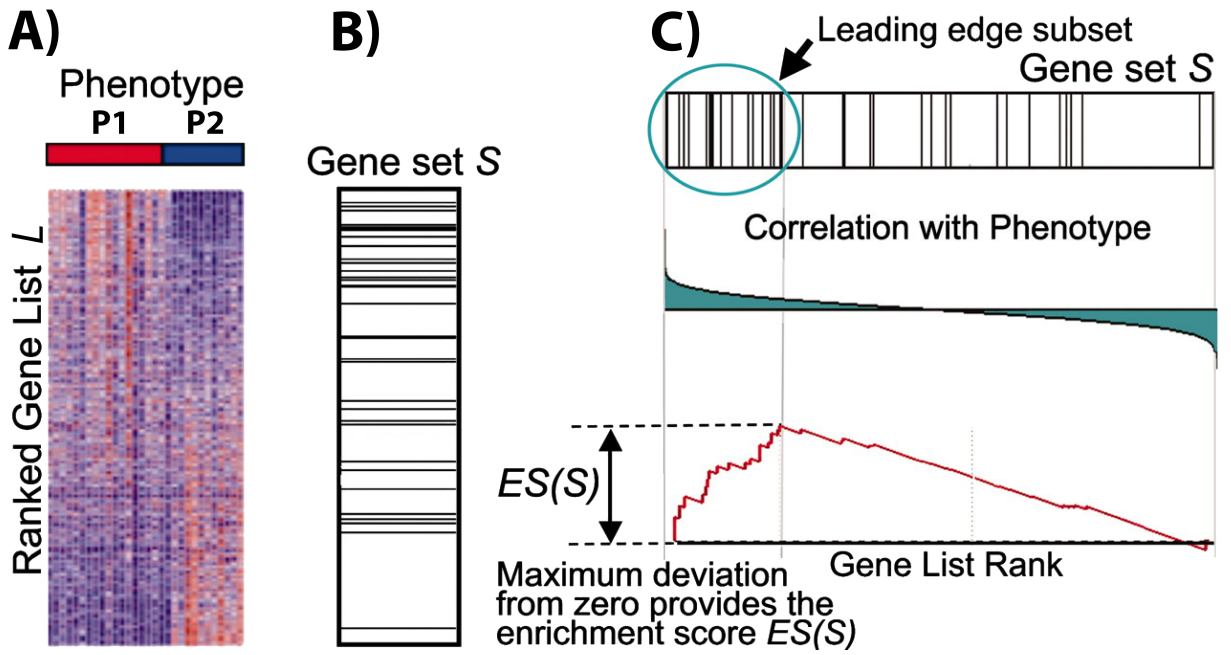


Figure 2.8: The GSEA method starts with a gene expression matrix where columns represents samples coming from two phenotypes (P1 and P2 in panel A), and rows represent genes. The correlation of gene expression values with the difference in phenotypes is computed, and an enrichment score is determined for each gene set (panel C) using a Kolmogorov-Smirnov statistic.

Hence, ES is computed by walking down L and increasing a running-sum statistic whenever a gene belonging to S is encountered, and decreasing it when a gene that does not belong to S is found, with the increment being proportional to the measure used to quantify the correlation of genes with the phenotype. An example of the running-sum statistic is shown in panel C) of Figure 2.8.

The resulting ES is the maximum deviation from zero of the running-sum statistic, corresponding to a weighted Kolmogorov-Smirnov-like statistic [51], and the genes contributing to the maximum deviation are referred to the *leading-edge subset*, representing the core genes contributing to the ES.

Statistical significance of the ES is estimated by a permutation based approach. The phenotype labels are permuted many times and fold changes are computed each time. At each iteration the ES of all pathways are computed, building the null distributions of ES for each pathway. The observed ES is compared with the null distribution, obtaining a nominal p-value by counting the number of times that randomly obtained ES scores are more extreme

than the observed score. It is important to note that the null distribution can be obtained by permuting phenotype labels as well as by permuting gene labels. The phenotype labels permutation approach is the more common, since it preserves gene-gene correlations.

When a single set is analyzed, the ES and its corresponding p-value are used as-is. If multiple sets are analyzed at the same time, each ES is normalized for each set by dividing the raw ES by the size of the set, obtaining a normalized ES (NES). The p-value of each ES is corrected for multiple comparisons by calculating the false discovery rate (FDR) [9, 10], in order to reduce the number of false positives.

Several improvements over the basic GSEA algorithm have been proposed in the course of the years. The two most notable are the modified GSEA proposed in [128] and [33]. The first extends GSEA in a number of ways: first, by fitting multiple regression models to the data in order to correlate gene expression to continuous covariates (e.g. age of the sample), and ordering by a coefficient of interest. Second, it replaces the Kolmogorov-Smirnov statistic with the van der Waerden statistic. This kind of statistic is chosen due to the properties related to small set sizes. Third, the permutation approach is replaced by a bootstrap approach [34].

The second notable modified GSEA approach takes the name of Gene Set Analysis (GSA). This approach replaces the Kolmogorov-Smirnov statistic with a *maxmean* statistic showing its greater statistical power, and introducing a re-standardization procedure on the data that allows to take into account, during the calculation of the statistic for each gene set, of the scores obtained by permutation of sample labels and of scores obtained by permutation of gene labels *at the same time*.

2.4.2 Topology aware methods

The methods described in the previous section share an important limitation: they perform the analysis without exploiting information regarding the topology of the pathway. This analysis paradigm disregards the fact that organisms are complex systems whose emerging phenotypes are the result of thousands of complex interactions happening between genes lo-

cated on various metabolic and signaling pathways. Topology aware methods, also referred to as *third generation pathway analysis methods* [82, 64] go beyond simple gene set analysis by incorporating topology information. In [82], we surveyed more than 20 methods that incorporate topology into the analysis of signaling pathways from the point of view of input data, mathematical models used, output format, and technical implementation details.

Most analysis methods accept as input either a list of gene identifiers, along with their measured fold changes and, if present, a p-value expressing the statistical significance of such fold changes. As in the case of gene-set based methods, some analysis methods rely on a selection of differentially expressed genes, like the over-representation approach described in the previous sections, based, in most cases, on an arbitrarily selected threshold either on the fold change or the p-value, and in some cases, of both.

Regardless of the choice of using filtered lists or not, analysis methods can use either the full information available, i.e. gene ID, fold change, and statistical significance, or any combination of these factors. For example, the TopoGSA [40] method uses the DE/non-DE label, scoring pathways based on the relative position of genes in each pathway. Other methods use all the genes and their fold changes (e.g. PARADIGM [117]), while the most recent implementation of the impact analysis [118] is able to incorporate all the three factors in the analysis.

The mathematical models employed by the various methods are much more heterogeneous than the ones used by ORA. Whereas ORA approaches rely mostly on Fisher's exact test, many different approaches are used by topology aware pathway analysis methods, from the choice of graph analyzed (e.g. typed interactions, directed or undirected) to the method using for assigning a score to the pathway representing the level of involvement in the phenotype comparison (e.g. graph measures, Bayesian network analysis), to the choice of assigning statistical significance to the score, and if so, what strategy (parametric or non-parametric) to assign to the score.

Finally, the output reported by the methods presents a certain variability. Although the

final result of a pathway analysis method is usually a list of pathways ranked by their involvement with the condition in analysis, some methods provide more information. The impact analysis provides information about the direction of the involvement, as well as information regarding the presence of *cascades of coherent perturbation propagation*, i.e. paths in a pathway where each node presents an expression changes coherent with the expression change of the nodes upstream, weighted by the type of interaction. Other methods provide information on sub-pathways that are deemed important in the phenomenon in analysis.

Impact Analysis

The impact analysis is the first pathway analysis method that incorporates the topology of pathways into the analysis. This method aims to overcome the limitations of ORA and FCS methods by integrating i) the number of DE genes in a pathway, ii) the fold change of those genes, and iii) the interactions among those genes into an *impact factor* representing how much the pathway in analysis is impacted in the condition in analysis. Eq. 2.3 shows the original definition of the impact factor for a pathway P.

$$IF(P) = log\left(\frac{1}{p}\right) + \frac{\sum_{g \in P} |PF(g)|}{\overline{|\Delta E|} \cdot N_{de}(P)} \tag{2.3}$$

The number of DE genes of a pathway is represented by a classical probabilistic term, such as Fisher's exact test, and it is captured by the first term of Eq. 2.3. The second term of this equation is the term that takes into account the topology of the pathway. The numerator is the sum of the *perturbation factors* (PFs) of the genes in the pathway, defined as follows for a gene g:

$$PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} \tag{2.4}$$

The PF of each gene is composed by a first term, $\Delta E(g)$, representing the measured change of the gene, and the summation term represents the effect that the genes upstream of g have on the PF of g itself.

Figure 2.9: Example of interaction among genes. Gene A and gene B are linked by an interaction of type *activation*, meaning that the product of gene A activates the product of gene B.

Particular importance goes to the term β_{ug}. This term is related to the type of interactions that each gene u upstream of g has on g itself. Users of the impact analysis can choose the value of β that best suits their knowledge on how a gene affects another gene when the two are connected by a certain type of interaction. For example, gene A can be upstream of gene B, and the type of interaction can be of type *activation*, as shown in Figure 2.9.

Default values for β have been determined by a panel of experts. For example, interactions such as *activation*, *expression*, *activation through phosphorylation* are assigned by default a value of $\beta = 1$, indicating that all the perturbation passes through those nodes with no modulation, while interactions like *repression* or *inhibition* are assigned a value of $\beta = -1$, indicating that the perturbation passing through those edges passes with reversed sign, i.e. it has an inverse effect on the gene downstream. Lastly, in Equation 2.3, the denominator of the second term is the average expression change of genes in the pathway, multiplied by the number of DE genes in the pathway.

The impact factor IF obtained from Equation 2.3 follows a $\Gamma(2, 1)$ distribution, and for each pathway P_i a p-value $p(P_i)$, expressing the probability of having, on P_i, *both* a number of DE genes higher than what can be observed just by chance and a perturbation value larger than what can be observed just by chance can be computed as $p(P_i) = (if + 1) \cdot e^{-if}$.

A second implementation of the impact analysis, presented in [107], involves the computation of a *perturbation accumulation* for a gene g with the formula in Equation 2.5.

$$Acc(g) = PF(g) - \Delta E(g) \tag{2.5}$$

In Equation 2.5, the term $PF(g)$ is the *perturbation factor* as described in Equation 2.4. Once the perturbation accumulation is computed for all the genes, a total accumulation T_{Acc} is computed for each pathway as sum of the Accumulation values for all the genes in the pathway. A p-value p_{Acc} expressing the probability of obtaining, just by chance, a value of T_{Acc} equal or more estreme than the observed value is computed with a resampling approach.

Similarly to the first implementation of the impact analysis described above, a p-value for each pathway P_i, combining classical over-representation analysis and the new perturbation accumulation p-value, is computed as follows:

$$p(P_i) = p_{DE} \cdot p_{Acc} \cdot (1 - ln(p_{DE} \cdot p_{Acc})) \tag{2.6}$$

CHAPTER 3 IMPROVEMENTS TO THE TOPOLOGICAL ANALYSIS OF SIGNALING PATHWAYS

Although the impact analysis incorporates a number of factors that are crucial for the analysis, such as the magnitude of the expression change for each gene, and the topology of the pathway, it still presents an important limitation: the results are highly dependent to the selection of DE genes used as input. This limitation has two important consequences. First, once the set of DE genes is determined, the contribution of those genes is considered proportional only to their values of differential expression, i.e. the expression *fold change* between phenotypes in analysis, completely discarding the information on the statistical significance that has already been computed for each gene. This implies that genes with marginal significance are given the same importance of genes that are highly significant. For example let us examine a real, publicly available, colorectal cancer dataset, obtained from GEO (GEO ID GSE4107) [52]. This dataset consists of expression profiling of 12 early onset colorectal cancer samples versus 10 normal samples, using the Affymetrix HG-U133 Plus 2.0 microarray platform, analyzing the genetic components behind the tumorigenesis of colorectal cancer. After normalization and pre-processing, we performed a moderated t-test with the *limma* R package to determine the significance of the fold change for each gene. False discovery rate (FDR) was then used to control the error for multiple comparisons. In order to select DE genes, a threshold on FDR corrected p-value, fold change, or both has to be chosen. In this case we used a threshold of 5% on the p-values.

After the selection of differentially expressed genes using a threshold of 5% on the p-value, gene contributions are considered equally important. For example, the contribution of the gene with ID 3725 (*jun proto-oncogene*, official gene symbol JUN), marked with the red box at the top of the list of genes is considered equal of the gene with ID 84612 (*par-6 family cell polarity regulator beta*, official gene symbol PARD6B), which is at the bottom of the list of DE genes, which has a significance barely above the threshold. The problem here is that the log fold changes of the two genes are comparable (1.53 for JUN and -1.21 for PARD6B), and if a method includes only the fold change in the analysis those two genes are considered as

Gene ID	logFC	corr.p-value
1843	2.4298886559	1.41E-008
3725	1.5311256684	1.15E-006
23645	1.429269302	2.42E-006
9510	3.9376626047	2.42E-006
...	...	...
51209	0.8953513664	0.049938805
84612	-1.2171622836	0.049938805
25966	0.4876491864	0.049938805
284273	-0.611286691	0.049938805
54537	-0.4122475786	0.0500210569
21	0.4725740362	0.050030946

Figure 3.1: After the selection of differentially expressed genes using a threshold of 5% on the p-value, gene contributions are considered equally important. For example, the contribution of the gene with ID 3725 (*jun proto-oncogene*, official gene symbol JUN), marked with the red box at the top of the list of genes is considered equal of the gene with ID 84612 (*par-6 family cell polarity regulator beta*, official gene symbol PARD6B), which is at the bottom of the list of DE genes, which has a significance barely above the threshold.

having approximately the same impact on the phenotype, while the statistical significance of the two fold changes, as expressed by the p-values, clearly indicates that the amount of trust that we can put in these two measurements is differs from one gene to the other. This common behavior indicates the need to take into account the statistical significance of a gene measured fold change alongside the fold change itself in order to obtain a reliable interpretation of the data.

The second consequence of the need to select *a priori* a list of DE genes based on arbitrary thresholds is that this represents an artificial truncation of the information available, as well as an unnecessary reliance on the upstream method for the selection of genes that produces highly variable results. It has been shown that the choice of threshold in the selection of genes as input of analysis methods severely affects the results [89]. Figure 3.1 shows an example of such phenomenon. The gene with ID 54537 (*family with sequence similarity 35, member A*, official symbol FAM35A) is barely *above* the significance threshold with a

p-value of 0.050021. And yet, this gene is considered as *not* interesting, unlike the gene with ID 284273 (*zinc binding alcohol dehydrogenase domain containing 2*, official symbol ZADH2), although the difference between the p-values of the two genes is less than 0.00001 and their fold changes have comparable values. This aspect indicates the need to develop methods that overcome this limitation, using the entire list of genes in the analysis.

3.1 Incorporating gene significance in the impact analysis of signaling pathways

In order to incorporate the statistical significance into the analysis we proposed the addition of a term α_g in the computation of the gene perturbation factor described in Equation 2.4.

The perturbation factor for a gene g is computed as follows:

$$PF(g) = \alpha_g \cdot \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \cdot \frac{PF(u)}{N_{ds}(u)} \tag{3.1}$$

When using the implementation of the impact analysis described in [107] the accumulation is computed as follows:

$$Acc(g_i) = PF(g_i) - \alpha_g \cdot \Delta E(g_i) \tag{3.2}$$

In [120] we proposed two alternatives for assigning the values of the weighting factor α_g.

The first alternative is defined as:

$$\alpha_g = -log\frac{P_g}{\alpha_t} \tag{3.3}$$

In this equation, the term P_g represents the significance value for the measured expression change of gene g, and α_t is the significance threshold chosen for selecting DE genes. The effect of this formulation of α_g is that genes whose significance is very close to α_t will be weighted less than genes that are more significant, since

$$\lim_{P_g \to \alpha_t} -log\frac{P_g}{\alpha_t} = 0 \tag{3.4}$$

For example, when using a threshold of 1%, a gene g_1 with a p-value of 0.0001 will result in a value of $\alpha_{g_1} = -log\frac{0.0001}{0.01} = 2$. Conversely, a gene g_2 with a p-value very close to the chosen threshold, for example 0.009 will result in a value of $\alpha_{g_2} = -log\frac{0.009}{0.01} \approx 0.0457$, thus contributing much less to the perturbation factor of the pathway it belongs.

This alternative for weighting genes has the disadvantage that it favors very small p-values, and it tends to infinity when p-values have the values of zero. Although this is an unlikely situation, at least theoretically, some pathway analysis packages return such values, especially when using empirical methods for the computation of the p-values. Therefore, we proposed a second alternative for the computation of α_g:

$$\alpha_g = 1 - \frac{P_g}{\alpha_t} \tag{3.5}$$

This expression, henceforth referred to as $1MR$ (1 Minus Ratio) does not present the disadvantage of the previous method, henceforth referred to as MLG (Minus LoG). With this expression the value of α_g is in the interval $[0, 1]$, and it will follow the same behavior, i.e. being close to zero when P_g tends to α_t and being close to 1 otherwise. However, this approach has the characteristics of compressing very small p-values (very significant) around 1. For example, a p-value of $1e^{-5}$ and a p-value of $1e^{10}$ will yield very similar values of α_g.

An inspection of the list of p-values is therefore necessary before choosing which method is suitable.

3.1.1 Cut-off free analysis

In [120] we proposed a method for performing impact analysis without the need of pre-selecting a list of DE genes. This method allows to use the entire list of genes expression values. Since it makes little sense to consider equally important genes at the top of the

list and genes at the bottom, this method builds on the concepts explained in Section 3.1 that allow to incorporate gene significance into the impact analysis of signaling pathways by weighting genes differently based on their p-value.

The two alternative approaches are the following, named respectively ALL_MLG (Equation3.6) and ALL_1MR (Equation3.7):

$$\alpha_g = -log\frac{P_g}{P_{max}} \tag{3.6}$$

$$\alpha_g = 1 - \frac{P_g}{P_{max}} \tag{3.7}$$

Since all the genes are now considered in the analysis, and there is no list of DE genes, the over-representation p-value that was used in Equation 2.3 cannot be computed and it is therefore excluded from the computation of the impact factor. Hence, the significance of a pathway is represented only by the total accumulation p-vale p_{Acc}.

We evaluated the improvements to the impact analysis on the colorectal cancer dataset GSE4107 already described at the beginning of this section [52]. The comparison of the top ranked pathways when using a list of DE genes (cut-off dependent analysis) is presented in Table 3.1 and when using a cut-off free analysis in Table 3.2. Besides the *Colorectal cancer pathway* itself, several other pathways are know to be related to colorectal cancer including: *PPAR signaling pathway* [99] and *Toll-like receptor signaling pathway* [121, 37]. Both with cut-off dependent and cut-off free the MLG model ranks the *Colorectal cancer* pathway better than the original SPIA method. In addition, with the cut-off dependent SPIA_MLG ranks better the *Toll-like receptor signaling pathway* and with cut-off free the ALL_MLG ranks *PPAR signaling pathway* better.

In addition to the single dataset comparison, in order to perform an objective evaluation of the improvements that we proposed in [120], we used the comparison framework introduced in [106]. This framework consists of 24 datasets describing multiple diseases: Alzheimer's

Table 3.1: Comparison of two models that incorporate gene significance (SPIA_MLG and SPIA_1MR) with the model without (SPIA) for the GSE4107 colorectal cancer dataset. The *Colorectal cancer pathway* and *Toll-like receptor signaling* pathway are both ranked better by SPIA_MLG. More over, the *PPAR signaling pathway* is rank better by SPIA_1MR.

SPIA		**SPIA_MLG**		**SPIA_1MR**	
Name	**adj.pv**	**Name**	**adj.pv**	**Name**	**adj.pv**
ECM-receptor interaction	0.00034	**Colorectal cancer**	0.181	ECM-receptor interaction	0.00034
Focal adhesion	0.00034	Dilated cardiomyopathy	0.181	Focal adhesion	0.00034
Small cell lung cancer	0.09067	Serotonergic synapse	0.181	Small cell lung cancer	0.04533
Glutamatergic synapse	0.17000	Bile secretion	0.181	Glutamatergic synapse	0.22667
VEGF signaling pathway	0.21371	Amphetamine addiction	0.181	Pathways in cancer	0.22667
Pathways in cancer	0.21371	Prion diseases	0.181	VEGF signaling pathway	0.22667
Systemic lupus erythematosus	0.21371	**Toll-like receptor signaling**	0.253	Pathogenic E. coli infection	0.42500
Pathogenic E. coli infection	0.31733	Protein processing in end. ret.	0.255	Systemic lupus erythematosus	0.42500
Chemokine signaling pathway	0.31733	Focal adhesion	0.348	**Colorectal cancer**	0.48960
Cytokine-cytokine rec. int.	0.31733	Cocaine addiction	0.420	Dilated cardiomyopathy	0.48960
Colorectal cancer	0.31733	Pathways in cancer	0.420	Type II diabetes mellitus	0.51927
African trypanosomiasis	0.31733	Systemic lupus erythematosus	0.435	**PPAR signaling pathway**	0.54400
Hepatitis C	0.43714	VEGF signaling pathway	0.435	Serotonergic synapse	0.54400
Staphylococcus aureus infection	0.43714	ECM-receptor interaction	0.435	Staphylococcus aureus infection	0.54400
...	...	...	...	...	...

Table 3.2: Comparison of two cut-off free models (ALL_MLG and ALL_1MR) with the model original model (SPIA) for the GSE4107 colorectal cancer dataset. The *Colorectal cancer pathway* and *PPAR signaling pathway* are both ranked better by ALL_MLG.

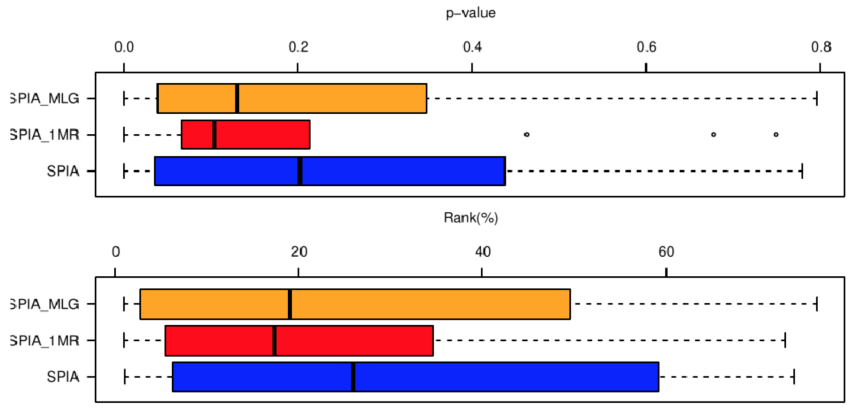
SPIA		**ALL_MLG**		**ALL_1MR**	
Name	**adj.pv**	**Name**	**adj.pv**	**Name**	**adj.pv**
ECM-receptor interaction	0.00034	Focal adhesion	0.174	Cytokine-cytokine rec. int.	0.000171
Focal adhesion	0.00034	Serotonergic synapse	0.174	Chemokine signaling pathway	0.000171
Small cell lung cancer	0.09067	**Colorectal cancer**	0.174	Focal adhesion	0.000171
Glutamatergic synapse	0.17000	ECM-receptor interaction	0.174	ECM-receptor interaction	0.000171
VEGF signaling pathway	0.21371	Dilated cardiomyopathy	0.174	Staphylococcus aureus infection	0.022833
Pathways in cancer	0.21371	Prion diseases	0.174	Systemic lupus erythematosus	0.022833
Systemic lupus erythematosus	0.21371	Parkinson's disease	0.174	Pathways in cancer	0.034250
Pathogenic E. coli infection	0.31733	Cocaine addiction	0.174	Small cell lung cancer	0.034250
Chemokine signaling pathway	0.31733	**PPAR signaling pathway**	0.174	Pathogenic E. coli infection	0.076111
Cytokine-cytokine rec. int.	0.31733	Bile secretion	0.174	**PPAR signaling pathway**	0.112091
Colorectal cancer	0.31733	Pathways in cancer	0.174	**Colorectal cancer**	0.112091
African trypanosomiasis	0.31733	Systemic lupus erythematosus	0.183	Hepatitis C	0.114167
Hepatitis C	0.43714	Renal cell carcinoma	0.190	Glutamatergic synapse	0.137000
Staphylococcus aureus infection	0.43714	Protein processing in end. ret.	0.209	Sulfur relay system	0.146133
...	...	...	...	...	...

Table 3.3: The list of data sets used for the evaluation of the performance of pathway analysis methods. For a more detailed description see [106]

Data set	Disease / Condition	Pathway
GSE1297	Alzheimer's disease	hsa05010
GSE5281_EC	Alzheimer's disease	hsa05010
GSE5281_HIP	Alzheimer's disease	hsa05010
GSE5281_VCX	Alzheimer's disease	hsa05010
GSE20153	Parkinson's disease	hsa05012
GSE20291	Parkinson's disease	hsa05012
GSE8762	Huntingon's disease	hsa05016
GSE4107	Colorectal Cancer	hsa05210
GSE8671	Colorectal Cancer	hsa05210
GSE9348	Colorectal Cancer	hsa05210
GSE14762	Renal Cancer	hsa05211
GSE781	Renal Cancer	hsa05211
GSE15471	Pancreatic Cancer	hsa05212
GSE16515	Pancreatic Cancer	hsa05212
GSE19728	Glioma	hsa05214
GSE21354	Glioma	hsa05214
GSE6956C	Prostate Cancer	hsa05215
GSE6956AA	Prostate Cancer	hsa05215
GSE3467	Thyroid Cancer	hsa05216
GSE3678	Thyroid Cancer	hsa05216
GSE9476	Acute myeloid leukemia	hsa05221
GSE18842	Non-Small Cell Lung Cancer	hsa05223
GSE19188	Non-Small Cell Lung Cancer	hsa05223
GSE3585	Dilated cardiomyopathy	hsa05414

disease, Parkinson's disease, Huntington's disease, colorectal cancer, renal cancer, pancreatic cancer, glioma, prostate cancer, thyroid cancer, acute myeloid leukemia, and non-small-cell lung cancer. The list of datasets is summarized in Table 3.3.

These diseases have been chosen because a pathway exists *describing the condition in analysis in the experiment that generated the dataset*. This pathway is considered as *target pathway*, i.e. the pathway that is most likely related to the condition. The evaluation works under the reasonable assumption that, when comparing two pathway analysis methods, the best method is the one that ranks the target pathway higher and with lower p-value. Hence, we compare our analysis strategies by comparing the distributions of ranks and p-values



	Method	p geomean	p med	% p.value<0.05	% q.value<0.05	rank mean	rank med
1	SPIA_MLG	0.0379	0.1310	28.57%	14.29%	27.3518	19.0368
2	SPIA_1MR	0.0300	0.1046	21.43%	14.29%	24.2791	17.3387
3	SPIA	0.0424	0.2021	28.57%	14.29%	32.7214	25.9420

Figure 3.2: **Comparison using the list of DE genes:** distribution of the rank and p-value of the target pathway over 14 data sets. Both methods that incorporate gene significance rank the target pathways better than SPIA and also assign them a more significant p-value.

of target pathways across the entire list of datasets. Figure 3.2 and shows the comparison between the impact analysis proposed in [107] and the two implementations that take into account the significance of individual genes, while Figure 3.3 shows the comparison between the two implementations that do not need a list of differentially expressed genes as input. In the table the results of the original method are in the column marked SPIA, whereas the two methods we proposed are in the columns SPIA_MLG and SPIA_1MR, for the alternatives in Equations 3.3 and 3.5, respectively. Both in terms of ranks and p-values the two models that incorporate gene significance (SPIA_MLG and SPIA_1MR) perform better than the method that does not incorporate it (SPIA), yielding lower p-values for the target pathway, hence offering a more accurate insight on the biological phenomenon. When evaluating the cut-off free models, all 24 datasets where used. Figure 3.3 shows that ALL_MLG model performs slightly better than ALL_1MR, but they are outperformed by the original impact analysis.

These results indicated that the introduction of the individual gene significance into the impact analysis of signaling pathways leads to more biologically relevant results and, as a consequence, it has the potential to lead to a deeper insights of the biological phenomena

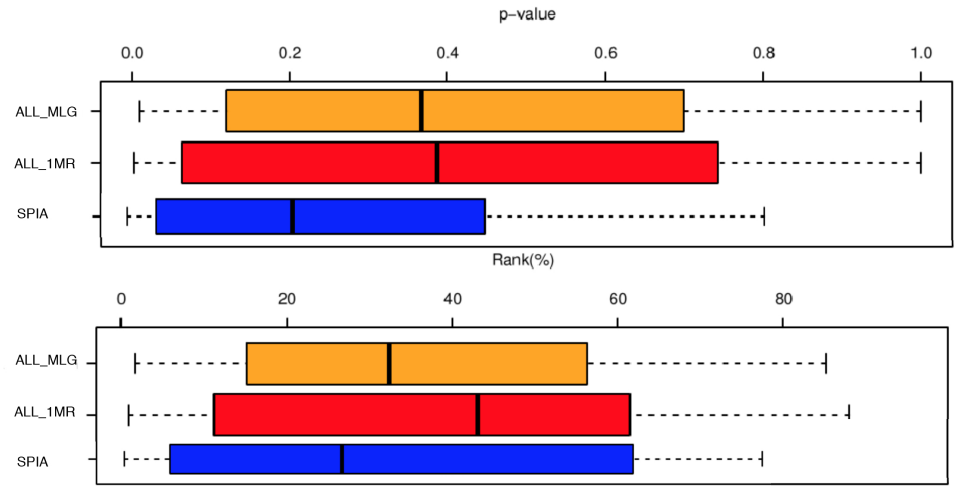


Figure 3.3: **Comparison of cut-off free analysis:** distribution of the rank and p-value of the target pathway over 24 data sets. Both methods perform similar in terms of rank and p-value with the ALL_MLG model performing slightly better. However, both methods perform worse than the original impact analysis.

involved in a certain condition.

