January 2018

Qualitative Change Detection Approach For Preventive Therapies

Cristina Mitrea
Wayne State University, cristina@wayne.edu

QUALITATIVE CHANGE DETECTION APPROACH FOR PREVENTIVE THERAPIES

by

CRISTINA FLORENTINA MITREA

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

2018

MAJOR: COMPUTER SCIENCE

Approved By:

Sorin Drăghici
Advisor Date

Monica Brockmeyer

Loren Schwiebert

Aliccia Bollig-Fischer

DEDICATION

In memory of my uncle, Robert Eghet, who inspired me to never stop learning.

ACKNOWLEDGEMENTS

First, I would like to thank my academic advisor and mentor, Dr. Sorin Drăghici, for his advice and guidance through the challenges of graduate school and beyond. I have learned much more than the specifics of a scientific domain and acquired skills that will serve me well throughout all my professional and personal life. I have gotten to know my strengths and weaknesses, and by being pushed to the limits of what I thought I could do, I achieved goals beyond my expectations.

Second, I would like to thank my dissertation committee members for their questions and comments that provided my work with the important practical aspect and strengthened the theoretical aspect by going through every step of the decision making to check its soundness. I specifically want to thank Dr. Bollig-Fischer who was my second mentor throughout my graduate studies and more so during the last three years. I learned from her: perseverance, to always be critical of my work, and what true collaboration means.

Third, I want to thank the past and present members of the Intelligent Systems and Bioinformatics Laboratory (ISBL) for their support and collaboration during graduate school. ISBL has been my home away from home. Sharing successes and bumps in the road created friendships that go beyond the lab. A big thank you to past and present members of the Bollig-Fischer lab for explaining cell culture protocols among many others. This lab is an amazing learning environment where complementary skills are best put to practice.

Last, but not least, I am forever grateful to my husband Radu, my parents Maria and Virgil, and my brother Florin. Even though my parents and brother are across an ocean, their trust, their constant encouragements and unconditional support of all my dreams as well as their understanding and patience for having to be far from them to achieve these dreams, have made all this possible. Radu is the one with whom I share all, successes and frustrations. He has patiently put up with all my stubbornness, perfectionism, compulsions, procrastination, and made it all better.

This has been such a journey!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Introduction to systems biology

In systems biology an organism is viewed as an integrated and interacting network of genes, proteins and biochemical reactions. These interactions give rise to new properties called emergent. These are properties that no single component possesses, such as motility. It is expected that studying biological processes at the systems level will bring a better understanding of the organism. The discovery of RNA interference, a process through which specialized molecules are used to prevent the activity of a gene, gave new possibilities to explore the function of genes. Before the 1990s, it was unthinkable to interrogate genomes and manipulate cells [3]. However, now we can not only interrogate the entire genome, but edit it [125, 18, 31, 14] and even synthesize new life forms [57].

To better understand systems biology we need to distinguish between two different methodological approaches usually associated with it. One is the holistic approach, where the focus is on the behavior of whole system and the other is the reductionist approach, where the focus is on the behavior of the most relevant lower-level components of the system. We share the view of Paul Nurse who considers scientific methodologies and questions to be reductionist, but he advocates for a more elaborate understanding of biological systems that takes into consideration that components interact and these interactions are constrained by overall functions acting at higher levels [120].

Another original perspective on systems biology is presented by Uri Alon. He underlines the importance of network motifs, simple and repetitive structures encountered in biological networks. In the quest for simplicity, a divide and conquer approach is the only way to uncover the principles behind the structure and behavior of a system [4]. The difference to other reductionist approaches is on the relevant low-level components which in this case are not genes, but network motifs.

1.2 Problem statement and background

Supported by intense research on specific diseases such as cancer, biological experiments are being performed at a higher rate than ever before, with increasing efficiency. However, at the end of a laboratory experiment there is no conclusion, but lots of data to analyze. Currently, data processing is mostly done with computers and specialized software. Truthfully, this is the most efficient approach if we consider the amount of information to process. In the case of bioinformatics, one popular data analysis method is pathway analysis, that takes as input high-throughput data comparing two phenotypes and a list of biological pathways with the goal of raking them based on the degree to which each they are impacted by the phenotype. This analysis technique, although sophisticated, does not consider the evolution of the phenomena in time and it does not readily allow the integration of different types of data.

1.2.1 Overview of systems biology approaches

An important step in developing novel therapeutics is the investigation of the effects of biochemical products on living beings and the environment. Traditional approaches for understanding the mechanisms underlying such effects were focused on single points of action, e.g. genes, and considered the observed effect as emerging from the properties of individual parts. These approaches disregarded any type of relationship among the points of action. However, living systems are complex and dynamic, and their reaction to external stimuli, such as the contact with toxic compounds, may be hard to predict by only looking at individual biological entities. A new paradigm emerged in the last 15 years, represented by approaches that consider the fact that genes, (or, in general, components of a biological systems) do not work independently from each other, but they are part of one or more *subsystems* carrying out specific tasks in the organism. These approaches leverage two types of information: the data coming from modern high-throughput technologies, suitable for identifying patterns in the change of thousands of genes, proteins or metabolites across multiple samples, and the

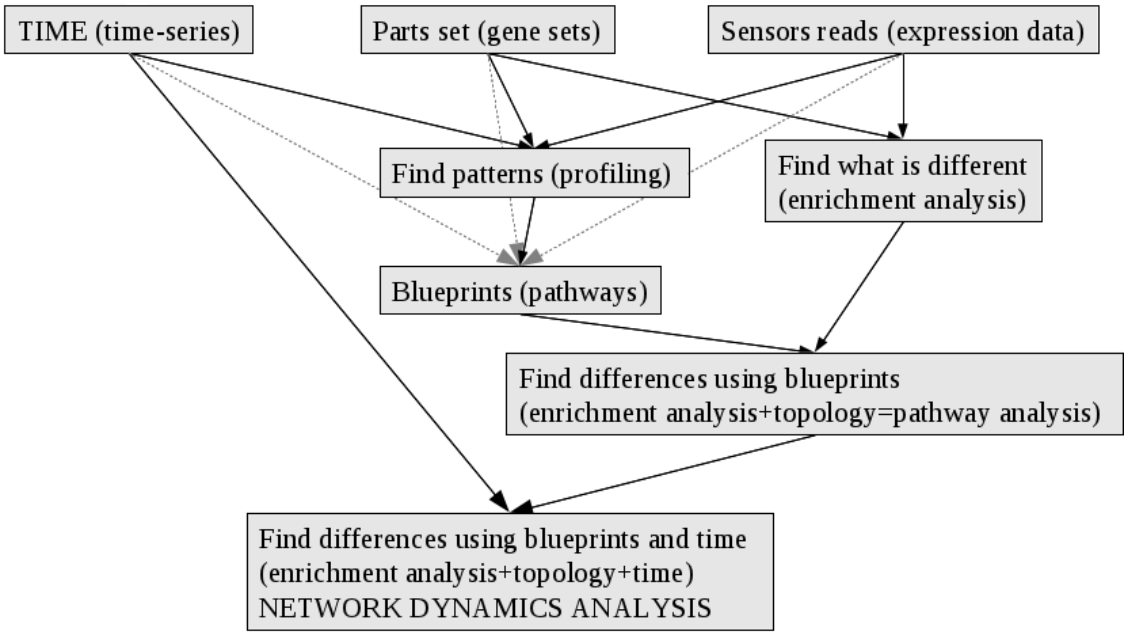

Figure 1.1: A graphical overview of analysis approaches for snapshot or time-course data in the context of biological systems. At the top level we display various elements of a system to be analyzed: i) the time, which is a parameter of the system's dynamics, ii) the components, which are the parts set, and iii) the sensor reads, which are composed of one or several snapshots of the components quantitative characteristics. The edges are links between inputs and analysis methods. For example, in the enrichment analysis we need a list of genes (parts set) and expression data for those genes (sensors reads).

availability of information regarding interactions among biological entities. In other words, these new approaches, which belong to the field of *systems biology*, focus on capturing *system-level* events happening in an organism, by exploiting high-throughput data and knowledge of the interactions.

Systems biology provides a conceptual basis for integrating the multitude of components and interactions underlying cellular processes. A summary of systems biology approaches can be seen in Fig. 1.1. Traditional approaches are focused mainly on individual components or pathways, network analysis of high-throughput data offers the opportunity to incorporate biological complexity by incorporating the connections between various components. New applications in drug discovery span a broad spectrum including modeling disease-altered networks, identification of drug-target interactions, and screening of chemi-

cal libraries [9]. Network and systems biology offer a novel way of approaching drug discovery by developing models that consider the global physiological environment of protein targets, and the effects of modifying them, without losing the key molecular details.

1.2.2 Change detection approaches for preventive therapies

In most, if not all, non-trauma health-care cases, pathological conditions are defined by phenotypic or clinical changes. For example, cancer is usually diagnosed after the patient experiences symptoms caused by significant transformations in their physiology. However, the progression from a healthy state to one of disease is gradual, happening over a period of time. This is particularly true in the case of conditions such as cancer or neurodegenerative disorders, for which the onset of the underlying pathology is believed to begin much earlier than the clinical, detectable onset [133, 80]. What if one could identify a departure from the healthy state well before a tumor is present, when changes can perhaps still be reversed? What if one could identify qualitative changes in the states of a biological system without even knowing what the states are? In this thesis, we propose a technique that aims at identifying such qualitative changes without a priori knowledge about the nature of the changes.

The goal is to develop an approach that can detect qualitative changes in the system, where a qualitative change is defined as a change that involves observable macroscopic phenotypical or clinical changes. We should emphasize that no known approach is available to tackle this type of problems. There are no clearly defined states or classes available a priori, so no supervised machine learning approaches can be used. We would like to be able to detect changes as they happen if possible, without massive amounts of partially redundant data collected beforehand, so no unsupervised methods could be used to extract common features and build clusters. Here we are looking at a system without having a reference set of genes, so no enrichment approach will be useful. Finally, there is no predefined phenotype, and therefore no gene set analysis methods can be employed either. What we would like to

achieve here is a method capable of: (1) monitoring the activity of a system by taking periodic measurements and (2) detecting when a specific system undergoes a qualitative change without prior knowledge about it.

In practical terms, the data to be analyzed is a time series of gene expression or any other sequential measurements of systemic states such as the one described in disease progression. Time-series data have been used in many ways, e.g. to infer information regarding regulatory mechanisms, the rate of change for a gene, the order in which genes are (de)activated, and the causal effects of gene expression changes [12]. Another common goal for the analysis of time-series data is to identify disease biomarkers presenting as a single gene or a network of genes [93].

Another challenge is to identify disease mechanisms for complex diseases when different types of data are involved in disease initiation and progression. For instance some complex diseases might have a genomic mechanisms, but also a metabolic component. Therefore high-throughput data analyses aimed at gaining insight into disease mechanisms should be developed for the various of types of biological data. That is not a trivial task as different types of biological data come from different technologies, quantify distinct biological features, and are often stored in different databases using various structures for the data. A specific case of such analysis is pathway analysis, where a disproportionate number of methods have been developed for gene expression data as opposed to metabolite data [111, 118]. Part of the challenge comes from the technology limitation. High-throughput gene expression data (tens of thousands, very often all genes out of approximately 25,000) has been readily available for more than two decades [169] as opposed to high-throughput metabolic data (hundreds to thousands of metabolites out of approximately 200,000) [134, 186, 20] that only recently reached real throughput [124]. In addition to the scarcity of the data, the versatile structure of the metabolic pathways further restricts the development of metabolic pathway analyses. As a consequence, most of the metabolic pathway analyses available use a metabolite set approach where the metabolites are considered to work independent of one another disre-

garding important information provided by the bio-chemical reactions [99]. There are only very few methods for metabolic pathway analysis that make use of more of the information a pathway contains [111, 118].

1.2.3 Our contributions

In this section, we present the summary of our contributions and the outline of this dissertation. These contributions include three major research projects, starting with a survey, a novel method for the analysis of time-course data in the context of biological systems, and a novel analysis integrating two different type of data with the goal of gaining insights into the mechanisms of an aggressive type of breast cancer.

- **A survey on topology-based pathway analysis approaches.** The goal of pathway analysis approaches is to identify pathways that are significantly impacted when comparing two phenotypes. Many current methods are based on algorithms that consider pathways as simple gene lists, dramatically under-utilizing the knowledge that such pathways are meant to capture. During the past years, a plethora of methods claiming to incorporate various aspects of the pathway topology have been proposed. These topology-based methods, sometimes referred to as "third generation", have the potential to better model the phenomena described by pathways. Although there are a large variety of approaches used for this purpose, no review was available to offer guidance for potential users and developers. The review we published in 2013 covers 22 such topology-based pathway analysis methods published in the last 15 years [111], and a follow-up was published in 2018 [118]. This work compares the methods based on input, output and mathematical models, and also identifies and discusses challenges faced by researches when developing a new topology-based pathway analysis method.

- **A qualitative change detection analysis.** In this thesis, a paradigm shift is proposed from treating disease to preserving the healthy state. A novel analysis method (QCD) was developed to detect intervals when a biological system undergoes qualita-

tive changes such as the transition from healthy to disease using time-course data and a network representing the biological system.

- **A metabolic pathway impact analysis analysis using the stoichiometry of the reaction.** This thesis proposes a novel analysis method aimed at identifying the metabolic pathways significantly changed between two phenotypes. The input is metabolite data from two different phenotypes and a list of metabolic pathways represented as sets of bio-chemical reactions. The analysis uses the change in metabolite concentrations between the two phenotypes and the stoichiometry of the bio-chemical reaction to evaluate the change between phenotypes at the reaction level. Then, the change at the reaction level is propagated from one reaction to another to evaluate the impact of the change between phenotypes at the pathway level. The result is a ranked list of the metabolic pathways given as the input.

1.3 Outline

The rest of this dissertation is organized as follows. In chapter 2, we present the survey of topology-based pathway analysis methods. In chapter 3, we present the novel analysis method aimed to detect qualitative changes in biological systems. In chapter 4, we present the novel analysis method for metabolic pathway analysis. Finally, in chapter 5, we draw some conclusions and outline future work.

CHAPTER 2: A SURVEY OF PATHWAY ANALYSIS APPROACHES

2.1 Introduction to pathway analysis

The goal of pathway analysis approaches is to identify pathways that are significantly impacted when comparing two phenotypes. Many current methods consider pathways as simple gene lists, dramatically under-utilizing the knowledge that such pathways are meant to capture. During the past years, a plethora of methods claiming to incorporate various aspects of the pathway topology have been proposed. Topology-based methods have the potential to better model the phenomena described by pathways. There is now a large variety of approaches used for this purpose, and a review is useful to offer guidance for potential users and developers. This chapter covers 22 such topology-based pathway analysis methods [111] and identifies an additional 12 methods developed in the past five years [118]. The methods are compared based on: type of pathways analyzed (e.g. signaling or metabolic), input (subset of genes, all genes, fold changes, gene p-values, etc.), mathematical models, pathway scoring approaches, output (one or more pathway scores, p-values, etc.) and implementation (web-based, standalone, etc.). This work identifies and discusses challenges, arising both in methodology and in pathway representation, including inconsistent terminology, different data formats, lack of meaningful benchmarks, and the lack of tissue and condition specificity.

The goal of the pathway analysis is to identify the signaling and metabolic pathways that are significantly perturbed in a given phenotype. There are several approaches that aim at accomplishing this goal in different ways. These approaches can be divided in two major categories: i) gene set enrichment analyses and ii) and topology-based analysis of pathways.

Gene sets consists of a list of genes while pathways incorporate the interactions between genes. The gene set has lost all the structure and the additional information captured by the original pathway. The comparison in Fig. 2.1 shows an example of how much of the important knowledge existent in pathway database is ignored when pathways are treated as simple gene sets.

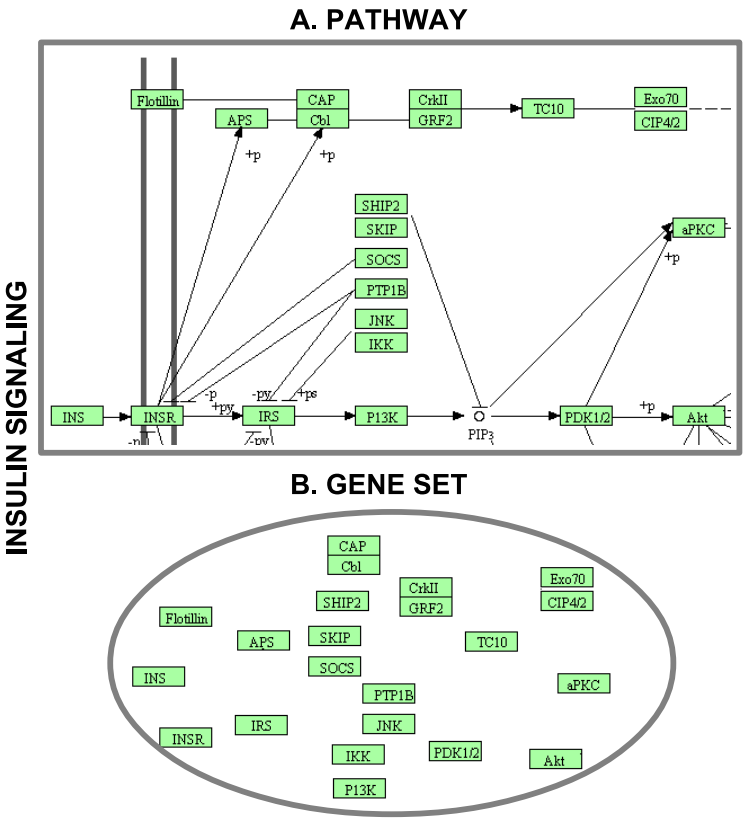

Figure 2.1: Gene sets versus pathways. Panel A shows a small part of the insulin signaling pathway from KEGG. Pathways contain important information regarding gene product (protein) localization, gene, protein or metabolite interactions and the type of these interactions (activation, repression, etc.), the direction of the signal propagation, to name a few. Panel B shows as a gene set the same small part of the insulin signaling pathway from KEGG. There are no interactions in the gene set. Also, any other structural information provided by the pathway is not present in a gene set. Considerable and important information from pathway databases is ignored when pathways are simplified and used as gene sets.

2.2 Input

This chapter focuses on pathway analysis methods that try to exploit some of the information contained in the pathway topology in order to identify the pathways that are significantly impacted in a condition under study. It describes 22 and categorizes an additional 12 topology-based pathway analysis methods designed to analyze either signaling pathways (see Fig. 2.2, top), or metabolic pathways (Fig. 2.2, bottom). In order to address this problem, any pathway analysis method will need: i) a collection of pathways capturing

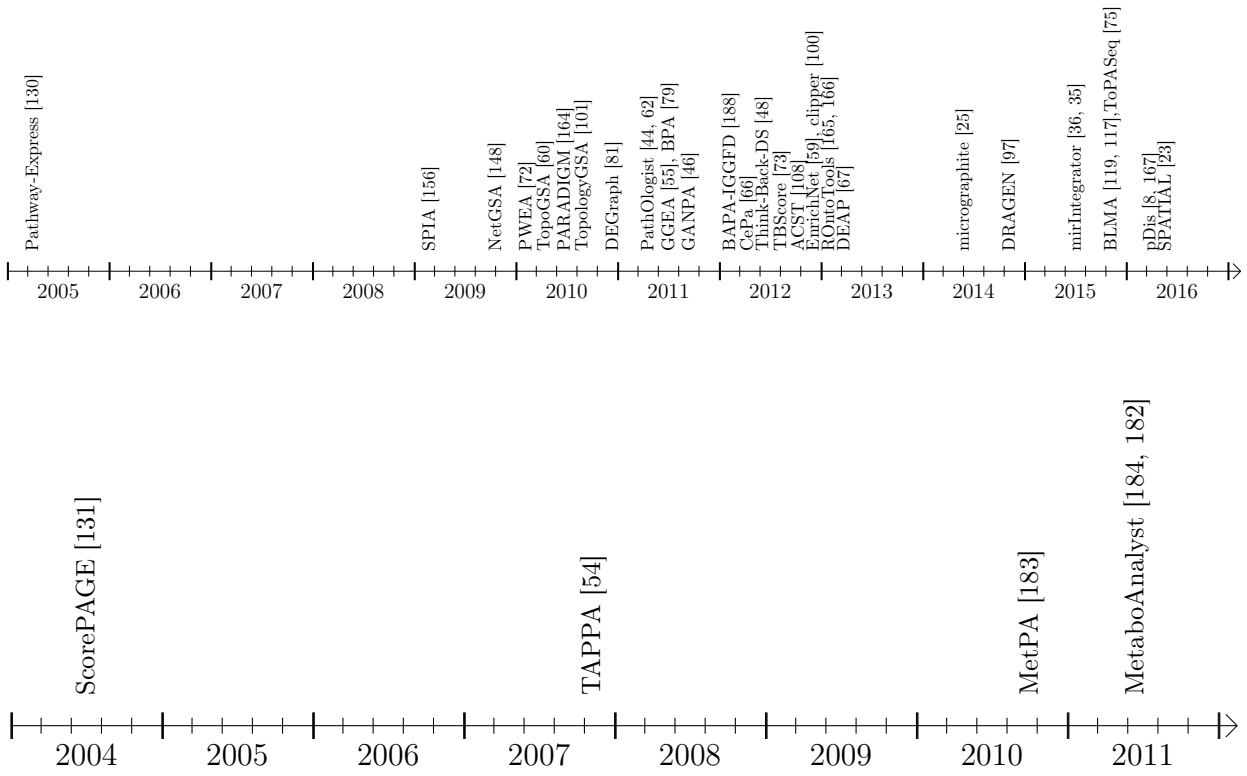

Figure 2.2: TOP : A timeline showing when various signaling pathway analysis tools (total 28) became available. Some of the methods can work with other pathway types, as well. Some of the methods use additional interaction information which can be from an in-house or public gene/protein interaction knowledge base. The commercial tools, iPathwayGuide and MetaCore are not included in this figure. BOTTOM : A timeline showing when various metabolic pathway analysis tools (total 4) became available.

our current knowledge about the interactions of genes, proteins, metabolites, or compounds in an organism (usually from a pathway database), and ii) experimental data in the form of measurements of gene expression, protein abundance, metabolite concentration, or copy numbers. The pathway data is accumulated, updated, and refined by amassing knowledge from scientific literature describing individual interactions or high-throughput experiment results. The experiment data is usually provided by measurements comparing two or more phenotypes such as treated vs. untreated, disease vs. healthy, or treated with drug A vs. drug B.

2.2.1 Experiment data for perturbation analysis

Typically, pathway analysis methods take as input data from high-throughput experiments, such as microarrays, next-generation sequencing, or proteomics. The input format is either a list of gene IDs or a list of such gene IDs associated with measured changes. These changes could be measured with different technologies and therefore can serve as proxies for different biochemical entities. For instance, one could use gene expression changes measured with microarrays, or protein levels measured with a proteomic approach, etc. Transcription data is often used to approximate the proteome, since high-throughput protein abundance data is not readily available.

Pathway analysis methods can use as input the list of all genes considered in the experiment together with their expression values. They can also take as input a subset of genes considered to be differentially expressed (DE) based on a predefined cut-off. The cut-off is typically applied on fold-change, statistical significance, or both. A selection based on both criteria can be performed easily if the data is displayed as a volcano plot, i.e. in a coordinate system that has fold changes on the x axis and the negative log of the p-value on the y axis. In such a plot, genes that have large absolute fold changes as well as significant p-values will appear in the top part of the plot, towards the sides. These methods use the list of DE genes and their corresponding fold-change values as input. Other methods use only the list of DE genes, without corresponding expression values, because their scoring methods are based only on the relative positions of the genes in the graph. Methods which use cut-offs are sensitive to the chosen threshold value, because a small change in the cut-off may drastically change the number of selected genes [112]. As a consequence, some genes with moderate differential expression may be lost, even though they might be important players in the impacted pathways [17]. Furthermore, the genes included in the set of DE genes can vary dramatically if the selection methods are changed. Hence, the results of pathway analyses based on DE genes may be vastly different depending on both the selection method as well as the threshold value [122]. On the other hand, methods which do not use a threshold

are more sensitive to the noise coming from the (very many) genes that do not change much between the two phenotypes, genes that are normally eliminated by the DE selection process. An approach used to address this issue while still using all gene measurements uses the individual p-values of each gene as weights [165].

The impact of time-sensitive changes on the underlying genetic networks is often associated with the observed outcome [74]. Gene expression time series capture a high level of detail and provides knowledge of the evolution of the processes under study, while static data is not able to capture such subtle events in a detailed enough way. Time series gene expression data is used by network discovery methods, which focus on deciphering new regulatory relations between biological components [123]. Gene expression time course data is usually analyzed using clustering algorithms [11, 7, 173, 91]. The analysis of time series expression data could be used for therapeutics development in: deciding the duration of adjuvant chemotherapy [19], selecting the drug dose [177] or designing co-treatment strategies for complex diseases [154].