3.2 Genetic algorithms for the estimation of individual gene contribution in the analysis of signaling pathways

In Section 3.1 we proposed two methods for the incorporation of individual gene significance into the impact analysis of signaling pathways. These approaches rely on the assumption that the contribution that a gene exercises on the pathways it belongs to is proportional to the statistical significance that is computed in a specific dataset. However, this might not be the case. For example, we can take the *Insulin signaling pathway* shown in Figure 3.4. If the insulin receptor (*INSR*, marked in red in the left side of the figure) is not present, the majority of the pathway is shut off, hence it is natural to assume that the INSR gene is more important than most of the other genes in that specific pathway. Conversely, if several genes are involved in a pathway but they only appear somewhere downstream, changes in their levels may not affect the given pathway as much. Moreover, some genes have multiple functions and are involved in several pathways but with different roles. For instance, the same *INSR* is also involved in the *Adherens junction* pathway as one of many tyrosine kinase receptors. However, if the expression of *INSR* changes, this pathway is not likely to be heavily perturbed because *INSR* is just one of many receptors on this pathway.

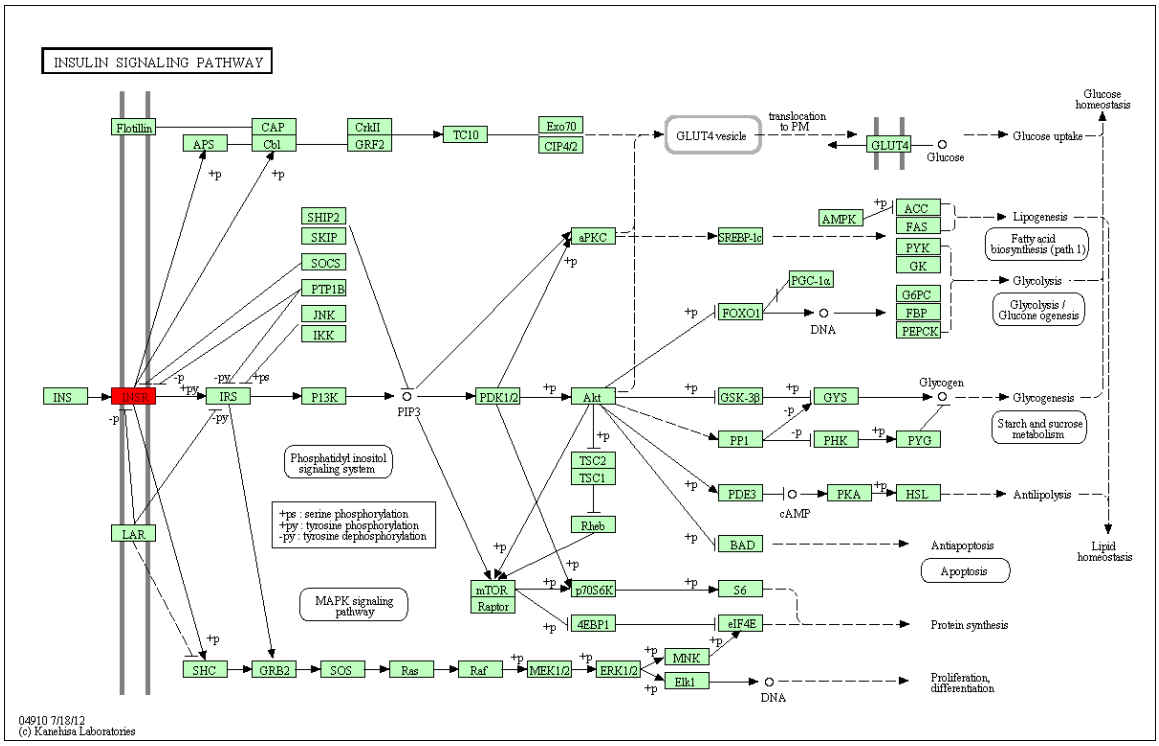


Figure 3.4: KEGG Insulin Signaling Pathway. The insulin receptor (INSR, in red, left side of the image) is the only entry point in this pathway and a big change in its expression will have a high impact on the entire pathway. *Source:* KEGG [60] - `http://www.genome.jp/kegg/pathway/hsa/hsa04910.html`

We consider all these factors to affect the importance of a gene and therefore its contribution to the given pathway. In principle, these contributions can be arbitrarily set beforehand, for example based on the type of gene product (i.e., transcription factor, transmembrane receptor, etc.), or with the approaches developed in 3.1. However, there is currently no basis for setting exact qualitative values for various gene types.

This is why in [119] we proposed an approach for estimating the contribution of individual genes that departs from such assumption. In this approach we used genetic algorithms in order to objectively compute these contributions, based on the performance of the impact analysis on a group of datasets representing a variety of conditions.

Genetic algorithms (GA) are search methods based on natural concepts such as natural selection, evolution and genetics. In its basic form, a GA involves the evolution of a fixed size population across generations, where each individual of the population represents a possible solution in the search space. Each individual is represented by a set of *genes*, and each gene

is represented in a way that depends by the implementation of the GA (e.g. binary string or floating point). The evolution of the population results in one or more of the individuals satisfying a certain criterion of the search, for example a local maximum or minimum. The evolution is led by a few key events: selection, crossover, and mutation. Selection is the process of elimination of the individuals that do not pass a fit-test. Several methods exist for selection, but the basic idea is to give preference to the better individuals. The key element in the selection is the way of determining which individual is better. For this purpose, an evaluation can come from an objective function that gives each individual a score that can serve, for example, for ranking the population. Crossover is the event in which two individuals A and B are chosen to be mated. The two sets of genes belonging to the two individuals are parted in the same way, and then two new individuals are constructed taking one of the partitions of A and one the partitions of B. Mutation, like crossover, is a way to explore different structures. This event represents a single, usually low-probability, random change in a gene. Selection, crossover, and mutation are applied across generations, and the average evaluation of the population increases. A stopping criterion is then applied (e.g. limit to the number of generations, threshold on the result of the evaluation function for the best individuals), and the best individuals are chosen as solutions. The use of GAs in bioinformatics is widespread, from applications in sequence alignment [42] to RNA structure prediction [116]. This technique, however, found little use in the context of regulatory pathway analysis. Most of the approaches apply genetic algorithms while trying to model the underlying, unknown, network [132], or to simulate the network dynamics [53]. However, to date there are no applications, to the best of our knowledge, of GAs to the analysis of regulatory pathways in the context of phenotype change, and our approach represented the first work in that direction.

We focused on the impact analysis described in Section 3.1 where the gene contribution can be captured by the factor α_g in Eq. 3.2.

Given a predefined set of pathways S_P, we obtained the set of unique genes U contained

in these pathways. We designed our individual as a vector of size equal to the size of the set U. Each gene, in the context of genetic algorithm, is represented by a floating point number between 0 and 1, representing the contribution of each gene $g \in U$.

The genetic algorithm has been implemented in the R Framework [110], adapting the *genalg* package, allowing for parallel execution of the evaluation of the individuals. Mutation chance has been set to 10%, while the selection method used was elitism of the top 20% of the population ranked by fitness. The type of crossover was single point. The parameters have been chosen according to the indications in [41].

The goal of the evaluation function is to capture the ability of the gene weights to model biological knowledge encoded in the given pathways and not any specific condition. In other words, the gene weights have to capture the gene importance related to the topology of the pathway, rather than computing the gene importance based on the data linked to a specific condition. The evaluation function is based on the validation of pathway analysis methods described in [106].

Given the *a priori* defined set of data sets DS with their associated target pathways, the evaluation function scores each individual by applying the impact analysis on each data set independently and recording the *normalized rank* of the target pathway associated with each data set. The return value of the evaluation function will be the average normalized rank of the target pathway over all data sets in DS. Hence, the lower the result of the evaluation function, the better the individual. Starting from the pool of datasets shown in Figure 3.4, we divided the pool into *train* and *test* groups to emulate two scenarios.

These two scenarios were chosen in a way that captures an ideal environment and the real environment. In the ideal environment, each one pathway of the 140 pathways available in the KEGG database would be associated with a dataset. Since this is not the case, in the real environment we have both pathways that are associated to a dataset and pathways that are not.

For both scenarios we selected the training and testing datasets in a similar fashion. First,

Table 3.4: The list of data sets used for the evaluation of the performance of pathway analysis methods. For a more detailed description see [106]

Data set	Disease / Condition	Pathway	Scen. 1	Scen. 2
GSE1297	Alzheimer's disease	hsa05010	train	train
GSE5281_EC	Alzheimer's disease	hsa05010	test	test
GSE5281_HIP	Alzheimer's disease	hsa05010	test	test
GSE5281_VCX	Alzheimer's disease	hsa05010	test	test
GSE20153	Parkinson's disease	hsa05012	test	test
GSE20291	Parkinson's disease	hsa05012	train	train
GSE8762	Huntingon's disease	hsa05016	-	test
GSE4107	Colorectal Cancer	hsa05210	train	train
GSE8671	Colorectal Cancer	hsa05210	test	test
GSE9348	Colorectal Cancer	hsa05210	test	test
GSE14762	Renal Cancer	hsa05211	-	-
GSE781	Renal Cancer	hsa05211	-	-
GSE15471	Pancreatic Cancer	hsa05212	train	train
GSE16515	Pancreatic Cancer	hsa05212	test	test
GSE19728	Glioma	hsa05214	-	test
GSE21354	Glioma	hsa05214	-	train
GSE6956C	Prostate Cancer	hsa05215	-	train
GSE6956AA	Prostate Cancer	hsa05215	-	test
GSE3467	Thyroid Cancer	hsa05216	-	train
GSE3678	Thyroid Cancer	hsa05216	-	test
GSE9476	Acute myeloid leukemia	hsa05221	train	-
GSE18842	Non-Small Cell Lung Cancer	hsa05223	-	test
GSE19188	Non-Small Cell Lung Cancer	hsa05223	-	train
GSE3585	Dilated cardiomyopathy	hsa05414	-	-

we performed the impact analysis, as implemented in [118] with the default set of $\alpha = 1$, on each data set. We next ordered the data sets based on the normalized rank of the target pathway and selected datasets starting at the top of the list. This approach allowed us to avoid data sets in which the target pathway was badly ranked, possibly indicating that those particular data sets contained bad data or they were not representative for that particular condition.

In the **real environment scenario**, based on the ordered list of data sets, we selected as *train* data sets the top five data sets that represent different conditions. The *test* set was chosen based on the conditions of five data sets selected as *train*. All the remaining data sets that study one of the five conditions was selected as *test*. The set of pathways for which the gene contributions will be estimated was chosen based on the *train* data sets. We selected five pathways representing the target pathway for the respective conditions. We then selected the top three pathways (based on the impact analysis results) on each of the

five *train* data sets to be part of the *train* set. This provided an additional 11 pathways, generating a train set that included a total of 16 pathways.

The set of genes used in this scenario was the set of genes that appear at least once in any of the pathway selected for training. We obtained a set of $1,355$ genes. As described in Section 3.1, each gene is associated with a parameter α_g. Therefore, our individual had a total of $1,355$ genes to be used in the genetic algorithm search. We choose the size of the population to be equal to the number of genes of an individual (1,355 individuals), and we performed 100 generations with a mutation rate of 1%. We applied an elitist selection to the population, after each evaluation, where the top 20% individuals in the list ranked by the evaluation function were passed to the next generation with no crossover.

In the **ideal environment scenario**, based on the same ordered list of data sets, we selected the top ranked data set for each condition as the *train*. Given that some of the conditions did not have at least two data sets associated to them and therefore no *test* data set could be selected, these conditions are removed from further analysis. Two other datasets relative to *renal cancer* had to be removed due to the excessive variability of the rank of the target pathway when the analysis was performed with default parameters. Hence, the *train* would contain eight data sets and the *test* eleven data sets (see Table 3.4). As this scenario would represent the ideal environment, we only selected for analysis the eight pathways associated as target pathways with the conditions in the *train* set. The total number of unique genes in these pathways was 372. We choose the size of the population to be 600, and we performed 100 generations with a mutation rate of 1%. We applied an elitist selection to the population, after each evaluation, where the top 20% individuals in the list ranked by the evaluation function were passed to the next generation with no crossover.

For both scenarios, at the end of the evolution of the population we extracted a random individual among the individuals with the best ranks (smallest value from the evaluation function) and we evaluated its performance in the test set. This yielded an average normalized rank of the individual on the test set of datasets. The evolution of the fitness function

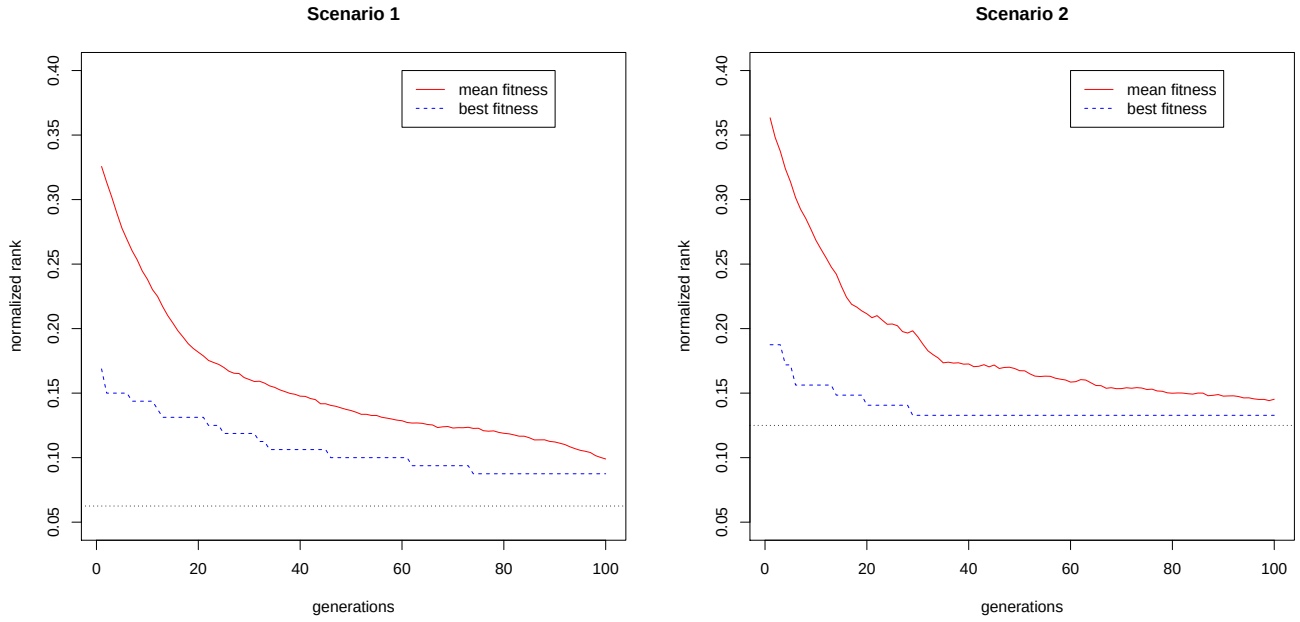


Figure 3.5: The evolution of the best and mean evaluations over the entire population at each generation. The evaluation function for one individual is the mean normalized rank over the training data sets. Because the evaluation function uses the normalized rank, the minimal value of the evaluation function is dependent on the total number of pathways evaluated (16 for Scenario 1 and 8 for Scenario 2). This minimal value achievable is show with a horizontal dotted line and is equal to 1/16 for Scenario 1 and 1/8 for Scenario 2. This value represents the case where the target pathways are ranked as first in all training data sets.

over all generation for each scenario is presented in Figure 3.5.

For both scenarios, the rank of the best individuals was better than the result obtained with default parameters. By default parameters, we refer to the original impact analysis method (see Equation. 2.5), where the weight of each gene was considered to be maximum $\alpha_g = 1$ for all gene. In order to assess if the performance of the individual was significantly better than a random choice, we used a bootstrap approach, generating the null distribution of the average normalized ranks, as described in [32]. We obtained this by creating 1,000 individuals with values of the gene weights randomly drawn from an uniform distribution with range $[0, 1]$. Each individual was then evaluated on the test set. This procedure yields a p-value, computed as the number of random individuals that obtain a score lower than the score of the best individual. This p-value represents the probability of getting a score lower than the best individual just by chance. We performed this procedure for the populations

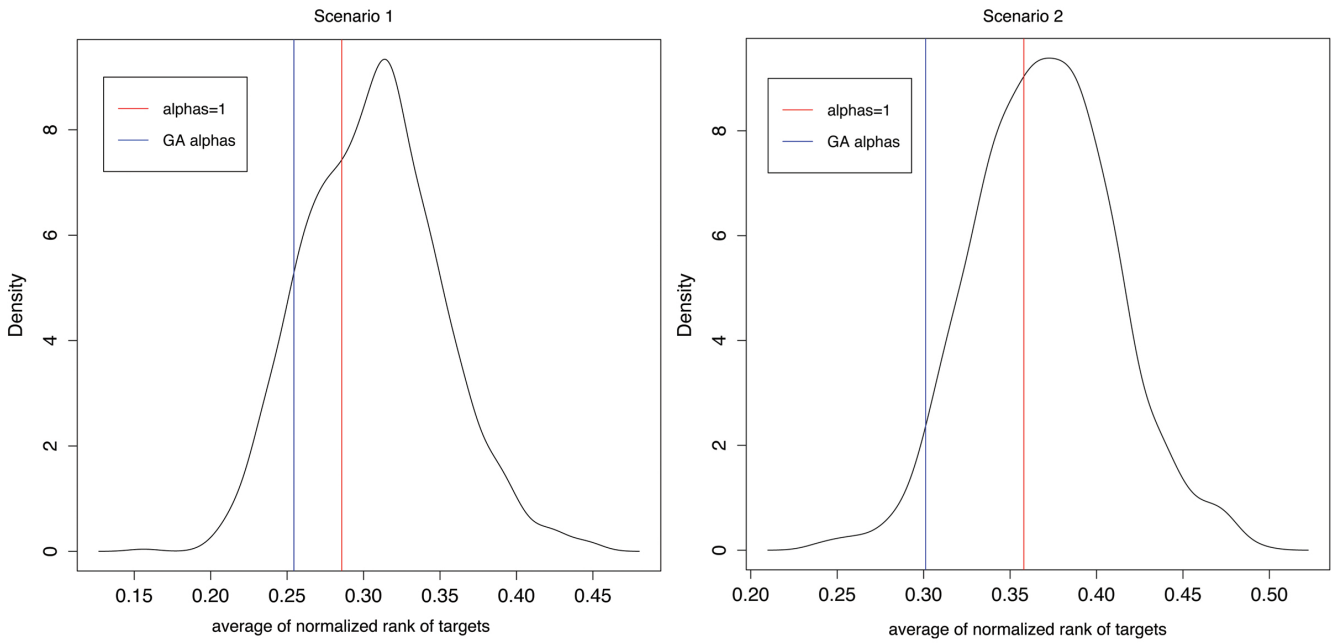


Figure 3.6: Null distributions of the average mean ranks of random individuals on the test sets. The left panel shows the distribution of the evaluation of random individuals on the *test* set associated with the selection of pathways in the first scenario, while the right panel shows the distribution of the evaluation of random individuals on the *test* set associated with the selection of pathways in the second scenario. The blue lines represent the value of the average normalized rank of the best individual in the two populations, while the red lines represent the average mean rank of the *default* individuals (all the αs equal to 1). The results show that the default values are reasonable but only slightly better than those provided by a random choice. In both cases, the values obtained after the GA search are significantly better than the mean of the random chance values.

obtained with both scenarios described above. The best individual of the population obtained from the first scenario achieved a p-value of 10.4% on the test set, while the best individual of the population obtained with the second scenario achieved a p-value of 3.3% on the test set. The distributions relative to the two bootstraps are shown in Figure 3.6. The left panel shows the distribution of the evaluation of random individuals on the *test* set associated with the selection of pathways in the first scenario, while the right panel shows the distribution of the evaluation of random individuals on the *test* set associated with the selection of pathways following the second scenario. The blue lines represent the value of the average normalized rank of the best individual in the two populations, while the red lines represent the average mean rank of the *default* weights.

These results show that in both cases the optimization reaches significantly better results

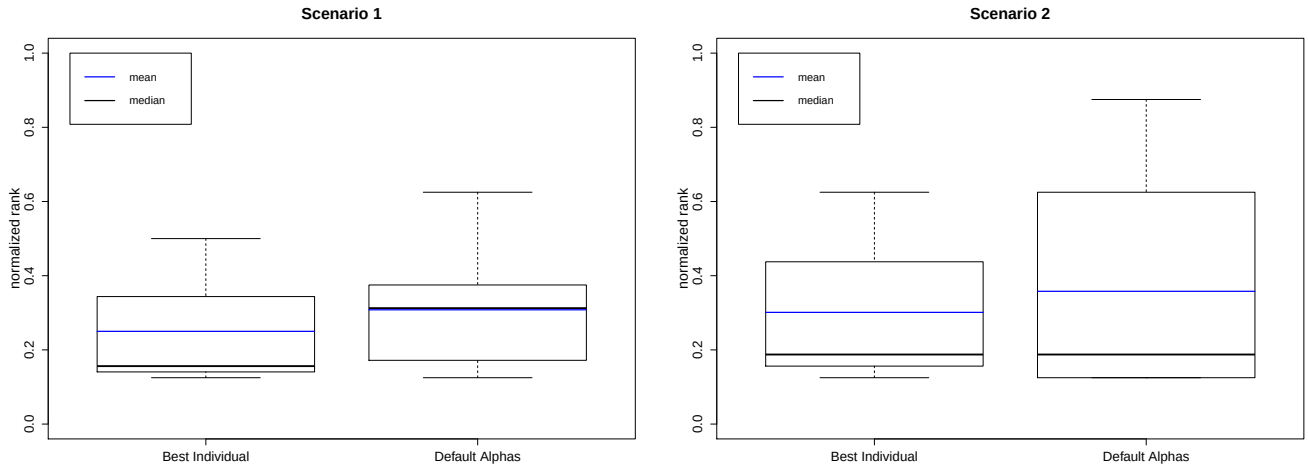


Figure 3.7: Normalized ranks of target pathways using parameters from best individuals (left side of each panel) and default parameters (right side of each panel). The left panel shows the comparison between the best individual of scenario 1 and default parameters in the test set from scenario 1, while the right panel shows the comparison between the best individual of scenario 2 and the default parameters in the test set from scenario 2. The blue line represents the mean of ranks, while the black line represents the median. In the left panel (scenario 1, real environment) the optimization procedure results in lower mean rank and lower median. In the right panel (scenario 2, ideal environment) the optimization procedure results in the lower mean rank, reduced variance, and the same median.

than the results obtained with the default set of parameters. In the test set of the *real environment* scenario a random choice would perform worse than the optimized parameters 89.6% of the times. In the test set of the *ideal environment* scenario a random choice would perform worse than the optimized parameters in more than 96% of the times. These values are also considerably better than those obtained with the default values.

Figure 3.7 shows the comparisons between the ranks obtained with best individuals and the ranks obtained with the default parameters for both scenarios when we perform the analysis in the respective test sets. The left panel shows the comparison performed in the test set of the first scenario. In this scenario the best individual outperforms the default parameters obtaining a lower mean (0.25 against 0.335), lower median(0.156 against 0.25), while there is no improvement in terms of variance (0.296 against 0.207). The right panel shows the comparison performed in the test set of the second scenario. In this scenario the best individual obtains a lower mean (0.301 agains 0.357), the same median (0.187), and a decreased variance (0.038 against 0.102).

The most important limitation associated with this framework is still related to the evaluation of the pathway analysis results. The evaluation only considered the rank of the target pathways. An individual was considered fitter than another one if it the average rank of the target pathways was lower. In reality, a more important distinction is between those individuals that *rank the target pathways as significant* and those that do not. An improvement in rank that still has the target pathways as not significant is not really an increase in accuracy, and therefore it should not be represented as an increased fitness. Conversely, a decrease in ranking within the significance range may not be a decrease in accuracy, and therefore should not always be penalize as a decrease fitness. However, using a step-like evaluation function based on the significance would have introduced abrupt changes that could have increased the difficulty of the genetic algorithm search.

Despite this limitation, the results obtained with this framework showed the effectiveness of evolutionary computation techniques in the optimization of parameters in bioinformatic applications. This framework is general enough to be applied to a multitude of methods for the analysis of biological pathways, where parameters are often chosen arbitrarily.

3.3 Estimating interaction efficiency of directly linked genes using microarray time series analysis

In Equation 2.4, the term β can be seen as a *weight* of the interaction between genes. In the standard implementation of the impact analysis, the value of β has been determined by a panel of experts. For example, an interaction of type *activation* has been assigned the value 1, while an interaction of type *inhibition* has been assigned the value -1. This term can be also seen as the *regulatory efficiency* of each interaction, i.e. the amount of impact that flows along that interaction from the gene upstream to the gene downstream the edge. The default values given to the various types of interaction present a number of limitations. First, out of the 25 different interactions present in KEGG and considered in the impact analysis), only 15 have a value assigned to them. These values are limited by the expertise of the panel of expert chosen at the time the impact analysis was developed. The fact, for example, that

the interaction *dephosphorylation* has been assigned a value of zero means that either no signal flows through that interaction or that there was not enough knowledge on that type of interaction to assign a value different from zero. Even when the β is different from zero, it is either $+1$ or -1, although there is no reason why it could be limited to those two values. Another big limitation is that the values of β are not assigned to a specific interaction, but to an *interaction type*. This means that all the interactions of the same type will have the same effect, disregarding the genes at the two ends. Lastly, the interactions are considered the same in *all the conditions*. It is intuitive to think that in some specific conditions some interactions may work differently. For example, in a certain disease two genes might not communicate, making the weight for that interaction zero for that specific disease. These limitations raise the need for a method for the assessment of the regulatory efficiency of gene to gene interactions in signaling pathways. In [27] we analyzed a number of methods for assessing such efficiencies by analyzing time series data. Time series data consists in measurements of gene expression of the same biological sample at sequential instants in time to profile the behavior of the gene as time progresses. This type of experiments allows researchers to trace the evolution of gene expression, for example, after a certain treatment. In our work we chose to test the ability of three different metrics to identify interactions that are already present in the KEGG signaling pathway database. The metrics were chosen based on specific features of gene expression time series. The first involved a comparison of the profiles of time series based on bit string matching. The second was a specific application of Dynamic Time Warping to detect similarities even if the time series are one the stretched and delayed version of the other. Finally, the third was a quantitative comparative analysis derived by a frequency domain representation of time series: the similarity metric is the correlation between dominant spectral components. These three approaches were tested on a real case study and a final comparison of them was performed using Information Retrieval benchmark tools.

3.3.1 Microarray data pre-processing and noise reduction

The data set analyzed was the result of an experiment of type expression profiling by array, performed at the Karmanos Cancer Institute. This experiment analyzed the expression profile of human mammal epithelial cells in response to the administration of HER2-specific small molecule kinase inhibitor (CP724,714). The resulting dataset consists of samples from a $24,527$ probe Illumina microarray, taken at intervals of three hours for a total of 45 h. The data are collected in a set of $24,527$ time series, each one consisting of 16 time points and one measurement per time point. The information about each time point also contains the p-value, to allow an assessment of the quality of the measurement. More details on the experiment that produced the dataset can be found in [16].

Data processing was performed in the R statistical framework, and pathways were retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG, `http://www.genome.jp/kegg/`) from the R library SPIA [107].

A first assumption that we made is that each time point depends only on the time point before. This aspect was taken in account in two out of the three similarity measures studied. A second assumption was made about the characteristic of the time series representing transcriptomic profiles, which feature a noisy and stochastic component [14, 35, 67, 100]. Therefore, the second assumption is that these time series are prone to random errors of two different origins: errors that are characteristic of the biological process of transcripts generation, and errors produced by the data extraction using various techniques and in particular microarrays. In order to limit the effect of such errors we applied a Savitzky-Golay filter, a noise reduction procedure [96].

Given these assumptions, the analysis proceeded as follows. First we selected the genes that had a p-value below the threshold of 0.05 after FDR correction for multiple comparisons for at least half of the elements of the time series. This empirical criterion was motivated by the trade off between a meaningful significance threshold and the desire to have a sufficient number of probes. The next step was the selection of differentially expressed genes. Several

statistical models for the identification of differentially expressed (DE) genes in time series have been proposed [104, 74, 59, 39, 102].

We chose to use the package *limma* [102]. The approach implemented by limma is based on a t-test and on a linear approximation of the behavior of individual genes, in which the coefficients of the adapted models describe the differences of RNA represented in the microarray. The limma algorithm was chosen to select DE genes because it was deemed the best fit for our dataset by a prior analysis performed on a very small subset of data, being able to select with a higher precision the genes that were over or under expressed.

Finally, we chose to focus only on the genes that had only one upstream gene. This was done in order to eliminate overlapping influences on genes, since not all the similarity methods were able to deal with this factor.

As stated in the above assumptions, even after such filtering, the microarray data are considerably affected by noise. Time series, regarded as dynamic functions, are made up of a signal component, which varies slowly over time, and a stochastic component with a much faster dynamic. To better understand the difference between speeds of transition in signal and noise, it is easier to transpose the problem in the frequency domain, using Fourier transforms of the signals. In general, from the biological evidence, the band of the signal is limited, while the band of the noise is virtually infinite. In other words, the signal has a low bandwidth, while the noise has a high bandwidth. Therefore, introducing in the process pipeline a low pass filter having a cut-off frequency properly tuned to the signal (a filter that stops high bandwidth signals, while accepting low bandwidth signals), it is be possible to reduce frequencies out of the band of interest, isolating almost completely the effective signal [93].

A Savitzky-Golay filter, based on the principle of local least squares fitting of a polynomial, has been applied for time series de-noising. The basic idea is to replace each data point by a weighted average of surrounding data points, which measure very nearly the same underlying value. In this way, averaging can reduce the level of noise without much biasing the result.