2.2.2 Pathway databases

An important component needed for identifying mechanisms of actions for biochemical components are biological pathways. These pathways are collections of molecular components and their interactions that represent the current existing knowledge about biological processes happening in various organisms. Many databases that describe the interactions between biological components have been developed and made available in the past 15 years.

Curated pathway databases that are publicly available are KEGG [121], NCI-PID [139], BioCarta [21], WikiPathways [128], PANTHER [107], and Reactome [84]. These curated knowledge bases are more reliable than protein interaction networks but do not include all known genes and their interactions. As an example, KEGG includes only about 5,000 human genes in signaling pathways. Protein-protein interaction data for human and some

model organisms is available from public databases among which are MIPS [106], DIP [180], BIND [10], HPRD [127], IntAct [69], and BioGRID [151].

Considering the wealth of information in these databases the information is still scattered due to the little consensus among them in terms of both the data structures used to store the data as well as the visual representation of the pathway information (see Fig. 2.3). For instance, in a KEGG signaling pathway nodes represent gene products and edges represent regulatory signals such as activation, inhibition, phosphorylation, etc. (see `http://www.genome.jp/kegg/document/help_pathway.html` for details). In a KEGG metabolic pathway the nodes represent biochemical compounds and edges represent chemical reactions. These chemical reactions are catalyzed by enzymes which are proteins encoded by genes. Hence, in a metabolic pathway genes are associated with edges. In

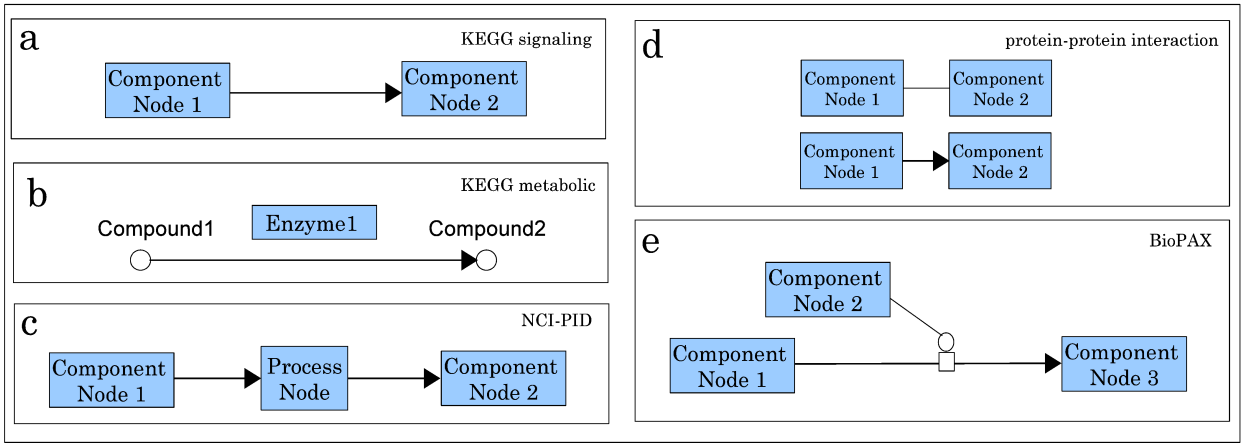

Figure 2.3: Comparison of representative graph models for molecular interactions as used by different pathway databases. Panel (a) presents the KEGG signaling pathway, where nodes represent genes/gene products and edges represent regulatory signals. Panel (b) presents the KEGG metabolic pathway, where nodes represent biochemical compounds and edges represent chemical reactions catalyzed by enzymes encoded by genes. Panel (c) presents the NCI-PID signaling pathway, where nodes fall in two categories: component nodes representing biomolecular components, or process nodes representing biochemical reactions or biological processes. Edges connect two biomolecular components through a biochemical reaction or a biological process. Panel (d) presents the protein-protein interactions, where nodes represent proteins and the interactions among them represent physical binding. Panel (e) presents the Biological Pathway Exchange (BioPAX), where nodes are physical entities and edges are conversions.

a REACTOME pathway, any two components linked by reaction nodes, where a component represents biochemical reactants such as metabolites and enzymes. In an NCI-PID signaling pathway nodes fall in two categories: component nodes representing biomolecular components, or process nodes representing biochemical reactions or biological processes. Edges connect two biomolecular components through a biochemical reaction or a biological process. Process nodes can have 3 states: positive regulation, negative regulation, or "involved in." (see `http://pid.nci.nih.gov/userguide/network_maps.shtml` for details). In a protein-protein interaction network nodes represent proteins and the interactions among them represent physical binding. These interactions can be inferred from two-hybrid assays and they may be either undirected (top), or directed from the bait protein to the prey protein (bottom). In the Biological Pathway Exchange (BioPAX) nodes are physical entities and edges are conversions. BioPAX format entities can represent complexes, DNA, proteins, RNA, small molecules, DNA regions or RNA regions. Conversions can represent biochemical reactions, complex assembly or degradation, transport or transport with biochemical reaction. This model provides a standard for pathway information to be available in machine readable format, thus easy to use for pathway analysis and to exchange between databases (see `http://www.biopax.org/release/biopax-level3-documentation.pdf` for details).

Various research groups have tried different strategies to address the challenge of modeling complex biomolecular phenomena. These efforts have lead to variation among knowledge bases, complicating the task of developing pathway analysis methods. There is currently no accepted standard for constructing pathways, and as pathway paradigms evolve to better represent the biology, pathway analysis methods evolve in parallel. Depending on the database, there may be differences in: information sources, experiment interpretation, models of molecular interactions, or boundaries of the pathways. Therefore, it is possible that pathways with the same designation and aiming to describe the same phenomena may have different topologies in different databases. As an example, one could compare the insulin signaling pathways of KEGG and BioCarta. BioCarta includes fewer nodes and emphasizes

the effect of insulin on transcription, while KEGG includes transcription regulation as well as apoptosis and other biological processes. Also, BioCarta includes the C-JUN transcription factor, which is missing from the KEGG representation.

Differences in graph models for molecular interactions are particularly apparent when comparing the signaling pathways in KEGG and NCI-PID. While KEGG represents the interaction information using the directed edges themselves, NCI-PID introduces "process nodes" to model interactions (see Fig. 2.3). Most pathway analysis methods are designed to use only one pathway graph model, which limits the user's possibilities. Developers are faced with the challenge of modifying methods to accept novel pathway databases or modifying the actual pathway graphs to conform to the method.

Pathway databases not only differ in the way that interactions are modeled, but their data are provided in different **formats** as well [29]. Common formats are Pathway Interaction Database eXtensible Markup Language (**PID XML**), KEGG Markup Language (**KGML**), Biological Pathway Exchange (**BioPAX**) Level 2 and Level 3, System Biology Markup Language (**SBML**), and the Biological Connection Markup Language (**BCML**) [16]. The NCI provides a unified assembly of BioCarta and Reactome, as well as their in-house "NCI-Nature curated pathways", in NCI-PID format [139]. In order to unify pathway databases, pathway information should be provided in a common format. XML is a flexible text format with increasing use for data exchange across different systems. However, XML is very low-level and lacks standard constructs to accurately describe biological phenomena. PID XML is both human- and machine-readable, and allows a platform-independent means of exchanging PID data. The BioPAX project is an effort to unify the format and exchange of pathway data, and has incorporated independent sources such as NCI, BioCarta, Reactome, and WikiPathways, UCSC, NIH, and others [22].

2.3 Analysis

For topology-based pathway analysis methods, the mathematical model describes how the graph and the experiment data are processed to compute a score for each pathway. The score quantifies the significance of changes in a (sub)pathway between the two phenotypes. This score may be a statistical significance or other non-statistical method-specific metric. The diversity of current topological based pathway analysis methods reflects the variety of mathematical models available for graphs. The output is typically a list of ranked (sub)pathways.

The input of a pathway analysis method is processed using mathematical modeling and statistical approaches that together define a scoring method. The goal of the scoring method is to compute a score for each pathway based on the graph model, resulting in a ranked list of pathways or sub-pathways. There are a variety of approaches to quantify the changes in a pathway. Some of the analysis methods use a hierarchically aggregated scoring algorithm, where on the first level, a score is calculated and assigned to each node or pair of nodes (component and/or interaction). On the second level, these scores are aggregated to compute the score of the pathway. On the last level, the statistical significance of the pathway score is assessed using univariate hypothesis testing. Another approach assigns a random variable to each node and a multivariate probability distribution is calculated for each pathway. The output score can be calculated in two ways. One way is to use multivariate hypothesis testing to assess the statistical significance of changes in the pathway distribution between the two phenotypes. The other way is to estimate the distribution parameters based on the Bayesian network model and use this distribution to compute a probabilistic score to measure the changes. See Fig. 2.4 for scoring algorithms categories that include: aggregated scoring, weighted gene set scoring and multivariate scoring.

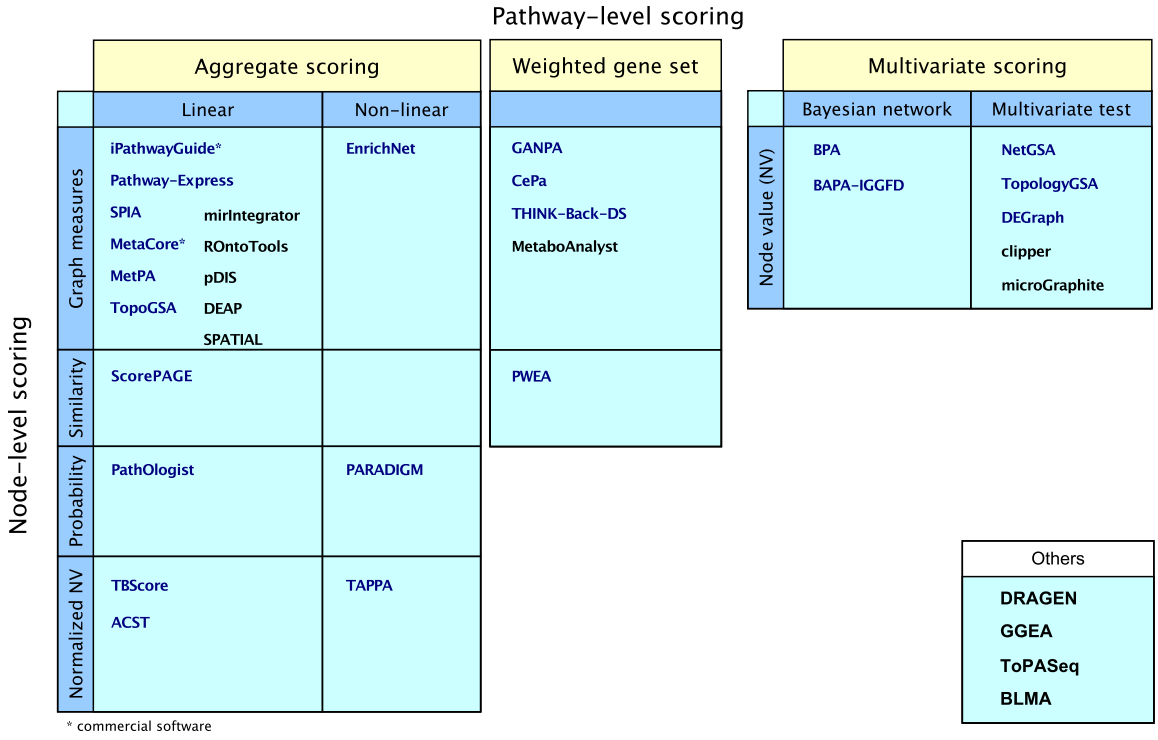

Figure 2.4: Comparison of the mathematical models of 34 pathway analysis methods. "Aggregate scoring" and "Weighted gene set" panels show methods that perform node-level scoring followed by pathway-level scoring performed either as an **aggregation of the node scores** or as a **weighted gene set analysis**, using the node scores as weights. The methods are divided according to their node-level scoring methods: graph measure techniques, similarity measurement techniques, probabilistic models, or using normalized node values based on node value and/or pathway structure. The "Multivariate scoring" methods use **multivariate scoring models** without node-level scoring. They use node values to directly compute a pathway score using Bayesian networks or applying multivariate hypothesis tests. The 22 methods displayed in blue were surveyed in the 2013 paper [111]. The methods displayed in black are the additional 12 methods surveyed in 2018 [118].

2.3.1 Graph models

Two major graph models are used to represent biological networks and pathways. The first model allows only one type of node, the biological component (i.e. a gene or protein), with the edges representing molecular interactions occurring between the nodes. In contrast, the second graph model places both components and interactions in nodes, requiring at least one type of node for each, and requiring each edge to connect a component with an

interaction node. Herein, we will refer to the first graph model category as "single-type" and to the second as "multi-type". Multi-type graph models are more complex than single-type, but they capture more pathway characteristics. For example single-type models are limited when trying to describe "all" and "any" relations between multiple components that are involved in the same interaction. Bipartite graphs are a particular case of multi-type graph models.

In most databases, pathways use the single-type graph model and the signaling and metabolic pathways from databases such KEGG and BioCarta are good examples. In signaling pathways, nodes are genes and edges describe various molecular interactions, which include activation/transcription/positive regulation, repression/blockage/negative regulation, (de)phosphorylation, binding/association. Metabolic pathways can be represented as either chemical networks or protein networks. In the chemical network representation, nodes are metabolites and edges are enzymes and/or substrates that catalyze the chemical reactions. In the protein network, the representation is reversed; nodes are enzymes and edges are metabolites. Among the surveyed methods which work with metabolic pathways only MetPA uses biochemical networks from KEGG. ScorePAGE and TAPPA use protein networks. Nevertheless, the most popular representation of metabolic pathways in public databases is the chemical network. In KEGG and BioCarta, the majority of edges in both metabolic and signaling pathways are directed, but binding between compounds is represented by undirected edges.

Protein-protein interaction (PPI) networks, constructed from interaction databases, use a single-type graph model. The nodes represent proteins and the edges depict their association/binding. Sometimes the edges are undirected, while some other times, the edges are directed to describe which protein was used as the bait and which one acted as the pray.

Reactome and NCI-PID are databases that use a bipartite graph model to represent pathways. Genes, metabolites, or molecular complexes are represented as component nodes, while interaction nodes define the chemical reactions or molecular processes that occur be-

tween the input and output component nodes. The edges, which connect a component node to an interaction node, specify the component's type of contribution to the reaction. These can be positive or negative regulation, among others.

The majority of analysis methods surveyed in this chapter use a single-type graph model. Some apply the analysis on a directed or un-directed single-type network built using the input pathway, while others transform the pathways into graphs with specific characteristics. An example of the later is TopologyGSA, which transforms the directed input pathway into an undirected decomposable graph, that has the advantage of being easily broken down into separate modules [90]. In this method, decomposable graphs are used to find "important" submodules - those which drive the changes across the whole pathway. For each pathway, TopologyGSA creates an undirected moral graph from the underlying directed acyclic graph (DAG) by connecting the parents of each child and removing the edge direction. The moral graph of a DAG is the undirected graph created by adding an (undirected) edge between all parents of the same node (sometimes called marrying), and then replacing all directed edges by undirected edges. The name stems from the fact that, in a moral graph, two nodes that have a common child are required to be married by sharing an edge. In TopologyGSA, the pathway moral graph is used to test the hypothesis that the underlying network is changed significantly between the two phenotypes. If the the research hypothesis is rejected, a decomposable/triangulated graph is generated from the moral graph by adding new edges. This graph is broken into the maximal possible submodules and the hypothesis is re-tested on each of them.

BPA is another method that implements pathway graph pre-processing. This method uses Bayesian networks to represent biological pathways. In Bayesian networks, random variables are assigned to each node of a DAG network and the edges represent the conditional dependencies between nodes. Before assigning the random variables, the pathway graph is checked for cycles. If the graph is not a DAG, Spirtes' method [150] is used to remove the cycles while the (in)dependency rules in the initial pathway graph are preserved.

Another example is BAPA-IGGFD, which is a method that simplifies pathway graphs by removing any edge representing interactions other than activation and inhibition. In addition, the pathways are pruned keeping only elements from three categories: signal receptors (including ligands) are at the beginning, transcription factors are usually at the end, and their direct regulators are in the middle. This pre-processing is motivated by noise reduction in the final scoring of genes that have a less important functional role in the pathway or belong to multiple pathways where they play different roles. BAPA-IGGFD [188] includes only an intuitive high-level description of this process is presented, without a detailed algorithm.

CePa uses a different method to modify the input pathways before the analysis. The NCI knowledge base is used as a source of NCI-Nature, BioCarta, Reactome, and KEGG pathways, which are provided in PID or short NCI-PID format. The pathway data is organized in the form of multi-type graphs, which are used to generate directed single-type graphs, where each node can represent one or multiple genes. A node in the generated graph is considered to be DE if any of its gene components is DE. Unfortunately, the details of how the original pathways are parsed to generate the new networks are not provided by the authors of CePa.

PathOlogist and PARADIGM are the two surveyed methods that use multi-type graph models. PathOlogist uses a bipartite graph model with component and interaction nodes. PARADIGM, conceptually motivated by the central dogma of molecular biology, takes a pathway graph as input and converts it into a more detailed graph, where each component node is replaced by several more specific nodes: biological entity nodes, interaction nodes, and nodes containing observed experiment data. The observed experiment nodes could in principle contain gene expression and copy number information. Biological entity nodes are DNA, mRNA, protein, and active protein. The interaction nodes are transcription, translation, or protein activation, among others. Biological entity and interaction node values are derived from these data and specify the probability of the node being active. These are the hidden states of the model.

2.3.2 Hierarchically aggregated scoring algorithms

These analysis approaches are detailed in Fig. 2.5. In this figure, the analysis is divided into three levels: node-level scoring, pathway-level scoring and significance assessment. All methods compute node level scores. One or both remaining levels may be skipped by certain approaches. PARADIGM is the only one that provides as direct output the node scores, rather than the pathway scores. These scores can be input into a gene set or pathway analysis algorithm, or a simple averaging function can be used to score the pathways and rank them, as in [164]. The rest of the methods go on to the second level where the scores of the pathways are calculated. Some methods stop at the second level, outputting the whole list of ranked pathways without evaluating their statistical significance, which is done by the remaining methods on the next level.

In the following few paragraphs we categorize and describe the surveyed methods based on their node level scoring model. Most of the surveyed analysis methods incorporate pathway topology information in the node scores. There are methods such as TAPPA and ACST that incorporate this information in the pathway scores. In TAPPA, the score of each node is the square root of the normalized log gene expressions (node value). ACST calculates the node level score using a sign statistic. The sign reflects the direction of the gene expression change between the phenotypes under study. This statistic can be a represented by a t-value or the log fold change of the gene expression. The statistic is standardized using a local mean and standard deviation.

The rest of the analysis algorithms use a variety of approaches to incorporate topology in the node level scores. We categorize them into methods that use graph measures (centrality), similarity measures, and probabilistic graphical models. TBScore is an exception that can not fall into either of these groups. TBScore weights the pathway DE genes based on their log fold change and the number of distinct DE genes directly downstream of them, using a depth-first search algorithm.

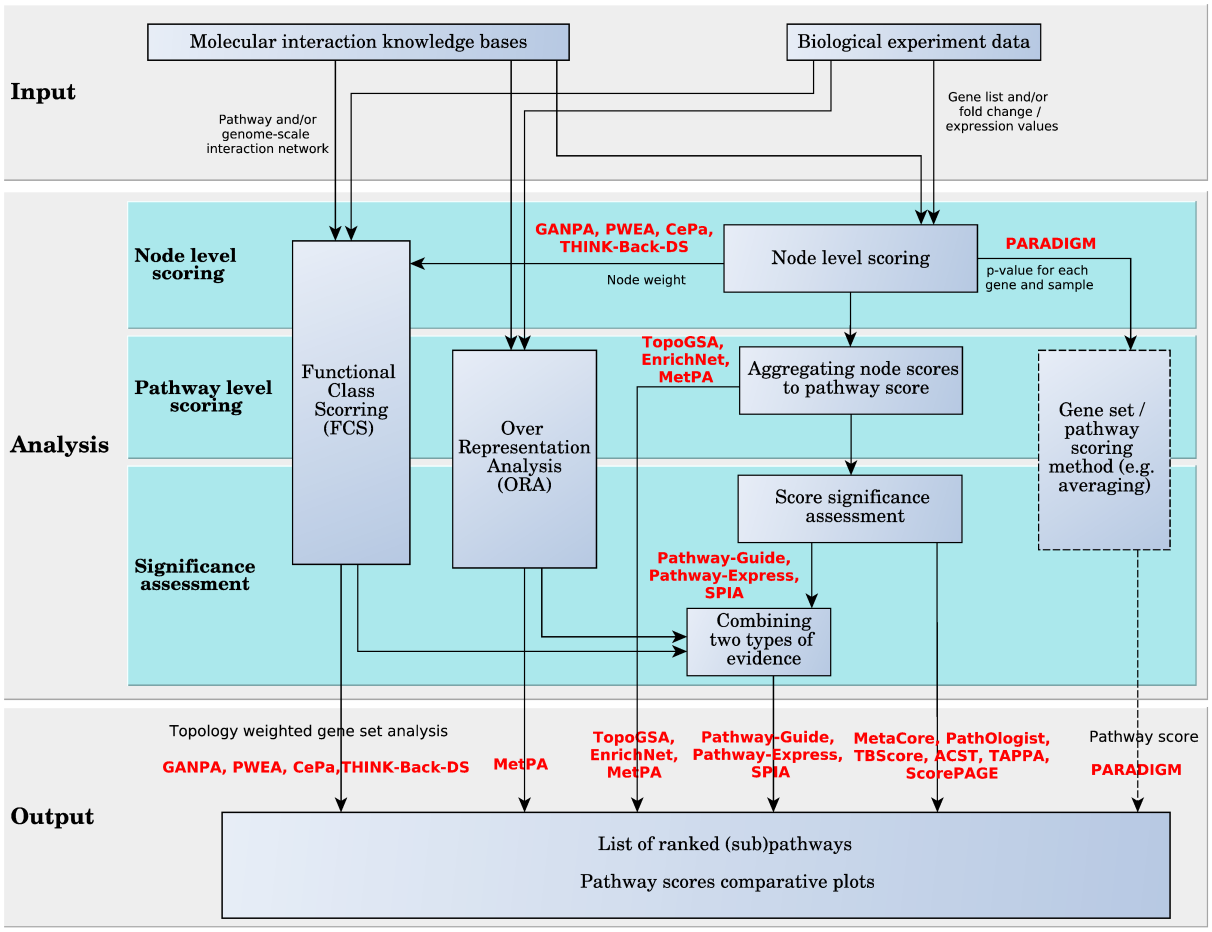

Figure 2.5: Diagram of pathway analysis scoring approaches for hierarchically aggregated scoring algorithms. The box with the dashed border indicates that the user can choose these options, but are not offered by the method implementation.

MetaCore, iPathwayGuide, Pathway-Express, SPIA, TopoGSA, CePa, EnrichNet, MetPA, THINK-BACK-DS, and GANPA use centrality measures or a variation of these measures to score nodes in a given pathway. Centrality measures describe the importance of a node relative to all other nodes in a network. There are several centrality measures that can be applied to networks of genes and their interactions and these are degree centrality, closeness, betweenness, and eigenvector centrality. Degree centrality accounts for the number of directed edges that enter and leave each node. Closeness sums the shortest distance from each node to all other nodes in the network. Node betweenness adds a layer of complexity to closeness; it measures the importance of a node according to the number of shortest paths that pass through it. Eigenvector centrality uses the network adjacency matrix of a graph to determine a dominant eigenvector; each element of this vector is a score for the corresponding node. Thus, each score is influenced by the scores of neighboring nodes. In the case of directed graphs, a node that has many downstream genes has more influence and receives a higher score.