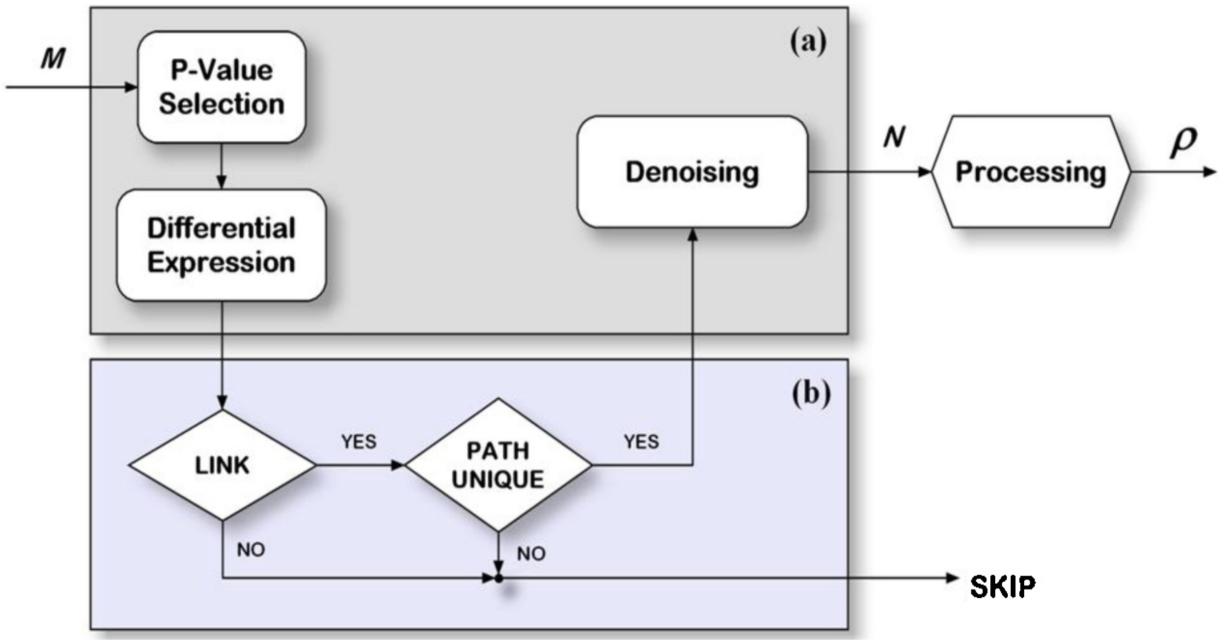


Figure 3.8: (a) Pre-conditioning procedure: The input is a matrix, M, whose rows are gene expression time series from microarray experiment. The first step of the process ("p-value Selection" block) produces a matrix of 15,744 time series with reduced noise. Then, after discarding time series that are not significant and/or not differentially expressed ("Differential Expression" block), 8524 times series are left. The "LINK" block checks the existence of a direct link between genes in a pathway. Finally, only 3116 time series expressing genes directly linked in a single pathway are arranged in the matrix N. This matrix is the input of the "Processing" block, which outputs the matrix column of the correlation values. (b) Selection of genes directly linked in pathways, and with unique upstream node.

Savitzky-Golay filters are well-adapted for data smoothing and the simplest type of digital filter replaces each data value by a linear combination of itself and some number of nearby neighbors [6, 96]. Figure 3.8 shows a block diagram representation of the experimental design of the study. The final de-noising block in Figure 3.8 is added to take into account the various noise components, which overlap randomly with the signal.

3.3.2 Similarity metrics

The pairwise comparison between selected genes is the next step in the processing of the time series. The most immediate and basic idea for quantifying the relationship between two random dynamics is the calculation of Pearson's correlation coefficient between them. In order to take into account the propagation time of biological signals, one could calculate the correlations between shifted versions of all the time series (from 0 to 9 h). If, for example, the signal from gene A takes approximately 3 h to get to gene B, and the two genes are linked

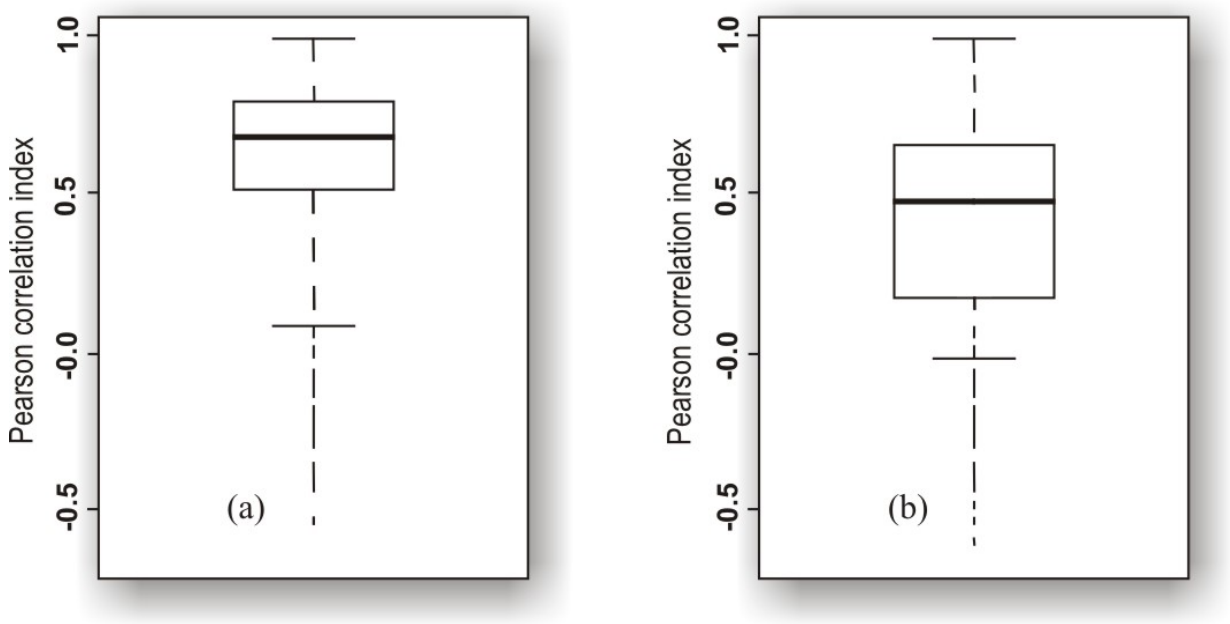


Figure 3.9: Boxplots of the distribution of Pearson correlations between genes directly connected by activation (a) and inhibition (b) interactions. We considered a detected"activation" as a correlation value greater than 0.5 and a detected "inhibition" as a correlation value smaller than -0.5.

by an activation edge, we expect the 3-hour shifted time expression profile of gene B to be positively correlated with the time expression profile of gene A. The results of this simple analysis showed its limitations. More specifically, the Pearson correlation index performed poorly in identifying the type of the regulation, being able to identify the relationship of activation only in 45% of the cases, and confusing inhibition interactions with activation in 50% of cases. Figure 3.9 shows the distribution of the correlation indices for activation (panel a) and inhibition (panel b) interactions.

Given the poor results of the classical correlation, we applied to the data three distinct methods for the computation of similarity between time series, in order to assess which one is the most effective: differential comparison, Dynamic Time Warping (DTW) and dominant spectral component (DSC).

Differential comparison.

Due to the nature of biological phenomena, the time series of a gene can be a translated, softened, or deformed version of the time series of another gene. However, if the two genes are related, we expect the *trend* of the series to be consistent. This trend could be summarized by

the monotonicity (e.g. increasing or decreasing) and concavity (or convexity) of the discrete function representing the series. Like in the previous section, we performed the comparison with the shifted versions of all the time series (shifted from 0 to 9 h). With this approach, the qualitative comparison of time series uses as parameters the differential information of first order, i.e. increase or decrease, and the differential information of second order, i.e. concavity or convexity, of the function representing the series. Each section joining two points in time is encoded with two bits: 00 if the difference between the extremes of that interval is positive, 11 if it is negative, and 01 if it is zero. The same process is repeated on the differences of differences, or second order differences. Since our series are composed of 16 time points the result of the differential of the first order will be a string of 30 bits. This is because for each pair of points (values in the time series) we have two bits describing the difference between them. Since there are 15 pairs, we will have 30 bits (15 bit pairs). Similarly, the result of the differential of second order will be a string of 28 bits, since there will be 14 "pairs of bit pairs". The comparison between genes directly connected is therefore reduced to a comparison between two strings of bits, properly shifted to capture any delays from 0 to 9 h. The two strings of bits were compared using the Hamming distance [44]. The similarity index is computed as the difference between 1 and the ratio among distance and total number of bits. For an instance the strings 1111000111 and 0011110111 have a Hamming distance equal to 4 and the similarity index is equal to$1 - 0.4 = 0.6$. In our case study a similarity index of 0.95 is heuristically used as the threshold to detect an effective link of activation type between two genes. Likewise, inhibition can be detected reversing all pairs of bits of type "00" and "11" in one of the compared strings.

Dynamic time warping.

Dynamic Time Warping (DTW), one of the discriminative algorithms based on pattern matching, is borrowed from the fields of voice and motion recognition and measures the similarity between two sequences that vary in time. It provides a measure of the similarity between two time series, assuming that the time pattern of one of them can be a

distorted, delayed, or stretched version of the other. Dynamic Time Warping is an excellent method for the discovery of matches between two time series, when one is a non-linear distortion of the other, with respect to the independent variable (typically the time). However, some constraints must be satisfied for the computation of the similarity between time series: monotonicity in the matches and the maximum limit of possible matches between adjacent elements of the sequence. In particular, this type of algorithm for time series gene expression has been shown to be an improvement with respect to other types of similarity measures [2]. The objective of DTW is to compare two time series $X = (x_1, x_2, \ldots, x_N)$ and $Y = (y_1, y_2, \ldots, y_N)$ of the same length $N \in \mathbb{N}$. In order to measure the similarity between these two sequences using DTW, we first construct an $N \times N$ matrix D, where the element $D_{i,j}$ corresponds to the squared distance, i.e. $d(x_i, y_j) = (x_i - y_j)^2$. Then we retrieve a path through this matrix that minimizes the total distance between X and Y, choosing progressively contiguous elements in D. The optimal path is the one that minimizes the warping cost $DTW(X, Y) = min(\sqrt{\sum_{k=1}^{K} w_k})$, where w_k is an element in a set of K contiguous elements of the D matrix. In other words, a stretching of the time axes of the sequences X and Y brings them as close as possible to each other and this minimum distance is considered as a pairwise similarity measure of the time series. Remarkably, despite of the large search space, this algorithm can be computed in $\mathcal{O}(N^2)$ time using dynamic programming, i.e. Bellman's equation [7].

Dominant spectral component.

The third method for assessing similarity is the dominant spectral component (DSC) approach. The model presented in [127] describes the spectral decomposition of expression profiles, allowing a comparison between frequency components of time series considered as signals, and this representation is capable of identifying links that are completely missed by the traditional linear correlation. Furthermore, using this method in the case of genes that are subjected to the influences of different upstream genes at the same time, it is possible to distinguish the influence that is due to different genes. The basic idea of this technique is

to decompose the time series $x(n), n \in \mathbb{N}$, in a set of sinusoids with variable amplitude and having various frequencies:

$$x[n] = \sum_{i=1}^{M} x_i[n] = \sum_{i=1}^{M} \alpha_i \cdot exp(\sigma_i n) \cdot cos(\omega_i n + \varphi_i) \tag{3.8}$$

The parameters amplitude α_i, damping factor σ_i, angular frequency ω_i, and phase shift φ_i, (with $i = 1, 2, ..., M$, where M is the number of spectral components), can be calculated with the autoregressive method [126]. These completely define the spectrum of the temporal expression profile of a gene. The similarity of the sequence $x[n]$ with another sequence $y[n]$ can be formulated again as the sum of the components of partial cross-correlation weighted with the relative energy of the components:

$$x[n] \circ y[n] = \sum_{i} \sum_{j} \sqrt{\frac{E_{x_i} E_{y_i}}{E_x E_y}} x_i[n] \circ y_i[n] \tag{3.9}$$

The symbol "$\circ$" represents the cross-correlation operator and E_x and E_y, E_{x_i}, and E_{y_i} are the values of total energy of a sequence (E_y and E_{x_i}) or of one of its components (E_{x_i} and E_{y_i}). These energy components are computed as follows:

$$E_x = \sum_{n=-\infty}^{+\infty} |x[n]|^2 \quad ; \quad E_y = \sum_{n=-\infty}^{+\infty} |y[n]|^2 \tag{3.10}$$

Eq. 3.10 shows how the correlation between two sequences can be split in a set of partial correlations, which contains more detailed information about the similarity between a couple of genes. This type of similarity is able to eliminate spectral components that are low in amplitude in at least one of the compared signals. Therefore, the phase and the noise components in amplitude spectrum, which ordinarily would make similarity recognition difficult, can be neglected. On the same grounds, the weighted partial components of correlation can be considered as a more reliable measure of the connection between genes. In particular, the heaviest weighted component resulting from the decomposition can be regarded as a metric to measure the correlation between genes. The corresponding non-weighted value is called

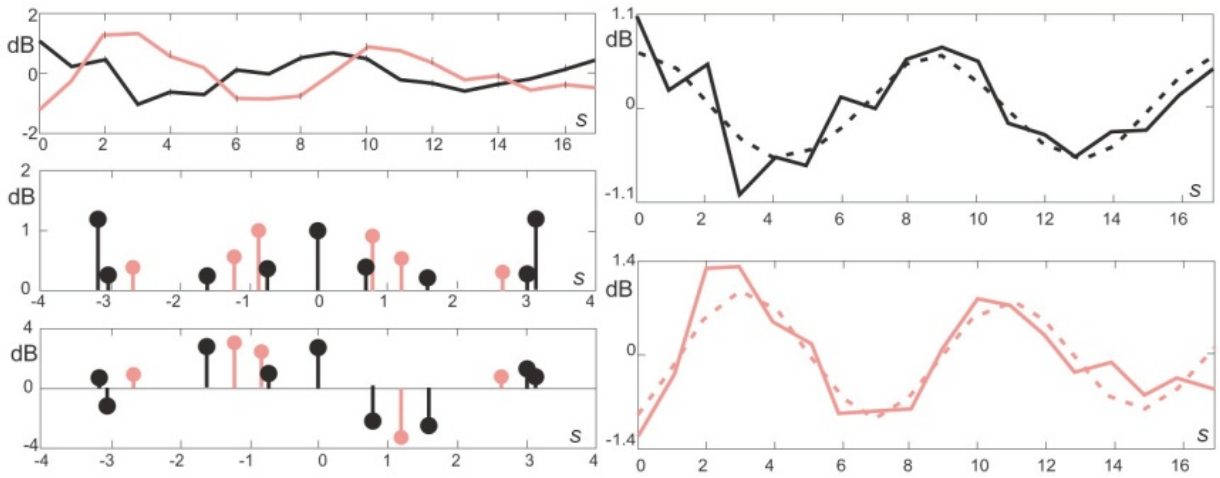


Figure 3.10: Time series (left, top panel), spectral amplitude and phase decomposition (left, middle and bottom panels), and waveforms corresponding to the dominant spectral components of two genes expressions values connected with an interaction of type "activation".

cross-correlation coefficient of the component itself. Figure 3.10 shows an example of the expression time series of two genes (red and black waveforms) connected by an activation inter- action (left, top panel), the corresponding spectral decomposition in amplitude and phase (left, middle and bottom panels) and the waveforms corresponding to the dominant spectral components (right pane dashed lines).

3.3.3 Performance comparison

The result of the similarity metrics can be evaluated with a method borrowed from information retrieval. The comparison between two series may have five different outcomes: (a) correlation detected when an appropriate link in the pathway exists between the two genes, i.e. $\beta \neq 0$, and therefore a true positive (TP), (b) no correlation detected when no link exists, i.e. $\beta = 0$, and therefore a true negative (TN), (c) correlation detected when no link exists, i.e. $\beta = 0$, and therefore a false positive (FP), (d) no correlation detected when a link exists, i.e. $\beta \neq 0$, and therefore a false negative (FN), and (e) a correlation is detected when a link exist, but the correlation and β are opposite in sign. In our case, the class of positive regulations is much larger than the class of negative regulations, so the Matthews Correlation Coefficient (MCC) [77] was used as a performance index (in addition to precision, recall and accuracy) because it is particularly suitable in the case of asymmetric binary

classifications with a significant difference of magnitude in the two classes [58]. The MCC essentially specifies the relationship between the predicted and the observed classification; it can take values in the range [-1;1]. The value -1 occurs if the prediction is opposite to the observed, +1 if the prediction is correct and 0 if the classifier performs a random choice. The MCC can be calculated directly from the confusion matrix using the formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.11}$$

If any of the factors in the denominator were zero, this ratio would be undefined. In this case the denominator can be set equal to 1, which will make MCC equal to 0 in that particular case. The other indices considered for the comparison of the similarity measures are the classical accuracy, precision, and recall. Figure 3.11 shows the performance, in terms of these indexes, for the three methods compared, plus the performance of the simple Pearson's correlation.

The Differential Comparison, DTE and DSC, applied to the subset of genes selected in the pre-processing phase, produced for each pair of directly linked genes a "similarity" value. These values were compared to the value of the β factor from the signaling pathways included in the SPIA package. These 132 pathways are extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database of pathways, and they cover 4253 unique genes, out of which 3116 were found in the microarray used in the experiment. These values refer to the time the experiment was performed.

The performances in terms of accuracy, precision, recall and Matthews Correlation Coefficient (MCC) were computed as described in detail in the paragraphs above. In particular we considered true positives those cases in which the detected correlation was greater than 0.75 in case of $\beta = 1$ and smaller than -0.75 in case of $\beta = -1$. The total number of direct links was considered the universal set of reference. The results are summarized in Figure 3.11. Each column represents a performance index with respect to the algorithm applied.

The histograms in Figure 3.11 show that the accuracy is almost the same for the three

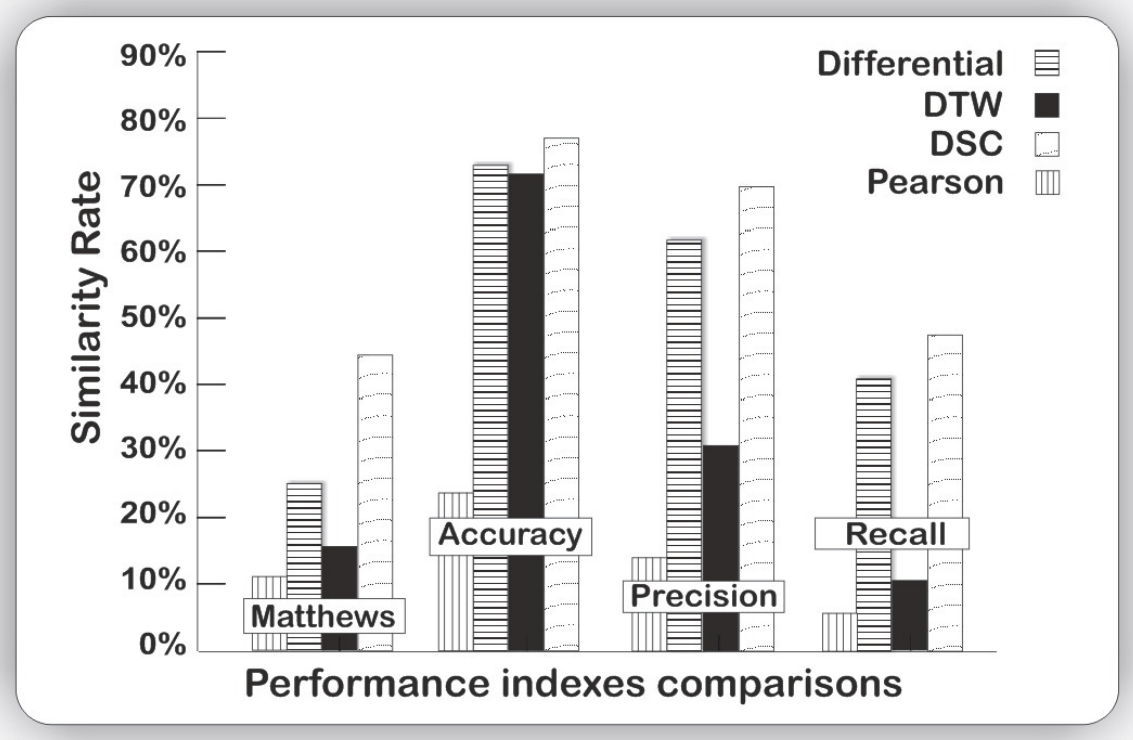


Figure 3.11: Barplots of performance comparison of similarity metrics, Differential comparison, DTW and DSC in the analysis of gene expression by microarray to assess the correspondence to regulatory efficiency in human signaling pathways.

metrics, which means that the overall number of true positives and true negatives is comparable among the three methods. DTW is the worst-performing similarity metric. This was expected because it typically stretches the values of data points to "adapt" two waveforms, and this feature in our experiment can result in a distortion producing low sensitivity and positive predicting values. The differential comparison metric, although computationally heavy, shows a capability of about 60% in terms of precision in classifying the links between genes starting from our dataset. DSC consistently achieves the best performance among the three methods according to all performance indexes: accuracy, recall, precision and Matthews correlation coefficient. Due to its features of analysis in the frequency domain, DSC processes all pairs of genes directly connected in the band of interest and thus it is particularly insensitive to noise effects. The superiority of the performance of DSC algorithm is evident in the MCC index, which is the most significant in our experiment, as already specified.

Furthermore, in principle DSC can be effectively used even in those situations in which a gene receives more than one signal (i.e. has more than one upstream gene). This is because in frequency domain it is possible to distinguish between the different components produced by signals with different frequencies.

These results indicated that DSC achieves the best performance in terms of accuracy, precision, recall, and MCC, because of the ability to filter noise and to recognize the frequency similarity in directly linked genes. The effectiveness of DSC in identifying relationships among the temporal expression profiles of genes makes it a suitable tool for a wide range of applications beyond the one we performed, such as network discovery.

CHAPTER 4 PATHWAY CROSSTALK

Many methods are available for the analysis of signaling pathways in the context of the interpretation of high-throughput biological data.

The common aspect of all these approaches is that they calculate a p-value that aims to quantify the significance of the involvement of the given pathway in the condition under study.

All existing methods treat the pathways as being independent, i.e. the analysis is performed individually for each pathway. Biological systems, in reality, are not just disconnected entities that operate separately one from each other, but, they are interconnected parts of the entire system that represents a whole organism. Signaling pathways can considerably affect each other through a "crosstalk" phenomenon related to genes that are shared among many pathways. Although it is intuitive that various pathways could influence each other, the presence and extent of this phenomenon have not been rigorously studied and, most importantly, there is no currently available technique able to quantify the amount of such crosstalk. In order to better describe the problem, let us focus momentarily on the simplest pathway analysis approach, the over-representation analysis (ORA) already described in Section 2.4.1. Let us consider the *Non-small-cell lung cancer* (NSCLC) pathway, that contains 54 genes. Given that ORA needs a "reference", we will assume to use an array with 3000 genes. Let us also assume that we are working with an experiment where 50 genes are found to be DE according to some criteria, 5 of which happen to belong to the NSCLC pathway. The ORA analysis, in the form of Fisher's exact test, yields a p-value of 0.00189, indicating that the pathway is significant in the given condition. Now let us also consider the *VEGF signaling pathway* (VEGF). This pathway contains 74 genes of which 26 are also on the NSLC pathway. These 26 common genes effectively "couple" the p-values of the two pathways as follows. If none of the 5 genes found to be DE on NSLC belong to VEGF, and if this pathway does not have any DE genes of its own, this pathway will have zero DE genes yielding a p-value of 1. However, if for instance 3 of the 5 DE genes are in common between

NSLC and VEGF, the p-value for VEGF will go down to 0.1276, not yet significant but much lower then before. Finally, if all 5 DE genes from NSLC happen to be among the genes in common with VEGF, the p-value for VEGF becomes 0.00748 which is now significant. Even though VEGF has been reported being involved in oncogenic processes [97], its role in other conditions such as intense exercise is well documented [80, 4, 36]. In a situation in which the VEGF pathway is highly impacted, for example in a condition such as exercise after a long period of inactivity, the coupling of those two pathways could possibly result in a situation in which the NSLC pathway shows as significantly impacted, therefore resulting as a false positive.

Clearly, the more common genes there are between two pathways, the higher the chance that the more DE genes will fall in the common set, and the tighter the coupling between their p-values will be. At one extreme, two pathways that contain the same set of genes (but perhaps a different layout describing different phenomena) will have a perfect coupling, yielding exactly the same p-values in all cases. At the other extreme, two pathways that have no genes in common will be completely uncoupled.

Given this, it is no surprise when analyzing many experiments, existing methods report as significant pathways that have little to do with the phenotype investigated (false positives - FPs), while pathways that are expected to be impacted are ranked lower and sometimes do not even reach the threshold of significance (false negatives - FNs).

Here, we show that all three major categories of pathway analysis methods (enrichment analysis, functional class scoring, and impact analysis) are severely influenced by crosstalk phenomena. Using real pathways and data, we show that in some cases pathways with significant p-values are not biologically meaningful, and that some biologically meaningful pathways with non-significant p-values become statistically significant when the crosstalk effects of other pathways are removed.

4.1 Crosstalk detection

In order to demonstrate the existence and assess the extent of crosstalk effects, we conducted a systematic exploration of this phenomenon. Identifying such effects in any number of specific real experiments would constitute only anecdotal evidence since the true amount of crosstalk between two given pathways in any given condition is not known. In order to demonstrate the existence and assess the extent of crosstalk effects, we designed and conducted the following systematic exploration of this phenomenon. We first constructed a reference set of genes from the union of all genes present on at least one KEGG signaling pathway (2963 genes at the time). Then, for each pathway P_i, we ran experiments as follows. We calculated the number n_i of DE genes that would make P_i significant at least at $\alpha = 0.01$ after a Bonferroni correction for multiple comparison. Henceforth, we will refer to this pathway as the "bait". We then used the reference set to pick n_i random genes from P_i and $100 - n_i$ genes that are not on P_i, and calculated the Fisher Exact Test significance of all other "prey" pathways, P_j. This essentially models a situation in which 100 genes are found to be DE, and these genes are such that the Fisher Exact Test will find the bait pathway P_i significant at 1% after the correction for multiple comparisons. Since the $100 - n_i$ genes that are not on P_i are randomly chosen among the reference set, no other pathway P_j should have more genes than expected by chance. Under these circumstances, the research hypothesis is true for the bait, while the null hypothesis is true for all other pathways. We repeated this selection 1,000 times for each pathway P_i, and each time we computed the Fisher Exact Test p-value [109], SPIA (impact analysis) [108], and GSEA [105] p-values for all pathways from the KEGG database [60]. With these results, we constructed the empirical distributions of the False Discovery Rate (FDR)-corrected p-values corresponding to each prey P_j. Under the null hypothesis, the p-values are expected to follow an uniform distribution, and to be independent between different pathways. In fact, the distributions of the p-values (see Figure 4.1) are significantly different from the uniform distribution (Kolmogorov-Smirnov goodness of fit p-values of the order of 10^{-16} in all cases). The distributions for all three

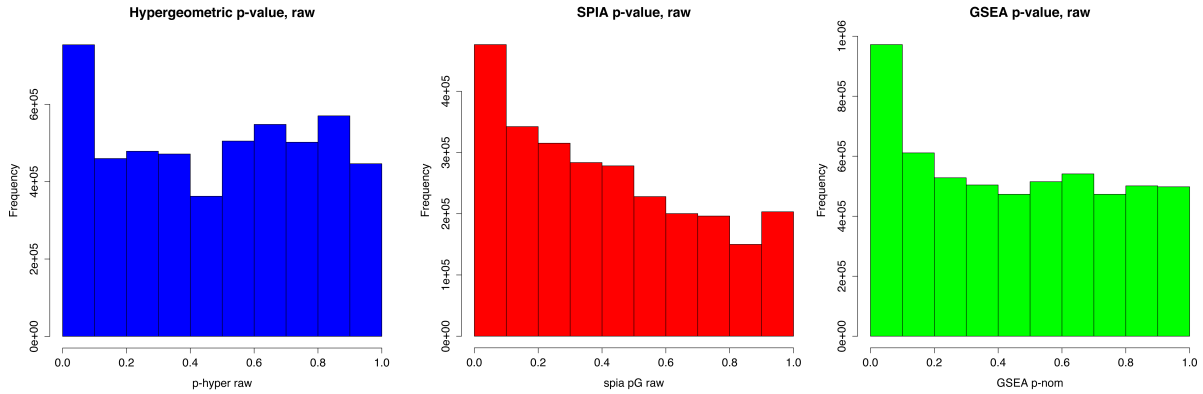


Figure 4.1: The distributions of the p-values obtained from the three analysis methods under the null hypothesis: Fisher's Exact Test (left), SPIA (middle), and GSEA (right). All three exhibit a significant departure from the expected uniform distribution (Kolmogorov-Smirnov p-values of the order of 10^{-16} in all cases). Notably, all methods yield a much higher than expected number of pathways with p-values lower than 0.1, i.e. false positives.

methods are severely skewed towards zero, showing that all methods produce a large number of false positives.

Furthermore, we observed much stronger crosstalk effects for specific pathway pairs (i, j): every time one of them is used as a bait, the p-value of the other one is pulled to values much lower than expected by chance, many times well below the significance threshold. All crosstalk effects can be represented in a crosstalk matrix (left panel in Figure 4.2). In this matrix, the elements $[i, j]$ represent the mean of the distribution of p-values for 1,000 random trials using pathway i as bait and pathway j as prey. This matrix is not symmetrical since the influence of pathway i on pathway j can be different from the influence of pathway j on i due to the different sizes of the two pathways and the presence of different numbers of DE genes in the non-shared portion of each pathway. The matrix shows strong crosstalk between several pathways (e.g. row 3 and columns 57 through 70).