In MetaCore, a measure similar to node betweenness is used to score genes. There is no peer-reviewed paper publicly available describing the details of the MetaCore pathway analysis method. We used the study by Dezso et al. [34] to uncover some of these details. In the method by Dezo et al., the DE gene list is overlapped with a global genome scale network containing all the interactions in the MetaCore knowledge base. A network, which is called condition specific shortest-path network (CSSPN), is built based on this overlap. In addition to DE genes, all genes which are on shortest paths that connect them in the global network are included in the CSSPN. For each pair of genes (g_i, g_j), where g_i is in the CSSPN and g_j is in the set of DE genes, two parameters N_{ij} and K_{ij} are computed. N_{ij} is the number of times g_i is part of the shortest-paths in the global network between g_j and every other gene in the CSSPN. K_{ij} is the number of times g_i is part of the shortest-paths in the global network between g_j and every other gene in the set of DE genes. It is assumed that the probability to observe these numbers just by chance, given the two sets of genes, the global network which

is of size N and the DE genes which is of size K, follows a hypergeometric distribution. Based on this distribution, K p-values are computed for each gene in the CSSPN and the minimum of these p-values is selected as the gene score. Using a predefined threshold on the false discovery rate (FDR) correction of the node scores, a subset of the CSSPN genes is selected. Further processing, at the pathway level, is applied to this list of selected genes.

iPathwayGuide, Pathway-Express, and SPIA use a perturbation factor, which takes into consideration the magnitude of all gene expression changes, the type of each gene, the direction and type of all gene interactions, as well as the efficiency with which the perturbation of each gene propagates to the downstream genes. The impact analysis models the flow of the signals in the pathways. In essence, the impact factor falls into the eigenvector centrality category of node scoring approaches. Although all three methods use the same impact analysis approach, there are slight differences between them. iPathwayGuide scores the pathways based on the impact factor as briefly described above In SPIA, the amount of differential expression is subtracted from the perturbation score of each node to focus on the amount of perturbation accumulated at any given node in order to separate the influence of experiment data and topology. iPathwayGuide is also able to exploit the p-values associated with each gene, as well as identify coherent perturbation cascades that represent putative mechanisms that explain all measured changes. All three methods combine the perturbation evidence with a classical enrichment (e.g. hypergeometric), or functional class scoring (e.g. GSEA) to calculate a global p-value. This corresponds to the joint probability of a pathway having the measured amount of perturbation, as well as the observed number of DE genes just by chance. TBscore has an interestingly similar approach in capturing the pathway perturbation, with the difference that DE genes with more connected downstream DE genes are considered more significant.

TopoGSA extracts a network from databases of protein interactions given a list of genes/proteins of interest. All four types of centrality measures and a fifth measure, called a "clustering coefficient" [172] are used to score the nodes in this network. Then, each

predefined pathway from a selected dataset is also scored using the same five measures, independently of the extracted network. Comparing the summarized node scores for each pathway with node scores from the extracted network allows the pathways to be ranked.

In CePa, node weights are computed using five centrality-based measures, and there is an extra case where all the node weights are assumed to be equal. The five measures are: in-degree, out-degree, betweenness, in-reach (length of longest shortest path that starts from the node), and out-reach (length of longest shortest path that ends at the node). CePa offers two options to assess the significance of pathways. One is based on the hypergeometric analysis using only node weights. The second is based on enrichment analysis and in addition to node weights, node scores are needed. Node scores are computed using a t-statistic. Pathway graphs in CePa can contain nodes representing one or multiple genes. In the case of single-gene nodes, the score is calculated based on the expression value of the corresponding gene. In the case of multi-gene nodes, the node score is the largest principal component of the expression values of the genes in the node.

EnrichNet uses a score similar to centrality closeness measures. This method calculates two distance vectors. The first vector contains distances between a list of input genes and a predefined pathway/gene set. The second vector contains the distance between the same input gene list and a background global set containing all pathways. A node score is computed as the distance between the node and all DE genes using a random walk with restart algorithm [185] through a genome scale molecular interaction network. The interaction network is represented by its weighted adjacency matrix, where weights are interaction strengths provided by the input knowledge base.

MetPA allows the user to select either the node betweenness or the out-node degree centrality measure for the node score. GANPA [46] uses the node degree measure as a weight or score for the gene. THINK-Back-DS uses a measure similar to closeness called density score to emphasize the DE genes which are in tight clusters.

ScorePAGE and PWEA use similarity measures in their node level scoring. Similarity measures estimate the coexpression, behavioral similarity, or co-regulation of pairs of components. Their values can be correlation coefficients, covariances, or dot products of the gene expression profile across time or sample. In these methods, the pathways with clusters of highly correlated genes are considered more significant. At the node level, a score is assigned to each pair of nodes in the network which is the ratio of one similarity measure over the shortest path distance between these nodes. Thus, the topology information is captured in the node score by incorporating the shortest path distance of the pair. In ScorePAGE, the correlation coefficient, covariance, or dot product is calculated for all gene pairs across their samples. PWEA uses the correlation coefficient to score node pairs. In this method, a score, called "Topological Influence Factor", or TIF, is assigned to each gene by exponentially averaging the score of all pairs that include the gene. As a consequence, a node involved in tight clusters of highly correlated genes has a higher score.

PARADIGM and PathOlogist incorporate the topology in the node level scoring using a probabilistic graphical model. In this model, nodes are random variables, and edges define the conditional dependency of the nodes they link. PARADIGM takes observed experiment data and calculates scores for all component nodes, in both observed and hidden states, from the detailed network created by the method based on the input pathway. For each node score, a positive or negative value denotes how likely it is for the node to be active or inactive, respectively. The scores are calculated to maximize the occurrence probability of the observed values. A p-value is associated with each score of each sample such that each node can be tagged as significantly active, significantly inactive, or not-significant. For each network, a matrix of p-values is output, in which columns are samples, and rows are component nodes.

PathOlogist is also based on a probabilistic graphical model. This method estimates the parameters of one or two distributions related to the up and/or down regulation of each gene using its expression values across all samples. These distributions are used to assign

a probability score to each gene in each sample, denoting how likely it is for the gene to be highly expressed. The method assigns two different scores to interaction nodes: (i) the "activity score", which is the probability that the parents of an interaction node (which are component nodes) are highly expressed, and (ii) the "consistency score", which is the probability that the interaction node is active and its children are expressed or inactive with unexpressed children.

In the following few paragraphs we describe how node scores are used to compute pathway scores. Many of the surveyed methods aggregate node level statistics to pathway level statistics using linear functions such as averaging or summation. The methods that use linear aggregation in this level of the analysis are: TopoGSA, MetaCore, MetPA, ScorePAGE, TBScore, ACST, PathOlogist, iPathwayGuide, Pathway-Express, and SPIA. The rest of the methods either use a nonlinear function to aggregate the node scores to pathway scores, like TAPPA, PARADIGM, and EnrichNet, or apply a gene set analysis method on the node scores, like GANPA, CePa, THINK-Back-DS, and PWEA.

In MetaCore, important genes are selected in the gene level scoring based on the list of DE genes and the network topology. At the pathway level, this method assumes that the number of selected genes that fall on a pathway is the pathway score and follows the hypergeometric distribution.

In TAPPA, the pathway score for each sample is a weighted sum of the product of all node pair scores in the pathway. The weight coefficient is 0 when there is no edge between a pair. For any connected node pair the weight is a sign function, which represents joint up- or down-regulation of the pair.

In ACST, pathway scores are calculated based on the position of node (gene) clusters for which the interaction types match the up- or down-regulation of genes. This uses the same concept of coherent signals used by iPathwayGuide. An edge (interaction) between 2 components in a pathway is called consistent if either (i) the pair has an inhibition interaction, and the directions of differential expression of the components is opposite, or (ii) the pair

has an activation interaction, and the direction of differential expression of the components is the same. All other interaction types are ignored. Maximal consistent graphs are defined as maximal sub-networks of the pathway in which all interactions are consistent. The score of each maximal consistent sub-graph is the summation of all node scores. The pathway score is the sum of the scores of all its maximal consistent sub-graphs. Node scores are t-statistics normalized by the distance from the sub-graph to the leaves of the pathway graph. The authors argue that the consistent sub-graphs close to the leaves of the pathway have a greater impact on the score of pathway rather than the clusters from the beginning of the pathway. This is somewhat different from the approach that iPathwayGuide, Pathway-Express, and SPIA follow. Although in these methods there is no explicit weighting based on the up- or down-stream position of a gene in a pathway, just because the perturbation of one gene is propagated following the signals described by the pathway, the perturbation of a gene somewhat near the entry point in a pathway will have more impact than the same amount of perturbation for a gene somewhere downstream on the pathway. Only time and additional testing will tell which of the two approaches manage to capture better the biological phenomena.

In EnrichNet, pathway scores measure the difference of the node score distribution for a pathway and a background network/gene set which consists of all pathways. At the node level, the distance of all DE genes to the pathway is measured and summarized as a distance distribution. The method assumes that the most relevant pathway is the one with the greatest difference between the pathway node score distribution and the background score distribution. The difference between the distributions is measured by the weighted averaging of the difference between the two discretized and normalized distributions. The averaging method down-weights the higher distances and emphasizes the lower distance nodes.

Methods such as iPathwayGuide, Pathway-Express, SPIA, and MetPA use two types of analysis to score the pathways. For each pathway, these methods calculate both a topology based score and a p-value from a gene set enrichment analysis measure, such as Fisher's exact

test, hypergeometric, or GlobalAncova. iPathwayGuide, Pathway-Express, and SPIA use the joint probability of observing the pathway perturbation, as well as the gene enrichment on a given pathway [40]. This model effectively combines the topology-based pathway score with the one based on enrichment to provide a single global pathway score. MetPA [183] also looks at both enrichment and topology, but does not assess the significance of the topology-based pathway scores and does not combine the two scores, and thus lacks a unique significance ranking. The most impacted pathways in MetPA are those with higher scores in both measures. It is not clear how to treat a trade-off between the two types of significance.

The pathway scoring techniques described so far in this section incorporate in-house analysis methods. A different direction is to design scoring techniques that incorporate existing gene set analysis methods, such as GSEA [153], GSA [43], or LRPath [138]. Pathway-level scores can be calculated using node scores which represent the topology characteristic of the pathway as weight adjustments to a gene set analysis method. PWEA, GANPA, THINK-Back-DS, and CePa use this approach and we refer to them as weighted gene set analysis methods. GSEA calculates the correlation coefficient of phenotype with gene expression (CC), GSA and LRPath use the t-test statistic in the computation of the node score. To compute the pathway score, PWEA adjusts the CC exponent of 0 or 1 in GSEA to CC^{TIF+1}, where TIF is the node weight described above. The node weights calculated by GANPA, THINK-Back-DS, and CePa are used to adjust CC or the t-statistic by multiplication, $node\ weight \times CC$ or $node\ weight \times t - statistic$. In CePa there is another option to use a hypergeometric analysis to calculate pathway scores. In this method, the node weights of DE nodes are summed up to the pathway level.

Some methods such as iPathwayGuide, Pathway-Express, SPIA, and ROntoTools offer the flexibility to integrate in the analysis any type of enrichment technique. Thus, the p-values provided by techniques such as GSEA, GSA, or PADOG [155] can be used instead of the p-values provided by simpler models such as hypergeometric.

In the following few paragraphs we describe how the pathway significance assessment is performed for the surveyed methods Pathway scores are intended to provide information regarding the amount of change incurred by the pathway between two phenotypes. However, the amount of change is not meaningful by itself since any amount of change can take place with a non-zero probability (i.e. the amount of change is only the effect size). An assessment of the *significance* of the measured changes is thus required, and should be done by analysis methods in the pathway significance assessment level.

Methods such as TopoGSA, MetPA, and EnrichNet, will output scores without any significance assessment, leaving it up to the user to interpret the results. This is problematic because the user does not have any instrument to help distinguish between changes due to noise or random causes, and meaningful changes, unlikely to occur just by chance and therefore, possibly related to the phenotype. The rest of the analysis methods perform a hypothesis testing for each pathway. The null hypothesis is that the value of the observed statistic is due to random noise or chance alone. The research hypothesis is that the observed values are substantial enough that they are potentially related to the phenotype. A p-value for calculated score is then computed and a user-defined threshold on the p-value is used to decide whether the the null hypothesis can be rejected or not for each pathway. Finally, a correction for multiple comparisons should be performed.

Typically, pathway analysis methods compute one score per pathway. However, methods such as PathOlogist and TAPPA compute the pathway score considering each sample separately. Therefore, for each pathway there is a population of scores that can be analyzed. This population combined with different sample features can provide various feature-specific analyses. There are two cases to be considered based on the qualitative or quantitative nature of the sample feature values. In the first case the sample feature is qualitative with binary values. For example, when samples are tagged corresponding to the two phenotypes, the significance assessment is done by testing whether the score distributions are the same in the two groups using two-sample rank-sum tests, such as the Mann-Whitney U test. If the

number of samples is high enough, the score distributions can be assumed to be normal. The null hypothesis here is that the two normal distributions have equal means and variances, the research hypothesis is that they are different. In the second case the sample feature is quantitative with continuous values. Two ways to identify significant pathways are implemented in this case. One way is to partition pathway scores into a known number of clusters, for example two, using k-means clustering. Cumulative distributions are calculated for each of the two classes. A logrank test [98], which is a non-parametric statistical test, can be performed to evaluate if the behavior of the variable is same in the two groups. Significant pathways are those that can be used to divide samples into groups with different characteristics. Another way to identify significant pathways in the case of continuous sample feature values is to find pathways whose scores are linearly correlated with the values of the feature. The null hypothesis in this case is that the correlation is zero, and a t-test is used.

For methods that calculate one score per pathway, the distribution of this score under the null hypothesis can be constructed and compared to the observed. However, there are often too few samples to calculate this distribution, so it is assumed that the distribution is known. For example, in MetaCore and many other techniques, when the pathway score is the number of DE nodes that fall on the pathway, the distribution is assumed to be hypergeometric. However, the hypergeometric distribution assumes that the variables (genes in this case) are independent, which is incorrect, as witnessed by the fact that the pathway graph structure itself is designed to reflect the specific ways in which the genes influence each other. Another approach to identify the distribution is to use statistical techniques such as the bootstrap method [42]. Bootstrapping can be done either at the sample level, by permuting the sample labels, or at gene set level, by permuting the the values assigned to the genes in the set.

To create the score distribution under the null hypothesis, iPathwayGuide, PathwayExpress, and SPIA methods use bootstrapping at the gene set level. For these methods, samples are drawn from the distribution of all DE genes and assigned to a gene set which

is different from the DE gene set but with equal number. The pathway score is computed assuming the new gene set as a decoy DE gene set. This procedure is repeated for a number of iterations. The scores resulting from these iterations estimate the distribution, which is then used to compute a p-value, and a pathway score is obtained by combining the gene set enrichment evidence with the topology-based p-value and applying Fisher's exact test. The final score is the FDR-adjusted p-value.

TBScore, the hypergeometric extension of CePa, and ACST calculate p-values using bootstrapping at the sample level by permuting the labels of the samples of the two phenotypes. In TBScore and CePa, an iterative procedure is then used to estimate the pathway score distribution under the null hypothesis. Correction for multiple comparison, again FDR, is used to compute the final pathway p-values. In ACST, after p-values are computed, a statistical technique called "resampling-based point estimator" is used to estimate the FDRs associated with the predefined threshold.

Weighted gene set methods surveyed in this chapter, PWEA, GANPA, the enrichment analysis extension implemented by CePa, and THINK-Back-DS, focus on providing a biologically meaningful topology-based adjustment to existing gene set analysis methods. Therefore the statistical assessment of pathway significance is provided by the already developed methods among which the most popular is GSEA (see Fig. 2.5).

2.3.3 Multivariate scoring algorithms

Multivariate scoring analysis methods mostly use multivariate probability distributions to score pathways and can be grouped into two categories. Methods in the first category use multivariate hypothesis testing, while methods in the second category are based on Bayesian network (see Fig. 2.6).

NetGSA, TopologyGSA, and DEGraph are methods based on multivariate hypothesis testing. These analysis methods assume the vectors of gene expression values in each (sub)pathway are random vectors with multivariate normal distributions. The network topol-

ogy information is stored in the covariance matrix of the corresponding distribution. For a network, if the two distributions of the gene expression vectors corresponding to the two phenotypes are significantly different, the network is assumed to be significantly impacted when comparing the two phenotypes. The significance assessment is done by a multivariate hypothesis test. The definition of the null hypothesis for the statistical tests and the techniques to calculate the parameters of the distributions are the main differences between these three analysis methods.

In NetGSA, it is assumed that the expression level of the genes (nodes in the network) obtained from experiments are correlated because of the interactions between them. In other words, the edges (interactions) of a graph (pathway) imply correlations. In order to compute the distribution parameters, the method defines a set of latent variables, which are the

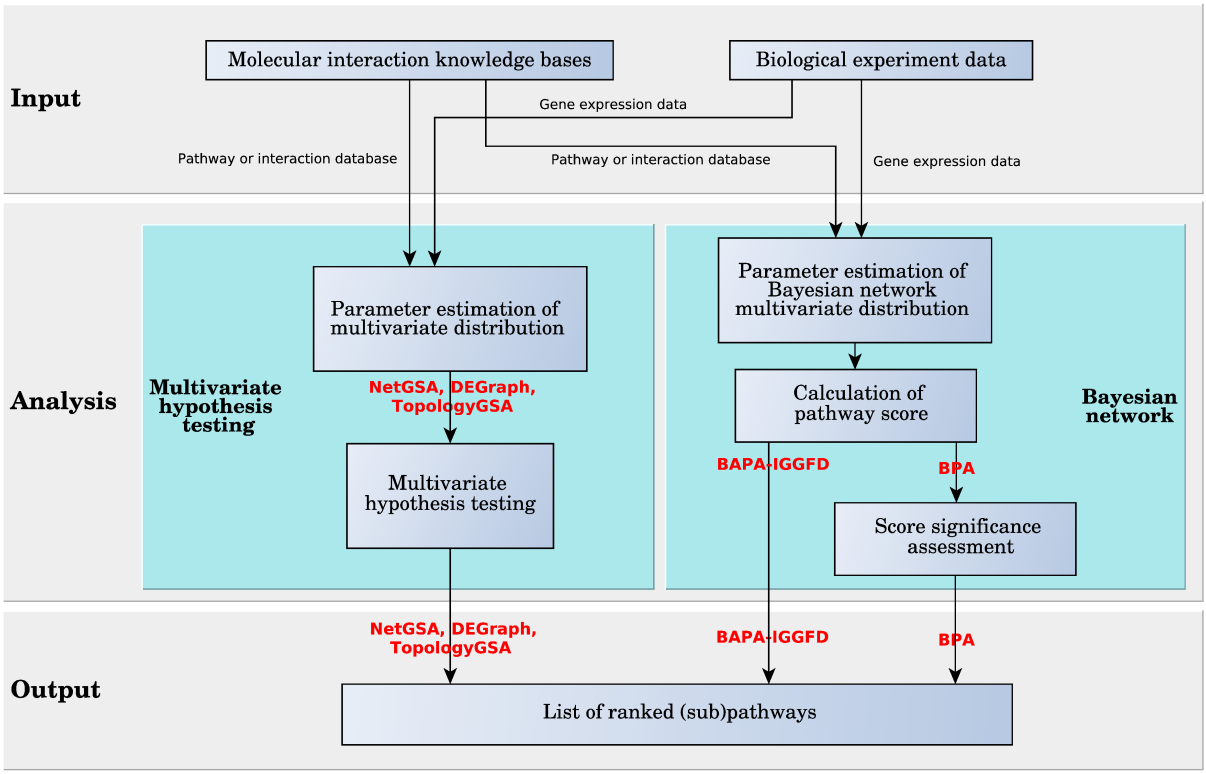

Figure 2.6: Diagram of pathway analysis scoring approaches for multivariate scoring algorithms.

uncorrelated gene expressions. The input correlated gene expression vector can be written in the form of the product of the vector of latent variables and the influence matrix. This matrix consists of the weights assigned to each edge measuring the strength of the interaction between two genes. The influence matrix and other parameters for the two phenotypes are computed based on linear mixed model theory [104]. The proposed hypothesis test, in this method, is to check whether a linear combination of the mean of the latent variables, called contrast vector, for the two cases are equal. The proposed contrast vector is computed based on the influence matrix and it is proved that the result includes the effects of all nodes inside a chosen network and excludes any outside effects, such as the correlation.

In TopologyGSA, the directed graph is converted into a moral undirected graph, detailed in section 2.3.1. The covariance matrices for each of the two phenotypes are estimated using the Iterative Proportional Scaling (IPS) algorithm [90] on the sample covariance for all pairs of genes. The two matrices are defined such that their inverses have zero elements corresponding to the missing edges. A set of two hypothesis tests are applied to compute the statistical significance of the impact on a given graph. The first test checks whether the concentration matrices, i.e., the inverses of the covariance matrices, in the two cases are equal. If this hypothesis is rejected, the graph is broken into the maximal possible submodules, and the hypothesis is retested on each. Based on the equality of concentration matrices, different statistical techniques are used in the second hypothesis test. The second test checks the significance of the influence of the graphs based on the equality of the means of the distributions.

DEGraph finds significant (sub)pathways by using a modified multivariate Hotelling's T^2-test hypothesis. The modification incorporates the topology of the network. The difference, referred to as shift, between the mean vectors of gene expression distributions corresponding to the two phenotypes is smoothed. A shift vector is defined to be smooth if the shift values of every two connected nodes are similar. The process of smoothing is done by removing the high frequency shift values according to the topology of the network. This is

achieved by filtering the shift by preserving only the first few components of the graph-Fourier basis of the shift vector. The graph-Fourier in DEGraph is applied by spectral analysis of the graph Laplacian [30], which resembles the Fourier decomposition of a function. The smoothed shift vector is used in the Hotelling's T^2-test to assess the statistical significance of a network. DEGraph also provides an algorithm that allows the exhaustive testing of all the sub-networks of the original network using a branch and bound algorithm.

BPA and BAPA-IGGFD are two methods based on Bayesian networks. In a Bayesian network, which is a special case of probabilistic graphical models, a random variable is assigned to each node of a directed acyclic (DAG) graph. The edges in the graph represent the conditional probabilities between nodes, so that the children are independent from each other and the rest of the graph when conditioned on the parents. In BPA, the value of the Bayesian random variable assigned to each node captures the state of a gene (DE or not). In contrast, in BAPA-IGGFD each random variable assigned to an edge is the probability that up or down regulation of the genes at both ends of an interaction are concordant with the type of interaction which can be activation or inhibition. In both BPA and BAPA-IGGFD, each random variable is assumed to follow a binomial distribution whose probability of success follows a beta distribution. However, these two methods use different approaches in representing the multivariate distribution of the corresponding random vector. BPA assumes that the random vector has a multinomial distribution, which is the generalization of the binomial distribution. In this case, the vector of the success probability follows the Dirichlet distribution, which is the multivariate extension of the beta distribution. Conversely, BAPA-IGGFD assumes the random variables are independent, therefore the multivariate distributions are calculated by multiplying the distributions of the random variables in the vector. It is worth mentioning that the assumption of independence in BAPA-IGGFD is contradicted by evidence, specifically in the case of edges that share nodes.

In BPA a discretized fold change profile is calculated for each gene. This represents the list of fold changes between every ordered pair of gene expression samples. The pair

elements come from each of the groups corresponding to the two phenotypes. These fold changes are discretized such that genes with values higher than 2 or lower than 0.5 are considered differentially expressed and the others are considered to have negligible changes. This profile is used as the observed data for the Bayesian network model. In BPA, given a set of parameters (success probabilities), the likelihood of observing a specific profile on the Bayesian network is assumed to have a multinomial distribution. Using the Bayes rule, the probability of observing the given profile without any assumption on the parameters is calculated. The parameters of the distributions are learned from the input data [113]. The network topology is incorporated in the distribution parameters and computation method by assuming that knowing the values of the parents' random variables, the children random variables are independent of the rest of the graph. A hypothesis testing is performed using the null hypothesis that the probability of seeing the observed data is the result of chance. Specifically, a set of observed data is generated in the bootstrapping analysis and its probability is compared with the the original observed data. The null distribution is approximated through randomization via bootstrapping. This randomization targets the structure of the Baeysian network (i.e. the relation between its nodes), which is more relevant than a simple bootstrapping in this case. Sampling with replacement is used when generating random data. An upper-tailed test is performed, with the p-value estimated by the percentage of random scores higher than the observed one. The process of generating the randomized samples is done by bootstrapping. A new fold change profile is generated by sampling with replacement from the original fold change profile.

In BAPA-IGGFD, based on the value of the fold change, discretized values of 0 or 1, corresponding to up- or down-regulation, are assigned to each node of the Bayesian network. For each predefined pathway, a vector of probabilities is computed as follows: 1) $\bar{\theta}_i$ for any parent-less gene g_i is the probability of g_i being up-regulated, 2) $\theta_{i|j}$ for any gene g_i which has an activator parent g_j is the probability that both genes are coherent in being up-regulated or down-regulated, and 3) $\phi_{i|j}$ for any gene g_i which has an inhibitor parent

g_j is the probability that the state of up or down regulation of the genes are opposite. The vector can be summarized as $\boldsymbol{\theta} = (\{\bar{\theta}_i | \forall g_i \text{ is parent-less}\}, \{\theta_{i|j} | \forall g_i \text{ has an activator parent}\}, \{\phi_{i|j} | \forall g_j \text{ has an inhibitor parent}\})$ which is called the parameter vector of the pathway. Each of these parameters are assumed to be independent from each other and follow the beta distribution both prior observing the microarray data and after its observation. The multivariate joint distributions of the parameter vector prior and posterior of the data observation are compared using symmetric Kullback-Leibler (SKL) divergence [89]. The pathways for which the prior and posterior distributions are dis-similar are assumed to be impacted more significantly between the two phenotypes. Because of the independence assumption, the distribution of the parameter vector is calculated by multiplying the beta distribution of each of the parameters. The variables of the distributions are calculated using PrimeDB database, or in other words, using the number of journal citations for an interaction type. We refer to $beta(\alpha, \beta)$ as the beta distribution with parameters α and β. For the prior distribution, it is assumed that $\bar{\theta}_i \sim beta(1,1)$, $\theta_{i|j} \sim beta(a_{i|j}, b_{i|j})$, and $\phi_{i|j} \sim beta(b_{i|j}, a_{i|j})$, where $a_{i|j}$ and $b_{i|j}$ are the number of journals citing the activation or inhibition between g_i and g_j, respectively. For the posterior distribution, it is assumed that $\bar{\theta}_i \sim beta(\bar{n}_i, n - \bar{n}_i)$, $\theta_{i|j} \sim beta(a_{i|j} + n_{i|j}, b_{i|j} + n - n_{i|j})$, and $\phi_{i|j} \sim beta(b_{i|j} + n_{i|j}, a_{i|j} + n - n_{i|j})$, where n is the total number of microarray experiments, n_i is the number of experiments in which g_i is up-regulated, and $n_{i|j}$ is the number of experiments in which the pairs g_i and g_j are concordant in up or down regulation. An extension to this method is proposed in which the variables of beta distributions are not calculated by the input as fixed numbers but are assumed to follow exponential distributions. In this case, the parameters of the exponential distributions are estimated from PrimeDB and the input data. It is claimed that this additional probability layer will lead to more robust results. For genes that have more than one parent the majority rule is used to calculate the distribution. The output of this method is the list of pathways scored by the SKL divergence. The lower the score is the more impacted the pathway is assumed to be.