We hypothesized that this crosstalk is due mostly to the genes that are in common between pathways. If this were true, we would expect to see a strong crosstalk between pairs of pathways that have many genes in common and a weak crosstalk between pathways that do not share any genes. In order to test this hypothesis, we calculated the Jaccard similarity index between all pairs of signaling pathways from KEGG. The Jaccard index is defined

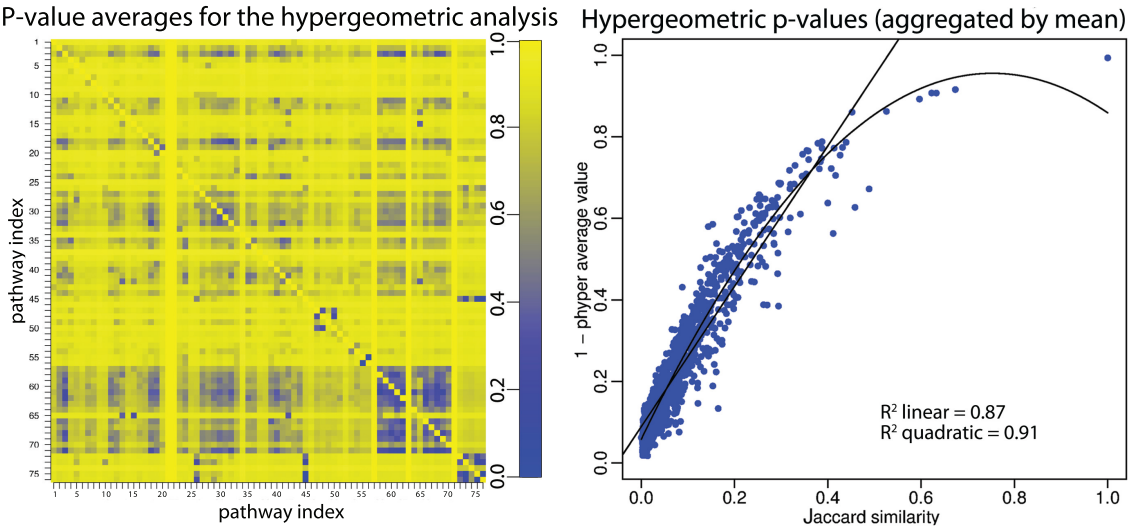


Figure 4.2: Pathway crosstalk in the Fisher Exact Test p-values. Left panel: a number of random genes were chosen from a "bait" pathway i such that its Fisher Exact Test p-value is 0.01. Other genes were chosen randomly from all other pathways (acting as preys), up to a constant number ($n = 100$). The elements $[i, j]$ where $i \neq j$ represent the mean of the distribution of p-values for 1000 random trials using pathway i as bait and pathway j as prey. The elements $[i, i]$ (on the diagonal) represent the classical Fisher Exact Test p-value of pathway i. The data show that a considerable number of pathways influence each other through a "crosstalk" of the p-values. For instance, row 3 of the matrix shows that when pathway 3 is chosen to be significant, several other pathways (e.g. columns 57 to 70) also tend to be significant (dark shades of blue represent significant p-values). Right panel: each point represents the average of the p-values of all the random trials for pairs with the same Jaccard index. The lines represent the fitting of linear and a quadratic models. Both models show a strong dependence between the p-value crosstalk and the Jaccard index. Similar results were obtained for GSEA and impact analysis (see Figures 4.3 and 4.4).

as $\frac{|P_i \bigcap P_j|}{|P_i \bigcup P_j|}$, and characterizes the overlap between two sets, relatively to the size of their union. Pathways that share many genes will have a large Jaccard index. The right panel in Figure 4.2 shows the relationship between the Fisher Exact Test p-values and the Jaccard index for all pathway pairs.

The data shows a very strong correlation between the two (Pearson correlation index of 0.87), which confirms our hypothesis that the crosstalk can be explained by the presence of genes that are involved in more than one pathway. Very similar results have been obtained for FCS analysis (GSEA) and for the impact analysis (SPIA) (see Figures 4.3 and 4.4). The Pearson correlation between the p-values provided by GSEA and the Jaccard indices of all KEGG pathways was 0.62, while in the case of SPIA the correlation was 0.83, which confirms our hypothesis that the crosstalk can be explained by the presence of genes that are involved in more than one pathway.

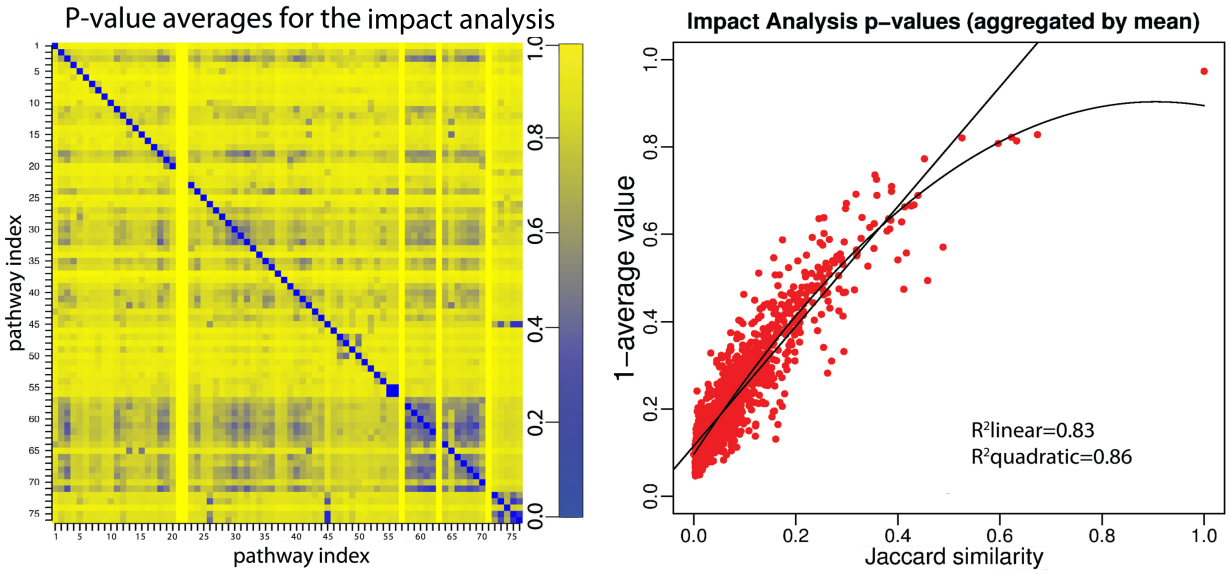


Figure 4.3: Pathway crosstalk in the impact analysis p-values. Left panel: a number of random genes were chosen from a "bait" pathway i such that its impact analysis p-value is 0.01. Other genes were chosen randomly from all other pathways (acting as preys), up to a constant number ($n = 100$). The elements $[i, j]$ where $i \neq j$ represent the mean of the distribution of p-values for 1000 random trials using pathway i as bait and pathway j as prey. The elements $[i, i]$ (on the diagonal) represent the impact analysis p-value of pathway i. The data show that a considerable number of pathways influence each other through a "crosstalk" of the p-values. Right panel: each point represents the average of the p-values of all the random trials for pairs with the same Jaccard index. The lines represent the fitting of linear and a quadratic models. Both models show a strong dependence between the p-value crosstalk and the Jaccard index. Similar results were obtained for GSEA and the classical ORA (see Figures 4.2 and 4.4).

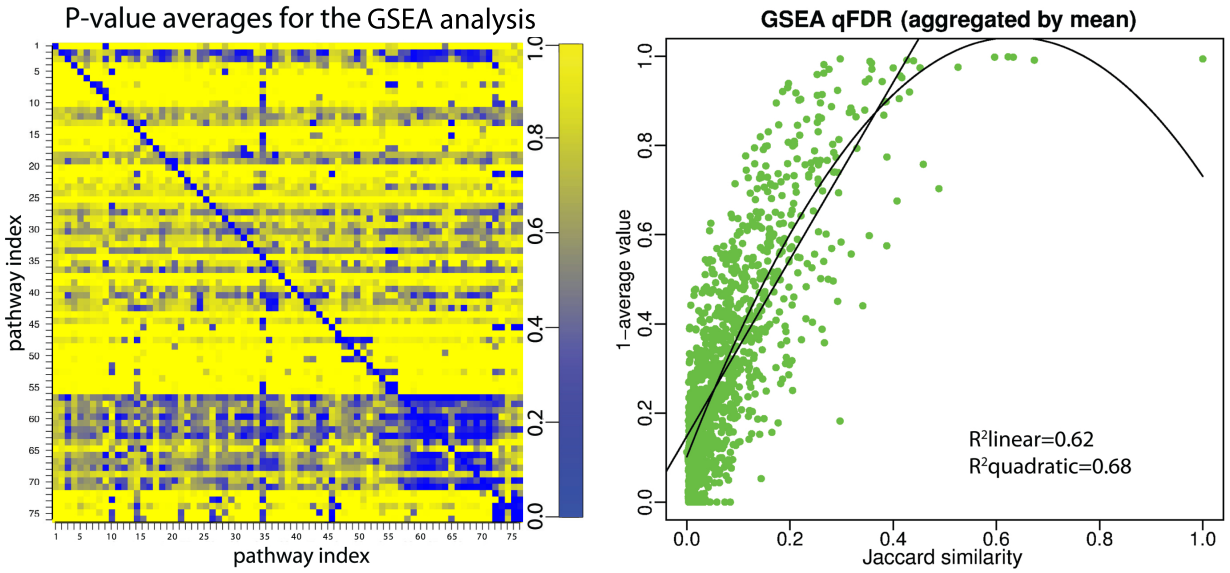


Figure 4.4: Pathway crosstalk in the GSEA p-values. Left panel: a number of random genes were chosen from a "bait" pathway i such that its GSEA p-value is 0.01. Other genes were chosen randomly from all other pathways (acting as preys), up to a constant number ($n = 100$). The elements $[i, j]$ where $i \neq j$ represent the mean of the distribution of p-values for 1000 random trials using pathway i as bait and pathway j as prey. The elements $[i, i]$ (on the diagonal) represent the GSEA p-value of pathway i. The data show that a considerable number of pathways influence each other through a "crosstalk" of the p-values. Right panel: each point represents the average of the p-values of all the random trials for pairs with the same Jaccard index. The lines represent the fitting of linear and a quadratic models. Both models show a strong dependence between the p-value crosstalk and the Jaccard index. Similar results were obtained for the classical ORA and impact analysis (see Figures 4.2 and 4.3)

4.1.1 Fat remodeling in obese mice

In addition to the simulated data, we analyzed an experiment investigating cellular and metabolic plasticity of white fat tissue (WAT), where the classical over-representation analysis (ORA) produced a number of false positives, and failed to rank highly pathways that were known to be involved in the given condition.

In this experiment, the chronic activation of WAT β-adrenergic receptors by certain physiological and pharmacological conditions transforms the tissue into one resembling brown fat, a thermogenic organ [43, 73, 84]. The dataset was obtained from a microarray analysis of white fat from mice treated with low dose (0.75 nmol/hr) CL 316,243 (CL) for 0, 3 and 7 days. The top 20 pathways ranked by ORA and their associated FDR-corrected p-values for the comparison between expression levels of genes at days 3 and 0 are shown in Table 4.1. In this figure, pathways highlighted in red represent pathways not related with the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, that they are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis. The three most significant pathways in the comparison between days 3 and 0 were *Parkinson's, Alzheimer's* and *Huntington's* diseases. The fourth pathway in the ranked list is *Leishmaniasis*. The first three pathways describe degenerative diseases of the central nervous system that have no connection to fat remodeling. *Leishmaniasis* describes the signaling involved in a disease spread by the bite of certain species of sand flies. Clearly, this pathway is also unlikely to give insights about the fat remodeling phenomenon. While other pathways such as *Phagosome* [87], *PPAR Signaling* [43], and *Cell cycle* [68], are definitely more related to the phenomenon of fat remodeling, their presence in the middle of a ranked list dominated by false positives (6 false positives in the 10 pathways significant at 1%) illustrates how these results do not describe the phenomenon in analysis, and they cannot be considered reliable.

rank	pathway	pval(FDR)
1	Parkinson's disease	$2.0e^{-06}$
2	Alzheimer's disease	$3.6e^{-06}$
3	Huntington's disease	$3.4e^{-05}$
4	Leishmaniasis	0.0003
5	Phagosome	0.0006
6	Cell cycle	0.0011
7	Oocyte meiosis	0.0016
8	Cardiac muscle contraction	0.0016
9	Toll-like receptor	0.0018
10	PPAR signaling pathway	0.0018
11	Chemokine signaling pathway	0.0154
12	Lysosome	0.0211
13	B cell receptor	0.0252
14	Systemic lupus erythematosus	0.0292
15	Compl. and coagulation cascades	0.0342
16	Cytokine-cytokine rec. inter.	0.0346
17	Chagas disease	0.0466
18	Progesterone mediated oocyte maturation	0.0530
19	Fc epsilon RI signaling pathway	0.0548
20	Leukocyte transendothelial migration	0.0548

Table 4.1: The results of the ORA analysis in the fat remodeling experiment for the comparison between days 3 and 0. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05. Pathways highlighted in red represent pathways not related with the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, that they are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis. The top four pathways are not related to fat remodeling. Although there are a number of pathways that are related to this phenomenon, the presence of many obvious false positives makes the results difficult to interpret.

A similar situation can be seen in the comparison between expression levels of genes at days 7 and 0. The results of the classical ORA are shown in Figure 4.2 (only the top 20 pathways are shown). The only significant pathways at the 5% level are *Parkinson's disease*, *Cell Cycle* and *Huntington's disease*. As discussed, *Parkinson's disease* has little to do with the tissue remodeling phenomenon. The *Cell Cycle* pathway is likely to be related to tissue remodeling [68], and *p53 Signaling* is known to be a central pathway in the response to cellular stress, including inflammation, and related to processes like cellular senescence and cell cycle [55], while *Hungtinton's disease* is a neurodegenerative condition that results in movement, thinking and psychiatric disorders. With four false positives in the top five pathways, the results of the classical ORA are distorted to the point of being useless.

These results might point to the conclusion that either the data are not reliable, or that the method used for pathway analysis is not able to correctly detect the underlying biological phenomenon, or even that the pathways available for the analysis are not representing correctly any of the processes involved in the phenomenon of fat remodeling. In order to understand the reason for the presence of so many obvious false positives in the results we looked at each of the three top pathways *Alzheimer's, Parkinson's*, and *Huntington's*. The sizes in genes of these three pathways are, respectively 155, 111, and 165, considering only genes that are in the microarray used for the experiment. We then looked at the common genes among these pathways, and we found that 80 genes are in the intersection between Alzheimer's and Parkinson's, 88 genes are in the intersection between Alzheimer's and Huntington's, and 87 genes are in the intersection between Parkinson's and Huntington's. Even more interesting, *79* genes are in common among the three intersections, showing that the three pathways share a common module that constitutes 50% of the Alzheimer's, 71% of the Parkinson's, and 47% of the Huntington's pathway. In other words, this common module constitutes a significant part of each of the three pathways in analysis. In terms of DE genes, the situation is even more extreme. The Alzheimer's pathway contains 32 DE genes, the Parkinson's 27, and the Huntington's 31. However, 26 DE genes are in common between

rank	pathway	p(FDR)
1	Parkinson's disease	$7.2e^{-06}$
2	Huntington's disease	$4.2e^{-05}$
3	Alzheimer's disease	0.0002
4	Cell cycle	0.0044
5	Cardiac muscle contraction	0.0087
6	p53 signaling pathway	0.0134
7	PPAR signaling pathway	0.0773
8	Gap junction	0.0920
9	Progesterone mediated oocyte maturation	0.0995
10	Oocyte meiosis	0.1327
11	Salivary secretion	0.1442
12	Cell adhesion molecules (CAMs)	0.2390
13	SNARE interactions in vesicular transport	0.2969
14	Prostate cancer	0.3837
15	Vasopressin-regulated water reabsorption	0.5111
16	Arrhythmogenic right ventricular cardiomyopathy	0.5111
17	Hedgehog signaling pathway	0.5174
18	Prion diseases	0.5420
19	Melanogenesis	0.5432
20	Pathways in cancer	0.5432

Table 4.2: The results of the ORA analysis in the fat remodeling experiment for the comparison between days 7 and 0. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05. Pathways highlighted in red represent pathways not related with the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, that they are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis.

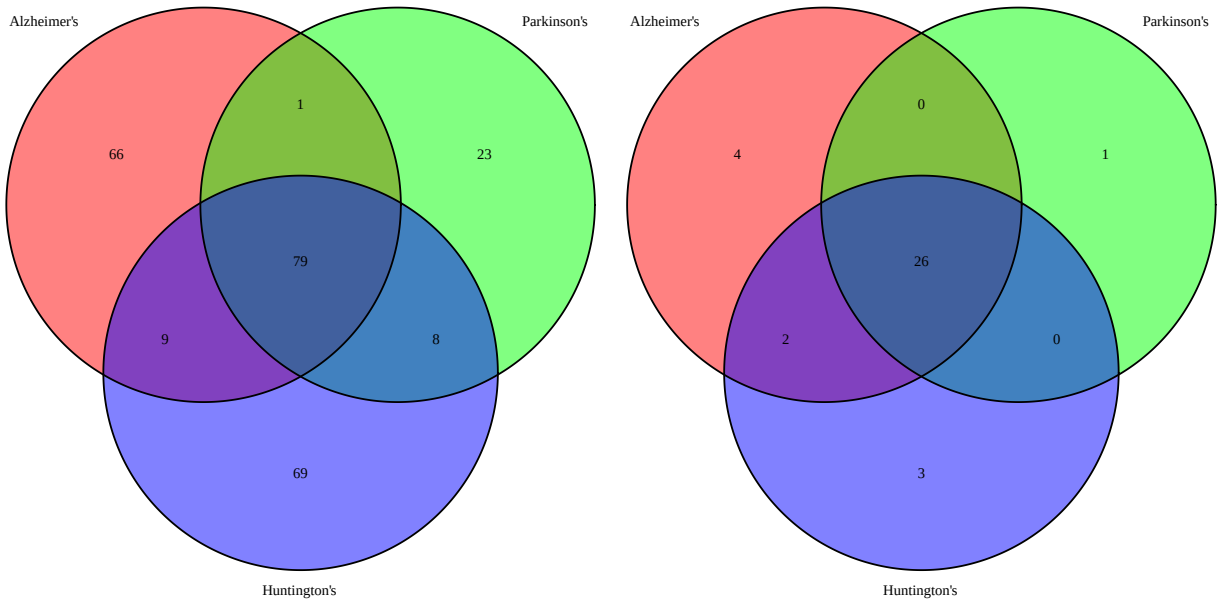


Figure 4.5: Common genes in the Alzheimer's (light red), Parkinson's (green), and Huntington's (blue) pathways. The left panel shows the intersections the three pathways, while the right panel shows the intersection among the DE genes belonging to each pathway. The intersection among DE genes indicates that a common mechanism among the three pathways is responsible for the phenotype, and the ORA is not able to correctly detect such mechanism, as it does not take into account crosstalk among pathways.

the three pathways, and only 4, 1, and 3 respectively are in only one of the pathways. Summarizing, the intersection of the three pathways at the top of both the lists of significant pathways contains **most of the differentially expressed genes** of those pathways. This situation is represented in Figure 4.5.

This indicates that the intersection itself would be the *meaningful biological mechanism* responsible for the phenotype observed, but the inability of ORA to account for overlap among pathways makes it impossible to identify such phenomenon.

4.2 Identification and correction of crosstalk effects

Recently, we proposed the first approach able to i) *detect crosstalk effects* when they exists, ii) *correct for them*, resulting in a more meaningful ranking among pathways in a specific biological condition, and iii) identify *novel functional modules* that can play an independent role and have different functions than the pathway they are currently located on. This method allows for a better understanding of individual experiment results, as well as a more refined definition of the existing signaling pathways for specific phenotypes. This

method takes as input a set of reference pathways and a list of genes that are DE in the given condition. The crosstalk analysis is composed of three steps: detection of crosstalk effects, identification of independent functional modules, and correction. The next sections describe in details these steps.

4.2.1 Detection of crosstalk effects: the crosstalk matrix

The main issue we are trying to address here is the fact that in the presence of overlapping pathways (i.e. for all pathways databases available today) crosstalk phenomena increase the probability of false positives, i.e. increase the number of pathways reported as significant but that in reality are not interesting (borrowing terminology from Brad Efron, we call pathways that have lesser biological significance "not interesting" even though they might be statistically significant with a large enough sample size). To better understand the approach we are going to present, let us briefly review the classical Fisher Exact Test approach described above. Figure 4.6a represents the contingency table used for assessing the significance of a pathway P_i by the classical over-representation (ORA) approach. The table divides genes as either being in the pathway or not, versus being considered DE or not DE (NDE); n_i represents the number of DE genes on P_i, while n represents the total number of DE genes, and m_i represents the number of NDE genes on P_i while m represents the total number of NDE genes. It follows that $n_i + m_i = |P_i|$ represents the number of genes on P_i, while with $n + m$ we represent the total number of genes.

The reasoning behind the ORA is that if the number of DE genes on a pathway is much higher than expected by chance, then the pathway is likely to be biologically interesting. In order to take into account the effect of the overlap on the significance of the two pathways we consider the effect of the removal of the overlapping part on the significance of the pathways. This is achieved as follows: let us consider two overlapping pathways P_i and P_j. With the notation $P_{i\setminus j}$ we define the set of elements in P_i excluding the intersection with P_j; in the same way, with the notations $n_{i\setminus j}+m_{i\setminus j}$ we represent the number of genes that are in pathway P_i but not in pathway P_j, and with $n_{i\setminus j}$ the number of DE genes that are on pathway P_i

	DE	NDE	Total
P_i	n_i	m_i	$n_i + m_i$
P_i^c	$n - n_i$	$m - m_i$	$(n+m) - (n_i + m_i)$
Total	n	m	$n+m$

(a) Standard over-representation approach contingency table; $n_i + m_i$ and $n + m$ represent, respectively, the number of genes belonging to pathway P_i and the total number of genes. n_i and n represent, respectively, the number of differentially expressed genes belonging to pathway P_i and the total number of DE genes.

	DE	NDE	Total
$P_{i\setminus j}$	$n_{i\setminus j}$	$m_{i\setminus j}$	$n_{i\setminus j} + m_{i\setminus j}$
$P_{i\setminus j}^c$	$n - n_{i\setminus j}$	$m - m_{i\setminus j}$	$(n+m) - (n_{i\setminus j} + m_{i\setminus j})$
Total	n	m	$n+m$

(b) Contingency table for over-representation approach taking in account the overlap between pairs of pathways; $P_{i\setminus j}$ represents the set of elements in P_i excluding the intersection with P_j; with the notations $n_{i\setminus j} + m_{i\setminus j}$ we represent the total number of genes that are in pathway P_i but not in pathway P_j, and with $n_{i\setminus j}$ the number of DE genes that are on pathway P_i but not in pathway P_j.

Figure 4.6: A comparison of the classical over-representation analysis (left) with the crosstalk matrix analysis proposed here (right).

but not in pathway P_j. We then consider the contingency table shown in Figure 4.6b, whose bottom margin is identical to that of Figure 4.6a.

With this contingency table, we compute for every pair of pathways $[i, j]$ the p-value of $P_{i\setminus j}$. Since this computation yields an $k \times k$ matrix, where k is the number of pathways, the results are most conveniently represented using a matrix visualized as a heat map of the negative log p-values, where each cell (i, j) of this matrix characterizes the significance of pathway P_i when we remove the effect of pathway P_j. The rows and the columns are ordered by the original p-values of the pathways, which are placed on the diagonal. We will refer to this matrix as the *crosstalk matrix*. This matrix is useful for identifying the effects of crosstalk among pathways.

An example of the crosstalk matrix can be found in Figure 4.7. We will refer to the part of the matrix above the horizontal significance threshold as the *significance strip*. The *non-significance* strip will be the part below the horizontal significance threshold. The *significance quadrant* will be the part of the significance strip to the left of the significance threshold. Using these terms, we can identify and discuss several interesting phenomena that are not

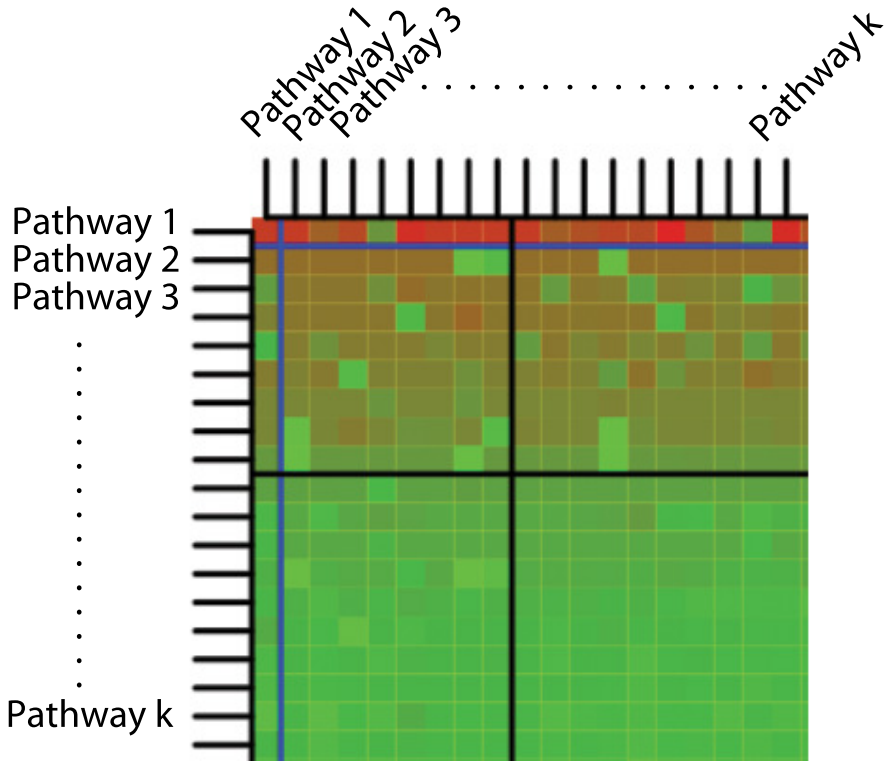


Figure 4.7: Example of a crosstalk matrix. On the diagonal we find the classical over-representation analysis, ordered by p-value. The blue line represents the 0.01 significance level, while the black line represents the 0.05 significance level. The p-values in the matrix have been log-transformed (base 10 log) and the sign of the result has been inverted. The color of the cell represents the p-value: bright red for p-values close to zero, bright green for p-values close to 1.

captured by any of the existing pathway analysis methods.

A first interesting case is when a pathway P_i is reported as significant by the classical analysis, but it *loses* its significance when the effect of another pathway P_j is removed. This is represented, in the crosstalk matrix, by a non significant p-value (green square) in the significance strip. In this case P_i is unlikely to be biologically meaningful, since its significance is most likely due to a crosstalk from P_j.

A second interesting case is when a pathway P_i that is *not* significant for the classical analysis *becomes* significant when the crosstalk effect of another pathway P_j is removed. This is represented in the crosstalk matrix by a significant p-value (red square) in the non-

significant strip. The meaning of this is that pathway P_j was *masking* the significance of P_i, indicating that a phenomenon likely to be biologically meaningful is happening in the part of P_i which is not in common with P_j.

A third and last interesting case is a symmetric (with respect to the diagonal) decrease in significance of pathways in the significance quadrant. This indicates the presence of an independent functional sub-module, common to both P_i and P_j, that is responsible for their significance. Note that the activity of this module is tightly related to the condition studied.

4.2.2 The maximum impact estimation: an expectation maximization technique for the assessment of the significance of signaling pathways in presence of crosstalk

The crosstalk matrix is a useful tool for the interpretation of the effect of crosstalk between pathways. However, the ultimate goal of the analysis of signaling pathways is to provide a meaningful ranking among pathways, as well as a p-value quantifying the likelihood that a certain pathway is involved in the phenomenon in analysis. Here, we developed a correction method for the ranking of pathways that takes into account the overlaps between pathways.

The main idea is that if there is no crosstalk, i.e. if each gene contributes to one and only one pathway, then there is no ambiguity in the ORA significance calculations. In such a case, if genes in a pathway are over-represented, the pathway is not a false positive caused by crosstalk. Our approach is therefore to infer an underlying *pathway impact matrix* where each gene contributes to one and only one pathway, hence is devoid of crosstalk, and then to perform the ORA using that impact matrix. Since this underlying pathway impact matrix is not observed directly, it is inferred through likelihood-based methods, and estimated using the expectation maximization (EM) algorithm. The corrected ranking is computed using ORA with the underlying pathway impact matrix, shown as follows.

	Y	P_1	P_2	P_3	...	P_k
g_1	1	0	1	1	$\cdots$	0
g_2	1	0	1	0	$\cdots$	0
g_3	1	1	0	0	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g_{n-1}	1	0	0	1	$\cdots$	0
g_n	1	0	1	0	$\cdots$	0
g_{n+1}	0	0	0	1	$\cdots$	0
g_{n+2}	0	1	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
g_{n+m-1}	0	1	0	0	$\cdots$	0
g_{n+m}	0	0	0	0	$\cdots$	0

Figure 4.8: Example of a DE/membership matrix; the column Y represents the indicator of differential expression of the various genes (1 for the n DE genes and 0 for the m NDE). Column P_j represents the membership indicator for pathway j. Row g_i describes gene i in terms of its differential expression and its membership to the various pathways.

Let us consider the DE indicator vector Y, representing the differential expression of genes, and the membership matrix X describing the membership of each gene in each one of k pathways $P_1 \dots P_k$. The vector Y is defined as follows:

$$Y_i = \begin{cases} 1 & \text{if } g_i \text{ is DE} \\ 0 & \text{if } g_i \text{ NDE} \end{cases}$$

and each cell $X_{i,j}$ of the matrix X is defined as follows:

$$X_{ij} = \begin{cases} 1 & \text{if } g_i \text{ belongs to } P_j \\ 0 & \text{if } g_i \text{ does not belong to } P_j \end{cases}$$

The matrix $Y|X$ obtained by combining the vector Y with the X matrix is shown in the example in Figure 4.8.

In many analysis methods, the membership matrix X is also interpreted as the *impact matrix*: if $X_{ij} = 1$, then gene g_i *impacts* pathway P_j. In ORA, for example, each gene is considered to have the same full impact on all pathways the gene belongs to. Crosstalk effects

result from the fact that a gene can belong to more than one pathway, but in principle, it can potentially have a different biological impact on each such pathway. Our aim is to identify the pathway where the biological impact of such a shared gene is maximum. We do so by estimating the maximum impact pathway using an expectation maximization approach as described in the following.