2.4 Output

Although the goal of the pathway analysis should be a ranked list of pathways as a unified output, not all tools reviewed in this chapter provide this. Some methods, such as MetPA provide a list of pathways with 2 p-values for each pathway, leaving the user to face the task of deciding which p-value to trust or how to deal with trade-offs between the two values. Among the methods that rank predefined pathways from public knowledge bases, some methods, such as TopologyGSA, DEgraph, NetGSA, and ACST, find "important" sub-pathways and rank the mixed list of pathways and sub-pathways. In PARADIGM, for each detailed network created by the method based on the input pathway, a matrix of p-values is provided as the output. In this matrix, columns are samples and rows are component nodes of the network. Each element of this matrix indicates how likely it is for the node to be in any of the three states comparing the two phenotypes: 1) significantly active, 2) significantly inactive, or 3) have an insignificant change. These scores can be used as substitutes for log fold changes and, as proposed in [164], can be input into a non-topology-based gene set analysis algorithm to rank the pathways. Other options to use these scores are either to apply a simple averaging or counting function on the scores of the significant genes to score the pathway, or they can be used to cluster the genes into groups with similar behavior. These clusters of genes can be used to further analyze different features assigned to samples to find group-specific features.

iPathwayGuide offers capability to identify so called "coherent chains of perturbation propagation", which are to be interpreted as putative mechanisms that are compatible with (and therefore could explain) all measured changes throughout the entire biological system investigated [2]. Even though unique among all other tools and potentially very useful, this capability is completely independent from the pathway ranking provided based on the perturbation and enrichment types of evidence. Therefore, it is possible for pathways that are significant not to contain such coherent signaling cascades, and conversely, pathways that may contain such cascades may not be significant.

In many input data sets, the samples are labeled based on different parameters. The parameters can have qualitative discrete values such as, tumor and normal tissue, or quantitative continuous values such as, survival time of the cell or drug concentration used to treat the tissue. For analysis methods that provide a pathway score for each sample, such as TAPPA and PathOlogist, the pathway activities can be interpreted based on the sample labels. NetGSA [149] offers more labeling options in addition to phenotype-based binary labels. The method provides simultaneous tests of multiple hypotheses based on these labels or temporal pathway score correlation to assess the significance of pathways. The rest of the pathway analysis methods compare the pathways using a single qualitative binary label corresponding to the two phenotypes. Methods such as TopoGSA, MetPA, and the hypergeometric extension of CePa calculate one score for each pair of input samples comparing the two phenotypes, while others provide one score for the whole input data set.

Some of the methods provide a graphical display of their results. This is primarily done for the analysis methods which have the ability to provide more than one score for each pathway. For example, analysis methods like TopoGSA have an additional option to compare the properties of the input dataset to predefined datasets corresponding to known functional processes from public databases in a comparative plot. As a result, a summary of network topological properties is displayed for all gene/protein sets in 2D and 3D plots. This functionality allows the user to visually identify an input similar to the original one, based on the plots or on a tabular ranking using a numerical score to quantify the similarity across all topological properties. Similarly, methods such as iPathwayGuide, Pathway-Express, SPIA, and MetPA which provide two scores (topology based and gene set enrichment) can use a 2D plot to illustrate the distribution of both scores for the analyzed pathways.

2.5 Implementation

The mathematical model for each analysis approach is independent of its implementation as a software package. Although the main strength of an approach lies in its algorithm,

its implementation can have an important role in reaching the full potential of that approach, as well as in gaining acceptance among the users. Practicality, user-friendliness, output format, and type of interface are all to be considered. Depending on the desired availability and intended audience, a software package may be implemented as standalone or web-based.

Web-based tools run the analyses on a remote server providing computational power and a graphical interface. On the user side, experiment datasets are uploaded, and on the server side, the tool performs the analysis. The results are displayed by the browser in the format provided by the tool. The output of most pathway analysis methods is a ranked list of pathways or sub-pathways. iPathwayGuide, MetPA, THINK-Back-DS and, EnrichNet are among the methods that have web-based implementations. The major advantage of web-based tools is that they are user-friendly and do not require a local installation.

Standalone tools need to be installed on local machines which often requires administrative skills. Advantages include instant availability that does not require internet access. Most standalone tools depend on full or partial copies of public pathway databases, stored locally, and need to be updated periodically. Methods like ScorePAGE, SPIA, TAPPA, PathOlogist, NetGSA, TopologyGSA, PWEA, ACST, BPA, and GANPA are in this category. Moreover, there are some methods available both as web-based and standalone, including Pathway-Express, MetaCore, TopoGSA, CePa, and PARADIGM. For PARADIGM, the web-based implementation is only available as part of TCGA while the standalone is available only as C++ source code that needs to be compiled and deployed locally. Another major advantage of standalone tools are the security and privacy of the experiment data.

The programming language and style used for implementation plays an important role in the acceptance of a method. Software tools that are neatly implemented, packaged, and available online are more appealing compared to those that do not have ready-to-use implementations. Many of the methods surveyed in this chapter are implemented in the R programming language and are available as software packages either from bioconductor.org, cran.r-project.org, or the author's website. Their popularity among biologists and bioinfor-

maticians is due to the fact that many bioinformatics dedicated packages are available in R. Pathway-Express (as part of ROntoTools), SPIA, TopoGSA, TopologyGSA, GANPA, DE-Graph, NetGSA, ACST, CePa, and ScorePAGE are among those methods. iPathwayGuide, Pathway-Express, TAPPA, and THINK-Back-DS have an implementation in Java, which provides a GUI with self explanatory functionality for users with less software development experience. This allows users to customize the graphical display of the results, using function-

Mathematical model and implementation for 34 topology-based pathway analysis methods						
Method name	**Graph model**	**Scoring method**	**Web/App**	**License**	**Language**	**Tool ref.**
ScorePAGE	Single-type, undirected	Hierarchical, similarity	App	N/A	R	on demand
MetaCore*	Single-type, directed	Hierarchical, graph measures	Web, App	Thomson Reuters	Java	[132]
Pathway-Express	Single-type, directed	Hierarchical, graph measures	Web, App	free**	Java, R	[39]
TAPPA	Single-type, undirected	Hierarchical, NNV	App	N/A	Java	N/A
PathOlogist	Multi-type, directed	Hierarchical, probability	App	CC-BY	MATLAB	[63]
iPathwayGuide*	Single-type, directed	Hierarchical, graph measures	Web	AdvaitaBio	Java	[1]
SPIA	Single-type, directed	Hierarchical, graph measures	App	GPL (>= 2)	R	[157]
NetGSA	Single-type, directed	Mutivariate, hypothesis test	App	GPL-2	R	[147]
PWEA	Single-type, undirected	Hierarchical, similarity	App	free**	C++	[71]
TopoGSA	Single-type, undirected	Hierarchical, graph measures	Web	free**	PHP, R	[61]
PARADIGM	Multi-type, directed	Hierarchical, probability	Web, App	free** (App) UCSC-CGB (web)	C	[163], [162]
TopologyGSA	Single-type, moral, undirected	Mutivariate, hypothesis test	App	AGPL-3	R	[102]
DEGraph	Single-type, undirected	Mutivariate, hypothesis test	App	GPL-3	R	[82]
MetPA	Single-type, directed	Hierarchical, graph measures	Web	free**	PHP, R	[181]
BPA	Single-type, DAG	Mutivariate, Bayesian network	App	free**	MATLAB	[78]
GANPA	Single-type, undirected	Hierarchical, graph measures	App	GPL-2	R	[47]
BAPA-IGGFD	Single-type, DAG	Mutivariate, Bayesian network	App	N/A	R	N/A
CePa	Single-type, directed	Hierarchical, graph measures	Web, App	GPL (>= 2)	R	[65], [64]
THINK-Back-DS	Single-type, directed	Hierarchical, graph measures	Web, App	free**	Java	[50],[49]
TBScore	Single-type, directed	Hierarchical, normalized node value (NNV)	N/A	N/A	N/A	N/A
ACST	Single-type, directed	Hierarchical, NNV	App	CC-BY	R	[109]
EnrichNet	Single-type, undirected	Hierarchical, graph measures	Web	free**	PHP	[58]
GGEA	Single-type, directed	Aggregate fuzzy similarity	App	Artistic-2.0	R	[55]
ROntoTools	Single-type, directed	Hierarchically aggregated	App	CC-BY	R	[165]
clipper	Single-type, directed	Multivariate analysis	App	AGPL-3	R	[100]
DEAP	Single-type, directed	Hierarchically aggregated	App	GNU LGPL	Python	[67]
DRAGEN	Single-type, directed	Linear regression	App	N/A	C++	[97]
ToPASeq***	Single-type, directed	Hierarchical and multivariate	App	AGPL-3	R	[75]
pDis	Single-type, directed	Hierarchically aggregated	App	free**	R	[8]
SPATIAL	Single-type, directed	Hierarchically aggregated	N/A	N/A	N/A	[23]
BLMA****	Single-type, directed	Hierarchically aggregated	App	GPL (≥2)	R	[117, 119]
microGraphite	Single-type, directed	Multivariate analysis	App	AGPL-3	R	[25]
mirIntegrator	Single-type, directed	Hierarchically aggregated	App	GPL≥3	R	[35, 116]
MetaboAnalyst	Single-type, directed	Hierarchically aggregated	Web	GPL (≥2)	R	[184, 182]

Table 2.1: Comparison of topology-based pathway analysis methods using criteria related to the mathematical model and implementation. The last 12 methods were surveyed in another study [118]. **Graph model** indicates whether the graph which is remodeled to be suitable for the scoring method is single-type or multi-type and whether it is directed or undirected. DAG stands for directed acyclic graph. **Scoring method** encloses the mathematical model used in the analysis to score nodes and graphs. **Web/App** indicates the existence of a web-based (Web) or standalone (App) implementation of the method. **License** represents the license under which the software is available. GPL - GNU General Public License, AGPL - GNU Affero General Public License, CC-BY - Creative Commons license. **Language** represents the programming language used for the implementation. **Tool ref.** points to the paper or url associated with the given tool.
* commercial methods; ** free for academic and non-commercial use; UCSC-CGB – the University of California Santa Cruz Cancer Genome Browser; N/A No publicly available implementation

alities such as zoom or rotation. CePa has a web-based implementation in Perl in addition to its R stand-alone package. The MATLAB programming language is used for implementation of methods like PathOlogist, BPA, and NetGSA in order to calculate more complex equations. Other programming languages like C and C++ are also used to implement pathway analysis methods such as PARADIGM and PWEA, which theoretically provide better speed and allow for efficient coding. A summary of the mathematical models and implementation details for the surveyed methods is presented in Table 2.1, where in addition to the 22 surveyed methods from [111] we add 12 extra methods surveyed in another study [118].

2.6 A summary of this chapter

Pathway analysis is a core strategy of many basic research, clinical research, and translational medicine programs. Emerging applications range from targeting and modeling disease networks to screening chemical or ligand libraries, to identification/validation of drug target interactions for improved efficacy and safety [9]. The integration of molecular interaction information into pathway analysis represents a major advance in the development of mathematical techniques aimed to evaluate systems perturbations in biological entities.

The important milestones in pathway analysis reflected by this survey are: the first pathway analysis method for metabolic networks [131], the first method for signaling pathway and the first method able to take into consideration the pathway topology [86, 40], the first application of topology-based multivariate hypothesis tests [148], and the first analysis able using multi-type graphs from heterogeneous sources [164]. In this chapter, analysis methods were compared according to types of input, scoring algorithms, results, and user accessibility. Each of these aspects presents its own particular challenges.

The validation of pathway analysis results is an important challenge researchers face when trying to develop such methods. While biologists are needed to verify the pathway analysis results, they depend on pathway analysis methods to support their hypotheses. Most efficient progress will occur with a high level of communication and collaboration between

experiment biologists, annotators, pathway designers, bioinformaticians, and computer scientists. As pathway knowledge becomes more complete, the challenge of leveraging this information to extract biological insight from high-throughput data will be redefined. Until then, advances will be incremental. Gold standard experiment data sets, designed to affect specific pathways in predefined ways, will be necessary to be able to assess the efficiency of new methods.

Another challenge we mentioned in this chapter is that the same biological pathways are represented differently from one pathway database to another. In particular, we pointed out the complications arising from inconsistent conversions for representing interactions among the different pathway databases, and the current efforts to address the problem through the creation of unified formats. However, none of the tools is compatible with all database formats, requiring either modification of pathway input or alteration of the underlying algorithm in order to accommodate the differences. As an example, a study by Vaske and others [164] attempts to compare SPIA [157] with their tool PARADIGM, by re-implementing SPIA in C, and forcing its compatibility with NCI-PID pathways. Grave implementation errors are present in the C version of SPIA, invalidating the comparison. A solution to overcome this challenge could be the development of a unified globally accepted pathway format. Another possible solution is to build conversion software tools that can translate between pathway formats. Some attempts exist to use BIO-PAX as the lingua franca for this domain.

Biological networks are divided in various categories containing complementary information. Signaling and signal transduction are captured by signaling pathways, while biochemical interactions are presented in metabolic pathways. In addition, the protein interaction knowledge bases contain different types of interaction information, complementary to the others. The majority of pathway databases are manually curated and change slowly, but they are evolving toward greater content and accuracy, with new prototype formats being proposed. There is no analysis method that takes advantage of the information stored in all

of these different sources. Few of the methods surveyed here use either signaling or metabolic pathways in addition to PPI networks. Promising developments include the integration of multiple component types and interaction types, each with specific properties. Although the information is less reliable, non-curated high-throughput protein interaction data is also proving useful, as protein interaction data can be used to support or filter results.

High-throughput technologies, developed for biological experiments, are improving in accuracy. However, they are still prone to error and the resulting data includes a significant amount of noise. In addition, these technologies produce various types of data among which are genome variations, mRNA level, metabolite concentration, or protein abundance. Each of these data types provides meaningful yet incomplete information regarding specific biological phenomena. The next challenge is to be able to integrate such diverse types of data.

Another challenge is the oversimplification that characterizes many of the models provided by pathway databases. In principle, each type of tissue might have different mechanisms so generic, organism-level pathways present a somewhat simplistic description of the phenomena. Furthermore, signaling and metabolic processes can also be different from one condition to another, or even from one patient to another. Understanding the specific pathways that are impacted in a given phenotype or sub-group of patients should be another goal for the next generation of pathway analysis tools.

Interpreting biological experimental data is also challenging due to inaccurate assumptions. For instance, most current pathway models show cascades of signals or biochemical processes next to one another, in time-agnostic diagrams. In reality, these phenomena happen over time, and often at different time scales. Furthermore, many data sets offer only a snapshot in time, at a particular moment. Almost by definition, such a frozen snapshot cannot properly capture and show the effect of successive events that take place over time.

The graphical scoring methods presented in this chapter are representative of the techniques available for future methods. We expect to see greater use of different types of

data, in addition to greater use of data mining and machine learning which will lead to more sophisticated topology-based pathway analysis methods.

It is important to (re-)state that the goal of this chapter was to survey the main topology-based techniques and methods available to identify the most significant pathways in a comparison between phenotypes. In other words, the goal was to identify, categorize and review these methods *without attempting to asses their performance*. A critical assessment and ranking will be the subject of a later publication. A natural tendency would be to try to use the various criteria used here to compare various methods and thus establish even a partial ordering. For instance, if method X using only one type of input (e.g. pathways from KEGG) while method Y uses two types of input (e.g. pathways from KEGG as well as PPI data), one might be tempted to conclude that method Y is somewhat more powerful than method X. Similarly, some methods use a subset of DE genes while others use the entire set of measured values. Again, it may be tempting to informally conclude that the later methods are more powerful since, they take more data into consideration or because they eliminate the need for a selection of DE genes. It is our opinion that such inferences and partial orderings are not advisable and should not be attempted based on the information presented in this chapter. A proper assessment of these methods should be focused on their ability to identify the pathways that are truly impacted in the given phenotypes, and not based on superficial characteristics or number of features of one type or another.

CHAPTER 3: QUALITATIVE CHANGE DETECTION

3.1 Overview of change detection methods

A biological system is characterized by a tendency to reach and maintain a state of homeostatic balance, considered to be a stable state. An alteration made by internal or external stimuli can trigger the system to transition from one stable state to another, referred to as a qualitative change. Notably, any of the system components taken in isolation may not vary dramatically; however, the system as a whole may undergo a qualitative change. Conversely, in a resilient system, important variations of one or a group of components may happen without necessarily involving a qualitative systemic change. Importantly, most systems have built-in tolerance mechanisms such that the response to a stimulus is delayed until the signal is perceived as real in order to filter noise and to conserve the energy necessary to undergo a systemic change.

In this chapter, we present a qualitative change detection (QCD) approach, an analysis method that uses sequential measurements as described by a time series (or by progressive disease stages), together with all known interactions described by biological networks, and that applies an impact analysis approach to identify the time interval in which the system transitions to a different qualitative state.

In the landscape of analysis methods for high-throughput data (see Fig. 3.1), the proposed method falls under the category of dynamic network analysis. Other methods in the same category aim to either identify significantly perturbed systems [95], time intervals with the highest difference in expression for each gene from a predefined set [152], dynamic network biomarkers using local network entropy [92], or time periods of differential gene expression using Gaussian processes [68]. However, all of these approaches perform comparisons between disease profiles and a reference profile (e.g. healthy). In the paradigm proposed here, none of these existing methods can be applied because the goal is to identify a transition to a

qualitatively different state without knowing the gene expression profile of the new state, and hence, without the ability to make a comparison between the control and disease phenotypes.

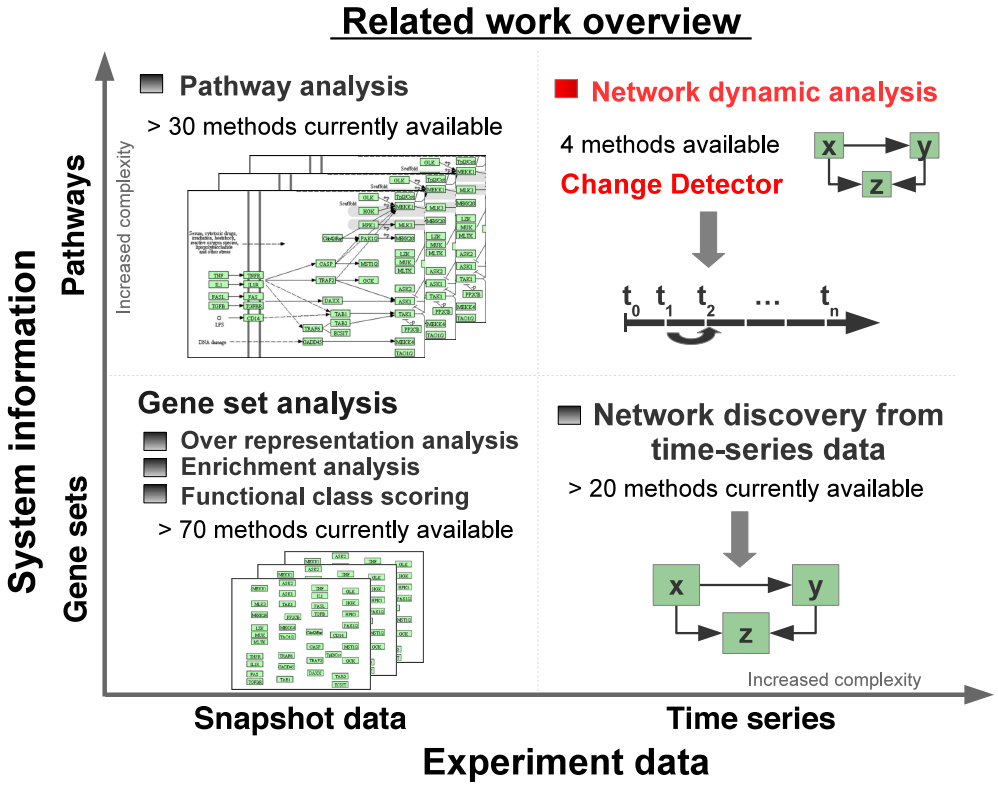

Figure 3.1: Overview of existing approaches as categorized by looking at the time component (horizontal axis) and the system information (vertical axis). From the time component perspective one can distinguish between two categories: snapshot data and time course data. Time course data is richer in information but also has increased complexity as opposed to snapshot data. From the system information perspective one could consider sets of genes together with their interactions (pathways) or without such interactions (gene sets). Pathways are much richer in information but also have increased complexity as opposed to gene sets. Based on these categories, the existing methods can be divided into the four groups shown, of which the gene set analysis is the most common, including more than 70 methods [70, 56]. Gene set analysis takes as input a collection of gene sets and a snapshot of expression data that compares two phenotypes and ranks the gene sets based on their relevance to the phenotype computed by the analysis. Pathway analysis has the same workflow as the gene set analysis but also takes into consideration the interactions between the genes as described by the topology of the pathways [87, 111]. Network discovery from time course data takes as input data collected at multiple time points and a set of genes and infers relations between the genes in the input set [123]. Network dynamic analysis is the most recent, has only 4 existing methods [96, 152, 92, 68], uses time series data and pathways, and aims to extract phenotype-related information related to changes in genes, pathways, or time intervals.

3.2 Methods

3.2.1 Qualitative change detection (QCD) method

Here, we propose a paradigm shift: instead of detecting the onset of disease, we would like to be able to detect the departure from the healthy state. The qualitative change detection (QCD) analysis presented here is able to detect intervals when a biological system undergoes qualitative changes such as the transition from healthy to disease.

The workflow of the analysis is summarized in Fig. 3.2. The input to QCD consists of: i) time-series data and ii) a network model of the biological system under study. The output is a list of time intervals when the systems transitions between qualitative states. The workflow of the analysis consists of the following steps:

1. Compare the status of the system between each pair of time points using an existing statistical method called **pathway impact analysis (IA)** [38, 156, 166, 165] and assess the levels of perturbation;

2. Separate large and small inter-state perturbations using a gamma mixture model fitted to the system perturbation by an expectation maximization algorithm;

3. Calculate the change interval(s) as the narrowest disjunct interval(s) of large changes.

In step 1 the perturbation of the system between all pairs of system states is computed utilizing IA. First, sequential states are assigned to the chronologically ordered time points or disease progression stages when the data were sampled. We then compare all pairs of systems states using IA [38], which was previously developed to evaluate the pathway impact when comparing two phenotypes; herein, we use it to calculate a system/pathway impact factor for each comparison of two system states (time points). The result will be a list of time intervals (comparisons) with their computed pathway perturbation factor.