Assuming that in a specific biological condition each gene distributes its impact differently to each pathway, we will consider the pathway to which each gene distributes the greatest fraction of its impact. We define a binary matrix Z that indicates, for each gene, the pathway that receives the biggest fraction of that gene's impact. For each gene g_i, the corresponding row $Z_i = [Z_{i1}, Z_{i2}, \ldots, Z_{ik}]$, where $Z_{ij} \in \{0, 1\}$, will have $\sum_{j=1}^{k} Z_{ij} = 1$, i.e. there is only one column in each row that has a non-zero element. This matrix Z is the *unknown underlying pathway impact matrix* referred to above; our goal is to estimate it.

Let us consider one row Z_i having a one in an unknown column j and zeros elsewhere. Since we don't know j, we compute the probability of each pathway to be the one where gene g_i gives the greatest fraction of its impact. To do this, we assume a non-negative vector of multinomial probabilities $\Pi = (\pi_1, \ldots, \pi_k)$ with $\sum_{j=1}^{k} \pi_j = 1$, defined by $\pi_j = p(Z_{ij} = 1|Y_i = 1)$. In other words, given a gene g_i that is DE, π_j is the probability that g_i gives the greatest fraction of its impact to P_j. Similarly, we also define $\Theta = (\theta_1, \ldots, \theta_k)$, where $\theta_j = p(Z_{ij} = 1|Y_i = 0)$ for the NDE genes.

Row i of the membership matrix X is denoted by X_i; this vector tells us which pathways gene i belongs to. Within the context of the probabilistic model described above, each row X_i can be interpreted as an observation of a gene with a given expression state Y that gives the greatest fraction of its impact to one of the pathways it belongs to. Therefore, for DE

genes we have $p(X_i = x_i|Y_i = 1, \Pi) = \Pi \cdot x_i'$. We further assume that the hidden matrix Z is consistent with the observed X, i.e., Z_{ij} can be 1 only when $X_{ij} = 1$; if $X_{ij} = 0$ then we must have $Z_{ij} = 0$ (a gene cannot contribute most to a pathway that it does not belong to). With this notation:

$$\begin{aligned} p(Z_i = z_i|X_i = x_i, Y_i = 1, \Pi) &= \frac{p(Z_i = z_i, X_i = x_i|Y_i = 1, \Pi)}{p(X_i = x_i|Y_i = 1, \Pi)} \\ &= \frac{I(z_i \cdot x_i' = 1) \cdot \Pi \cdot z_i'}{\Pi \cdot x_i'} \end{aligned} \tag{4.1}$$

where $I(.)$ is the indicator function. For example, if $x_i = (11001)$ and g_i is a DE gene, then the conditional distribution of Z_i is given by:

$$\begin{aligned} &p(Z_i = (10000)|X_i = x_i, Y_i = 1, \Pi) = \pi_1/(\pi_1 + \pi_2 + \pi_5) \\ &p(Z_i = (01000)|X_i = x_i, Y_i = 1, \Pi) = \pi_2/(\pi_1 + \pi_2 + \pi_5) \\ &p(Z_i = (00100)|X_i = x_i, Y_i = 1, \Pi) = 0 \\ &p(Z_i = (00010)|X_i = x_i, Y_i = 1, \Pi) = 0 \\ &p(Z_i = (00001)|X_i = x_i, Y_i = 1, \Pi) = \pi_5/(\pi_1 + \pi_2 + \pi_5) \end{aligned} \tag{4.2}$$

This yields a vector of conditional probabilities $c_i = (c_{i1}, c_{i2}, ..., c_{ik})$ for each row Z_i of DE genes, where $c_{ij} = p(Z_{ij} = z_{ij}|\ X_i = x_i)$ as defined above. Once those probabilities are estimated, we can produce a most likely matrix Z by assigning each gene to the pathway with the highest probability of receiving the biggest fraction of the impact of the gene. Specifically, $z_{ij} = 1$ when $\max_s\{c_{is}\} = c_{ij}$; $z_{ij} = 0$ otherwise.

If there were no crosstalk, each gene would contribute to a single pathway, the matrix X and the matrix Z would be equal, and they would have only one element equal to 1 in each row. In this case, π_j could be estimated as the number of DE genes belonging to the pathway divided by the total number of DE genes. The probabilities π and θ could be estimated as follows:

$$\widehat{\pi}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{4.3}$$

$$\widehat{\theta}_j = \frac{\sum_{i=n+1}^{n+m} x_{ij}}{m} \tag{4.4}$$

In the presence of crosstalk, however, it is not possible to compute Π and Θ directly from X. A likelihood-based estimation can be used instead.

The log-likelihood of observing the membership matrix X given the gene expression vector Y is then:

$$logL = \sum_{i=1}^{n+m} log(p(X_i|Y_i; \pi_1, \pi_2, \pi_3 \ldots \pi_k, \theta_1, \theta_2, \theta_3 \ldots \theta_k)) \tag{4.5}$$

Equation 4.5 is written under the assumption of conditional independence of rows of X; i.e., under the reasonable assumption that the pathway to which a gene i gives most of its impact does not depend on the pathway to which another gene j impacts the most. In other words, the split of the fractions of the impact of a gene does not depend the splits of the impact of other genes.

This assumption, together with the observation that the DE genes do not depend on

θ's and that the NDE genes do not depend on π's, allows us to compute the likelihood by separating the matrix in two sub-matrices: $X|Y = 1$, representing the sub-matrix of the DE genes, and $X|Y = 0$, representing the sub-matrix of the NDE genes:

$$\begin{aligned} logL &= \sum_{i=1}^{n} log(p(X_i|Y_i = 1, \Pi)) + \sum_{i=n+1}^{m+n} log(p(X_i|Y_i = 0, \Theta)) \\ &= \sum_{i=1}^{n} log(\Pi \cdot X_i') + \sum_{i=n+1}^{m+n} log(\Theta \cdot X_i') \end{aligned} \tag{4.6}$$

In this formula, the (row) vector Π represents the probability of the $i-th$ DE gene to give the greatest fraction of its impact to a specific pathway, X_i is the $i-th$ row of the membership matrix X, and X_i' represents its transpose. The dot-product $\Pi \cdot X_I'$ produces a scalar representing the probability $P(X_i = x_i|Y = 1, \Pi)$, i.e. the probability of observing the i-th row of the matrix X_i given the fact that gene i is DE. The same notation has been used for the dot-product $\Theta \cdot X_i'$.

In the following, we will only work with the first term to illustrate how to estimate Π. Θ can be estimated from $X|Y = 0$ in a similar fashion.

There is no closed form solution for the maximization of Eq. 4.6. However, we can use the Z matrix as a hidden variable for the estimation of the parameters Π. The log joint conditional likelihood for the DE part of the matrix can be written as:

$$
\begin{aligned}
logJL^{DE} &= log(p(X, Z|Y = 1, \Pi)) \\
&= \sum_{i=1}^{n} log(p(X_i, Z_i|Y_i = 1, \Pi)) \\
&= \sum_{i=1}^{n} log(I(Z_i^{DE} \cdot (X_i^{DE})' = 1) \cdot Z_i^{DE} \cdot \Pi) \\
&= \sum_{i=1}^{n} \left(log(I(Z_i^{DE} \cdot (X_i^{DE})' = 1)) \cdot \sum_{j=1}^{k} z_{i,j}^{DE} \cdot log(\pi_j) \right) \\
&= \sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE}) + \sum_{i=1}^{n} \sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}
\end{aligned} \tag{4.7}
$$

We use an expectation maximization (EM) approach to maximize the log likelihood in Equation 4.5 by maximizing the joint log likelihood defined in Equation 4.7. The EM is an iterative algorithm that starts with an initial guess for Π, denoted with Π^0; each iteration is a mapping between Π^t and Π^{t+1}. The superscript indicates the index of the iteration. We choose to initialize each element of the vector as follows:

$$
\pi_j^0 = \frac{\sum_{i=1}^{n} x_{i,j}}{\sum_{i=1}^{n} \sum_{h=1}^{k} x_{i,h}}, \; j \in \{1 \dots k\} \tag{4.8}
$$

This initializes each value π_j with the ratio between the number of DE genes in pathway j and the sum over the matrix X. This initialization is consistent with the model described in Equation 4.3.

Each iteration of the EM algorithm is composed by two steps: the expectation step and the maximization step; during the expectation step we compute the expectation of the log joint conditional likelihood in Equation 4.7 with respect to the posterior $p(Z_{i,j}^{DE}|X_i^{DE}, \Pi^{old})$:

$$
\begin{aligned}
&E\left(\sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE}) + \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right) \\
&= E\left(\sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE})\right) + E\left(\sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right) \\
&= E\left(\sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right) \qquad (4.9)
\end{aligned}
$$

The term $E\left(\sum_{i=1}^{n} log(\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE})\right)$ is equal to 0 because the term $\sum_{j=1}^{k} z_{i,j}^{DE} \cdot x_{i,j}^{DE}$ is equal to 1 for the consistency of Z with X.

The derivation of the non zero term of the expectation is as follows:

$$
\begin{aligned}
E\left(\sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot z_{i,j}^{DE}\right) &= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot E\left(z_{i,j}^{DE} | X_{i,j}^{DE}, \Pi^{old}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot p(z_{i,j}^{DE} = 1 | X_i^{DE}, \Pi^{old}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot \frac{p(z_{i,j}^{DE}, X_i^{DE} | \Pi^{old})}{\sum_{r=1}^{k} p(z_{i,j}^{DE}, X_i^{DE} | \Pi^{old})} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{k} log(\pi_j) \cdot \frac{x_{i,j}^{DE} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,j}^{DE} \cdot \pi_r^{old}} \qquad (4.10)
\end{aligned}
$$

The maximization of the expectation with respect to Π, subject to the constraint that $\sum_{j=1}^{k} \pi_j = 1$, is obtained with the Lagrange multiplier method as follows:

$$\frac{d[\sum_{j=1}^{k} log(\pi_j) \sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} + \lambda((\sum_{j=1}^{k} \pi_j) - 1)]}{d\pi_h} = 0, \forall h \in \{1 \dots k\}$$

$$\frac{\sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h} + \lambda = 0, \forall h \in \{1 \dots k\} \tag{4.11}$$

We can write a systems of equations over all the possible values of h in order to compute λ.

$$\begin{cases} \frac{\sum_{i=1}^{n} \frac{x_{i,1} \cdot \pi_1^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r}}{\pi_1}^{old} + \lambda & = 0 \\ \vdots & \\ \frac{\sum_{i=1}^{n} \frac{x_{i,k} \cdot \pi_k^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_k} + \lambda & = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} \frac{x_{i,1} \cdot \pi_1^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} & = -\lambda \cdot \pi_1 \\ \vdots & \\ \sum_{i=1}^{n} \frac{x_{i,k} \cdot \pi_k^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} & = -\lambda \cdot \pi_k \end{cases}$$

Summing left and right sides we obtain:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} = -\lambda \sum_{j=1}^{k} \pi_j \tag{4.12}$$

Since $\sum_{j=1}^{k} \pi_j = 1$, we can write:

$$\lambda = -\sum_{j=1}^{k}\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} \tag{4.13}$$

We substitute λ in 4.11 and use an iterative process in which a new π value is calculated at each step:

$$\begin{aligned}
\frac{\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h^{new}} + \lambda = 0, \forall h \in \{1 \dots k\} \\
\frac{\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h^{new}} - \sum_{j=1}^{k}\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}} = 0, \forall h \in \{1 \dots k\} \\
\frac{\sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\pi_h^{new}} = \sum_{j=1}^{k}\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}, \forall h \in \{1 \dots k\} \\
\pi_h^{new} = \frac{\sum_{i=1}^{n} \frac{x_{i,h} \cdot \pi_h^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}{\sum_{j=1}^{k}\sum_{i=1}^{n} \frac{x_{i,j} \cdot \pi_j^{old}}{\sum_{r=1}^{k} x_{i,r} \cdot \pi_r^{old}}}, \forall h \in \{1 \dots k\}
\end{aligned} \tag{4.14}$$

Since the sum over each row is 1, if we invert the order of the summations at the denominator in the last row of Equation 4.14, the value of the denominator becomes n. This, in other words, means that each value π_h^{new} is the sum of column h over the number of DE genes.

The algorithm stops when the distance between two consecutive vectors $||\Pi^{(t)} - \Pi^{(t-1)}||$ is less than the quantity $\frac{||\Pi^{(1)} - \Pi^{(0)}||}{100}$, i.e. the distance between the first two vectors divided by 100. At the end of the steps of the EM algorithm we obtain the matrix C from which we

can obtain the most probable Z given the condition under study: for each row, we assign the value 1 to the cell with the highest probability, and 0 to all the others. This is equivalent to saying that each gene gives its full impact to the pathway with the highest π value.

4.2.3 Independent functional modules detection.

The maximum impact estimation procedure alone is not be able to identify overlapping modules responsible for the entire significance of other pathways, as in the situations represented by case 3 in the section describing the crosstalk matrix. In such cases the overlap should be considered as a separate pathway that is more likely to be biologically meaningful in the condition under analysis. An additional step is needed in order to correctly deal with this situation. In this additional step, we extract certain significant overlaps from the list of pathways, and include them in the list as *independent functional modules*. An independent functional module is a module for which there is evidence of an activity independent of the pathways it resides in, for the given condition. If an independent module is found in more than one, possibly unrelated, conditions, this module is considered as a *candidate novel pathway*.

A module must satisfy certain conditions in order to be treated as an independent functional module. Let us assume that we are analyzing the overlap between the pathways P_i and P_j; the first condition is that both pathways are significant (after FDR correction for multiple comparisons) at a certain threshold α. The threshold alpha is the significance threshold chosen by the user. Typical values for this threshold are 0.01 and 0.05. This condition limits the search to the significance quadrant of the crosstalk matrix. The second condition is that the overlap $P_{i \bigcap j}$ itself must be significant at α (after FDR correction). The set of pathways over which we perform the correction for multiple comparison is the set of original pathways,

without the two pathways P_i and P_j, and with the inclusion of the pathways $P_{i\setminus j}$, $P_{j\setminus i}$, and the module $P_{i\bigcap j}$. If the original list of pathways contained k pathways, the correction of the significance of the module is computed on a list that contains $k+1$ pathways. The third condition is that the sub-pathways obtained by removing the overlap from both original pathways, indicated by $P_{i\setminus j}$ and $P_{j\setminus i}$, must *not* be significant at α (after FDR correction). If we denote with $p(P)$ the p-value of a generic pathway P, then the conditions can be summarized as follows:

1. $p(P_i) < \alpha, p(P_j) < \alpha$
2. $p(P_{i\bigcap j}) < \alpha$
3. $p(P_{i\setminus j}) \geq \alpha, p(P_{j\setminus i}) \geq \alpha$

This pairwise procedure might yield modules that are similar one to each other, for example in cases where a module is contained in three or more pathways. That could be solved with a three-way or n-way search, but we opted for another approach for limiting the number of new modules. Once all interesting pairwise modules are created, we test for similarity among modules. The index used for similarity is a modified Jaccard Similarity index mJS defined as follows:

$$mJS = \frac{|M_1 \bigcap M_2|}{min(|M_1|, |M_2|)} \tag{4.15}$$

where M_1 and M_2 are two modules obtained with the search criteria explained above. We merge any two modules similarity is greater that a certain threshold st. Once the modules are merged, the similarity among all the modules (including the newly created one) is computed again, and the merging procedure is applied again until there are no more modules that can

be merged.

This newly created modules are removed from all pathways with which they overlap, and this list of modified pathways is used in the EM procedure. For the datasets analyzed during this work, we used an *st* threshold of 0.25.

The value 0.25 for the module selection procedure was selected by calculating all modules for all datasets we analyzed with different thresholds in the $[0, 0.4]$ range (with a difference of 0.025 between thresholds). The results are shown in Figure 4.9. As it can be seen in the figure, the number of modules found in all datasets shows a plateau in the $[0.1, 0.375]$ range.

It has to be noted that the goal of the module detection process is not to compute the exact significance of each module, but to estimate the change of the significance of a pair of pathways when the intersection among them is removed. If this change is big enough, and the intersection's significance is comparable to the one of the original pathways, we assume that the module is the responsible for the significance of the two parent pathways, and we modify the list of pathways accordingly. When the list of pathways is modified with the addition of the newly discovered modules, the correction for multiple comparisons is performed on the new augmented list, estimating the significance of pathways and modules appropriately. If there are n new modules added to the original list of k pathways, there will be $k + n$ tests and we correct for $k + n$ multiple comparisons.

After applying the module discovery and the EM approach, the result is a modified membership matrix that can be used to perform the desired type of analysis. This matrix now includes three types of pathways: i) original pathways as found in the literature, ii) novel functional modules that are impacted in the given condition independently from the

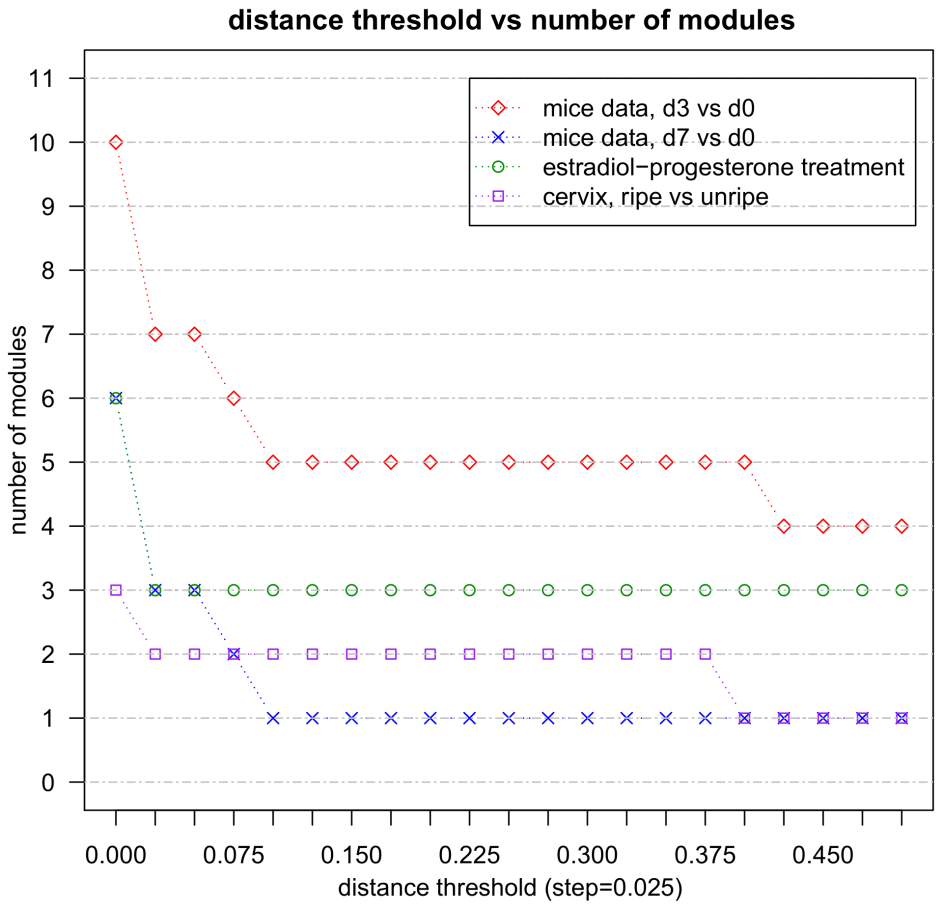


Figure 4.9: Number of modules obtained when changing the threshold distance under which two modules are considered similar enough to be joined. All datasets showed a plateau in the $[0.1, 0.375]$ range indicating that the number of modules found does not depend on the choice of the threshold for a wide range of threshold values.

pathways they belong to, and iii) the pathways from which such independent modules have been removed. If the same independent module is found in several conditions, in other words if this module is active independently from its parent pathways in several different phenotypes, such a module should be considered a good candidate for a novel pathway.

4.3 Results

We applied our approach to a number of real experiments: the fat remodeling treatment on obese mice described in Section 4.1.1, an experiment investigating cervical ripening [49], an experiment investigating the effect of various types of hormones on the endometrium of healthy, post-menopausal women who underwent hysterectomy [46], and an experiment investigating gene expression in Alzheimer's disease [15].

4.3.1 Fat remodeling in obese mice

Comparison between expression levels at day 3 versus day 0.

In order to correct for the crosstalk effects we started by computing the *crosstalk matrix* as described in Section 4.2.1. The analysis of the matrix corresponding to the comparison between day 3 and day 0 illustrates some interesting examples of crosstalk effects. Figure 4.10 represents a detail of the entire matrix. In this figure, the high significance of *Parkinson's* (bright red in row 1, column 1) disappears when the crosstalk due to *Alzheimer's* is eliminated (green in row 1, column 2). This indicates that *Parkinson's* is a false positive, since its significance is due exclusively to genes from *Alzheimer's*. Furthermore, the high significance of *Alzheimer's* (bright red in row 2, column 2) *also* disappears when the crosstalk effect of *Parkinson's* is eliminated (green in row 2, column 1). This means that *Alzheimer's* significance is also due only to the genes in common with *Parkinson's*. Essentially, the analysis tells us that the genes in common between the two pathways are activated independently of ei-

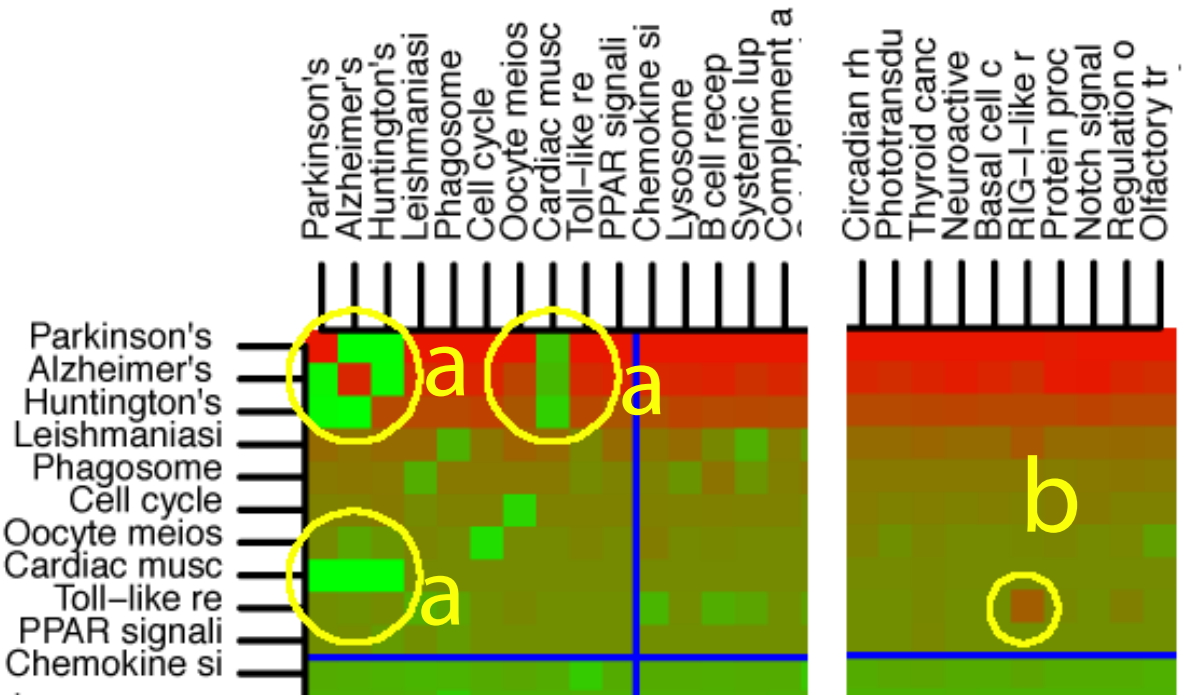


Figure 4.10: Detail of the crosstalk matrix: comparison between days 3 and 0 in the CL treatment. Areas marked with a correspond to functional modules that are activated independently from the pathways they belong to. The cell marked with b corresponds to a specific part of the *TLR* pathway that is responsible for the immune response to host genetic material. cells on the diagonal contain the p-values of the classical ORA, ordered from the most significant one to the least significant one. The cell $P_{i,j}$ contains the p-value of pathway P_i after the effect of P_j is removed. The color of each cell represents the p-value: bright red for p-values close to zero, bright green for p-values close to 1.

ther pathway, which suggests that these genes constitute an independent functional module. The same phenomenon involves the *Cardiac Muscle Contraction* and *Huntington's disease* pathways. The same independent functional module is responsible for the changes shown in areas marked with a in Figure 4.10.

An inspection of these genes and their signaling mechanisms reveals that this module is composed by genes present in mitochondria, organelles involved in all pathways above. The fact that this module is strongly activated in this fat remodeling experiment that is not related to any of the above conditions (Alzheimer's, Parkinson's, Huntington's), suggests that this should be considered as an independent pathway, dedicated to mitochondrial activity.

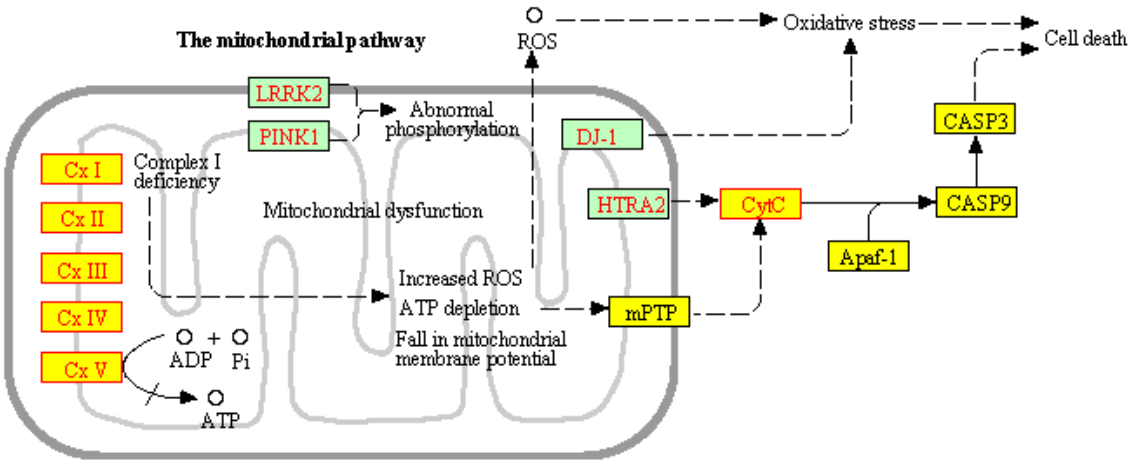


Figure 4.11: Mitochondrial activity pathway. This independent functional module is responsible for the incorrect identification of the pathways *Parkinson's disease*, *Alzheimer's disease*, *Huntington's disease*, and *Cardiac Muscle Contraction* by the classical ORA.

Figure 4.11 shows a representation of this new pathway.

In order to investigate the involvement of mitochondria in this condition, epididymal white fat of control and CL-treated (CL-7d) mice were stained with fluorescent Alexa-555 conjugated to streptavidin and imaged by spinning disc confocal microscopy. Mitochondria were stained with fluorescent Alexa-555 conjugated to streptavidin, and imaged in whole mount by confocal microscopy. Figure 4.12 shows a comparison between the control (left) and CL-treated mice (right). The right panel of this figure shows a massive generation of new mitochondria after 7 days of treatment, demonstrating *in vivo* that indeed, the mitochondrial pathway is central in this experiment.

Another very interesting phenomenon can be observed in Figure 4.10 (circle *b*). Here, *Toll-like Receptor Signaling* (*TLR*) pathway becomes *more significant* when the *Rig-I Like Receptor Signaling* (*RLR*) pathway (not significant on its own) is removed. The *TLR* pathway is the generic pathway involved in the immune response. The *RLR* pathway is the antiviral innate immunity pathway, which includes the mechanisms specifically aimed at the

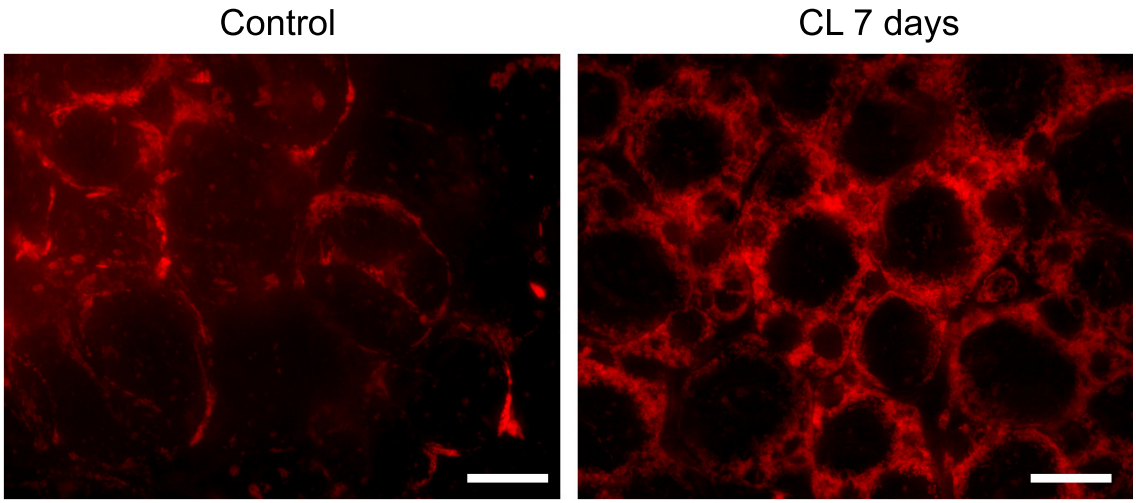


Figure 4.12: Epididymal white adipose tissue of a control mouse (left) and a mouse treated with CL for 7 days (right). Treatment with CL for 7 days triggered massive mitochondrial biogenesis, demonstrating *in vivo* that indeed, the mitochondrial pathway is central in this experiment. The white bar represents a 20 microns length.

detection of exogenous DNA or RNA. In essence, the crosstalk analysis tells us that in the fat remodeling experiment, the immune system has been activated but this immune response is *not* due to the presence of foreign genetic material. This is exactly what happens here. The CL treatment causes the death of some white fat cells [43]. In turn, this causes an immune response in which macrophages are required to dispose of the dead cells [72]. Such subtle distinctions between various triggers that activated the immune response are not possible with any classical analysis methods, and it is remarkable that a data analysis method was able to provide this type of insight.