The pathway impact analysis takes as input signaling networks (pathways) and a list of genes with their respective changes between two states of a system (e.g. condition

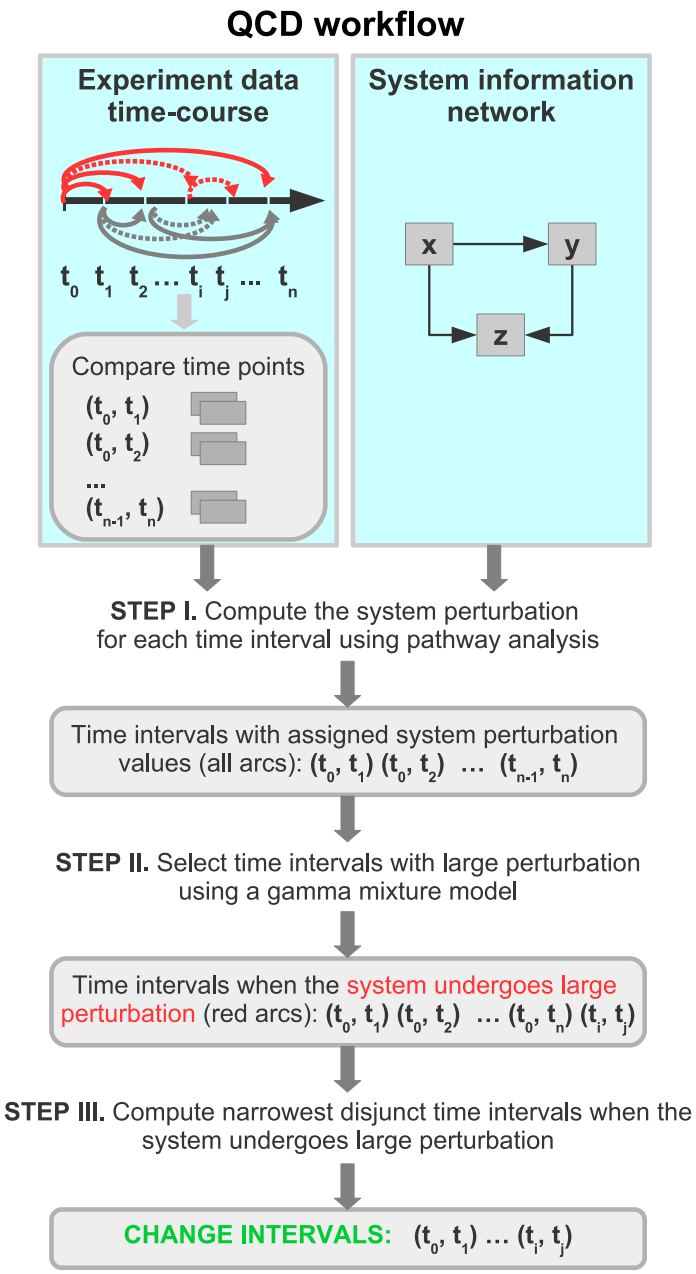

Figure 3.2: Workflow of the QCD method. The algorithm takes as input time series data and network(s) that models the biological system. The time series data is used to compare every pair of time points (time interval). In STEP I, a pathway impact analysis is used to compute a perturbation score for each comparison. In STEP II, an expectation maximization algorithm is employed to identify the parameters of a gamma mixture model and select the interval(s) when the system/pathway/network experienced a large perturbation. In STEP III, change intervals are selected by identifying the overlap of the set of intervals with large system perturbation and selecting the narrowest disjunct time intervals.

vs. control). In a typical signaling pathway, nodes represent genes or gene products and edges represent signals, such as activation or repression, directed from one node to another. The goal of IA is to identify the pathways significantly impacted in a given phenotype by analyzing all measured expression changes for all genes, as well as all of their interactions, as described by each pathway. This type of analysis incorporates two types of evidence, which taken together estimate the disruption on a pathway when comparing two phenotypes. The first type is evidence given by the perturbation analysis. The magnitude of expression change (log fold-change) and the pathway structure are used to compute a perturbation factor for each gene (eq. 3.1). The gene perturbation factors are summed up to the pathway level to account for the observed pathway perturbation.

$$PF(g) = \Delta E(g) + \sum_{u \in US(g)} \beta_{ug} \cdot \frac{PF(u)}{\#DS(u)} \tag{3.1}$$

where $PF(g)$ is the perturbation factor for gene g, $US(g)$ is the set of genes directly upstream of g, β_{ug} is the strength of interaction between u and g, $DS(g)$ is the set of genes directly downstream of g, $\Delta E(g)$ is the log fold change in expression for g, and # denotes set cardinality.

For the perturbation analysis, we sum the absolute value of the gene perturbation factors (eq. 3.2) so that the up-regulation and down-regulation do not cancel each other.

$$PF(P) = \sum_{g \in P} |PF(g)| \tag{3.2}$$

where $PF(P)$ is the perturbation factor for pathway P and $|\cdot|$ denotes the absolute value operator.

We use the all-gene approach, without gene weights; therefore, since we do not select differentially expressed genes, the enrichment part cannot be computed. The pathway perturbation factors are positive values with 0 marking no perturbation — the higher the value, the larger the pathway perturbation. We work under the assumption that the pathway

perturbation factors follow a gamma distribution with mode = 0 when the pathway is not perturbed.

In step 2, the distribution of the pathway perturbation factors is modeled using a gamma mixture model (see Fig. 3.3). The hypothesis states that if there is a change interval the system state comparisons will yield a mix of large and small system perturbations. Small system perturbations are expected when comparing system states before and after the change interval. Large system perturbations are expected when comparing system states before the change interval to states after the change interval. Therefore, a mixture of two gamma distributions is used: one for the comparisons in which the system is unperturbed (i.e., the null hypothesis) and another for comparisons in which the system is perturbed. The mixture model will be initialized with two distributions having the mode equal to the minimum and maximum of the perturbation factors. The mixture model fitting will provide two distributions that best fit the data together with a percentage that estimates how much of the observed data comes from each of these two distributions. If any of the distributions has a percentage of less than 10%, the QCD analysis considers that there is only one distribution and, therefore, there is no significant change, and no change interval.

If both distributions fitted contribute more than 10%, the goodness of fit is then evaluated by computing the percentage of overlap between the observed and fitted distributions of system perturbations (see Fig. 3.3, overlap). Other statistical approaches (the Kolmogorov-Smirnov test and the Kullback-Leibler divergence) are also used to evaluate the goodness of fit and results are presented in section 3.5.1. If the mixture contains more than 10% of either of the distributions, the intersection of the two distributions is used as the threshold to select comparisons with large system perturbations. Comparisons that yield a pathway perturbation factor higher than this threshold will be marked as having a large system perturbation.

An important requirement is to demonstrate that the approach does not report false positive changes in random data or in cases in which there are no changes in the organism.

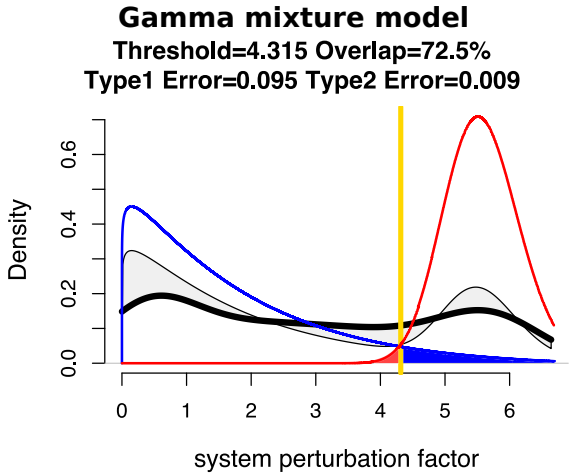

Figure 3.3: The fit of a mixture of two gamma distributions (blue and red lines) to the observed perturbation values of the system as computed for all pair-wise comparisons (thick black line). The fitted mixture distribution is marked by the thinner black line. The difference between the fitted and observed data is shaded in light gray. A goodness of fit measure is the overlap calculated as the ratio between the intersection and union of the areas under the observed data (thick black line) and fitted model (thinner black line). A perfect fit would yield an overlap of 100%. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). If a second distribution is found to be present (red), the threshold used to distinguish between small and large system perturbations will be the yellow vertical line. Under these circumstances, the blue area under the blue line is the Type 1 error and the red area under the red line is the Type 2 error.

Section 3.5.3 includes the results obtained with controls only, as well as results obtained with random data. These results show that the proposed approach does not report falsely significant changes.

In step 3, change intervals are computed as the overlap of comparisons with a large system perturbation using an algorithm based on the definition in subsection 3.2.2. The algorithm takes as input a list of comparisons with an assigned system perturbation value and a predefined system perturbation threshold computed in step 2 described above. The algorithm iterates over the list of comparisons and identifies the start and end points of change intervals as points that have at least one comparison that shows a large perturbation (higher than the threshold) starting or ending in the respective points, and no large perturbation comparisons start or end in between those points. The output is a list of change intervals

described by their start and end points. Note that the change interval does not have to be a comparison that shows a large system perturbation by itself.

3.2.2 Change interval formal definition

Notations:

$N =$ the number of time points

$S =$ the set of states

$p =$ the set of perturbation values

$CI =$ the set of change intervals

$pcut =$ the perturbation threshold

Definitions:

$S = \{S_i \mid i \in \{0, \ldots, N-1\}\}$

$p = \{p_{ij} \mid i, j \in \{0, \ldots, N-1\}, i < j\}$, where p_{ij} is the system perturbation value when comparing S_i and S_j

$CI = \{(x, y), x, y \in \{0, \ldots, N-1\}, x < y\}$, that satisfy the following conditions

$\forall i, j \in \{x, \ldots, y\}, (i, j) \neq (x, y)$ and $p_{ij} \leq pcut$

(i) $\exists i \in \{0, \ldots, y-1\}$ such that $p_{iy} > pcut$

(ii) $\exists j \in \{x+1, \ldots, N-1\}$ such that $p_{xj} > pcut$

(iii) x is the max value to satisfy the above conditions for a given y

(iv) y is the min value to satisfy the above conditions for a given x

3.2.3 Meta-states statistical validation

To better understand the phenomenon under study, after the detection of a change interval, the states of the system before and after a change interval should be analyzed to gain insight regarding the state of the system before and after a qualitative change. To describe this analysis, the situation in which there is a single change interval will be considered, as in the *E. coli* flagellum building dataset. In this case, the system is considered to be stable before and after the change interval. In this context, we group the states in which the system

is stable into meta-states. We define a meta-state (see Fig. 3.4A) as a group of consecutive states that satisfy the following two conditions:

1. All comparisons between states within a meta-state have a small system perturbation;
2. All comparisons between states from a meta-state to states outside the meta-state (excluding the states in the change interval) have a large system perturbation.

In the above, definition, the "small" and "large" perturbations, are defined based on the threshold between the two gamma distributions computed in the previous step and shown as the yellow line in Fig 3.3.

Note that all comparisons between the states within a change interval and the meta-states immediately before and immediately after it may have a small system perturbation. This is because, during the change interval, the system is in transition between the two meta-states; therefore, its state during the transition is a mix of the two meta-states that may not be qualitatively different from either of them.

Based on the detected change interval, groups of sequential system states can form potential meta-states (see panel B in Fig. 3.4). Panel A in Fig. 3.4 shows the ideal results of all comparisons between all states involved in these meta-states. In essence, all comparisons within each potential meta-state should show a small system perturbation while all comparisons between a meta-state time point and a time point outside the meta-state (excluding the change interval) should show a large system perturbation.

To validate each observed potential meta-state, a statistical approach is applied to evaluate how closely it meets the conditions of a theoretical meta-state. The validation of the potential meta-state is described for the *E. coli* flagellum building dataset. The data was sampled at 21 time points (system states S0–S20) and the change interval was detected as (S6–S10). In this case, there are two potential meta-states: MS1, which contains the states before the change interval (states from S0 to S6), and MS2, which contains the states after the change interval (states from S10 to S20). To investigate the potential meta-states,

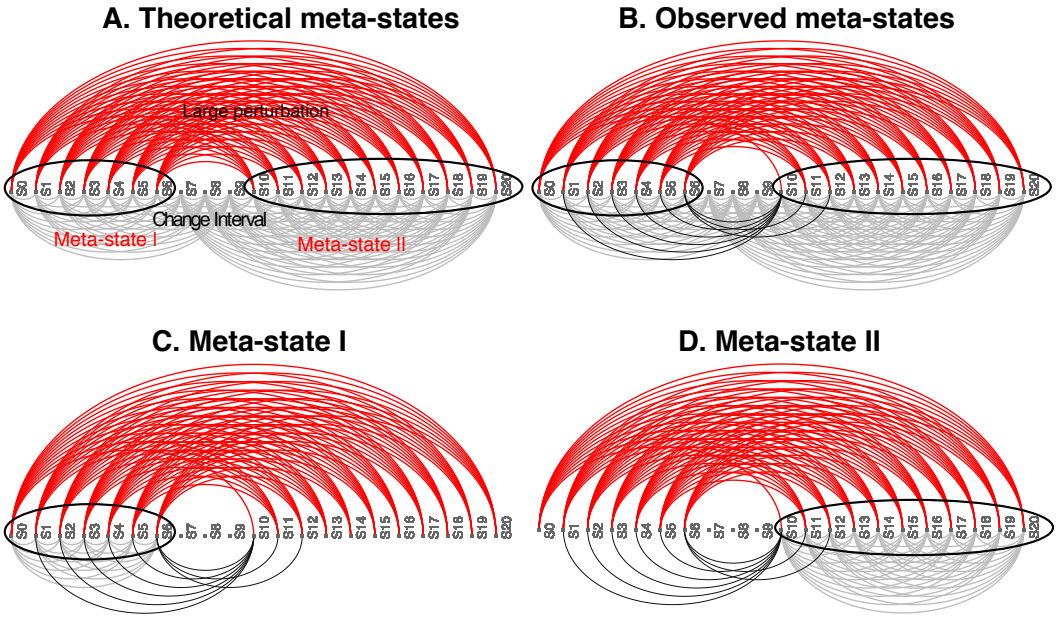

Figure 3.4: Meta-states in the *E. coli* flagellum building case study. Arc plots show possible comparison between time-points (states): comparisons with large system perturbation are red, and comparisons with small system perturbation are gray or black. Panel A: the expected arc plots of two theoretical meta-states (groups of states in the black ellipses) relative to the detected change interval (S6–S10): all comparisons within each potential meta-state should show a small system perturbation while all comparisons between a meta-state time point and a time point outside the meta-state (excluding the change interval) should show a large system perturbation. Panel B: the actual arc plot showing the observed large perturbation (red) vs small perturbation comparisons (gray and black) for all possible state comparisons. Black arcs show comparisons between states of potential meta-states (groups of states in the black ellipses) to states outside the potential meta-state (excluding the change interval) that show a small perturbation. Panel C: the arc plot shows the observed comparisons for potential meta-state I (S0–S6, states in the black ellipse). Black arcs show comparisons between states of potential meta-state I to states outside it (excluding the change interval) that show a small perturbation. Panel D: the arc plot shows the observed comparisons for potential meta-state II (S10–S20, states in the black ellipse). Black arcs show comparisons between states of potential meta-state II to states outside it (excluding the change interval) that show a small perturbation.

all comparisons (arcs) (see Fig. 3.4B) are considered from the perspective of the meta-state definition above. For MS1, all comparisons between the states S0 to S6 should yield only small system perturbations. In addition, all comparisons between any states in MS1 and any states outside MS1 (not including the change interval) should involve large perturbations. With these considerations, all comparisons involving MS1 states can be assigned a binary value: either consistent or inconsistent with the expectations above. For each potential meta-state, a statistic is computed as the number of time intervals with status consistent with the status (large/small) assigned in the corresponding theoretical meta-state (meta-state definition). Under the null hypothesis, in which there are no groups of system states that form meta-states, the probability that a comparison is consistent or not should be 0.5. Based on this framework, a binomial model is used to calculate a p-value for the statistic computed for each of the groups of states that are potential meta-states (see Fig. 3.4C and D):

$$X \sim binom(n, 0.5), \tag{3.3}$$

where n is the number of trials and 0.5 is the probability that the status of a comparison is consistent with the meta-state definition.

The p-value computed for the potential meta-state characterizes the amount of evidence indicating the existence of a true meta-state (comparisons consistent with the definition, vs. inconsistent comparisons). A significant p-value lower than a predefined threshold would confirm the identification of a true meta-state. In our case studies, most p-values were significant at a 1% threshold (see details in section 3.5.1).

3.2.4 Synthetic data parameters

For the *E. coli* flagellum building and *B. subtilis* sporulation, the gene expression synthetic data were generated using the interactions described by the biological network and Hill functions for protein accumulation (eq. 3.4) and decay (eq. 3.5) with a rate of $\alpha = 0.005$.

Given that $X \rightarrow Y$ denotes that transcription factor X regulates gene Y

$$Y(t) = Y_{st}e^{-\alpha \cdot t}, \text{ protein accumulation} \tag{3.4}$$

$$Y(t) = Y_{st}(1 - e^{-\alpha \cdot t}), \text{ protein decay} \tag{3.5}$$

where Y_{st} is the steady state expression level for gene Y, α is the decay rate for protein Y, t is time and Y(t) is the expression level for gene Y at time t.

For the third case study, *C. elegans*, data were generated using a step function for the $X_1 = FLP$ neuron and a constant function (0) for the $X_2 = ASH$ neuron The following formula describes the change in voltage over time for the $Y = AVD$ neuron:

$$dY/dt = f(0.5 \cdot X_1 + 0.5 \cdot X_2 > K_Y) - Y \tag{3.6}$$

The following formula describes the change in voltage over time for the $Z = AVA$ neuron:

$$dZ/dt = f(0.5 \cdot (X_1 + X_2) + 0.4 \cdot Y > K_Z) - Z \tag{3.7}$$

Constants 0.5 and 0.4 are the strengths of the synaptic connections, and K_Y and K_Z are the activation thresholds.

For the real gene expression, microarray data were downloaded from the GEO database. The CEL files downloaded from GEO for the real gene expression were processed using custom R scripts (R version 3.1.2). Data pre-processing (background correction and normalization) was performed using the threestep function from the affyPLM (version 1.42.0) R package. Gene IDs were matched to gene symbols using the respective annotation packages from R: org.Sc.sgd.db (yeast), org.Dm.eg.db (fruit fly), org.Mm.eg.db and moe430a.db (mouse), org.Hs.eg.db and hgu133plus2.db (human). Gene expression level at a specific time-point was computed as the average of the replicates for the specific time point when

replicates were available. The ROntoTools 1.6.1 R package was used for impact analysis. The mixtools 1.0.3 R package was used for the mixture model analysis.

3.3 Results

The analysis of eight well-studied phenomena was performed with the proposed method (QCD) for seven model organisms using both synthetic and real data. To assess the ability of QCD to detect qualitative changes, results were compared to prior knowledge of the phenomenon under study. QCD uses system knowledge, as described by a known gene signaling network or a map of neurons and their synaptic connections, as well as sequential measurements of the system components (genes or neurons). Data were obtained by measuring either the mRNA level of the genes involved in the system, in the case of real data, or generated based on equations describing the model of each organism, in the case of synthetic data.

The results of the analyses show that QCD can reliably identify the time interval during which a biological system goes from one qualitative state to another in response to organism development or to a shift in environmental conditions. We evaluate the method using phenomena that involve major physiological changes. We also evaluate the method for phenomena involving more subtle, yet important changes. Major physiological changes analyzed using synthetic data are *E. coli* flagellum building [5, 85] and *B. subtilis* sporulation [5, 45]. The subtle change analyzed using synthetic data is *C. elegans* avoidance reflex [5, 26]. Major physiological changes analyzed using real gene expression data are yeast sporulation [28] and fruit fly pupariation [15]. More subtle changes analyzed using real gene expression data involve fruit fly ethanol exposure [88].

QCD was compared with an existing method developed by Liu and colleagues used to detect network biomarkers and the pre-disease state (herein abbreviated DNBM) [92]. In addition to the six datasets mentioned above, we also ran QCD on the two datasets from the Liu et al study. The first dataset is derived from a mouse study of exposure to a toxic

gas (carbonyl chloride). Using these data QCD identified one qualitative change, before the exposure became lethal, preceding the pre-disease state detected by Liu et al. The second dataset contains data describing the progression of human hepatocellular carcinoma. Using these data, QCD identified a qualitative change from a benign stage (control) to a pre-malignant stage (high-grade dysplastic nodules), also preceding the pre-disease state detected by the Liu et al study.

3.3.1 Bacterium flagellum building

When in an environment lacking nutrients, the *E. coli* bacterium initiates the process of building a flagellum that will provide the motility necessary for finding an environment rich in nutrients.

We analyzed the process of building the *E. coli* flagellar motor, using synthetic data and the flagellum building network [85] (see Fig. 3.5A). Previous studies describe this network as a multi-output coherent type 1 feed-forward loop (C1-FFL) [85, 146]. A C1-FFL is a network in which one gene activates another and, together, they activate another gene or (groups of) genes in the multi-output networks [146, 110].

The flagella building network is a generalization of the C1-FFL. In essence, the flagella building network is a multi-output C1-FFL in which the exact timing of the sequence of steps is controlled by the different activation thresholds (see the edge labels in Fig. 3.5A). These thresholds ensure that all the elements of the flagellum are built in a specific order so that it can properly assemble (e.g. the base of the structure must be in place before all other elements, etc.). Due to the different activation thresholds, a reverse order of the activation thresholds for $flhDC$ and $fliA$ yields a first-in first-out (FIFO) order in the gene transcription. This is typical of sensory transcription networks as a mechanism used to filter out (not react to) noise containing false positive signals of short duration.

Gene expression data was generated for the flagellum building network for a period of 10 hours using a continuous function that models the protein accumulation and param-

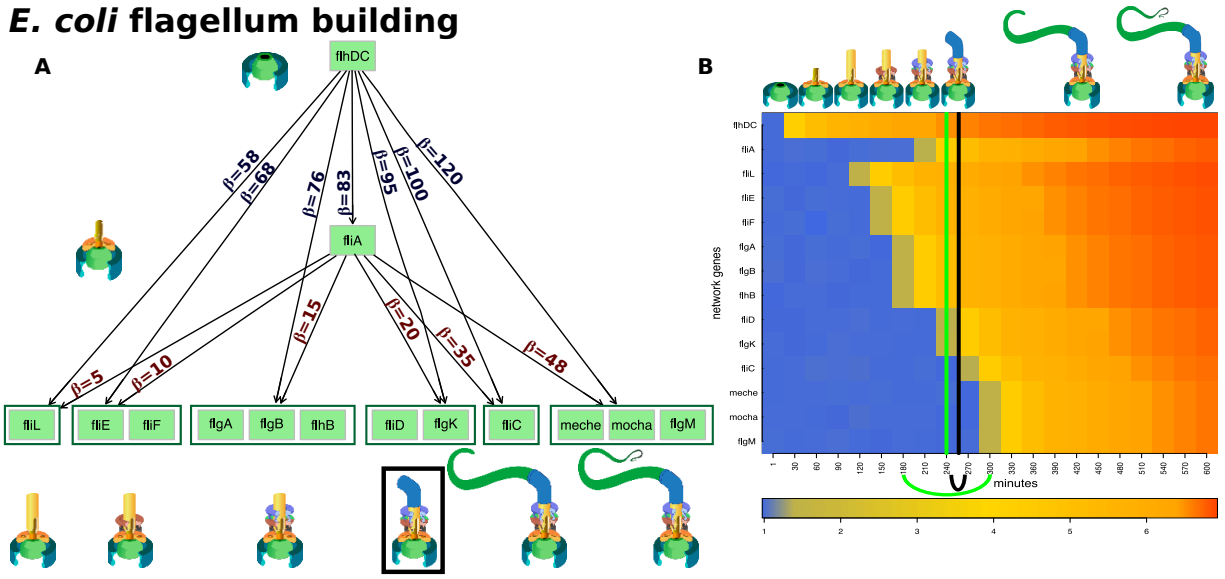

Figure 3.5: The input and results of the qualitative change detector (QCD) for the *E. coli* flagella building phenomenon. Panel A: The multi-output coherent type 1 feed-forward loop (C1-FFL) network that describes the flagellum building, together with the activation thresholds (β on the edge) for each of the six groups of genes (dark green boxes) [5, 85]. The flagellum building is depicted in the cartoons matching the activation of each group of genes. The black box denotes building the flagellum hook which is the point of no return in this process and hence the real change interval that we aim to discover. Panel B: The heatmap of the sampled data (input to QCD), and the real change interval (black arc below the heatmap and black vertical line positioned in the center of the interval) as described by literature. The change interval detected by QCD is shown by the green arc below the heatmap and the green vertical line positioned in the center of the interval (very close to the black line showing the actual point of no return). The stages of the flagella building are presented as cartoons in chronological order on the top part of the figure.

eters from previous studies [5, 85]. Samples were taken every 30 minutes leading to a gene expression time course dataset with 21 time points. Panel B in Fig. 3.5 shows the evolution of gene expression over time for the genes involved in this phenomenon.

Importantly, the organism commits to building the flagellum when the first hook of the flagellar motor starts to be built ($fliA$ reaches the threshold to regulate the next group of genes, $fliD$ and $flgK$) [85]. This is an important check point in the flagella building process as the assembly of the following component can still be halted if necessary [171]. However, after this checkpoint, the bacterium commits to building the flagellum (see the top of panel B in Fig. 3.5). For these reasons, the interval between 240 and 270 minutes can

be considered the boundary that separates the two qualitatively different states: with and without flagellum. The goal of our approach is to find this interval without any knowledge about the phenomenon and with knowledge only from the gene expression data and the network of the system.

The *E. coli* flagellum construction is controlled by two transcription factors, $flhDC$ and $fliA$ (see Fig. 3.5A). The master regulator $flhDC$ activates $fliA$ and there is an OR relationship through which these two master regulators activate the other genes in the network (12 genes). The genes are part of 6 groups: (i) $fliL$, (ii) $fliE$ and $fliF$, (iii) $flgA$, $flgB$, $flhB$, (iv) $fliD$, $flgK$, (v) $fliC$, (vi) *meche*, *mocha* and $flgM$.