We then applied the proposed Module Detection and Maximum Impact Estimation described in Sections 4.2.2 and 4.2.3 to the data. The corrected p-values obtained after correction are shown in Figure 4.3.

The ranking based on these crosstalk corrected p-values is greatly improved. The most significant pathway is now the newly discovered mitochondrial pathway shown in Figure4.11

rank	pathway	pval(FDR)
1	Mitochondrial Activity	$8.1e{-}10$
2	Phagosome	$9.3e{-}09$
3	Cellcycl+Oocyteme	$5.8e{-}08$
4	PPAR signaling pathway	0.001
5	Compl. C.C.+Systemic L.E.	0.002
6	* Cytok.-cytok. rec. int.	0.043
7	Toll-like receptor signaling	0.051
8	MAPK signaling pathway	0.115
9	B-cell receptor signaling	0.145
10	Lysosome	0.187
11	Nat. killer cell med. cytotox.	0.187
12	* Cell cycle	0.229
13	Calcium signaling pathway	0.229
14	Cell adhesion molecules	0.258
15	NOD-like receptor signaling	0.258
16	Vasc. smooth muscle contr.	0.424
17	Dilated cardiomyopathy	0.424
18	* Oocyte meiosis	0.432
19	Type I diabetes mellitus	0.432
20	Wnt signaling pathway	0.476

Table 4.3: The results of the ORA analysis in the fat remodeling experiment for the comparison between days 3 and 0 after (right) correction for crosstalk effects. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05. Pathways highlighted in red represent pathways not related with the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, that they are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis. Pathways ranked 1, 3, and 5 are modules that are functioning independently of the rest of their pathways in this particular condition. Starred pathways are pathways edited by removing such modules. Note the lack of any obvious false positive above the significance threshold(s).

and validated by the in-situ hybridization shown in Figure 4.12. The new p-values also indicate the *Phagosome* pathway as one of the pathways related to this phenomenon [87]. Third in the list is an independent module shared by *Cell Cycle* and *Oocyte Meiosis*. This can be thought of as a pathway related to the creation of new cells. Finally, the true involvement of *PPAR signaling* pathway in the phenomenon of fat remodeling has been previously demonstrated [43]. After removing the influence of the mitochondrial crosstalk, the *Parkinson's*, *Alzheimer's*, and *Huntington's* pathways are not significant anymore (now ranked 60^{th}, 61^{st} and 54^{th}, respectively). Also, after removing the crosstalk from *Phagosome*, *Leishmaniasis* is not significant anymore (now ranked 62^{nd}).

Comparison between day 7 and day 0.

Similarly to the results obtained for the comparison of expression levels at day 3 versus day 0, shown in Figure 4.2, the top pathways for the original ORA are *Parkinson's disease, Alzheimer's disease*, and *Huntington's disease*, diseases that have little to do with the tissue remodeling phenomenon. We computed the *crosstalk matrix* as described in Section 4.2.1. Figure 4.13 represents a detail of the entire matrix. The areas marked with *a* highlight the same phenomenon present in the matrix corresponding to the comparison between days 3 and 0 of the same experiment. The significance of the pathways *Parkinson's disease*, *Huntington's disease*, *Alzheimer's disease*, and *Cardiac Muscle Contraction* is entirely due to the same mitochondrial activity pathway shown in Figure 4.11. The greatly enhanced mitochondrial activity in the treated tissue was validated in vivo by in-situ hybridization (see Fig. 4.12). This shows additional evidence towards the activation of this independent pathway in this condition.

We then applied the proposed Module Detection and Maximum Impact Estimation described in Sections 4.2.2 and 4.2.3 to the data. The ranking obtained with the p-values corrected for crosstalk is shown in Fig. 4.4 and is greatly improved. The most significant pathway is the mitochondrial pathway, showing that greatly enhanced mitochondria activity continues to be the most important difference between the treated and untreated cells

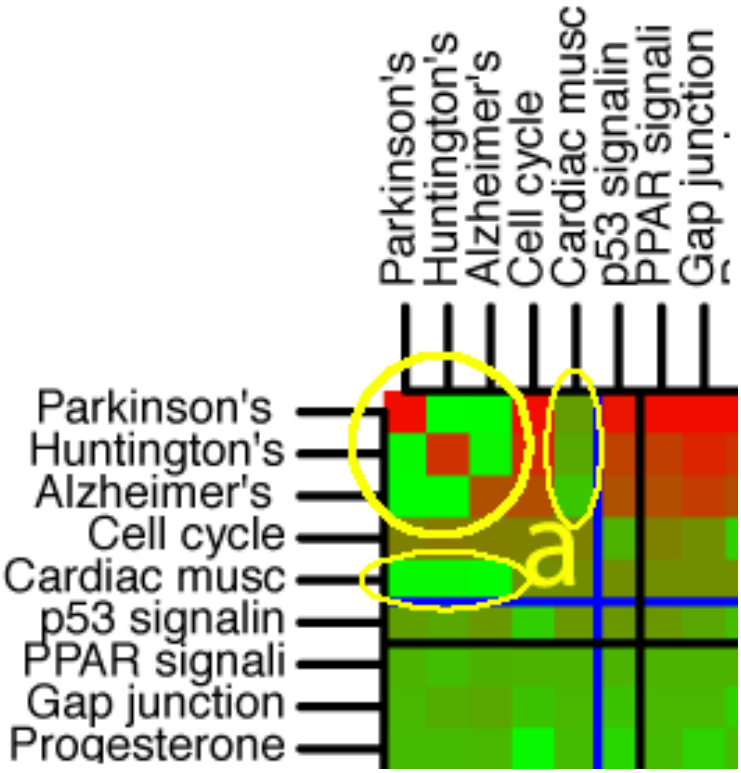


Figure 4.13: Detail of the crosstalk matrix for the comparison between days 7 and 0 in the same treatment. The areas marked with *a* correspond to the *Mitochondrial activity* pathway shown in Fig. 4.11, the same pathway that was found to be activated in the dataset associated with the comparison of expression levels at days 3 and 0.

even after 7 days. This in turn suggests that the tissue underwent a long-lasting remodeling phenomenon, in addition to a number of transitory phenomena such as cellular death and phagocytosis (note that the *Phagosome* pathway, significantly impacted after 3 days is not significant anymore after 7 days).

The pathway ranked second is *Arrhythmogenic Right Ventricular Cardiomiopathy*. While this pathway was treated here as a false positive due to lack of literature evidence linking it specifically to tissue remodeling, the module reported by the method includes genes related to *desmosomes*, cell structures responsible for certain types of cellular adhesion [65] which may also be relevant here. Fourth and fifth pathways in rank are, respectively, the *PPAR Signaling* pathway and the *Cell Adhesion Molecules* pathway, both closely related to the phenomenon of fat remodeling [43].

4.3.2 Cervical ripening

The second data set analyzed was obtained from a recent study that investigated the transcriptome of uterine cervical ripening in human pregnancy before the onset of labor at term [49]. The tissue analyzed is the human uterine cervix, the lower part of the uterus extending from the isthmus of the uterus into the vagina. This tissue is mainly composed of smooth muscle and extracellular matrix, which consists of collagen, elastin, proteoglycans, and glycoproteins [69, 115]. The uterine cervix has an essential function in the maintenance of pregnancy and also in parturition [47, 49, 48]. Cervical ripening is a critical component of the common terminal pathway of parturition, which includes the extensive remodeling of the cervix [49]. Disorders of cervical ripening can lead to premature or protracted cervical change, complicating term (e.g. protracted dilatation or arrest of dilatation) or preterm gestations (e.g. premature cervical dilation in the second trimester) [49]. The state of cervical ripening has traditionally been assessed by clinical examination (Bishop score or its modifications [13]), which includes the digital examination of the cervix for its consistency, dilatation, effacement, and position. This method has also been used to predict the likelihood that a patient would go into spontaneous labor.

rank	pathway	pval(FDR)
1	Mitochondrial Activity	$2.3e{-}08$
2	Arr. right ventr. cardiom. (ARVC)	0.001
3	Cell cycle	0.001
4	PPAR signaling pathway	0.015
5	Cell adhesion molecules (CAMs)	0.019
6	Melanogenesis	0.019
7	Vascular smooth muscle contr.	0.080
8	p53 signaling pathway	0.125
9	Pathways in cancer	0.562
10	SNARE inter. in vesicular transp.	0.562
11	Chagas disease	0.575
12	Long-term potentiation	0.575
13	Phagosome	0.588
14	Vasopressin-reg. water reabs.	0.765
15	Hedgehog signaling pathway	0.765
16	Dorso-ventral axis formation	0.765
17	Intest. imm, netw. for IgA prod.	0.784
18	Wnt signaling pathway	0.984
19	ECM-receptor interaction	0.984
20	Phototransduction	0.984

Table 4.4: The results of the ORA analysis in the fat remodeling experiment for the comparison between days 7 and 0 after (right) correction for crosstalk effects. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05. Pathways highlighted in red represent pathways not related with the phenomenon in analysis, while pathways highlighted in green are those for which we know, with reasonable confidence, that they are involved in the given phenomenon. The white background indicates pathways for which we do not have conclusive information on their involvement (or lack of) with the phenomenon in analysis. Starred pathways are pathways edited by removing such modules. Note the lack of any obvious false positive above the significance threshold(s). The mitochondrial activity pathway (validated in vivo) is reported as the most significant pathway even after 7 days, suggesting permanent tissue remodeling. The *Phagosome* pathway, significantly impacted after 3 days (see Fig. 4.3) is not significant anymore after 7 days, consistent with the transitory nature of cellular death and phagocytosis. The four false positives present in the results of ORA (shown in Figure 4.2) have been removed. The *Arrhythmogenic Right Ventricular Cardiomyopathy* pathway is reported as a false positive here, but the DE genes located on this pathway are involved in cell adhesion, which may be a relevant phenomenon here.

The goal of this experiment was to examine the relationship between human cervical ripening and the cervical transcriptome, aiming to improve our understanding of the biology of cervical ripening at term. This study included pregnant women who underwent elective C-section at term with an unripe (n=11) or ripe cervix (n=11). Cervical biopsies were obtained from these women trans-vaginally, from the anterior lip of the uterine cervix following C-section. Microarray analysis was performed on RNA isolated from these cervical tissue specimens using Affymetrix GeneChip HGU133Plus2.0 arrays [49].

On this dataset we performed the comparison between gene expression levels from cervical tissues obtained from women with an unripe (n=11) or ripe cervix (n=11) using the classical ORA. The results are shown in Fig. 4.14a. Pathways with a p-value smaller than 0.05 after FDR correction were *Focal adhesion, ECM-receptor interaction, Amoebiasis, Cell adhesion molecules (CAMs), Small cell lung cancer*, and *Dilated cardiomyopathy.*

There is plenty of experimental evidence that biological processes described by the pathways *Focal Adhesion, ECM-Receptor Interaction*, and *Cell Adhesion Molecules* are related to cervical ripening. The relation between these pathways and the phenomenon in analysis was revealed by studies on humans and animals showing the involvement of extra-cellular matrix metabolism and cell adhesion molecules in cervical ripening [69, 70, 75, 114, 115]. However, the pathway *Amoebiasis* describes the biological process of infection from a parasite that invades the intestinal epithelium. Amoeba infection involves the parasite attachment to the intestinal mucus layer, followed by disruption and death of host epithelial cells. This process is completely unrelated to the physiological condition of cervical ripening in term pregnancy. The same is true for the *Small Cell Lung Cancer* pathway. Clearly, the top ranked pathways include some describing complex phenomena that are unrelated to the studied condition. Also, the significant pathways known to be involved in the process of cervical ripening are somewhat general pathways describing cellular interactions.

The analysis of the crosstalk matrix shown in Figure 4.16 shows that there is a independent functional module among the top three pathways in the ranking. This novel module includes

rank	pathway	p(fdr)
1	Focal adhesion	$1.1e{-}08$
2	ECM-receptor interaction	$1.1e{-}08$
3	Amoebiasis	$1.2e{-}06$
4	Cell adhesion molecules	0.009
5	Small cell lung cancer	0.015
6	Dilated cardiomyopathy	0.015
7	Viral myocarditis	0.066
8	TGF-beta signaling path.	0.098
9	Prion diseases	0.1555
10	Leukocyte transend. migr.	0.1869
11	Pathways in cancer	0.1869
12	Nat. killer c. med. cytotox.	0.2202
13	Malaria	0.2202
14	Adherens junction	0.3711
15	Arr. right ventr. cardiom.	0.3711
16	Calcium signaling pathway	0.3712
17	Cholinergic synapse	0.6605
18	Vascular smooth muscle contr.	0.6605
19	Glutamatergic synapse	0.6969
20	HTLV-I infection	0.6969

(a) Top 20 pathways reported by ORA before correction for crosstalk. Pathways like *Amoebiasis* and *Small Cell Lung Cancer* are not related to this phenotype.

rank	pathway	p(fdr)
1	Integrin mediated ECM Signal.	$2.90e{-}13$
2	Cell adhesion molecules	0.0041
3	Dilated cardiomyopathy	0.0041
4	Leukocyte transend. migr.	0.0134
5	TGF-beta signaling pathway	0.2228
6	Endocrine/other f.r. Ca reabs.	0.5791
7	Insulin signaling pathway	0.9182
8	Alzheimer's disease	1
9	Vascular smooth muscle contr.	1
10	Glutamatergic synapse	1
11	Mineral absorption	1
12	Nat. killer cell mediated cytotox.	1
13	Calcium signaling pathway	1
14	Complement and coag. casc.	1
15	MAPK signaling pathway	1
16	HTLV-I infection	1
17	* * Focal adhesion	1
18	* ECM-receptor interaction	1
19	* * Amoebiasis	1
20	Small cell lung cancer	1

(b) The top 20 pathways reported by ORA after the crosstalk analysis. After the correction neither *Amoebiasis* nor *Small Cell Lung Cancer* are significant anymore. At the same time, *Cell Adhesion Molecules* and the *Integrin-mediated ECM Signaling* have an increased significance. Starred pathways are pathways edited by removing such module.

Figure 4.14: The results of the ORA for the cervical ripening experiment, before (left) and after (right) the correction for crosstalk effects. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05.

the genes present in the interaction between the cellular transmembrane protein integrin and three important ECM components, collagen, laminin, and fibronectin. The KEGG pathways involved in the identification of this pathway are *Focal adhesion*, *ECM-receptor interaction*, and *Amoebiasis*. Henceforth, we will refer to this pathway, shown in Figure 4.15, as the *Integrin-Mediated ECM Signaling*.

Very interestingly, the independent functional module found in this condition is, in fact, the exact same module found in the hormone treatment experiment described below in Section 4.3.3. Interestingly, the KEGG pathways involved in the identification of this functional module are slightly different between the two phenotypes. While in this phenotype this module was found from the interaction of *Focal adhesion*, *ECM-receptor interaction* and *Amoebiasis*, in the hormone treatment the last pathway is replaced by *Pathways in Cancer*. The fact that the same module was found to be activated and statistically significant in two different phenotypes, from the interaction of different sets of canonical pathways, further supports the idea that this module describes an independent mechanism and should therefore be considered as an independent pathway.

Further analysis of the crosstalk matrix shows that the *Small Cell Lung Cancer* loses significance when the crosstalk effects of the first three pathways are removed (bright green loss of significance in first 3 columns of row 5 in Fig. 4.16). This allows us to conclude that it is a false positive in the classical ORA, with its ORA significance due exclusively to crosstalk effects.

The ranking of pathways with the p-values corrected for crosstalk by our analysis is shown in Fig. 4.14b. The first pathway is *Integrin-mediated ECM Signaling* with an FDR corrected p-value of $2.9e^{-13}$. *Cell Adhesion Molecules* is now the second in ranking, with an FDR corrected p-value of 0.004. The false positives in the classical ORA results, *Amoebiasis* and *Small Cell Lung Cancer*, are not significant anymore. The biological significance of the pathway *Dilated Cardiomiopathy* may be linked to the fact that 10%-15% of the uterine cervix is constituted of smooth muscle, and cervical ripening involves alterations of this component.

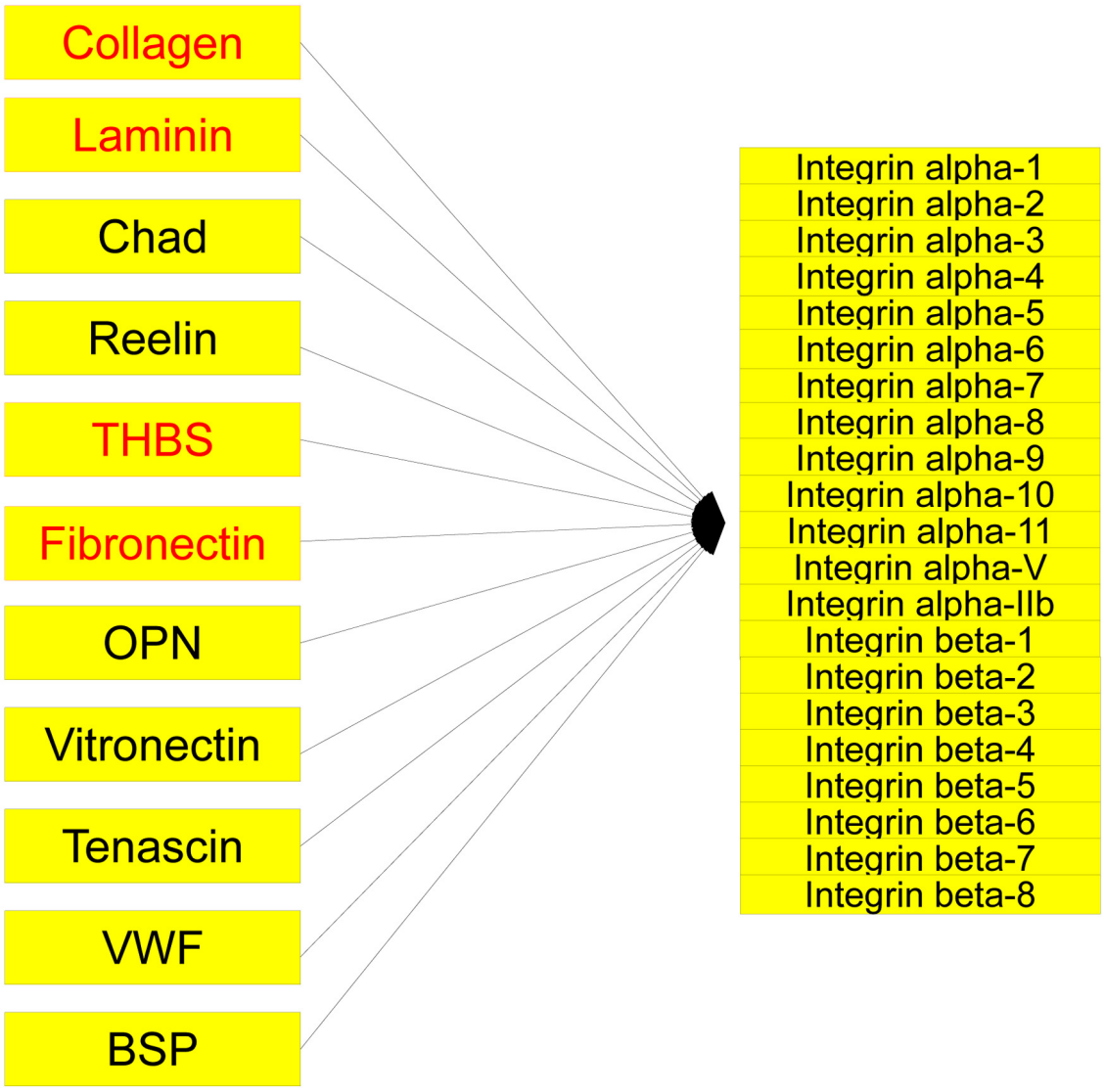


Figure 4.15: The novel *Integrin-Mediated ECM Signaling*. This new module was found to be independently activated and statistically significant in two different conditions: hormone treatment of post-menopausal women and cervical ripening in normal pregnancies. Genes shown in red were found to be differentially expressed in the hormone treatment experiment.

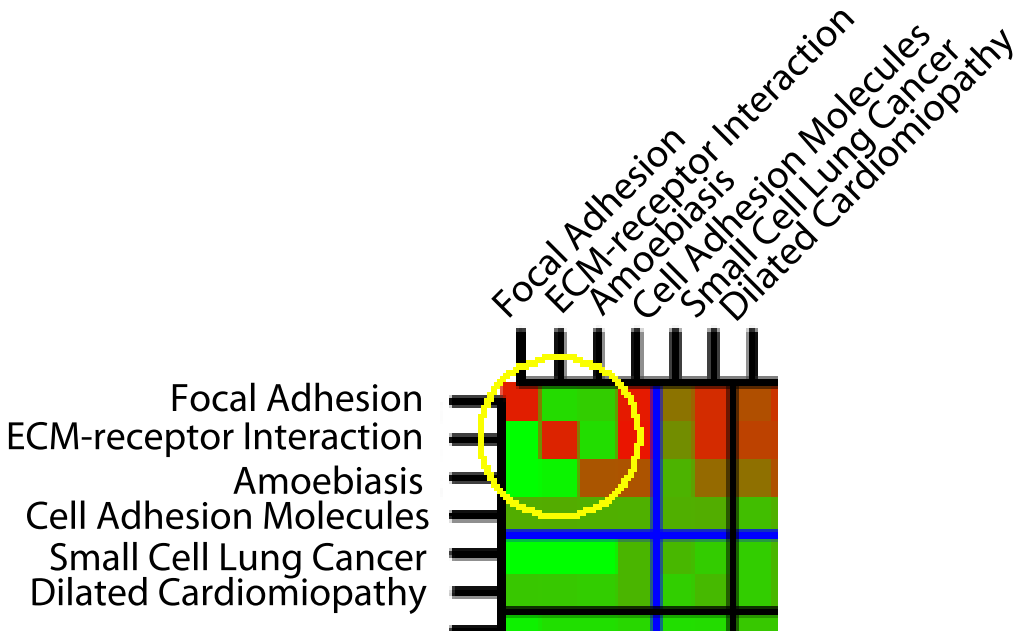


Figure 4.16: Details of the crosstalk matrix of the cervical ripening experiment. The circle highlights the evidence for an independent module involving pathways *Focal Adhesion, ECM-Receptor Interaction*, and *Amoebiasis*. The bright green loss of significance of *Small-Cell Lung Cancer* in columns 1-3 shows that this pathway was a false positive in the ORA since its significance was due only to the crosstalk from the first 3 pathways.

The last significant pathway at the 5% significance threshold is *Leukocyte Transendothelial Migration*. Although human and animal studies [49] have shown that cervical ripening does not require activation of a typical inflammatory response, and influx of inflammatory cells into the cervix, the significance of this pathway may reflect the beginning of later inflammatory events typical of parturition [113, 125].

4.3.3 Estrogen treatment on post-menopausal women

This dataset was produced by an experiment investigating the effect of various types of hormones on the endometrium of healthy, post-menopausal women who underwent hysterectomy [46]. Hormone therapy has been used for the treatment of conditions associated with menopause [86]. Estrogen replacement therapy has been proven useful against the insurgence of collateral effects of the post-menopausal syndrome [17, 50, 123]. However, the administration of estrogens only has been shown to increase the incidence of endometrial carcinoma [131]. Therefore, in addition to estrogen, progestins are now given to menopausal women. Although the risk of endometrial cancer is reduced with the addition of progestins,

the incidence of other forms of cancer seems to increase when progestin is administered with estrogen. Initiatives like the Million Women Study (`http://www.millionwomenstudy.org/`) and the Women Health Initiative (`http://www.nhlbi.nih.gov/whi/`) showed that hormone replacement therapy can increase the risk of lung and breast cancer [23, 25]. In this context, it is interesting to compare the effects of various combinations of hormones at the transcriptome level [46].

Here, we illustrate our analysis method on the comparison of the expression levels of genes from samples treated with estrogen (E2) plus medroxyprogesterone acetate (MPA) versus normal samples. The classical over-representation analysis (ORA) finds the following pathways significant at the 5% level after FDR correction: *ECM receptor interaction, Focal Adhesion, Pathways in Cancer, Small Cell Lung Cancer, Axon Guidance, Prostate Cancer, and Jak-STAT Signaling*. These results are shown in Figure 4.17a.

The E2+MPA treatment is known to be associated with certain type of cancer including non-small-cell lung cancer (NSCLC) [24]. Hence, the presence of *Pathways in Cancer* is justified, even though its identification as significant does not help understand the specific mechanism that might be active here. However, the set of significant pathways include small-cell lung cancer (SCLC) which is *not* known to be associated with this treatment and fail to include the NSCLC which *has been* linked to it [24]. *Prostate Cancer* is also unlikely to be related to this specific treatment given that this treatment is administered to women, rather than men. Like in the previous case, the presence of false positives and the presence of pathways describing general cellular adhesion processes (focal adhesion and ECM-receptor interaction) does not help with the understanding of the underlying phenomenon.

The analysis of the *cross-talk matrix* shows some interesting cases of cross-talk effects. The first case is an example of a module shared among pathways that is responsible for their significance. The module in common between the first four pathways in the ranked list describes the interactions among integrins and collagen, laminin, and fibronectin. The second case is shown in Fig. 4.18; this detail of the cross-talk matrix shows the row correspond-

rank	pathway	p(FDR)
1	ECM-rec. interaction	0.0343
2	Focal adhesion	0.0401
3	Pathways in cancer	0.0401
4	Small cell lung cancer	0.0401
5	Axon guidance	0.0401
6	Prostate cancer	0.0401
7	Jak-STAT signaling pathway	0.0401
8	Progest.-med. oocyte mat.	0.0951
9	Adipocytokine signaling	0.0951
10	Melanoma	0.1208
11	Graft-versus-host disease	0.1291
12	Reg. of actin cytoskeleton	0.2020
13	Aldosterone-reg. Na reabs.	0.2020
14	Oocyte meiosis	0.2168
15	Long-term depression	0.2174
16	mTOR signaling pathway	0.3048
17	Nat. killer cell med. cytotox.	0.3185
18	Vibrio cholerae infection	0.3225
19	SNARE inter. in ves. trans.	0.3699
20	Salivary secretion	0.3699

(a) The top 20 pathways reported by the classical ORA before correction for crosstalk. The NSCLC, known to be linked to this treatment [24] is not identified by the classical method, while the SCLC, which showed no increase in incidence in the treatment group [24], appears as significant. The significance of *Pathways in Cancer* is consistent with the putative link between hormone treatments and higher incidence of some types of cancer but offers no explanation or insight into the underlying mechanisms.

rank	pathway	p(FDR)
1	Jak-STAT signaling pathway	$5e{-}09$
2	Integrin Mediated ECM Sign.	0.0001
3	Axon guidance	0.0036
4	Vascular sm. muscle contr.	0.0070
5	Aldosterone-reg. Na reabs.	0.0190
6	Adipocytokine signaling	0.0326
7	Nat. killer cell med. cytotox.	0.0344
8	Regulation of actin cytosk.	0.1403
9	Compl. and coag. cascades	0.3413
10	Adherens junction	0.3413
11	SNARE interac. in ves. trans.	0.4842
12	Circadian rhythm - mammal	0.5074
13	Lysosome	0.6552
14	Protein proc. in endopl. ret.	0.7182
15	Vibrio cholerae infection	0.7182
16	* * * Focal adhesion	0.9844
17	Type I diabetes mellitus	1
18	Phagosome	1
19	Huntington's disease	1
20	Cell cycle	1

(b) The top 20 pathways reported by ORA after the correction for crosstalk effects. The correction method removed *Pathways in Cancer*, *SCLC*, and *Prostate Cancer* from the list of significant pathways, increasing the significance of pathways offering more insights such as *Jak-STAT signaling pathway* and the new *Integrin mediated ECM signaling* module. A star before the name of the pathway means that a module overlapping with other pathways has been removed from the pathway.

Figure 4.17: Results of ORA for the estrogen treatment experiment, before (left) and after (right) the correction for crosstalk effects. All p-values are FDR-corrected. The lines show the significance thresholds: blue - 0.01, yellow - 0.05.

ing to the *Graft-Versus-Host Disease* pathway, which became significant after removal of associated pathways in multiple cases, including *Cell adhesion molecules (CAMs)*, *Leishmaniasis*, *Intestinal immune network for IgA production*, and *Asthma*. This happens because all shared genes between *Graft-Versus-Host Disease* and the others are all non-DE genes in this condition. In other words, the DE genes present in *Graft-Versus-Host Disease* pathway are specific to the pathway itself. Among those, two particularly interesting ones are PRF1 (perforin 1) and GZMB (granzyme B), both of which play important functional roles in the natural killer (NK) cell-mediated cytolysis. Consistent with this, the *Graft-Versus-Host Disease* pathway is highlighted as being significantly affected by the E2+MPA treatment in the crosstalk matrix, not due to other interactions but due to *genes specific to NK cell-mediated cytotoxicity*. It is remarkable that the results of this type of analysis allowed the identification of a module, composed by genes belonging to the *Graft-Versus-Host Disease* pathway, that is impacted by the hormone treatment, and whose importance was masked by crosstalk effects with other pathways. This module is relevant in the condition studied, and treating it separately would provide a more accurate understanding of the underlying biological phenomenon. However, since the activity of this module was not identified yet in another condition, nor do we have an independent in vivo validation for this phenotype, we do not have enough evidence to propose this as an independent pathway at this time.