QCD compares all system states (time points) to each other using a pathway impact analysis. In essence, the state of the system at each time point is compared to the state at all other time points using a pathway impact analysis [38] that takes into consideration all gene expression changes, the position of each gene on the pathway (Fig. 3.5A), and the type and direction of every interaction to determine if the state of the system was altered. The result of this impact analysis is a set of system perturbation factors that quantify the system perturbation. To determine the significant system perturbations, we assume there are two types of intervals: i) those with large perturbations between the states involved, and ii) those with small perturbations caused only by random fluctuations. We then use an expectation maximization algorithm to fit a gamma mixture model of two distributions to the perturbation factors (see Fig. 3.6). The intersection of the two distributions will be the optimal threshold that can be used to separate the large perturbations from the small perturbations as presented in Fig. 3.6A. Using this approach, we assign a "large" or "small" perturbation status to each comparison. Panel B in Fig. 3.6 shows all the state comparisons considered, in which the gray and black arcs show small perturbations and the red arcs show large perturbations between the states of the system at those time points.

In essence, most of the comparisons between any time point earlier than 180 mins and any time point after 300 mins show large perturbations (exceptions are marked by the

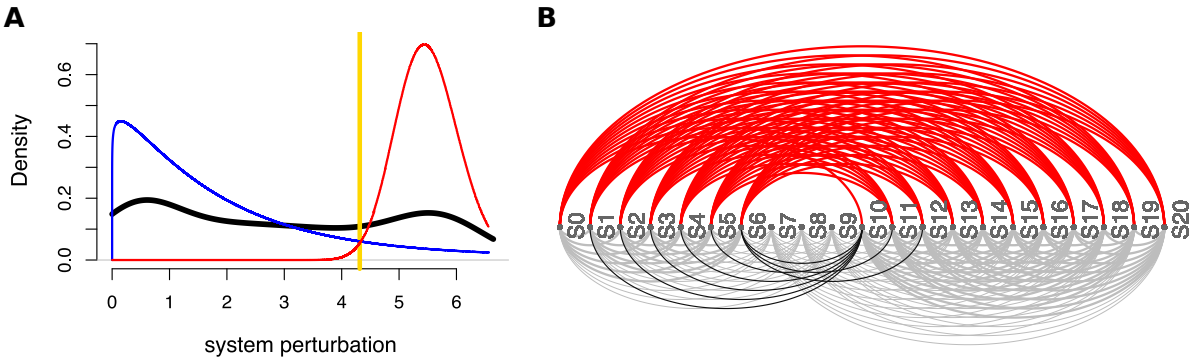

Figure 3.6: Identifying the qualitative change interval for the *E. coli* flagella building phenomenon. Panel A: Identifying state comparisons involving large perturbations. The black line shows the observed density of the perturbation values for all pairwise comparisons of system states. We assume that some comparisons will be characterized by large perturbations, while others by small perturbations. A mixture of two gamma distributions are fitted to the observed data to yield the distributions of large (red) and small (blue) perturbation whose mixture best fits the observed data (red and blue lines). The intersection point (yellow vertical line) is the optimal threshold used to distinguish between the large and small perturbations. Panel B: The arcplot of all comparisons performed by QCD between all pairs of system states. Red arcs, above the x axis, represent comparisons that show a large perturbation, while gray arcs, below the x axis, represent comparisons with a small perturbation. All the comparisons between states in the intervals S0–S6 and S10–S20 are associated with small perturbations. At the same time, the vast majority of all possible comparisons between any state in the interval S0–S6 and any state in the interval S10–S20 are associated with large perturbations. The black arcs are comparisons between a state in the interval S0–S6 and a state in the interval S10–S20 that are associated with small system perturbations. The smallest interval of overlapping large perturbation intervals, the interval between S6 and S10, is the detected change interval.

black arcs). This suggests that a qualitative change of the system occurs between 180 and 300 mins, which is indeed the case. The real change takes place between 240 minutes, when $fliD$ and $flgK$ expression begins, and 270 minutes, when $fliA$ starts to regulate the next group of genes and the building of the first hook of the flagellar motor begins.

The identification of a change interval should be followed by an analysis of the states of the system before and after a change interval in order to gain insight into the system transition. Without loss of generality, we will consider the situation in which there is a single change interval, as in this dataset. Furthermore, we also assume that the system is in

a stable state before and after the change interval. Under these circumstances, we can group the states in which the system is stable into meta-states.

A meta-state is a group of consecutive states where all comparisons between states within a meta-state have a small perturbation and all comparisons between states from a meta-state to states outside it (excluding the states in any change intervals) have a large perturbation.

The results shown in panel B of Fig. 3.6 suggest that states S0–S6 might form a meta-state, MS1. Similarly, the states S10–S20 might define a second meta-state, MS2. To investigate these potential meta-states, all comparisons (arcs) were studied from the perspective of the above definition of a meta-state. From this perspective, all these comparisons can be either consistent or inconsistent with the expectations noted above. This is a binary choice, and under the null hypothesis in which there are no meta-states, the probability that a comparison is consistent or not should be 0.5. Based on this framework, a binomial model can be used to calculate a p-value characterizing the amount of evidence that indicates the existence of a true meta-state (comparisons consistent with the definition vs. inconsistent comparisons). More details can be found subsection 3.2.3. Groups of states with significant p-values will be reported as meta-states.

In this case, both groups of states identified by QCD had highly significant p-values: $p = 5.44 \times 10^{-19}$ for meta-state 1 (S0–S6) and $p = 3.61 \times 10^{-28}$ for meta-state 2 (S10–S20).

3.3.2 Bacterium sporulation

When deprived of food, the *B. subtilis* bacterium turns into a spore, a robust structure able to survive in an environment lacking nutrients. This is a crucial feature that ensures the bacterium's survival in an environment scarce in food in which it cannot survive in its active form.

Compared to the *E. coli* flagellum-building network, which includes only activation signals, the network controlling sporulation also includes repression signals (Fig. 3.7A). This

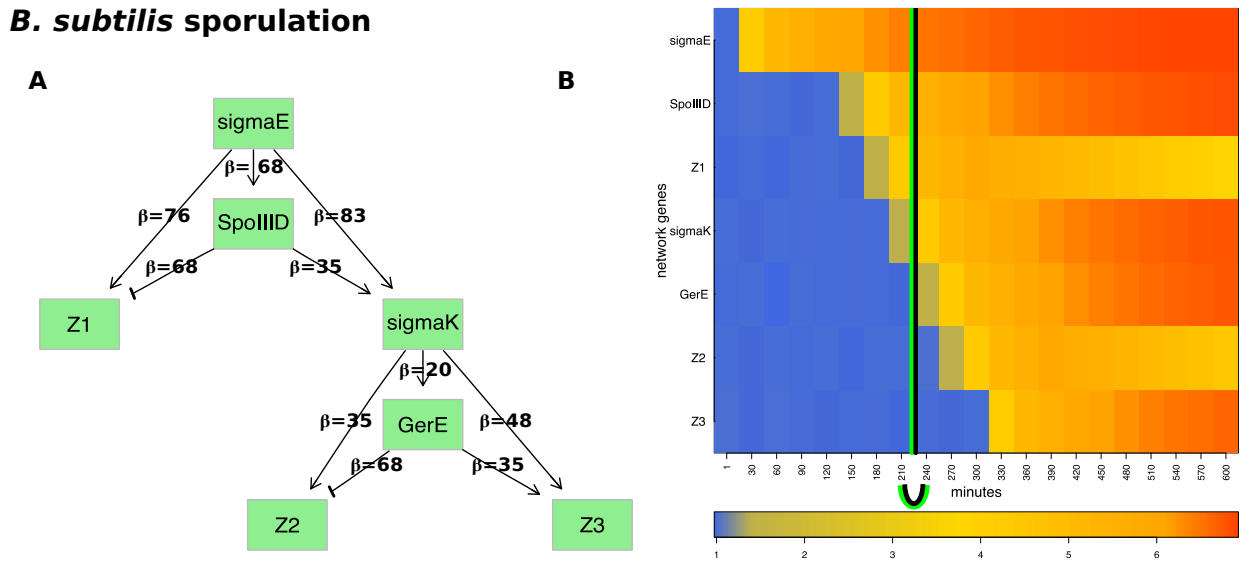

Figure 3.7: The input and results for QCD for *B. subtilis* sporulation. Panel A shows the sporulation network. The genetic network represented by the two coherent type 1 feed-forward loops (C1-FFLs) and two incoherent type-1 FFLs (I1-FFLs) that describe the sporulation network as reported in previous studies [5, 45]. Panel B shows the heatmap of the sampled data. The real change interval is shown by the black arc below the heatmap (black vertical line positioned at the average of the interval limits) as described by literature. The change interval detected by the proposed method is shown by the green arc (green vertical lines positioned at the average of the interval limits), match perfectly with the actual timing of these events.

network has a hierarchical structure that consists of four transcription factors: $sigmaE$, $sigmaK$, $GerE$, $SpoIIID$ and three groups of genes $Z1$, $Z2$, $Z3$. This network is comprised of two network motifs, each of them represented by two networks. The two coherent feed-forward loops (C1-FFLs) aim at $sigmaK$ and $Z3$, respectively, while the two incoherent type-1 feed-forward loops (I1-FFLs) are centered around $Z1$ and $Z2$, respectively. The C1-FFLs are denoted as coherent because their central genes, $sigmaK$ and $Z3$, respectively, receive activation signals from both genes upstream of each. Specifically, $sigmaK$ receives activation signals from both $SpoIIID$ and $sigmaE$, while $Z3$ receives activation signals from both $GerE$ and $sigmaK$. In contrast, the incoherent network is characterized by a gene that receives one activation and one repression signal from the two genes immediately upstream of the target gene. For example, $Z1$ is activated by $sigmaE$ but repressed by $SpoIIID$.

The gene expression data was sampled from the spore formation network for a period of 10 hours using a continuous function that models the protein accumulation and parameters observed in previous studies [5, 45]. Samples were taken every 30 minutes leading to a gene expression time course dataset with 21 time points.

Importantly, the organism commits to the spore formation when the second suppressor ($GerE$) is expressed (4h = 240 min) [45]. In turn, $GerE$ is regulated by $sigmaK$ which also regulates the communication between the mother cell and the spore through a checkpoint that is crucial for the formation of viable spores. Hence, the true interval of change is the interval between 210 minutes, when $sigmaK$ shows the first change in expression, and 240 minutes, when $GerE$ shows the first change in expression.

Our method was applied on the sporulation network and the synthetic gene expression dataset obtained by the above sampling. In these data, QCD identified one change interval (210 – 240 min) (Fig. 3.7B). The detected interval exactly matches the time interval between the time when the spore formation starts ($GerE$ is being expressed) and up to the moment when the next group of sporulation genes ($Z2$) is activated.

We also evaluated the two groups of system states: before the change interval (0 – 210 min) and after the change interval (240 – 600 min), as potential meta-states MS1 and MS2, respectively. The p-value for each was highly significant: $p = 2.31 \times 10^{-19}$, for MS1, and $p = 6.23 \times 10^{-32}$, for MS2. These p-values validate the hypothesis that these are true meta-states. Interestingly, these meta-states can be mapped to the rod-shaped bacterium form and the endospore form, respectively, while the detected change interval can be associated with the process of spore formation. These results are consistent with previous studies and interpretations [45]. Before the change interval, the bacterium preserved most of its initial characteristics, while after this interval, the bacterium assumed most of the characteristics of an endospore. During the change interval, the system exhibited characteristics of both "spore" and "no spore" states.

3.3.3 Worm avoidance reflex

A phenomenon involving more subtle changes is the nociception reflex. Nociception is a sensory process that allows the detection of harmful stimuli and activates a reflex response to move a part of the body or the whole body away from the stimulus. Nociceptors are present in fish, worms, and fruit flies, among others, and help trigger an avoidance reflex such as a backward movement. In the roundworm (*C. elegans*), the avoidance reflex network is composed of two parallel receptor neurons that communicate with two sequential command neurons (Fig. 3.8A).

The *C. elegans* avoidance reflex network is a generalization of the C1-FFL in the form of a multi-input C1-FFL. As previously described, C1-FFL is a network of three nodes in which one node activates another and, together, they activate another node [146, 110]. In multi-input C1-FFL networks, the initial activation is performed by multiple nodes or groups of nodes rather than by just one node. ASH is the main nociceptor and triggers avoidance behavior in response to harmful stimuli such as the nose touch and volatile chemicals. FLP is a sensory neuron triggered by painful, heat-related stimuli or mechanical stimuli, such as a harsh nose touch, that initiates the nematode's backward movement. AVD is a command interneuron that functions as a modulator for backward locomotion induced by a head touch. Neurons AVA and AVD drive the worm's backward movement.

Neuronal signal data was generated for the avoidance reflex network over a period of 8 milliseconds, using a continuous function that models the signal processing and parameters observed in previous studies [5, 26]. Samples were taken every millisecond leading to a time course dataset with eight time points.

The nematode commits to the backward movement at 3ms, which is the moment the nose touch (FLP - spiking function) reaches the threshold to trigger the second command interneuron (AVD). The movement starts at 5ms when the AVA neuron starts firing [26]. The two time-points mark the 3 to 5ms time interval which is the real change interval. Using these data, QCD identified the narrower 4ms to 5ms interval (Fig. 3.8B).

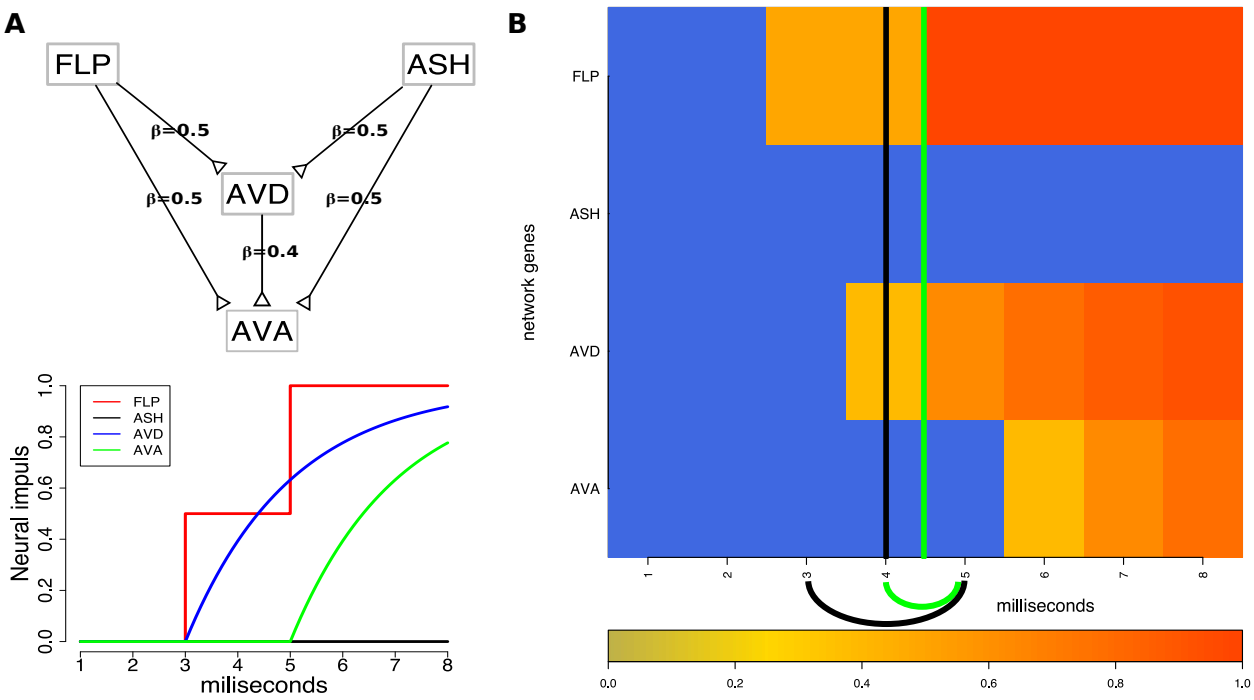

Figure 3.8: The input and results for QCD for the *C. elegans* avoidance reflex. Panel A, top: The network that describes the avoidance reflex network as presented in previous studies [5, 26] is a multi-input coherent type 1 feed-forward loop (C1-FFL) with two inputs. Synaptic weights are marked by the β values on the edges. Panel A, bottom: The signal dynamics of the avoidance reflex network. Panel B: The heatmap of the sampled data (which is the input to QCD) and the real change interval shown here by the black arc below the heatmap (black vertical line positioned in the center of the interval) as described by literature. The change interval detected by the proposed method and shown by a green arc below the heatmap (vertical lines positioned in the center of the interval), matches almost perfectly with the actual timing of these events.

In addition, the two groups of system states, before and after the change interval, were evaluated as potential meta-states. The p-values for the two groups of states are highly significant: $p = 4.28 \times 10^{-4}$ for meta-state 1 and $p = 4.28 \times 10^{-4}$ for meta-state 2. In summary, in the case of the avoidance reflex, the detected change interval is a transition between "no movement" and "moving backward" meta-states.

Results of the first three case studies, for which we used synthetic data, proved that QCD can be quite accurate. However, in practice, the data from real biological experiments can be very noisy. In order to investigate the capabilities of this approach to detect the

correct change interval from **real gene expression data**, we used datasets collected from three different experiments: yeast sporulation, fruit fly metamorphosis, and acute ethanol exposure (see Fig. 3.9, Fig. 3.10 and Fig. 3.11). All data are available in the public domain in the Gene Expression Omnibus (GEO) [13, 41]. Again, we chose different phenomena and different model organisms for a thorough method evaluation.

3.3.4 Baker's yeast sporulation

Starvation for nitrogen and carbon sources (high stress) induces meiosis and spore formation in diploid yeast (*S. cerevisiae*) cells. Stress-tolerant haploid spores are formed through cell division (meiosis) within the mother cell. This is a qualitative and obvious physiological change in yeast cells adapting to their environment. The sporulation process has been thoroughly studied and is well understood [28], which makes it a good candidate on which to validate QCD.

We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database as a source for the biological networks describing the studied phenomena. The regulation of autophagy pathway (KEGG ID: sce04140) describes the phenomena involved in sporulation. This pathway consists of mechanisms involved in processing internal and external stresses including nutrient availability. As a result, regulation of autophagy is essential for survival because it is used to maintain important cellular functions when environmental conditions change.

The QCD method was applied on the regulation of autophagy pathway and gene expression data from the yeast sporulation study by Chu et al. (GSE27, [28]). Panel A in Fig. 3.9 shows this pathway, as well as the genes measured in this experiment, marked in red. The experiment spanned 11.5 hours and data were collected at seven unequally spaced time points (0, 0.5, 2, 5, 7, 9, and 11.5 hours). The experiment was designed such that the sampling captures all known stages of the biological process. Sporulation is divided into four major stages: early, middle, mid-late, and late [28].

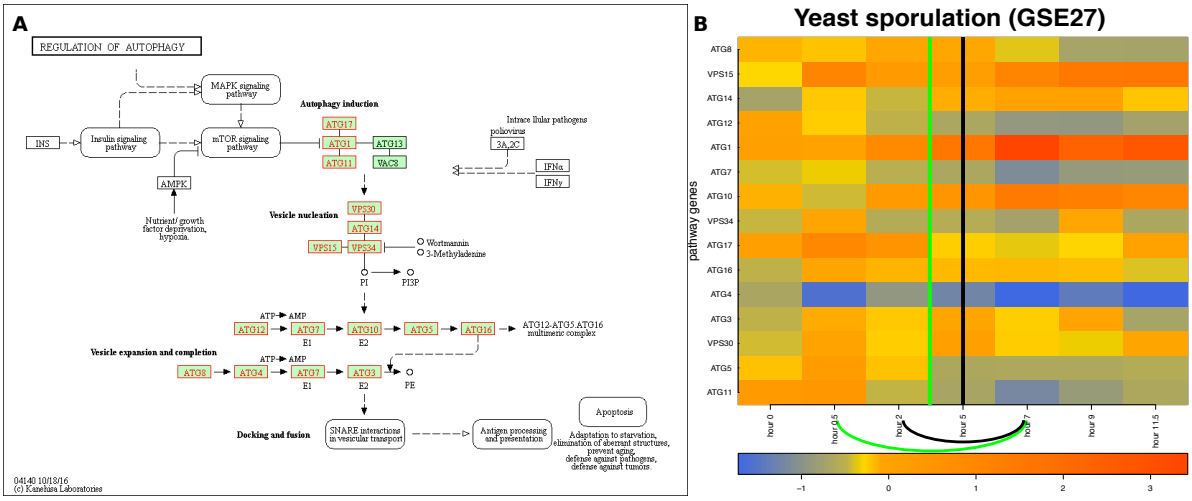

Figure 3.9: The input and results for QCD for yeast sporulation. The input is the regulation of autophagy pathway from KEGG (sce04140), in Panel A, and gene expression data from the GEO dataset GSE27, in Panel B. The data captures the sporulation phenomenon, specifically the transition from diploid cells through meiosis to the spore cells. Panel B shows the heatmap of the time course (0 to 11.5 hours) for the measured KEGG pathway genes (in red), with the change interval detected for the phenomenon (green arc and the green vertical line in the center of the interval (0.5 – 7h)), as well as the real change interval (black arc and the black vertical line in the center of the interval (2 – 7h)).

The commitment to sporulation starts in the middle stage (2 – 5h) and spans the mid-late stage (meiosis II phase, 5 – 7h) [28]. Therefore, the true change interval for this phenomenon is 2h to 7h. As observed by Chu et al., the transition phase ends after the mid-late stage. This study also showed that one of the first discernible steps of spore morphogenesis occurs after the meiosis II spindles are formed, which makes the late phase a stable one. Also, the middle-late phase is still part of the change interval as previous studies reported that the middle-late phase includes the major cytological events of sporulation [103, 114]. Panel B of Fig. 3.9 displays the measured changes of the genes on the regulation of autophagy pathway over the time course noted above.

In this case, QCD identifies a qualitative change in the interval from 0.5h to 7h, which includes the real change interval (2h to 7h) and starts one time point earlier. The change interval is the transition that separates the two potential meta-states (active state

and spore state). The active and spore potential meta-states have p-values of $p = 0.062$ and $p = 0.0195$, respectively.

3.3.5 Fruit fly metamorphosis

Three major states — egg, larva and pupa — occur during the development of the fruit fly. The larvae typically pass through three molting stages (instars) during which they shed various body elements and form new ones. Importantly, the third molting stage the larvae pupate and become adults, which marks the completion of the metamorphosis process.

The QCD method was applied on the Hedgehog signaling pathway from KEGG (pathway ID: dme 04340) and data publicly available for the metamorphosis of *D. melanogaster* (GSE3057, [15]). The Hedgehog signaling pathway, named after the signaling molecule Hedgehog (Hh), has a crucial role in organizing the body plan for the fruit fly during development. Panel A in Fig. 3.10 shows this pathway as well as the genes measured in the metamorphosis experiment (in red in this figure). The experiment started 18 hours before pupariation, spanned 30 hours, and was sampled at nine time points, two prior to pupariation (-18 hours and -4 hours), and the other seven time points equally spaced over 12 hours after the actual pupariation (0h, 2h, 4h, 6h, 8h, 10h, 12h).