After applying our the module detection and maximum impact estimation to the dataset, the results become more helpful in providing insights about the specific underlying mechanisms, as shown in Fig. 4.17b. The first pathway in the ranked list is the *Jak-STAT signaling pathway*. Indeed, there is evidence that estrogen treatments impact such pathway through interaction with the suppressor of cytokine signaling (SOCS2) [71]. The second pathway is a new pathway, based on the module common between *Focal Adhesion, ECM-receptor Interaction*, and *Pathways in Cancer* (see the left panel of Fig. 4.18). In this figure, within the significant quadrant, the symmetric pattern that can be observed between the three pathways above and *Pathways in Cancer* indicate the presence of a functional module that responds

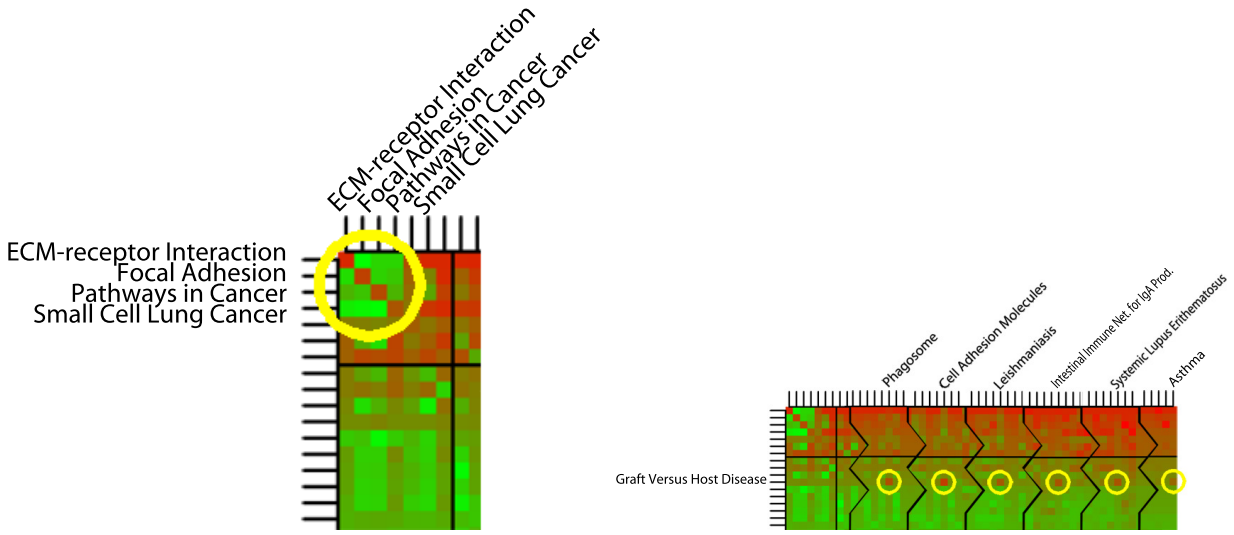


Figure 4.18: Detail of the crosstalk matrix of the estrogen treatment. Left panel: the circle highlights an example of a common module that is responsible for the significance of an entire group of pathways. The common module between the pathways *ECM-Receptor Interaction, Focal Adhesion, Pathways In Cancer*, and *Small Cell Lung Cancer* describes the interaction between *integrin* and *collagen, laminin,* and *fibronectin*. Henceforth, we will refer to this module as the *Integrin-mediated ECM signaling* pathway (see Fig. 4.15). Right panel: row corresponding to the pathway *Graft-Versus-Host disease*. The pathway becomes significant after the removal of specific pathways, highlighted by the yellow circles. The set of pathways includes *Phagosome, Cell adhesion molecules (CAMs)*, *Leishmaniasis*, *Intestinal immune network for IgA production, Systemic Lupus Erithematosus*, and *Asthma*. This indicates a situation in which the genes specific to *Graft-Versus-Host disease* are related to the phenomenon in analysis, but their significance is *masked* by the presence of crosstalk with other pathways.

specifically to the hormone treatment. Interestingly, this pathway is the same pathway that has been shown to be active in a completely different phenotype, the cervical ripening experiment described in Section 4.3.2, and it is the *Integrin-Mediated ECM Signaling* described in Fig. 4.15. This pathway is responsible for the significance of the top four pathways in Fig. 4.17a.

As in the experiment studying cervical ripening, this novel pathway is composed of genes present in the interaction between the cellular transmembrane protein integrin and three important ECM components, collagen, laminin, and fibronectin, all of which appeared as differentially expressed in hormone treatment compared to the control. This is interesting because the ECM-receptor interaction carries two major functions: the first is to transduce extracellular signals into the cell for regulation of downstream pathways possibly through focal adhesion complex, and the second function is to provide structural support to resident cells; the binding between integrins and collagen, laminin, fibronectin is involved in the second process. Collagen, a major component of the ECM, forms fibers and attaches to the cell surface through binding with integrins and fibronectins. Collagen is also present in the basement membrane with laminin, forming a thin sheet of fibers that underlies the epithelium [3]. Previous studies have shown that collagen, laminin, and fibronectin participate in regulating normal development of mammalian mammary tissues [11]. They also play an important role in cancer progression possibly through ECM remodeling, which leads to alterations in cell adhesion and tumor cell motility. Consistent with this, enhanced attachment of estrogen-dependent breast cancer cells to the substrate containing ECM components (collagen I and IV, laminin, fibronectin) was observed with E2 treatment [81]. More evidence was provided in recent studies using mouse mammary epithelial cells, where the expression of estrogen receptor alpha (ESR1) was greatly down-regulated by integrin-mediated interaction with collagen-IV and laminin, rather than effects of growth factors such as insulin [88]. Consistently to the previous findings, our method finds that it is the module describing the interaction between integrin and collagen, laminin, and fibronectin (rather than the interaction

between ligands and their receptors) that is affected specifically by the hormone treatment, a striking pattern unlikely to be detected by classical over-representation analysis.

A similar pattern was observed between pathways *Prostate Cancer* and *Focal Adhesion*, where the removal of a common submodule caused loss of significance in both pathways. A close investigation of the *Focal Adhesion* pathway revealed that its downstream signaling cascade is regulated by two types of extracellular signals, the ECM components that interact with integrins, and the growth factors (EGF) that bind to the transmembrane GF receptor (EGFR). Although a number of DE genes belong to the *ECM-Receptor Interaction* pathway, it is the EGFR-induced signaling cascade that is involved in both *Prostate cancer* and *Focal adhesion*, which contains at least two downstream pathways that responded specifically to the E2+MPA treatment. The first one is the canonical Wnt cascade, during which the transcription factor beta-catenin gets activated by PI3K-AKT (phosphatidylinositol 3 kinase-V-Akt murine thymoma viral oncogene homolog) mediated signals, and translocate into the nucleus for downstream gene regulation [85]. The other is the classical *MAPK* (Mitogen-Activated Protein Kinase) *pathway*, also known as the RAF-MAP2K-MAPK pathway, where RAF, MAP2K, and MAPK represent the three key serine/threonine-specific protein kinases present in the cascade [129]. What is also noticeable is that in both cases, while *Wnt Signaling* pathway and *MAPK Signaling* pathway both contain sub-pathways other than the two highlighted here, such as the *Wnt5-induced non-canonical Wnt pathway* or *JNK-p38-mediated MAPK pathway*, only the canonical Wnt cascade and the classical MAPK cascade are associated with both *Prostate Cancer* and *Focal Adhesion*, among which a number of important genes are DE under the hormone condition, such as PTEN (phosphatase and tensin homolog), a tumor suppressor that regulates PI3K-AKT signaling pathway, MAPK, one of the three key protein kinases in the MAPK pathway, and AR (androgen receptor), an oncogene that plays an important role in MAPK-regulated cell proliferation [45, 92]. Indeed, estradiol has been shown to activate beta-catenin-mediated Wnt pathway through inhibition of its partner GSK3 in the rat hippocampus, which releases beta-catenin and allows its

nuclear translocation [18]. More functional evidence was provided using human colon and breast cancer cells, in which estrogen receptor (ER) and beta-catenin were found to participate in the same multi-protein complex, whose interaction gets enhanced with the presence of estrogen [66]. Since both beta-catenin and ER function as transcription factors, it is possible that the role of beta-catenin in this complex is to recruit additional co-activators and chromatin remodeling factors that interact with ER for downstream transcriptional regulation [66]. Estrogen has been demonstrated to induce cell proliferation through increased phosphorylation of MAPK cascade, with the mechanistic link between estrogen and MAPK signaling lying in a partner of ER, the PELP1 (proline, glutamate and leucine rich protein 1, the modulator of non-genomic activity of estrogen receptor) protein [124]. PELP1 forms a complex with ER and Src family of tyrosine kinases as a scaffold protein, which is enhanced by E2, further induces activation of MAPK kinases and affects ER-mediated transcription [124]. Consistent with these studies, our method detected a module shared between *Prostate cancer* and *Focal adhesion*, the EGFR-induced canonical Wnt and classical MAPK cascade, which is responsible for significance of both pathways.

4.3.4 Alzheimer's disease

We analyzed the data set produced by an experiment investigating the correlation between gene expression values "*with MiniMental Status Examination (MMSE) and neurofibrillary tangle (NFT)*" in subjects with Alzheimer's disease [15]. Figures 4.19a and 4.19b show the comparison between the results of the classical ORA and the results of the crosstalk analysis. At the top of the results of the ORA we find *Huntington's, Alzheimer, Parkinson's, Glutamatergic Synapse, and Arrhythmogenic right ventricular cardiomyopathy.* In this list, *Alzheimer's* is the obvious true positive, *Huntington's, Parkinson's, and Glutamatergic Synapse* are definitely related to the phenomenon, being involved in neurodegenerative diseases, while *Arrhythmogenic Right Ventricular Cardiomyopathy* is clearly a false positive. The cross-talk analysis reports, as only significant pathway, the module composed by the intersection between the *Alzheimer's, Parkinson's, and Huntington's* pathways.

Rank	Title	p-value(fdr)
1	Huntington's disease	$3.49 \cdot 10^{-06}$
2	Alzheimer's disease	$3.49 \cdot 10^{-06}$
3	Parkinson's disease	$3.49 \cdot 10^{-06}$
4	Glutamatergic synapse	0.00342933
5	Arrhythm. right ventr. cardiom.	0.0110
6	Circadian rhythm - mammal	0.0995
7	Dopaminergic synapse	0.1322
8	Long-term depression	0.1625
9	Calcium signaling pathway	0.1922
10	Retrograde endocann. signaling	0.1922

(a) Results of the ORA analysis of the GSE1297 data set using KEGG as a reference database. While related to neurodegenerative diseases, the pathways Huntington's and Parkinson's are not true positives. The pathway *Arrhythmogenic right ventricular cardiomyopathy* is not related to the phenomenon.

Rank	Title	p-value (fdr)
1	Alzheime+Parkinso+Huntingt	$2.44 \cdot 10^{-07}$
2	Arrhythm. right ventr. cardiom.	0.1826
3	Glutamatergic synapse	0.3789
4	GABAergic synapse	0.6135
5	ECM-receptor interaction	0.6135
6	Circadian rhythm - mammal	0.6135
7	Gap junction	0.8626
8	Phosphat. signaling system	1
9	Axon guidance	1
10	Serotonergic synapse	1

(b) Results of the crosstalk analysis of the GSE1297 data set using KEGG as a reference database. The crosstalk analysis is able to extract a functional module from the three neurodegenerative disease pathways that rank at the top of the ORA list. Genes found in this module are related to the phenomena of oxidative phosphorylation and cytochrome oxidase, highly related to Parkinson's disease. The pathway *Arrhythmogenic right ventricular cardiomyopathy* is not significant anymore.

Figure 4.19: The results of the ORA analysis in the GSE1297 experiment before (left) and after (right) correction for crosstalk effects. All p-values are FDR-corrected. The blue line shows the 0.05 significance threshold.

The DE genes in this module consist are related to the phenomena of oxidative phosphorylation and cytochrome oxidation. There is evidence [78, 90, 130] that these mechanisms are indeed central in Alzheimer's, and the crosstalk analysis was able to pinpoint the functional sub-pathway that is responsible for the phenotype, eliminating the false positive present in the classical analysis list.

4.3.5 Alzheimer's Disease - Reactome database

In order to show the ability of our method to work with different databases, we analyzed the dataset produced in [15] against the set of pathways from the Reactome database [57]. The results of the crosstalk analysis are shown in Figures 4.20a (for the ORA) and 4.20b (for the crosstalk analysis). In this case, the crosstalk analysis compacts the pathways that are at the top of the ORA result. Those pathways are all related to Alzheimer's disease [76, 91], and the crosstalk procedure of building the functional module that is involved in the phenomenon highlights the close interaction among them. The only false positive of the ORA result is *Regulation of Insulin Secretion*. This pathway describes signaling involving pancreatic beta cells and it is not related to brain cells. This pathway is not significant anymore after the

Rank	Title	p-value (fdr)
1	Tca Cycle/Respiratory Electron Transport	$5.22 \cdot 10^{-09}$
2	Respiratory Electron/Atp Synthesis	$1.92 \cdot 10^{-07}$
3	Respiratory Electron Transport	$2.94 \cdot 10^{-05}$
4	Gaba Synthesis Release Reuptake and Degradation	$2.94 \cdot 10^{-05}$
5	Neurotr. Release Cycle	$2.94 \cdot 10^{-05}$
6	Neuronal System	$2.94 \cdot 10^{-05}$
7	Glutamate Neurotr. Release Cycle	0.0006
8	Transmission Across Chemical Synapses	0.0006
9	Formation of Atp by Chemiosm. Coup.	0.0143
10	Regulation of Insulin Secretion	0.0160
11	Norepinephrine Neurotr. Rel. Cycle	0.0160
12	Protein Folding	0.0183
13	Integration of Energy Metabolism	0.0184
14	Prefoldin Mediated Transfer of Substrate to Cct Tric	0.033
15	Darpp 32 Events	0.0374

(a) Results of the ORA analysis of the GSE1297 data set using Reactome as a reference database. The top pathways are related to Alzheimer's Disease. The pathway *Regulation of Insulin Secretion* describe the signaling events involving pancreatic beta cells, and it is not related to brain cells.

Rank	Title	p-value (fdr)
1	Respiratory electron/atp synthesis + Respiratory electron tr. + Tca Cycle	$3.85 \cdot 10^{-07}$
2	Gaba Synth. + Glutamate neuron. + Neurona system + Neurotrans. + Transm. Chemical Synapses	0.0001
3	* Tca Cycle and Respiratory Electron Transport	0.0004
4	Prefoldi + Protein	0.1601
5	Glucose Metabolism	0.7029
6	Hemostasis	1
7	Metabolism of Nucleotides	1
8	Nuclear Signaling by Erbb4	1
9	Biological Oxidations	1
10	* Neuronal System	1
11	Axon Guidance	1
12	Smooth Muscle Contraction	1
13	G Alpha Z Signalling Events	1
14	Mitotic G2 G2 M Phases	1
15	Base Free Sugar Phosphate Removal	1

(b) Results of the crosstalk analysis of the GSE1297 data set using Reactome as a reference database. The crosstalk analysis groups the pathways related to Alzheimer's Disease. The pathway *Regulation of Insulin Secretion* is not significant anymore.

Figure 4.20: The results of the ORA analysis in the GSE1297 experiment using Reactome as reference database, before (left) and after (right) correction for crosstalk effects. All p-values are FDR-corrected. The blue line shows the significance thresholds of 0.05.

correction for crosstalk. It has to be noted that there is no *Alzheimer's* specific pathway in the Reactome database. However, the crosstalk analysis was able to identify highly related pathways, providing a more concise result list with no obvious false positives.

CHAPTER 5 CROSSTALK PACKAGE USER GUIDE

This chapter contains the user guide for the Crosstalk R package for analysis and correction of crosstalk effects in the analysis of signaling pathways. The analysis implemented in this package include the detection and quantification of crosstalk effects via computation of the crosstalk matrix described in Section 4.2.1, the maximum impact estimation described in Section 4.2.2, and the independent functional module detection procedure described in Section 4.2.3. In addition to the method, the package contains the data from the fat remodeling treatment experiment in obese mice described in [43]. This chapter describes how to format the main components of crosstalk analysis: the pathway knowledge, the experimental data, and the pathway analysis method.

5.1 Pathway data

Before any activity is performed, the package must be loaded on the current environment.

Default pathway data for *Homo sapiens* and *Mus muscles* from KEGG are provided with the package. They can be loaded with the function `getPathways` by specifying `cached` as source. The following example loads the cached pathway data for *Mus musculus* and shows the content of the first pathway. Each pathways in the object returned by the function `getPathways` is represented by a character vector whose elements are the genes belonging to the pathway.

```
> paths <- getPathways(organism='mmu', pathSource='cached')
> paths$pathways[1]
$mmu03008
 [1] "102641332" "19384"     "19428"     "75471"     "103573"    "97112"
 [7] "17724"     "17725"     "21453"     "100862468" "20826"     "55989"
[13] "67134"     "14113"     "237730"    "66181"     "52530"     "68147"
[19] "245474"    "73736"     "14000"     "98956"     "195434"    "72554"
[25] "213895"    "59028"     "102614"    "117109"    "208366"    "227522"
```

```
[31] "54364"     "66161"     "67724"     "69961"     "74097"     "24127"
[37] "24128"     "69237"     "230737"    "237107"    "30877"     "100019"
[43] "67459"     "104444"    "434234"    "66932"     "170722"    "245610"
[49] "53319"     "83454"     "237082"    "56488"     "67973"     "27993"
[55] "102462"    "73674"     "105372"    "21771"     "217995"    "72515"
[61] "217109"    "213773"    "216987"    "110816"    "225348"    "269470"
[67] "12995"     "13000"     "13001"     "230082"    "224092"    "101592"
[73] "74778"     "71340"     "57815"     "67045"     "102216272" "16418"
[79] "14791"     "66711"     "68272"     "67619"
```

The `getPahtways` textttfunction allows for three alternatives for obtaining the pathway knowledge, controlled by the parameter `pathSource`. The first one, i.e. using the value `cached`, is shown in the example above, and it loads data already cached in the system. The second alternative is to use the value `spia`, in which case the pathways are retrieved from the `SPIA` package. Lastly, if the parameter is set to `ronto` the pathways are retrieved from the package `ROntoTools`.

5.2 Experimental data

The experimental data has to be provided as a logical vector whose names are the IDs of all the genes that were screened in the experiment. Each element of these vectors takes the value `TRUE` if the gene was considered *interesting* (e.g. differentially expressed, or DE) in the experiment, and `FALSE` otherwise. In the `crosstalk` package we included an example of data from an experiment investigating cellular and metabolic plasticity of white fat tissue (WAT), where the classical over-representation analysis (ORA) produced a number of false positives, and failed to rank highly pathways that were known to be involved in the given condition [43, 29]. The data can be loaded as follows.

```
> data(micedata)
```

The first few elements of the data object show the format of the data.

```
> head(micedata)
```

```
 18000   16423 653016   12266   27370 319991
 FALSE   FALSE  FALSE    TRUE   FALSE  FALSE
```

In this specific example, gene IDs are *Entrez* gene IDs. It is important to note that the package does not require a specific type of ID, as long as both IDs in the experimental data object and in the pathway data are consistent with each other.

5.3 Crosstalk matrix

Once the input data (experimental data and pathway knowledge) is loaded users can quantify the amount of crosstalk among pairs of pathways by computing the *crosstalk matrix* with the function`crosstalk`, as described in Section 4.2.1. The following example code computes the matrix.

```
> objectCrossTalk <- crossTalk(dedata = micedata,
+         pathway.data=paths$pathways, path.titles = NULL,
+         shortTitles = TRUE, thresholds=c(0.01, 0.05))
```

The `crosstalk` function accepts the following parameters: `dedata`, containing the list of genes involved in the experiment, in the format described in the previous section, the `pathway.data` parameter, containing the pathway knowledge, an optional argument containing the titles of the pathways, the `shortTitles` parameter, indicating if pathway titles should be truncated, and the parameter `thresholds`, indicating the thresholds used in the visualization by the `ctHeatmap` function

The object resulting from the `crosstalk` function is an object with three elements: the `heatmap` object, containing the over-representation p-values and the thresholds for the visualization, the `overlap` object, containing overlap information among pairs of pathways, and the `crescentGenes` object, containing information on genes that belong to a pathway but not to the intersection with other pathways.

The crosstalk matrix can be plotted using the function `ctHeatmap`, and the results can be seen in Fig. 5.1. In this function, the parameter `heatmap` is a matrix containing the p-values of pathways obtained by the procedure explained in Section 4.2.1. The two parameters

`threshold1p` and `threshold5p` represent the two thresholds to be plotted on the map. The parameter `title` controls the text on top of the map, and the parameter `cex.axis.matrix` controls the size of axes labels. This parameter is necessary when printing the heatmap directly to pdf, as the size of labels differs from the size visualized in R plots.

```
> ctHeatmap(heatmap = objectCrossTalk$heatmap$map,
+         threshold1p=objectCrossTalk$heatmap$thr1p,
+         threshold5p=objectCrossTalk$heatmap$thr5p,
+         title=paste(c("p-hyper(raw) ", 'ctTest'),
+         collapse=""),
+         cex.axis.matrix=0.3)
```

5.4 Identification of independent functional modules

The second step of the analysis is the identification of independent functional modules identification described in Section 4.2.3 performed by the `addModules` function. The result of this function is an object with the same format as the pathway data object obtained by the `getPathways` function.

```
> modList <- addModules(pathwayData = paths$pathways, dedata = micedata,
+         pathwayTitles = paths$titles, thresholds = c(1e-6,1e-2), 0.25)
```

The first parameter of the `addModules` function is the pathway knowledge. The second parameter, `dedata`, represents the experimental data. The third and optional parameter `pathwayTitles` contains the name of each pathway. It is represented by a character vector of the same length of `pathwayData`. The parameter `thresholds` contains two thresholds. Modules are searched within all pathways whose FDR-corrected p-value is below the first threshold, and then the same process is repeated for pathways whose FDR-corrected p-value is between the first and the second threshold. If only one threshold is needed, the two values can be set as the same. Finally, the `distanceThreshold` parameter contains the similarity threshold used to merge modules, as explained in Section 4.2.3. Two modules that are closer (in terms of Jaccard distance) than this parameter are merged together.

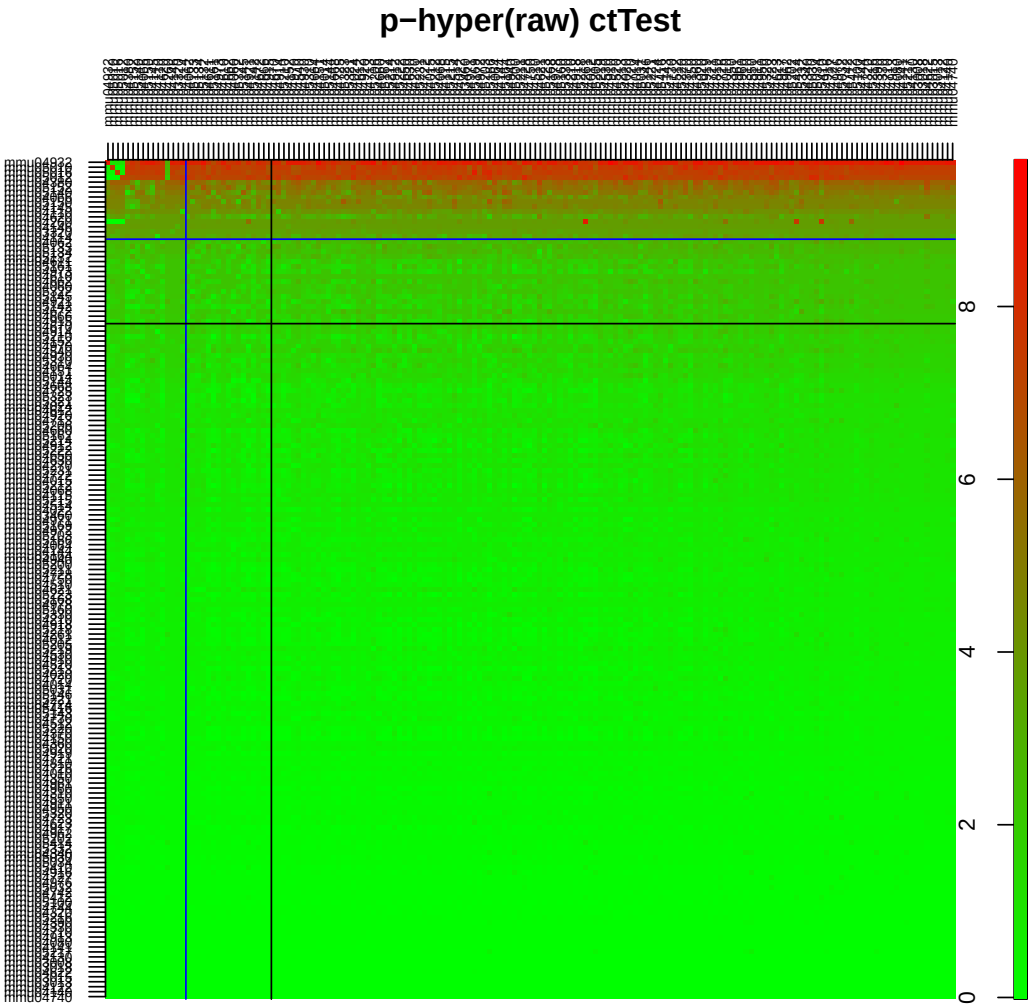


Figure 5.1: Crosstalk matrix. The color of each cell represents the p-value: bright red for p-values close to zero, bright green for p-values close to 1. Cells on the diagonal contain the p-values of the classical ORA, ordered from the most significant one to the least significant one. The cell $P_{i,j}$ contains the p-value of pathway P_i after the effect of P_j is removed. The horizontal and vertical lines represent the thresholds chosen.

The object returned by this function is a list of pathways that includes: i) original pathways as passed as input, ii) novel functional modules that are impacted in the given condition (described by `dedata`) independently from the pathways they belong to, and iii) the pathways from which such independent modules have been removed.

5.5 Maximum impact estimation

Lastly, the maximum impact estimation is obtained by using the `mieMatrices` function on the data.

```
> MIEMatrices <- mieMatrices(deData = micedata, pathwayData = modList,
+         xThreshold=.01,  stopThreshold=1e-2)
```

The parameters of the `mieMatrices` function are the following. `deData` represents the experimental data. `pathwayData` represents the pathway knowledge. `stopThreshold` is related to the iterative nature of the maximum impact estimation procedure. At each iteration the method computes a vector of probabilities (see Section 4.2.2 for details), and when the distance between two successive vectors is smaller than stopThreshold the function stops. The `xThreshold` parameter controls the pathways for which the maximum impact estimation is performed. The value of this parameter represents a p-value. Pathways with FDR-corrected ORA p-value above the threshold are not considered in the maximum impact estimation.

The result of the `mieMatrices` function is an object that contains the two maximum impact matrices. These matrices can be used to compute the crosstalk-corrected over-representation p-values, with the function `computeORA` as follows:

```
> orares <- computeORA(micedata, miematrices = MIEMatrices,
+         pathwaydata = modList)
```

The format of the result is the following:

```
> head(orares)
[1] "mmu04260+mmu04932+mmu05010+mmu05012+mmu05016"
[2] "mmu04145+mmu05140+mmu05152"
```

```
[3] "mmu04146"

[4] "mmu04110+mmu04114"

[5] "mmu03320"

[6] "mmu04066"
```

	ID	pvalues	p.adj.fdr.
module1	mmu04260+mmu04932+mmu05010+mmu05012+mmu05016	5.372178e-10	9.293868e-08
module2	mmu04145+mmu05140+mmu05152	1.527643e-04	1.258026e-02
mmu04146	mmu04146	2.181548e-04	1.258026e-02
module3	mmu04110+mmu04114	3.503756e-04	1.515375e-02
mmu03320	mmu03320	1.358556e-03	4.700603e-02
mmu04066	mmu04066	2.512884e-03	7.245483e-02

	de	size
module1	26	89
module2	9	29
mmu04146	15	73
module3	9	32
mmu03320	12	61
mmu04066	10	49

The pathways belonging to each module are in the `modList` object and can be visualized as follows.

```
> modList$module1
[1] "mmu04260+mmu04932+mmu05010+mmu05012+mmu05016"
> modList$module2
[1] "mmu04145+mmu05140+mmu05152"
> modList$module3
[1] "mmu04110+mmu04114"
```

And the DE genes belonging to the module:

```
> orares['module1','genes']
 [1]  70316  66945  11950  69875 230075  12867  12869  12859  12862  12865
[11]  11947  67680  67273  66043  22272  66142  22273  66445  66576  66594
[21]  66694  67003  78330 227197 226646  66152
```

CHAPTER 6 CONCLUSIONS

The identification of biological processes involved with a certain phenotype, such as a disease or drug treatment, is the goal of the majority of life sciences experiments. Pathway analysis methods are used to interpret high-throughput biological data to identify such processes by incorporating information on biological systems to translate data into biological knowledge. In this thesis we identified a number of issues affecting existing pathway analysis methods, and we proposed a number of approaches addressing these issues, allowing for a better understanding of phenotype mechanisms.

In the first part of the thesis we developed methods to tackle a number of issues with the most widely pathway analysis method that takes into account the topology of each pathway, the impact analysis. The first issue is that the current implementation of the impact analysis does not take into account the statistical significance of individual genes in the analysis. Without this kind of information, genes with marginal significance are considered as being as important as genes with high significance values, potentially introducing noise in the analysis. The first method developed in this thesis allows for incorporation of gene significance in the analysis, and allows to take full advantage of all the measured gene expression changes, rather than relying on arbitrarily set thresholds to focus on a subset of genes. In addition to that, we assessed the performance of a number of methods to detect the efficiency of signal propagation on signaling pathways. Lastly, we developed an objective method for assessing individual gene contributions by using a genetic algorithm approach. This method is the first method that tackles an important issue in pathway analysis: the objective estimation of the parameters of pathway analysis methods. Most methods for pathway analysis include a number of parameters that are often set based on trial and error, by analyzing a small number of datasets, real or simulated. The genetic algorithm framework developed here is the first to use an extensive collection of real datasets to estimate these parameters. In this work the framework was used to assess the individual contribution of genes to the pathways they belong to, and such contributions were used in the impact analysis of signaling pathways, and

the results show the effectiveness of evolutionary computation techniques in the optimization of parameters in bioinformatics applications. Also, this framework is general enough to be applied to all pathway analysis methods.