Panel B of Fig. 3.10 shows the measured changes of the genes on this pathway over the time course described above. Puparium formation is triggered at the end of the third instar larvae stage that occurred during this experiment in the interval from -4 hours to 0 hours, and is marked by a high peak of the steroid hormone 20-hydroxyecdysone [15]. A second peak of the steroid hormone 20-hydroxyecdysone occurs roughly at the 10-hour time point and triggers the transformation from prepupa to pupa [15]. Puparium formation represents the onset of metamorphosis; therefore, the real change interval for this case study is indeed from -4 hours to 0 hours. The QCD method identifies one change interval from -18 hours to 0 hours. Notably, the third instar larvae stage, which starts 24 hours before pupariation and lasts until 0 hours (prepupae phase starts), is not a stable state in which the organism

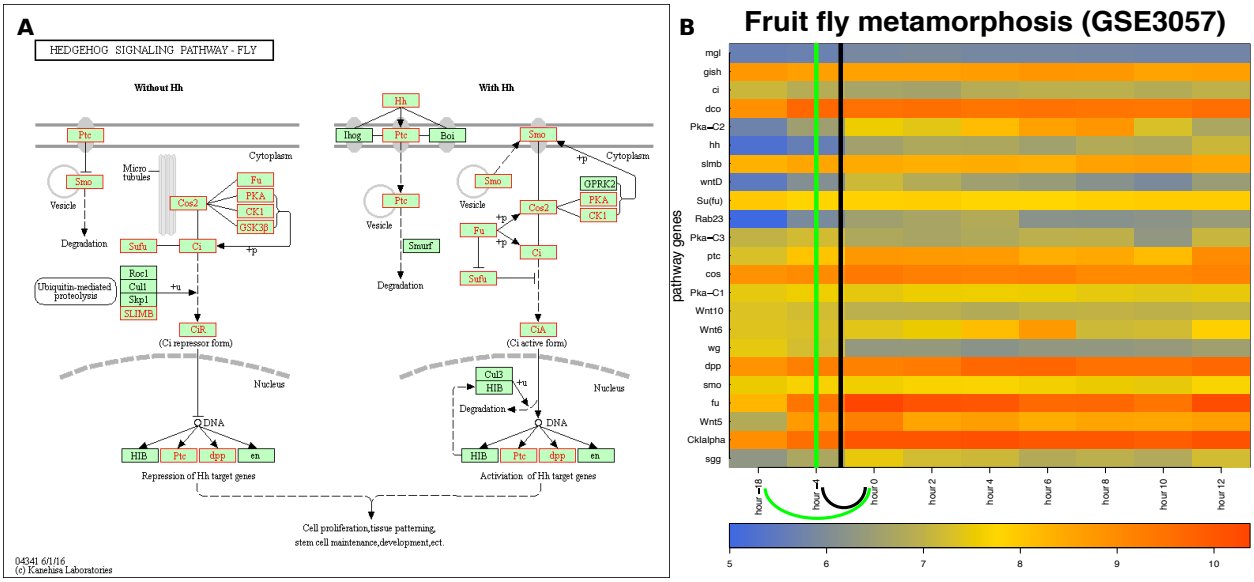

Figure 3.10: The input and results for QCD on fruit fly metamorphosis (pupariation). The input is the Hedgehog pathway from KEGG (dme04340), in Panel A, and gene expression data from the GEO dataset GSE3057, in Panel B. The data captures the pupariation phenomenon, specifically transition from the end of the larva stage through the prepupa stage and to the beginning of the pupa stage of the fruit fly. Panel B shows the heatmap of the time course (-18 to 12 hours) for the measured KEGG pathway genes (in red), with the change interval detected for the phenomenon (green arc and the green vertical line in the center of the interval (-18 – 0h)), as well as the real change interval (black arc and the black vertical line in the center of the interval (-4 – 0h)).

(fruit fly) exists. Therefore, the QCD not only correctly identifies the qualitative transition from larva to pupa, but it also shows the organism is in a continuous transition during the third instar larvae stage. The second change in this experiment (prepupa to pupa) arguably perturbs the system less than the first one since both prepupa and pupa are part of the pupal stage.

Notably, in this case study the change takes place at the beginning of the time course. To determine potential-meta-states relative to this change interval, we selected the only state before the change interval (-18h) as the potential meta-state 1 and all states after the change interval (0h–12h) as potential meta-state 2. These two meta-states are characterized by highly significant p-values: $p = 7.81 \times 10^{-3}$ and $p = 3.73 \times 10^{-9}$, respectively.

3.3.6 Fruit fly acute ethanol exposure

The fruit fly has been used as a model to study drug addiction. In the fruit fly, drug addiction produces physiological effects similar to those observed in mammals because the cellular neuronal mechanism that mediate the signals from the chemical compounds found in these drugs is conserved across these species.

To apply the QCD method, we used the Hedgehog signaling pathway (KEGG ID: dme04340) and the acute ethanol exposure data available from GEO (GSE18208) and described by Kong et al. [88]. The Hedgehog signaling pathway was chosen for its capability to model major mechanisms involved in fruit fly development, including its adaptive mechanisms. Panel A in Fig. 3.11 displays this pathway, as well as the genes measured in this

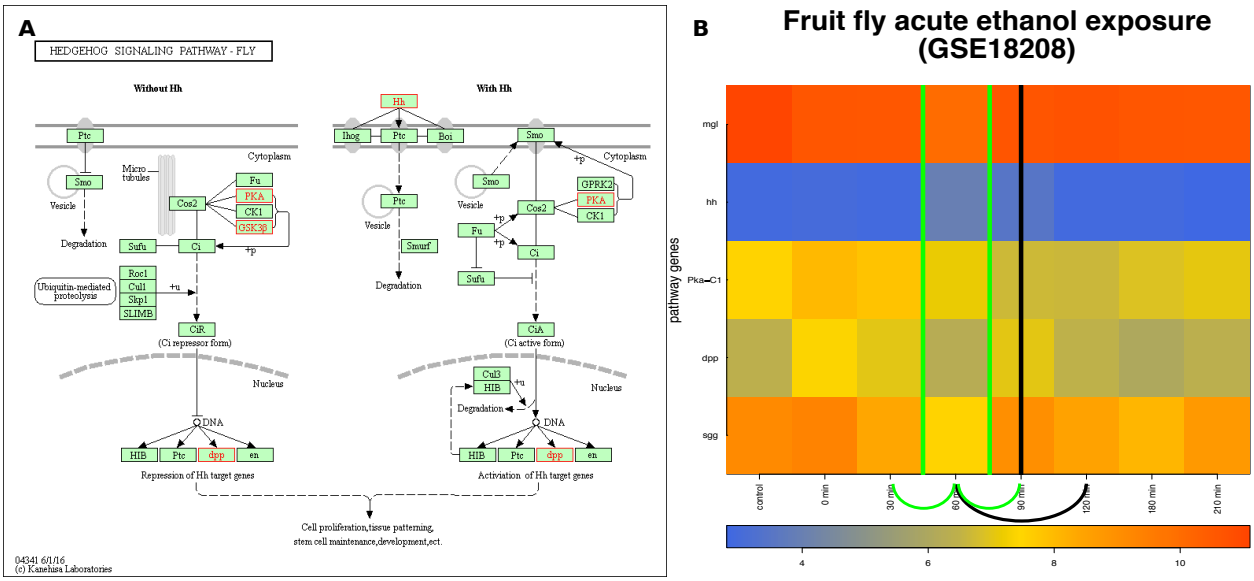

Figure 3.11: The input and results for QCD on fruit fly ethanol exposure. The input is the Hedgehog pathway from KEGG (dme04340) in Panel A, and gene expression data from GEO GSE18208, in Panel B. The data captures the acute ethanol exposure phenomenon, specifically transition from the "sober" stage through the "'drunk" stage and back to the "sober" stage. Panel B shows the heatmap of the time course (control, 0 to 3.5 hours) for the measured KEGG pathway genes (in red), with the change interval detected for the phenomenon (green arc and the green line in the center of the intervals (0.5 – 1h) and (1 – 1.5h), as well as the real change interval (black arc with a black line in the center of the interval (1 – 2h)).

experiment, marked in red. Panel B of Fig. 3.11 shows the measured changes of the genes on this pathway over the time course from the biological experiment. The experiment spanned 3.5 hours (210 minutes) of recovery after a 30-minute ethanol exposure, sedating up to 75% of the flies. Samples were taken at eight time points. The time points include one control, before exposure, one at 0 hour, right after exposure and every 30 minutes after that up to 3.5 hours; the missing data point at 2.5 hours (150 min) was not provided in the dataset. This experiment's treatment conditions included exposure to humidified air or ethanol vapor (60%) for 30 minutes, and then recovery for up to 210 minutes [88]. The recovery period from ethanol sedation has been reported by another study to be approximately between 40 minutes and 2 hours [144], which is the real change interval. Based on this recovery time, by the end of this experiment (210 minutes), the fruit flies should recover from the effects of ethanol exposure. In the GSE18208 dataset 40 minutes was not one of the sampled time points; therefore, to mark the real change interval, we used the very next time point available in the dataset, the one-hour time point.

The intuitive physiological transitions expected for these data are from no exposure (sober) to exposure to ethanol (drunk) and back to fully recovered (sober). However, the drunken state is temporary, since it is followed by recovery. Because of this transition, we expected two change intervals, from sober to drunk and from drunk to sober. Furthermore, the initial and end states (sober before exposure and sober after recovery) were expected to be very similar from a gene expression point of view. In other words, the sober state is the same in the initial and final state in this case, as opposed to the flagellum building case where the initial and final states, with and without flagellum, are obviously different.

The ethanol exposure has a delayed effect at the gene level. According to Kong et al., the expression of immunity genes increased after ethanol exposure in the time range from 0.5 hours to 1.5 hours [88]. Because of this delayed effect, we did not expect the biggest changes between the control and 0 hours but rather between the control and some later time point(s).

The QCD results on these data have shown that the biological system indeed goes through two qualitative changes, and the change intervals are: 0.5 hours to 1 hour and 1 hour to 1.5 hours, matching the expected transitions from a sober state to a drunken state and then back to the sober state. The effects of the ethanol exposure appear to peak at the 1-hour time point. Based on the change intervals and the return of the system to its initial state, there are two groups of states that may form meta-states. These potential meta-states consist of the following time points: control, 0 hour, 0.5 hours, and 1.5 hours to 3.5 hours, for meta-state 1, and the 1-hour time point for meta-state 2. The distribution of the significant and non-significant transitions yielded a highly significant p-value, $p = 1.37 \times 10^{-5}$, for meta-state 1, but a non-significant p-value ($p = 0.22$) for meta-state 2. This result is probably due to the small number of comparisons involving the single time point included in meta-state 2.

3.3.7 Mouse exposure to phosgene

Carbonyl chloride (phosgene) is a toxic compound used for the production of materials such as plastics and rubber. Exposure to carbonyl chloride produces irreversible lung injury and potentially life-threatening pulmonary edema that manifest within a day. Early intervention, within one hour of exposure, has been reported to be effective for the treatment of carbonyl chloride exposure [141]. However, the damage inflicted by exposure to carbonyl chloride is progressive with the most significant physiological effects reported to occur between four and 12 hours after exposure [142]. Due to the high toxicity of carbonyl chloride, the organism will not return to its pre-exposure like state, yet it will be in a different state (injured or most likely lethally injured) state at 72 hours after exposure. In mice, by the 12^{th} hour after exposure, a mortality rate of 50-60% was reported, which increased to 60-70% by the 24^{th} hour [143].

In the study by Sciuto et al. [143], mice were exposed to 32 mg of phosgene per cubic meter for 20 minutes and samples were collected from lung tissue at nine time points: untreated (0 hours), 30 minutes, 1 hour, and 4, 8, 12, 24, 48, and 72 hours after exposure.

We applied QCD to study this phenomenon, using these data, and the chemokine signaling pathway from KEGG (mmu04062) as the network/map of the biological system. The chemokine signaling pathway was chosen because it describes the signaling mechanisms of an inflammatory response and such mechanisms are intimately involved in the response to the exposure to a toxic gas.

Panel A in Fig. 3.12 shows the chemokine signaling pathway as well as the measured genes marked in red. Panel B of Fig. 3.12 shows the measured changes of the genes on this pathway over the time course of the biological experiment. The QCD method identified one qualitative change in the interval of 0.5 hours to 1 hour which corresponds to the time interval for the initiation of latent effects of the toxic gas exposure. In other words, QCD identified an interval during which damage is treatable [141].

The change interval identified by QCD was then used to group the states (time points) before (0 hour to 0.5 hours) and after (1 hour to 72 hours) into potential meta-

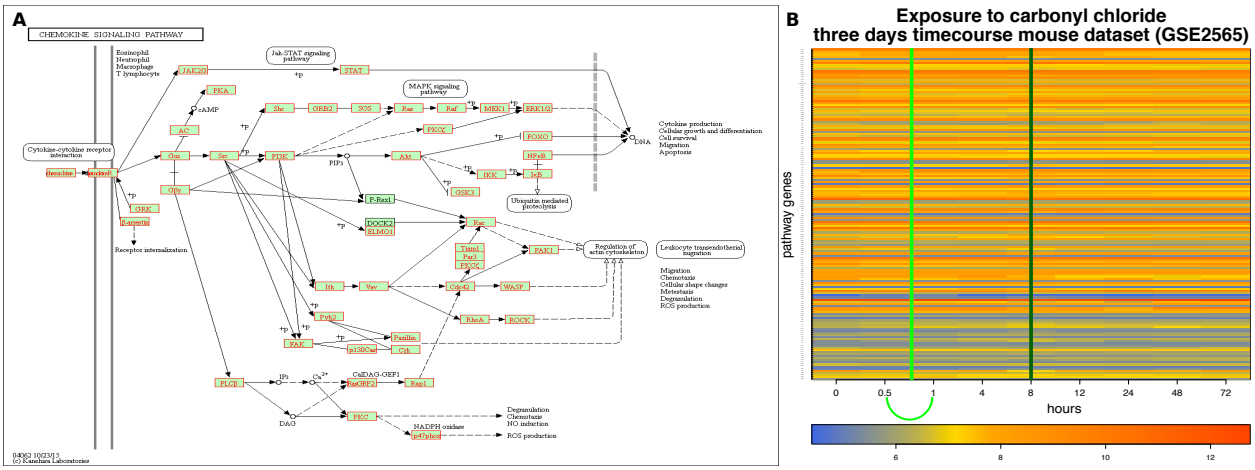

Figure 3.12: The input and results for QCD on mouse toxic gas exposure. The input is the chemokine signaling pathway from KEGG (mmu04062), in Panel A, and gene expression data from GEO GSE2565, in Panel B. The data captures the exposure to phosgene phenomenon, specifically the transition from not exposed through the progressive damage of the exposure up to lethality. Panel B shows the heatmap of the time course (0 to 72 hours) for the measured KEGG pathway genes (in red), with the change interval detected for the phenomenon (green arc and the green line in the center of the interval (0.5 – 1h)). The dark green vertical line (8h) marks the pre-disease state detected by the DNBM method).

states. These groups of states can be labeled as "before exposure" and "long-lasting damage" based on the organism's physiology for the respective states. The groups of states were then evaluated for statistical significance and the p-value of the "before exposure" group was not significant ($p = 0.696$), while the "long-lasting damage" group had a highly significant p-value ($p = 9.39 \times 10^{-4}$), which makes it a true meta-state. This result may suggest that the control (no exposure) and the 30-minute after exposure system states are not similar enough to yield a statistically significant meta-state. This is very likely to be true since there is still damage inflicted at the 30-minute time point even though it is still treatable.

We compared the results of QCD in this case to the results of an existing method developed to detect network biomarkers and the pre-disease state (DNBM) [92]. The DNBM takes as input both the high-throughput data and the large network of protein-protein interactions for the organism under study. The output of DNBM is a pre-disease state in the form of a sample or list of samples from the data. The hypothesis is that a subset of the large network, termed the leading network, is the first to change toward the disease state, which makes its components and structure causally related with the disease. The DMBM models the change in gene expression over time as a Markov process. Then, a state-transition-based local network entropy (SNE) is used as a general, early measure of upcoming transitions by estimating the resilience of the network. The SNE is a Shannon-type entropy [145], intended to quantify the change in state for the biological network.

Notably, the DNBM identifies one single (pre-disease) state prior to the onset of disease, while the proposed QCD identifies a change interval of transition to disease, which can be much more informative regarding the disease evolution, as well as providing an opportunity for therapeutic intervention. In addition, in the case of the QCD, the impact analysis approach may provide a better evaluation of the system's impact than the network entropy. At the same time, a reinforcement of the impact by comparing every two time points may provide a better approximation of the change onset. Therefore, evaluating the systemic change between every two time points results in the early-detection property.

3.3.8 Human hepatitis C virus infection progression to liver cancer

Hepatocellular carcinoma (HCC) is a common liver cancer that can be the result of an infection with the hepatitis C virus (HCV). The progression from HCV infection spans multiple disease stages before reaching HCC, as reported by Wurmbach et al. [179]. We used the data from this study to identify qualitative changes for this phenomenon. The dataset (GSE6764, [179]) contains gene expression collected from 75 samples (48 patients) and covers eight progressive stages of HCV induced HCC: four no-cancer stages including no HCV/control, cirrhosis, low-grade dysplastic, and high-grade dysplastic, and four cancer stages including very early HCC, early HCC, advanced HCC, and very advanced HCC. Normal liver control is used as the initial stage and stages are ordered by disease progression.

To apply QCD on these data, we used the viral carcinogenesis pathway from KEGG (hsa05203) as the network/map of the biological system. The viral carcinogenesis pathway describes the signaling mechanisms involved in inflammatory responses such as the one triggered by HCV. Panel A in Fig. 3.13 shows this pathway as well as the genes measured in this experiment marked in red. Panel B of Fig. 3.13 shows the measured changes of the genes on this pathway over the different disease stages from the biological experiment.

From these data, the QCD identified one qualitative change (change interval) from stage zero (control), a benign state to stage three (high-grade dysplastic), the last of the four benign states and a state in which treatments are effective. The group of states before the change interval was considered as potential meta-state one (MS1) and contains only the control state. The group of states after the change interval was considered as potential meta-state two (MS2) and contains five states: high grade dysplastic nodules, very early HCC, early HCC, advanced HCC, and very advanced HCC. In essence, the analysis identified the transition from the benign state (first meta-state) to the cancerous state (second meta-state). The p-values of these meta-states were $p = 0.031$ for MS1 and $p = 3.05 \times 10^{-5}$ for MS2.

For this case study, the DNBM detected the pre-disease state at the fifth stage, very early HCC, which is the first malignant stage. The existent DNBM detected the start of

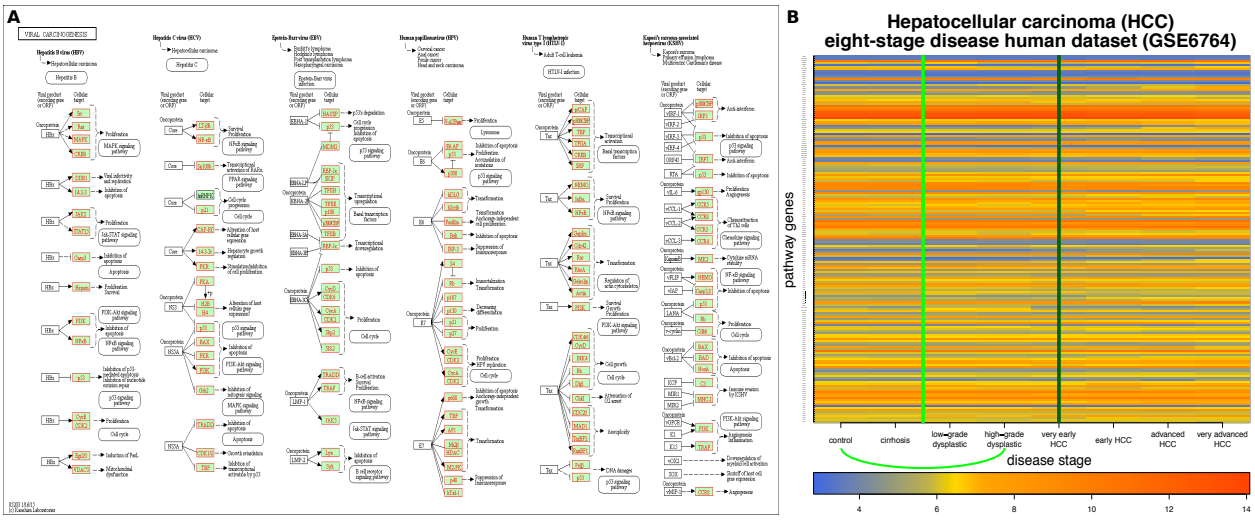

Figure 3.13: The input and results for QCD on human hepatitis C virus (HCV) to hepatocellular carcinoma (HCC) progression. The input is the viral carcinogenesis pathway from KEGG (hsa05203), in Panel A, and gene expression data from GEO GSE6764, in Panel B. The data captures the progression from human HCV to HCC, specifically the transition from control (healthy) through the progressive stages of liver damage up very advanced HCC. Panel B shows the heatmap of the disease progression (control to very advanced HCC) for the measured KEGG pathway genes (in red), with the change interval detected for the phenomenon (green arc and the green line in the center of the interval (control – high-grade dysplastic nodules)). The dark green vertical line (very early HCC) marks the pre-disease state detected by the DNBM method).

the malignant state while our proposed QCD method detected the transition from benign to malignant. These results (see Tables 3.2 and 3.3) show the applicability of this method in developing preventive therapies. Identifying the genes that change within the change interval could lead to the identification of very early markers for disease and potential targets for disease prevention. A detailed description of the results of the QCD analysis at each step of the analysis workflow for all eight datasets is included in section 3.5.1.

In the case of disease progression, once a change interval is identified one should start the therapeutic intervention as early as possible within the change interval. For example, in the case of the HCV to HCC progression that could be any time up to the high-grade dysplastic stage.

Results summary					
Organism	**Phenomena**	**Data source**	**Time points (# of samples)**	**Detected change interval**	**Real change interval**
E. coli	Flagellum building	mathematical model	0–600min (21) equally spaced	180–300min	240–270min
B. subtilis	Sporulation	mathematical model	0–600min (21) equally spaced	210–240min	210–240min
Worm	Avoidance reflex	mathematical model	1–8 ms (8) equally spaced	4–5ms	3–5ms
Yeast	Sporulation	GSE27	0–11.5h (7)	0.5–7h	2–7h
Fruit fly	Pupariation	GSE3057	-18–12h (9)	-18–0h	-4–0h
	Acute ethanol exposure	GSE18208	untreated 0h treated 0–3.5h (8)	0.5–1h & 1–1.5h	1–2h
Mouse	Phosgene exposure	GSE2565	0–72h (9)	0.5–1h	8h *
Human	HCV induced HCC progression	GSE6764	8 stages (8)	control–high-grade dysplastic	very early HCC *

Table 3.2: Summary of the results for the analysis of synthetic (simulation) and real data for various phenomena in model organisms. From left to right, the columns of the table show the organism, the phenomenon studied, the data source (simulation or GEO dataset), the duration of the simulation or experiment, the number of measurements, the time interval reported by the algorithm as including a qualitative change, and the actual time interval in which the phenomenon was simulated (first three rows) or actually took place (next five rows). * denotes the results of the existing method [92].

Meta-states results summary		
	p-value Meta-state I	**p-value Meta-state II**
E coli flagellum building	5.44×10^{-19}	3.61×10^{-28}
B subtilis sporulation	2.31×10^{-19}	6.23×10^{-32}
Worm avoidance reflex	4.28×10^{-4}	4.28×10^{-4}
Yeast sporulation	0.062	0.019
Fruit fly pupariation	7.81×10^{-3}	3.73×10^{-9}
Fruit fly alcohol exposure	1.37×10^{-5}	0.227
Mouse carbonyl chloride exposure	0.696	9.39×10^{-4}
Human HCV induced HCC progression	0.031	3.05×10^{-5}

Table 3.3: Summary of the results establishing the significance of the meta-states. In each case-study two potential meta-states were identified, relative to the time-course data or sequential series of system states provided as input. Meta-state1 consists of the group of states before the change interval. Meta-state2 consists of the group of states after the change interval. For each case study and each potential meta-state, we calculate a statistic as the number of time-intervals with status consistent with the status assigned in the corresponding theoretical meta-state. Here, we show the p-values (one tail, grater) computed for this statistic as it follows a binomial distribution with a theoretical likelihood of success of 50%. From a total of 16 meta-states: 13 are significant at a threshold of 5%.

To further evaluate the potential of the proposed method to detect changes as they occur, we ran the method on data from only the first three stages of the disease progression. DQC detected a change interval from the first (control) to the third stage (low-grade dysplastic), showing that a systemic qualitative change is happening and can be detected at a very early stage, as soon as the disease process has started.

3.4 Discussion

Disease prevention and early detection are two major healthcare objectives that contribute to improving quality of life. Currently, early detection of complex diseases is achieved only after the physiological traits of the phenotype are present, when existing treatments may be ineffective. Chronic disease, a particular case of complex disease, is generally detected in the late stage of a relatively slow, progressive process. Representative examples that affect a large number of people are heart disease, cancer, and neurodegenerative disorders. It is a real challenge for people with these diseases to maintain a good quality of life after diagnosis. Understanding when the transition to disease occurs is a good first step towards interrupting the process and maintaining the healthy state.

To maintain the healthy state, one needs to monitor the biological system and measure the gene expression or any parameters the system has in order to assess how much the system is changing. The moment a qualitative change occurs, either cumulative or sudden, a change interval emerges. For instance, in the case of the eight stages of HCC, a qualitative change occurs from control to high-grade dysplasia. A cirrhotic liver is characterized by the presence of scar tissue due to long-term damage. In an attempt to replace the damaged cells in the cirrhotic liver, clusters of newly formed cells can occur in the scar tissue. Dysplastic (abnormally grown) nodules found in the liver are typically identified in cirrhotic livers. Low-grade dysplastic nodules (LGDN) cells are larger than the normal liver cells [170]. High-grade dysplastic nodules (HGDN) cells are smaller than the normal liver cells and have a greater nucleus-to-cytoplasm-size ratio [170]. The difference between HGDNs and very early HCC is

the stromal invasion present in the latter [140]. A study on the LGDNs and HGDNs in HCC development concluded that LGDNs together with large regenerative nodules, should be monitored with ultrasound, while HGDNs should be preventively treated due to their high malignant risk [24]. Taken together, these data support the qualitative change identified by QCD from a low malignant risk stage of the liver disease to a high risk stage and close precursor to the malignant stage of very early HCC.