In the second part of this thesis we addressed an issue related to how overlap among pathways affects the results of pathway analysis methods. Pathway analysis methods are used to interpret high-throughput biological data to identify biological processes involved with a certain condition, by calculating a p-value that aims to quantify such involvement. We showed that, although these p-values were thought to be independent, this is not the case, and that many pathways can considerably affect each other's p-values through a "crosstalk" phenomenon. We showed that all three major categories of pathway analysis methods (enrichment analysis, functional class scoring, and topology-based methods) are severely influenced by crosstalk phenomena. Using real pathways and data, we showed that in some cases pathways with significant p-values are not biologically meaningful, and that some biologically meaningful pathways with non-significant p-values become statistically significant when the crosstalk effects of other pathways are removed. We developed an approach able to *detect* and *correct* crosstalk effects, as well as *identify independent functional modules*. We assessed this novel approach on data from five real experiments coming from four phenotypes involving two species. In all cases, this approach was able to eliminate most false positives, as well as correctly identify as significant pathways that had been biologically proven to be involved in the given condition, yet not found to be significant by the classical analysis. We also found several independent functional modules including a *mitochondrial activity* module active in different stages of fat remodeling in mice, and an *integrin-mediated ECM signaling* found to be involved in hormone treatment in post-menopausal women and cervical ripening in pregnant women. Interesting, the latter module was extracted independently from the crosstalk interactions of two different groups of pathways, in the two conditions analyzed.

This approach is a departure from the current paradigm that considers the pathways as static models, independent of the phenotype. In the view proposed here, various spe-

cific modules, or sub-pathways, can be dynamically linked to specific conditions. When such independent functional modules are identified in independent conditions, such as the *integrin-mediated ECM signaling*, these modules could be considered as candidate new pathways.

REFERENCES

[1] Biocyc, pathway/genome databases and pathway tools software.

[2] J. Aach and G. M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, 2001.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Molecular Biology of the Cell 4th edition. *Garland Science*, 2002.

[4] S. Amaral, P. Papanek, and A. Greene. Angiotensin II and VEGF are involved in angiogenesis induced by short-term exercise training. *American Journal Of Physiology-Heart And Circulatory Physiology*, 281(3):H1163–H1169, Sep 2001.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, May 2000.

[6] S. Badaloni and M. Falda. Coping with uncertainty in temporal gene expressions using symbolic representations. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, 11–20. Springer, 2010.

[7] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1st edition, 1957.

[8] L. Beltrame, E. Calura, R. R. Popovici, L. Rizzetto, D. R. Guedez, M. Donato, C. Romualdi, S. Drăghici, and D. Cavalieri. The biological connection markup language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics*, 27(15):2127–2133, 2011.

[9] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society B*, 57(1):289–300, 1995.

[10] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple

testing under dependency. *Annals of Statistics*, 29(4):1165–1188, August 2001.

[11] S. D. Berry, R. D. Howard, and R. M. Akers. Mammary localization and abundance of laminin, fibronectin, and collagen IV proteins in prepubertal heifers. *J Dairy Sci*, 86(9):2864–74, 2003.

[12] BioCarta. BioCarta - Charting Pathways of Life. http://www.biocarta.com.

[13] E. H. Bishop. Pelvic Scoring For Elective Induction. *Obstetrics and gynecology*, 24:266–268, Aug. 1964.

[14] W. J. Blake, M. Kærn, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[15] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences of the United States of America*, 101(7):2173–2178, 2004.

[16] A. Bollig-Fischer, T. G. Dewey, and S. P. Ethier. Oncogene activation induces metabolic transformation resulting in insulin-independence in human breast cancer cells. *PLoS One*, 6(3):e17959, 2011.

[17] S. Campbell and M. Whitehead. Estrogens For Menopausal Flushing. *British Medical Journal*, 1(6053):104–105, 1977.

[18] P. Cardona-Gomez, M. Perez, J. Avila, L. M. Garcia-Segura, and F. Wandosell. Estradiol inhibits GSK3 and regulates interaction of estrogen receptors, GSK3, and beta-catenin in the hippocampus. *Mol Cell Neurosci*, 25(3):363–73, 2004.

[19] A. Carracedo and P. Pandolfi. The pten–pi3k pathway: of feedbacks and cross-talks. *Oncogene*, 27(41):5527–5541, 2008.

[20] D. Cavalieri, D. Rivero, L. Beltrame, S. I. Buschow, E. Calura, L. Rizzetto, S. Gessani, M. C. Gauzzi, W. Reith, A. Baur, R. Bonaiuti, M. Brandizi, C. De Filippo, U. D'Oro, S. Drăghici, I. Dunand-Sauthier, E. Gatti, F. Granucci, M. Gündel, M. Kramer, M. Kuka, A. Lanyi, C. J. Melief, N. van Montfoort, R. Ostuni, P. Pierre, R. Popovici,

E. Rajnavolgyi, S. Schierer, G. Schuler, V. Soumelis, A. Splendiani, I. Stefanini, M. G. Torcia, I. Zanoni, R. Zollinger, C. G. Figdor, and J. M. Austyn. DC-ATLAS: a systems biology resource to dissect receptor specific signal transduction in dendritic cells. *Immunome Research*, 6(1):10, 2010.

[21] A. Ceol, A. C. Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. Mint, the molecular interaction database: 2009 update. *Nucleic acids research*, gkp983, 2009.

[22] E. Cerami, B. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research*, 39(Database issue):D685–D690, 2011.

[23] R. T. Chlebowski, G. L. Anderson, J. E. Manson, A. G. Schwartz, H. Wakelee, M. Gass, R. J. Rodabough, K. C. Johnson, J. Wactawski-Wende, J. M. Kotchen, J. K. Ockene, M. J. O'Sullivan, F. A. Hubbell, J. W. Chien, C. Chen, and M. L. Stefanick. Lung Cancer Among Postmenopausal Women Treated With Estrogen Alone in the Women's Health Initiative Randomized Trial. *Journal of the National Cancer Institute*, 102(18):1413–1421, Sept. 2010.

[24] R. T. Chlebowski, A. Schwartz, H. Wakelee, G. L. Anderson, M. L. Stefanick, J. E. Manson, J. W. Chien, C. Chen, J. Wactawski-Wende, and M. Gass. Non-small cell lung cancer and estrogen plus progestin use in postmenopausal women in the Women's Health Initiative randomized clinical trial – Chlebowski et al. 27 (18): CRA1500 – ASCO Meeting Abstracts. *Journal of Clinical Oncology*, 27(185), 2009.

[25] R. T. Chlebowski, A. G. Schwartz, H. Wakelee, G. L. Anderson, M. L. Stefanick, J. E. Manson, R. J. Rodabough, J. W. Chien, J. Wactawski-Wende, M. Gass, J. M. Kotchen, K. C. Johnson, M. J. O'Sullivan, J. K. Ockene, C. Chen, and F. A. Hubbell. Oestrogen plus progestin and lung cancer in postmenopausal women (Women's Health Initiative trial): a post-hoc analysis of a randomised controlled trial. *Lancet*, 374(9697):1243–1251, Oct. 2009.

[26] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2014.

[27] M. R. Del Sorbo, W. Balzano, M. Donato, and S. Drăghici. Assessing co-regulation of directly linked genes in biological networks using microarray time series analysis. *Biosystems*, 114(2):149–154, 2013.

[28] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer, I. Vastrik, G. Wu, P. D'Eustachio, C. Schaefer, J. Luciano, et al. The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9):935–942, 2010.

[29] M. Donato, Z. Xu, A. Tomoiaga, J. G. Granneman, R. G. MacKenzie, R. Bao, N. G. Than, P. H. Westfall, R. Romero, and S. Drăghici. Analysis and correction of crosstalk effects in pathway analysis. *Genome Research*, 23(11):1885–1893, 2013.

[30] S. Drăghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichiţa, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.

[31] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

[32] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1):1–26, 1979.

[33] B. Efron and R. Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

[34] B. Efron and R. J. Tibshirani. *An Introduction to Bootstrap*. Chapman and Hall, London, UK, 1993.

[35] A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.

[36] K. Fabel, K. Fabel, B. Tam, D. Kaufer, A. Baiker, N. Simmons, C. Kuo, and T. Palmer. VEGF is necessary for exercise-induced adult hippocampal neurogenesis. *European Journal Of Neuroscience*, 18(10):2803–2812, NOV 2003.

[37] M. Fukata and M. T. Abreu. Role of toll-like receptors in gastrointestinal malignancies. *Oncogene*, 27(2):234–243, 2008.

[38] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[39] C. S. Gillespie, G. Lei, R. J. Boys, A. Greenall, and D. J. Wilkinson. Analysing time course microarray data using bioconductor: a case study using yeast2 affymetrix arrays. *BMC research notes*, 3(1):81, 2010.

[40] E. Glaab, A. Baudot, N. Krasnogor, and A. Valencia. TopoGSA: network topological gene set analysis. *Bioinformatics*, 26(9):1271–1272, 2010.

[41] D. E. Goldberg, K. Deb, and J. H. Clark. Genetic algorithms, noise, and the sizing of populations. *IlliGAL report 91010*, 1991.

[42] C. Gondro and B. Kinghorn. A simple genetic algorithm for multiple sequence alignment. *Genet. Mol. Res*, 6(4):964–982, 2007.

[43] J. G. Granneman, P. Li, Z. Zhu, and Y. Lu. Metabolic and cellular plasticity in white adipose tissue I: effects of beta3-adrenergic receptor activation. *American Journal Of Physiology-Endocrinology And Metabolism*, 289(4):E608–616, 2005.

[44] R. W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[45] G. Han, G. Buchanan, M. Ittmann, J. M. Harris, X. Yu, F. J. Demayo, W. Tilley, and N. M. Greenberg. Mutation of the androgen receptor causes oncogenic transformation of the prostate. *Proc Natl Acad Sci U S A*, 102(4):1151–6, 2005.

[46] P. Hanifi-Moghaddam, B. Boers-Sijmons, A. H. A. Klaassens, F. H. van Wijk, M. A. den Bakker, M. C. Ott, G. L. Shipley, H. A. M. Verheul, H. J. Kloosterboer, C. W. Burger, and L. J. Blok. Molecular analysis of human endometrium: short-term tibolone signaling differs significantly from estrogen and estrogen plus progestagen signaling. *Journal of Molecular Medicine-jmm*, 85(5):471–480, May 2007.

[47] S. S. Hassan, R. Romero, R. Haddad, I. Hendler, N. Khalek, G. Tromp, M. P. Diamond, Y. Sorokin, and J. Malone. The transcriptome of the uterine cervix before and after spontaneous term parturition. *American Journal of Obstetrics & Gynecology*, 195(3):778–786, Sept. 2006.

[48] S. S. Hassan, R. Romero, A. L. Tarca, C.-L. Nhan-Chang, P. Mittal, E. Vaisbuch, J. M. Gonzalez, T. Chaiworapongsa, R. Ali-Fehmi, Z. Dong, N. G. Than, and C. J. Kim. The molecular basis for sonographic cervical shortening at term: identification of differentially expressed genes and the epithelial-mesenchymal transition as a function of cervical length. *American Journal of Obstetrics & Gynecology*, 203(5):472.e1–472.e14, 2010.

[49] S. S. Hassan, R. Romero, A. L. Tarca, C.-L. Nhan-Chang, E. Vaisbuch, O. Erez, P. Mittal, J. P. Kusanovic, S. Mazaki-Tovi, L. Yeo, S. Drăghici, J.-S. Kim, N. Uldbjerg, and C. J. Kim. The transcriptome of cervical ripening in human pregnancy before the onset of labor at term: Identification of novel molecular functions involved in this process. *The Journal of Maternal-Fetal & Neonatal Medicine*, 22(12):1183–1193, Dec. 2009.

[50] B. E. Henderson, R. K. Ross, A. Paganinihill, and T. M. Mack. Estrogen Use and Cardiovascular-disease. *American Journal of Obstetrics & Gynecology*, 154(6):1181–1186, June 1986.

[51] M. Hollander, D. A. Wolfe, and E. Chicken. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

[52] Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah. A susceptibility gene set for early

onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clinical cancer research*, 13(4):1107–1114, 2007.

[53] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI—a complex pathway simulator. *Bioinformatics*, 22(24):3067–3074, 2006.

[54] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[55] S. P. Hussain and C. C. Harris. p53 biological network: at the crossroads of the cellular-stress response pathway and molecular carcinogenesis. *Journal of Nihon Medical School*, 73(2):54–64, Apr. 2006.

[56] B.-H. Jiang and L.-Z. Liu. Pi3k/pten signaling in angiogenesis and tumorigenesis. *Advances in cancer research*, 102:19–65, 2009.

[57] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. REACTOME: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–432, 2005.

[58] G. Jurman and C. Furlanello. A unifying view for performance measures in multi-class prediction. *arXiv preprint arXiv:1008.2908*, 2010.

[59] A. A. Kalaitzis and N. D. Lawrence. A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. *BMC bioinformatics*, 12(1):180, 2011.

[60] M. Kanehisa, S. Goto, S. Kawashima, Y. Okunom, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(Database Issue):277–280, Jan 2004.

[61] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo, and A. R. Pico. Wikipathways: building research communities on biological pathways.

Nucleic acids research, 40(D1):D1301–D1307, 2012.

[62] P. Khatri and S. Drăghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[63] P. Khatri, S. Drăghici, G. C. Ostermeier, and S. A. Krawetz. Profiling gene expression using Onto-Express. *Genomics*, 79(2):266–270, 2002.

[64] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012.

[65] J. L. Klessner, B. V. Desai, E. V. Amargo, S. Getsios, and K. J. Green. EGFR and ADAMs cooperate to regulate shedding and endocytic trafficking of the desmosomal cadherin desmoglein 2. *Molecular biology of the cell*, 20(1):328–337, Jan. 2009.

[66] A. P. Kouzmenko, K. Takeyama, S. Ito, T. Furutani, S. Sawatsubashi, A. Maki, E. Suzuki, Y. Kawasaki, T. Akiyama, T. Tabata, and S. Kato. Wnt/beta-catenin and estrogen signaling converge in vivo. *J Biol Chem*, 279(39):40255–8, 2004.

[67] J. E. Ladbury and S. T. Arold. Noise in cellular signaling pathways: causes and effects. *Trends in biochemical sciences*, 37(5):173–178, 2012.

[68] Y.-H. Lee, A. P. Petkova, E. P. Mottillo, and J. G. Granneman. In vivo identification of bipotential adipocyte progenitors recruited by Beta3-adrenoceptor activation and high-fat feeding. *Cell Metabolism*, 15(4):480–491, Apr. 2012.

[69] P. C. Leppert. Anatomy and physiology of cervical ripening. *Clinical obstetrics and gynecology*, 38(2):267–279, June 1995.

[70] P. C. Leppert, J. M. Cerreta, and I. Mandl. Orientation of elastic fibers in the human cervix. *American Journal of Obstetrics & Gynecology*, 155(1):219–224, July 1986.

[71] K. C. Leung, N. Doyle, M. Ballesteros, K. Sjogren, C. K. W. Watts, T. H. Low, G. M. Leong, R. J. M. Ross, and K. K. Y. Ho. Estrogen inhibits GH signaling by suppressing GH-induced JAK2 phosphorylation, an effect mediated by SOCS-2. *PNAS*, 100(3):1016–1021, Feb. 2003.

[72] M. Li, D. F. Carpio, Y. Zheng, P. Bruzzo, V. Singh, F. Ouaaz, R. M. Medzhitov, and A. A. Beg. An Essential Role of the NF-kappa B/Toll-Like Receptor pathway in Induction of Inflammatory and Tissue-Repair Gene Expression by Necrotic Cells. *The Journal of Immunology*, 166(12):7128–7135, 2001.

[73] P. Li, Z. Zhu, Y. Lu, and J. G. Granneman. Metabolic and cellular plasticity in white adipose tissue II: role of peroxisome proliferator-activated receptor-alpha. *American Journal Of Physiology-Endocrinology And Metabolism*, 289(4):E617–626, 2005.

[74] P. Ma, W. Zhong, and J. S. Liu. Identifying differentially expressed genes in time course microarray data. *Statistics in Biosciences*, 1(2):144–159, 2009.

[75] M. S. Mahendroo, A. Porter, D. W. Russell, and R. A. Word. The parturition defect in steroid 5alpha-reductase type 1 knockout mice is due to impaired cervical ripening. *Molecular endocrinology (Baltimore, Md.)*, 13(6):981–992, June 1999.

[76] T. J. Marczynski. GABAergic deafferentation hypothesis of brain aging and Alzheimer's disease revisited. *Brain research bulletin*, 45(4):341–379, 1998.

[77] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[78] P. Mecocci, U. MacGarvey, and M. F. Beal. Oxidative damage to mitochondrial DNA is increased in Alzheimer's disease. *Annals of neurology*, 36(5):747–751, 1994.

[79] H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, et al. The panther database of protein families, subfamilies, functions and pathways. *Nucleic acids research*, 33(suppl 1):D284–D288, 2005.

[80] M. Milkiewicz, M. Brown, S. Egginton, and O. Hudlicka. Association between shear stress, angiogenesis, and VEGF in skeletal muscles in vivo. *Microcirculation*, 8(4):229–241, AUG 2001.

[81] R. Millon, F. Nicora, D. Muller, M. Eber, C. Klein-Soyer, and J. Abecassis. Modu-

lation of human breast cancer cell adhesion by estrogens and antiestrogens. *Clin Exp Metastasis*, 7(4):405–15, 1989.

[82] C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa, and S. Drăghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4:278, 2013.

[83] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-11α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, Jul 2003.

[84] E. P. Mottillo, X. J. Shen, and J. G. Granneman. Role of hormone-sensitive lipase in beta-adrenergic remodeling of white adipose tissue. *Am J Physiol Endocrinol Metab*, 293(5):E1188–97, 2007.

[85] A. T. Naito, H. Akazawa, H. Takano, T. Minamino, T. Nagai, H. Aburatani, and I. Komuro. Phosphatidylinositol 3-kinase-Akt pathway plays a critical role in early cardiomyogenesis by regulating canonical Wnt signaling. *Circ Res*, 97(2):144–51, 2005.

[86] H. D. Nelson, L. L. Humphrey, P. Nygren, S. M. Teutsch, and J. D. Allan. Postmenopausal hormone replacement therapy - Scientific review. *Jama-journal of the American Medical Association*, 288(7):872–881, Aug. 2002.

[87] S. L. Newman, J. E. Henson, and P. M. Henson. Phagocytosis of senescent neutrophils by human monocyte-derived macrophages and rabbit inflammatory macrophages. *The Journal of Experimental Medicine*, 156(2):430, Aug. 1982.

[88] V. Novaro, C. D. Roskelley, and M. J. Bissell. Collagen-IV and laminin-1 regulate estrogen receptor alpha expression and function in mouse mammary epithelial cells. *Journal of Cell Science*, 116(Pt 14):2975–86, 2003.

[89] K.-H. Pan, C.-J. Lih, and S. N. Cohen. Effects of threshold choice on biological conclu-

sions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25):8961–8965, 2005.

[90] W. D. Parker, C. M. Filley, and J. K. Parks. Cytochrome oxidase deficiency in Alzheimer's disease. *Neurology*, 40(8):1302–1302, 1990.

[91] W. D. Parker, J. Parks, C. M. Filley, and B. Kleinschmidt-DeMasters. Electron transport chain defects in Alzheimer's disease brain. *Neurology*, 44(6):1090–1090, 1994.

[92] H. Peterziel, S. Mink, A. Schonert, M. Becker, H. Klocker, and A. C. Cato. Rapid signalling by androgen receptor in prostate cancer cells. *Oncogene*, 18(46):6322–9, 1999.

[93] J. G. Proakis and D. G. Manolakis. *Introduction to digital signal processing*. Prentice Hall Professional Technical Reference, 1988.

[94] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[95] L. H. Saal, P. Johansson, K. Holm, S. K. Gruvberger-Saal, Q.-B. She, M. Maurer, S. Koujak, A. A. Ferrando, P. Malmström, L. Memeo, J. Isola, P.-O. Bendahl, N. Rosen, H. Hibshoosh, M. Ringnér, Å. Borg, and R. Parsons. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant pten tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences*, 104(18):7564–7569, 2007.

[96] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[97] D. Senger, V. L, B. Lf, N. Ja, Y. Kt, Y. Tk, B. B, J. Rw, D. Am, and D. Hf. Vascular-Permeability Factor (VPF, VEGF) In Tumor Biology. *Cancer And Metastasis Reviews*, 12(3-4):303–324, SEP 1993.

[98] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504,

2003.

[99] I. Shureiqi, W. Jiang, X. Zuo, Y. Wu, J. B. Stimmel, L. M. Leesnitzer, J. S. Morris, H.-Z. Fan, S. M. Fischer, and S. M. Lippman. The 15-lipoxygenase-1 product 13-s-hydroxyoctadecadienoic acid down-regulates PPAR-δ to induce apoptosis in colorectal cancer cells. *Proceedings of the National Academy of Sciences*, 100(17):9968–9973, 2003.

[100] M. L. Simpson, C. D. Cox, M. S. Allen, J. M. McCollum, R. D. Dar, D. K. Karig, and J. F. Cooke. Noise in biological circuits. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, 1(2):214–225, 2009.

[101] G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[102] G. K. Smyth. *Limma: linear models for microarray data*, 397–420. Springer, New York, 2005.

[103] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Suppl 1):D535–D539, 2006.

[104] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.

[105] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding of The National Academy of Sciences of the Unites States of America*, 102(43):15545–15550, 2005.

[106] A. L. Tarca, S. Drăghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136, 2012.

[107] A. L. Tarca, S. Drăghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2009.

[108] A. L. Tarca, S. Drăghici, P. Khatri, S. S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero. A novel signaling pathway impact analysis (SPIA). *Bioinformatics*, 25(1):75–82, 2009.

[109] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.

[110] R. D. C. Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005.

[111] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9):2129–2141, 2003.

[112] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. Discovering statistically significant pathways in expression profiling studies. *Proceeding of The National Academy of Sciences of the USA*, 102(38):13544–13549, 2005.

[113] B. C. Timmons, A.-M. Fairhurst, and M. S. Mahendroo. Temporal changes in myeloid cells in the cervix during pregnancy and parturition. *J Immunol*, 182(5):2700–2707, Mar. 2009.

[114] N. Uldbjerg, G. Ekman, A. Malmström, K. Olsson, and U. Ulmsten. Ripening of the human uterine cervix related to changes in collagen, glycosaminoglycans, and collagenolytic activity. *American Journal of Obstetric and Gynecology*, 147(6):662–666, Nov. 1983.

[115] N. Uldbjerg, U. Ulmsten, and G. Ekman. The ripening of the human uterine cervix in terms of connective tissue biochemistry. *Clinical obstetrics and gynecology*, 26(1):14–26, Mar. 1983.

[116] F. Van Batenburg, A. P. Gultyaev, and C. W. Pleij. An APL-programmed genetic al-

gorithm for the prediction of RNA secondary structure. *Journal of Theoretical Biology*, 174(3):269–280, 1995.

[117] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[118] C. Voichiţa, M. Donato, and S. Drăghici. Incorporating gene significance in the impact analysis of signaling pathways. *Proceedings of the International Conference on Machine Learning Applications (ICMLA)*, Dec. 2012.

[119] C. Voichiţa, M. Donato, and S. Drăghici. A genetic algorithms framework for estimating individual gene contributions in signaling pathways. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, 650–657, Cancun, Mexico, 20-23 June 2013. IEEE.

[120] C. Voichiţa, M. Donato, and S. Drăghici. Incorporating gene significance in the impact analysis of signaling pathways. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, 126–131, Boca Raton, FL, USA, 12-15 Dec. 2012. IEEE.

[121] E. Wang, Z.-R. Qian, M. Nakasono, T. Tanahashi, K. Yoshimoto, Y. Bando, E. Kudo, M. Shimada, and T. Sano. High expression of toll-like receptor 4/myeloid differentiation factor 88 signals correlates with poor prognosis in colorectal cancer. *British journal of cancer*, 102(5):908–915, 2010.

[122] W. A. Warr. Chembl. an interview with john overington, team leader, chemogenomics at the european bioinformatics institute outstation of the european molecular biology laboratory (embl-ebi). *Journal of computer-aided molecular design*, 23(4):195–198, 2009.

[123] N. S. Weiss, C. L. Ure, J. H. Ballard, A. R. Williams, and J. R. Daling. Decreased Risk of Fractures of the Hip and Lower Forearm With Post-menopausal Use of Estrogen. *New England Journal of Medicine*, 303(21):1195–1198, 1980.

[124] C. W. Wong, C. McNally, E. Nickbarg, B. S. Komm, and B. J. Cheskis. Estrogen

receptor-interacting protein that modulates its nongenomic activity-crosstalk with Src/Erk phosphorylation cascade. *Proc Natl Acad Sci U S A*, 99(23):14783–8, 2002.

[125] R. A. Word, X.-H. Li, M. Hnat, and K. Carrick. Dynamics of cervical remodeling during pregnancy and parturition: mechanisms and current concepts. *Seminars in reproductive medicine*, 25(1):69–79, Jan. 2007.

[126] H. Yan. *Signal processing for magnetic resonance imaging and spectroscopy*, volume 15. CRC Press, 2002.

[127] L. K. Yeung, H. Yan, A. W.-C. Liew, L. K. Szeto, M. Yang, and R. Kong. Measuring correlation between microarray time-series data using dominant spectral component. In *Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29*, 309–314. Australian Computer Society, Inc., 2004.

[128] J. M. Zahn, R. Sonu, H. Vogel, E. Crane, K. Mazan-Mamczarz, R. Rabkin, R. W. Davis, K. G. Becker, A. B. Owen, and S. K. Kim. Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS genetics*, 2(7):e115, 2006.

[129] W. Zhang and H. T. Liu. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res*, 12(1):9–18, 2002.

[130] X. Zhu, A. K. Raina, H.-g. Lee, G. Casadesus, M. A. Smith, and G. Perry. Oxidative stress signalling in Alzheimer's disease. *Brain research*, 1000(1):32–39, 2004.

[131] H. K. Ziel. Estrogens Role In Endometrial Cancer. *Obstetrics and Gynecology*, 60(4):509–515, 1982.

[132] W. Zou and V. V. Tolstikov. Pattern recognition and pathway analysis with genetic algorithms in mass spectrometry based metabolomics. *Algorithms*, 2(2):638–666, 2009.

ABSTRACT

SYSTEMS BIOLOGY APPROACHES FOR THE ANALYSIS OF HIGH-THROUGHPUT BIOLOGICAL DATA

by

MICHELE DONATO

May 2016

Advisor: Dr. Sorin Draghici

Major: Computer Science (Bioinformatics)

Degree: Doctor of Philosophy

The identification of biological processes involved with a certain phenotype, such as a disease or drug treatment, is the goal of biology experiments. Pathway analysis methods are used to interpret high-throughput biological data to identify such processes by incorporating information on biological systems to translate data into biological knowledge. Current methods share a number of limitations. First, they do not take into account the individual contribution of each gene to the phenotype in analysis. Second, most of the methods include parameters of difficult interpretation, often arbitrarily set. Third, the results of all methods are affected by the fact that pathways are not independent, but communicate through a phenomenon referred to as *crosstalk*. Crosstalk effects heavily influence the results of pathway analysis methods, adding false positives and false negatives, making them difficult to interpret. We developed methods to address these limitations by i) allowing for incorporation of individual gene contributions, ii) developing objective methods for the estimation of parameters of pathway analysis methods, and iii) developing an approach able to detect and correct for crosstalk effects. We show on real and simulated data that our approaches increase specificity and sensitivity of pathway analysis, allowing for a more effective identification of the processes and mechanisms underlying biological phenomena.

AUTOBIOGRAPHICAL STATEMENT

Michele Donato

Education

- Ph.D. Computer Science, Wayne State University, Detroit MI, USA, May 2016.
- M.S. Computer Science, Wayne State University, Detroit MI, USA, October 2015.
- M.S. Computer Engineering, University of Pisa, Pisa, Italy, October 2006.
- B.S. Computer Engineering, University of Pisa, Pisa, Italy, October 2006.

Selected peer-reviewed publications

[1] Adib Shafi, Michele Donato, and Sorin Drăghici. A systems biology approach for the identification of significantly perturbed genes. In *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, September 09-12 2015.

[2] Maria Rosaria Del Sorbo, Walter Balzano, Michele Donato, and Sorin Drăghici. Assessing co-regulation of directly linked genes in biological networks using microarray time series analysis. *Biosystems*, 114(2):149–154, 2013.

[3] Michele Donato, Zhonghui Xu, Alin Tomoiaga, James G Granneman, Robert G MacKenzie, Riyue Bao, Nandor G Than, Peter H Westfall, Roberto Romero, and Sorin Drăghici. Analysis and correction of crosstalk effects in pathway analysis. *Genome Research*, 23(11):1885–1893, 2013.

[4] Călin Voichiţa, Michele Donato, and Sorin Drăghici. A genetic algorithms framework for estimating individual gene contributions in signaling pathways. In *Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 650–657, Cancun, Mexico, 20-23 June 2013. IEEE.

[5] Cristina Mitrea, Zeinab Taghavi, Behzad Bokanizad, Samer Hanoudi, Rebecca Tagett, Michele Donato, Călin Voichiţa, and Sorin Drăghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4:278, 2013.

[6] Călin Voichiţa, Michele Donato, and Sorin Drăghici. Incorporating gene significance in the impact analysis of signaling pathways. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, pages 126–131, Boca Raton, FL, USA, 12-15 December 2012. IEEE.

[7] Luca Beltrame, Enrica Calura, Razvan R Popovici, Lisa Rizzetto, Damariz Rivero Guedez, Michele Donato, Chiara Romualdi, Sorin Drăghici, and Duccio Cavalieri. The biological connection markup language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. *Bioinformatics*, 27(15):2127–2133, 2011.

[8] Michele Donato and Sorin Drăghici. Signaling pathways coupling phenomena. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pages 1–6, Barcelona, Spain, 18-23 July 2010. IEEE.