To further investigate the results of our analysis in the case of HCC progression, we identified the differentially expressed (DE) genes (absolute $\log_2$ fold change greater than 1) when comparing the control to high-grade dysplasia and the control to very advanced HCC. The total number of measured genes is 20,156. In the control versus high-grade dysplasia comparison, there are 149 DE genes, while in the control versus very advanced HCC comparison, there are 1,355 DE genes, which is almost an order of magnitude higher. This suggests that using the differentially expressed genes across the change interval, as opposed to the genes that differ between the control and very advanced HCC, offers a more focused analysis. In essence, the comparison across the narrowest change interval targets the genes involved in the initial tumor formation, rather than all genes that change as a consequence of the cancer.

The number of common DE genes among the two comparisons is 80, representing 53% of the initial 149 genes. We downloaded the curated list of cancer genes available in the cancer gene census [53] (accessible at: http://cancer.sanger.ac.uk/census). This list is presented together with the catalogue of somatic mutations in cancer (COSMIC) [51] (accessible at: http://cancer.sanger.ac.uk/cosmic). We used this list of cancer genes to filter the 80 common genes to obtain a cancer gene set. The result consists of two genes: *CHEK2* and *FAT1* (see section 3.5.2 for the expression profile). These genes are highly relevant to the condition under study considering *CHEK2* mutations have been linked to various cancers [160, 37]; it has also been shown to be a mediator of a tumorigenic mechanism in

HCC [115]. Furthermore, *FAT1* has been shown to have an oncogenic role in HCC [126, 161], and it has been identified as a biomarker in multiple cancers [32, 168].

The viral carcinogenesis pathway from KEGG was used to identify the change interval for the HCV-induced HCC progression. We also used this pathway to filter the 80 common genes and to obtain a "viral carcinogenesis" gene set, which contains genes from the pathway that change at the onset of the disease. The result consists of two early growth response genes: *EGR2* and *EGR3* (see section 3.5.2 for the expression profile). *EGR2* has been shown to be an apoptosis promoter gene [159], which is downregulated by miRNAs in cancer [178, 94]. *EGR3* has been shown to be involved in a number of cancers and in the regulation of the immune response [77, 136, 129, 27], and this gene has recently been linked to HCC when it was used to inhibit the growth of tumor cells [187].

3.5 Overall assessment of the QCD results

3.5.1 QCD analysis results and the corresponding meta-states

The workflow of the QCD analysis consists of the following three steps: (i) compare the status of the system between each pair of time points using a pathway impact analysis [38, 156, 166, 165] and assess the levels of perturbation by computing a system perturbation factor; (ii) separate large and small inter-state perturbations using a gamma mixture model fitted to the system perturbation by an expectation maximization algorithm; (iii) calculate the change interval(s) as the narrowest disjunct interval(s) of large changes. Figures 3.14, 3.15, 3.16, 3.17, 3.18, 3.19, 3.20, 3.21 show the results at each of these steps when applying QCD on 8 case studies.

Assessing the goodness of fit

After the second step (fitting the gamma mixture model), we compute several statistics to evaluate the goodness of fit between the observed perturbation factors and the fitted mixture model. In Panels A in Figures 3.14, 3.15, 3.16, 3.17, 3.18, 3.19, 3.20, 3.21, KLD is the Kullback-Leibler divergence computed between the density of the observed perturbation

and the density of the fitted mixture model. The KLD is a non symmetrical measure, therefore we compute both the KLD_o, o-observed first, and the KLD_f, f-fitted first, values. Values closer to 0 indicate higher similarity between the distributions. Our results show 7 out of 8 cases with KLDs less than 0.2. We used the KLD method from the R package LaplacesDemon 16.0.1 to compute this measure.

The KS_p is the p-value of the Kolmogorov-Smirnov test between the observed perturbation values and a sample of 10,000 values from the fitted mixture model. The null hypothesis in this case is that the samples come from the same distribution. A high p-value (close to 1), tells us that there is no evidence that the two distributions are significantly different. For instance, in the *E. coli* case, the p-value of 0.42 indicates that there is no significant difference between the observed distribution and the fitted one. Typical thresholds for rejecting the null hypothesis are 0.01, 0.05 and 0.1, and our results show p-values higher than 0.39, with 6 out of 8 values higher than 0.8. The ks.test method from the R package stats 3.1.2 was used to compute these p-values.

Another measure of the goodness of fit is the ratio between the intersection and union of the areas delimited by the observed and fitted density lines. We refer to this ratio as the overlap between the observed and fitted distributions. An overlap of 100% would mean a perfect match. The minimum overlap value on our case studies is 68.97% and the maximum is 89.5%.

In Panels A in Figures 3.14, 3.15, 3.16, 3.17, 3.18, 3.19, 3.20, 3.21, Type 1 and Type 2 errors are computed under the null hypothesis when there are no change intervals and all system perturbations are small system perturbations. The lower the type 1 and type 2 errors the more reliable the results. The maximum type 1 error is 0.18, and the minimum 0.012 with most of them (5/8) case studies having a less than 0.1 type 1 error. The maximum type 2 error is 0.067, and the minimum 0.009 with all case studies having a less than 0.1 type 2 error.

Identifying meta-states

The states of the system before and after a change interval should be analyzed to gain insight regarding the state of the system before and after a qualitative change. To describe this analysis, we will consider the situation in which there is a single change interval, as in the *E. coli* flagellum building data set. We consider that in this case the system is stable before and after the change interval. In this context, we group the states in which the system is stable into meta-states. We define a meta-state (see Fig. 3.14, panel C) as a group of consecutive states that satisfy the following two conditions, using a system perturbation threshold previously computed: (i) all comparisons between states within a meta-state have a small system perturbation; (ii) all comparison between states from a meta-state to states outside it (excluding the states in the change interval) have a large system perturbation.

For any given change interval, we consider the groups of states before and after the change interval as potential meta-state. However, not every group of states before and after a change interval has to form a meta-state. In order to identify those groups of states that do form a meta-state from those that do not, we analyze each such group individually, from the perspective of the definition above. Given a potential meta-state, any individual comparison between two individual states can be either consistent or inconsistent with the definition above. Under the null hypothesis, in which there is no meta-state, the probability that a comparison is consistent or not should be 0.5. This a priori probability of consistent/not consistent comparison was verified by a number of simulations (100,000) with random data where the mean and median of this probability were 0.501 and 0.5013, respectively. A large number of comparisons consistent with the definition of a meta-state will constitute evidence for the existence of such meta-state. A binomial distribution can be used for each meta-state to compute a p-value from the observed number of consistent comparisons for the given meta-state. From a total of 16 meta-states (2 for each of the 8 data sets included here), 11 were significant at a threshold of 1%, two were significant at 5%, and one at the 10% significance level. This suggests that most of the time, the organism transitions from a stable state to

another stable state. Groups of states that do not form statistically significant meta-states may be due to a phenomenon that is still evolving, or simply to a low number of time points, which reduces the number of comparisons available, and therefore the statistical power of the test employed. An example of the latter situation is the ethanol exposure experiment in which the only meta-state eligible for consideration included a single time point (S3).

Figures 3.14, 3.15, 3.16, 3.17, 3.18, 3.19, 3.20, 3.21, show the potential meta-states with their ideal comparisons according to the definition (panel C in each figure), as well as the observed comparisons (panel D in each figure). Panels E and F in each figure show the meta-states considered, together with their respective p-values.

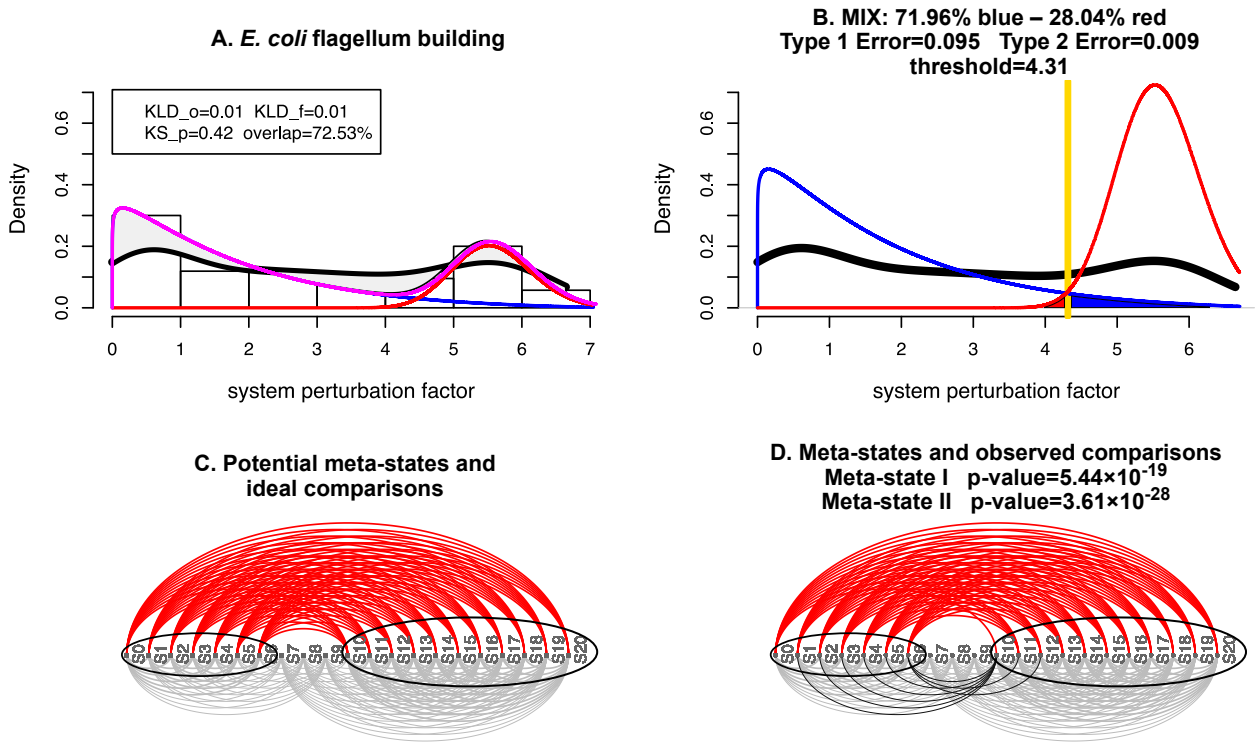

Figure 3.14: The results of QCD on synthetic data of the *E. coli* flagellum building. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case).

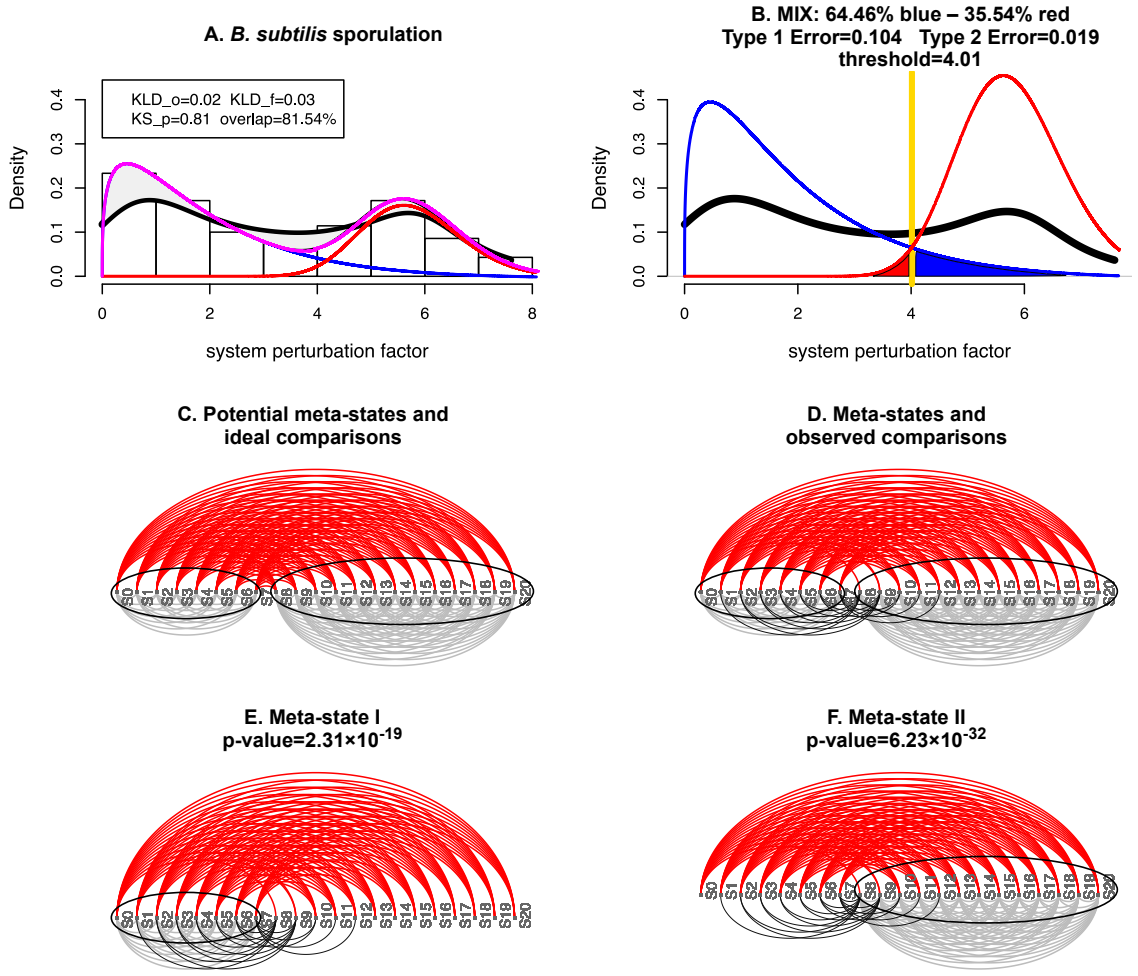

Figure 3.15: The results of QCD on synthetic data of the *B. subtilis* sporulation. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case). Panel E shows the comparisons considered for meta-state I. Panel F shows the comparisons considered of meta-state II.

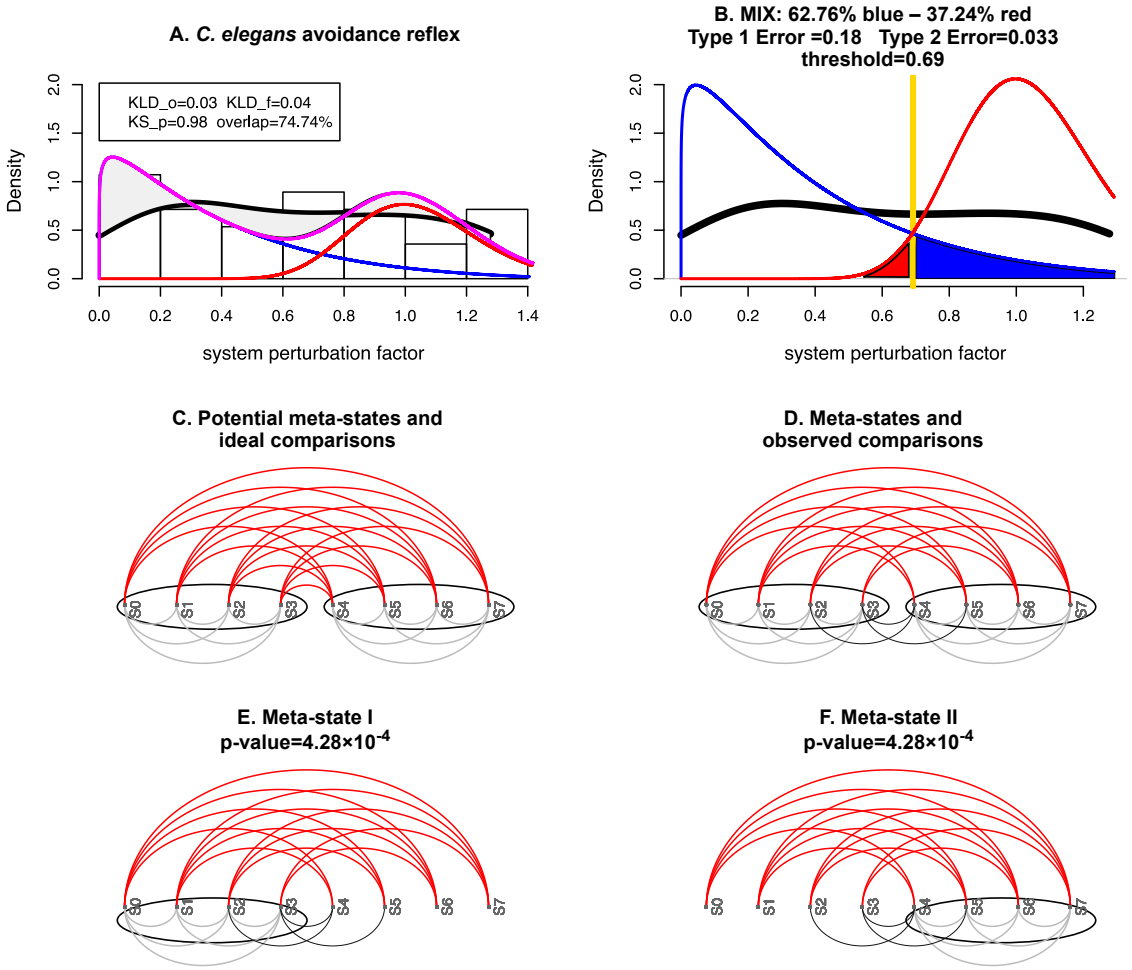

Figure 3.16: The results of QCD on synthetic data of the *C. elegans* avoidance reflex. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case). Panel E shows the comparisons considered for meta-state I. Panel F shows the comparisons considered of meta-state II.

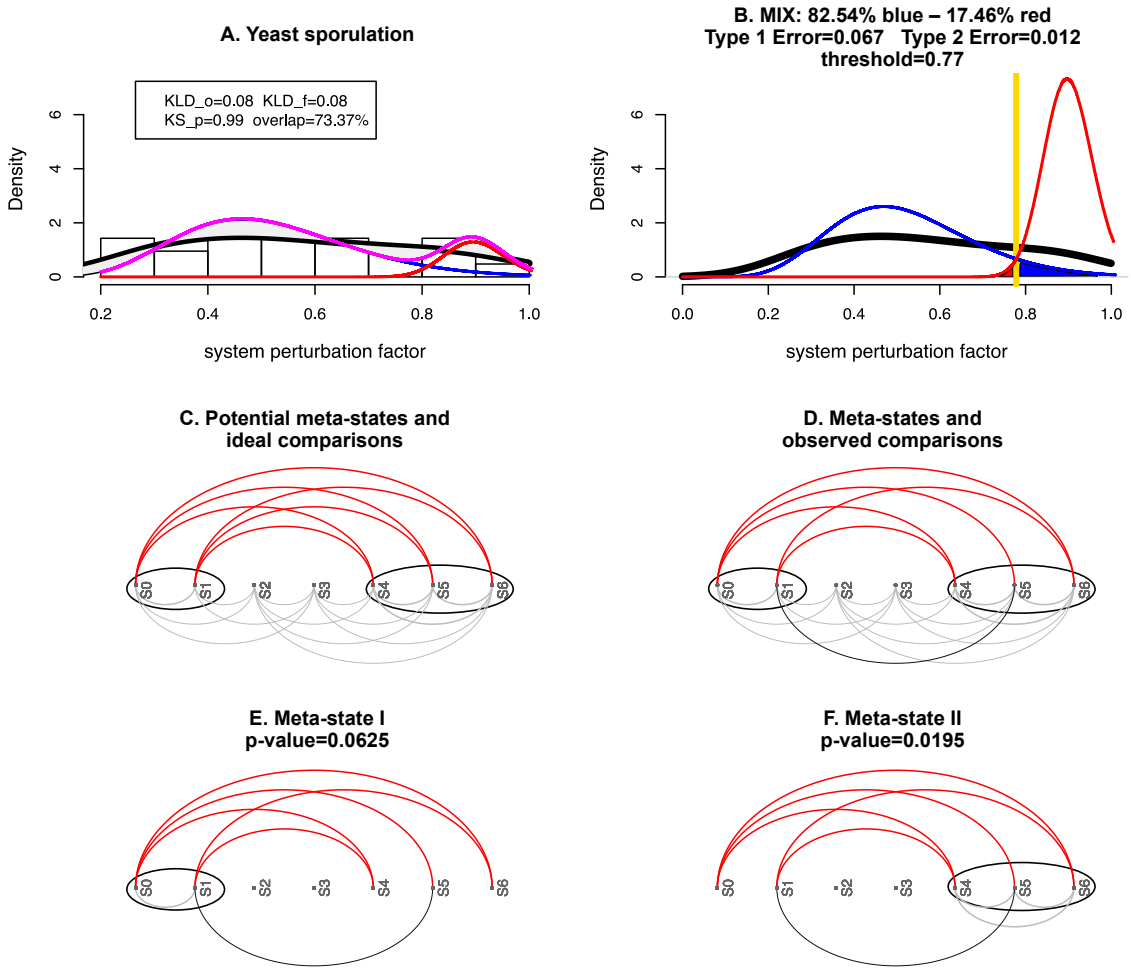

Figure 3.17: The results of QCD on real data of the yeast sporulation. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case). Panel E shows the comparisons considered for meta-state I. Panel F shows the comparisons considered of meta-state II.

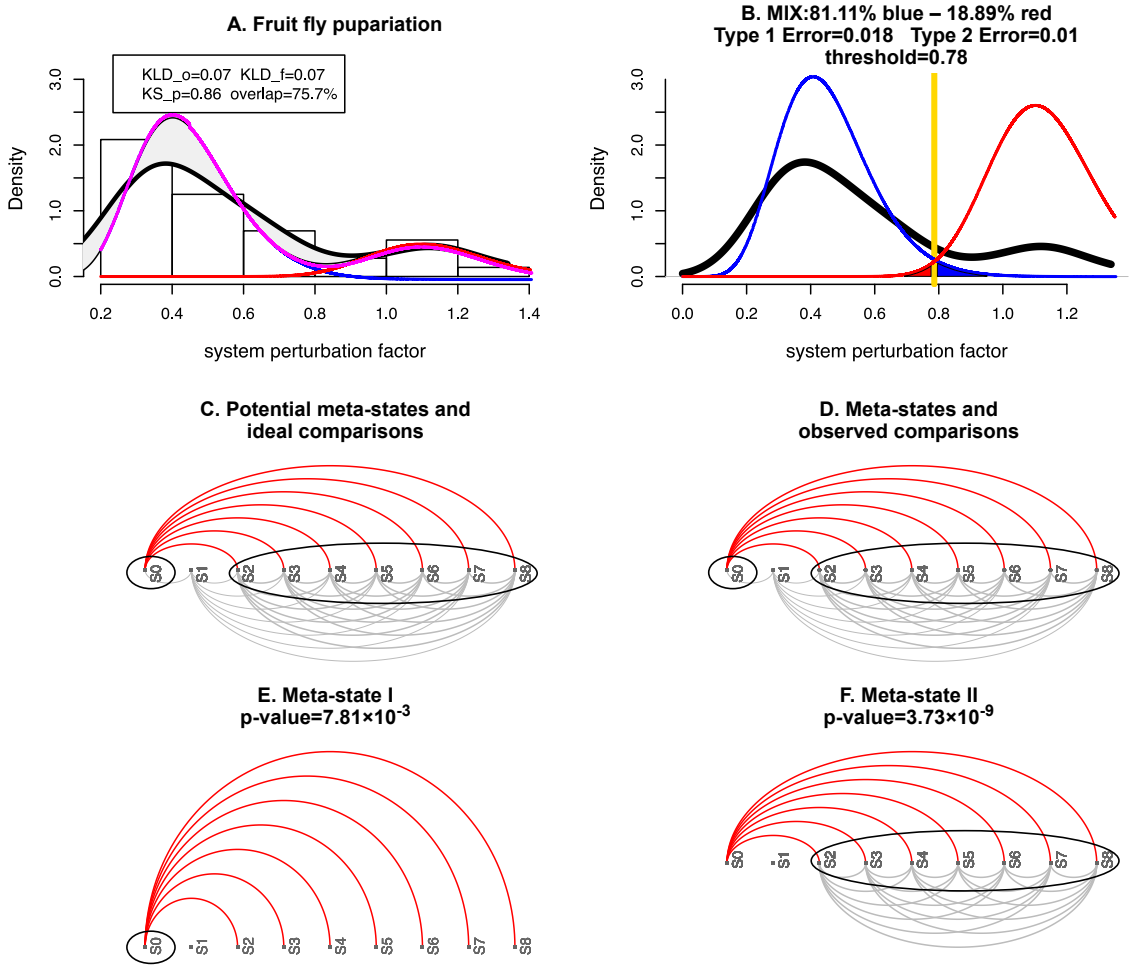

Figure 3.18: The results of QCD on real data of the fruit fly pupariation. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case). Panel E shows the comparisons considered for meta-state I. Panel F shows the comparisons considered of meta-state II.

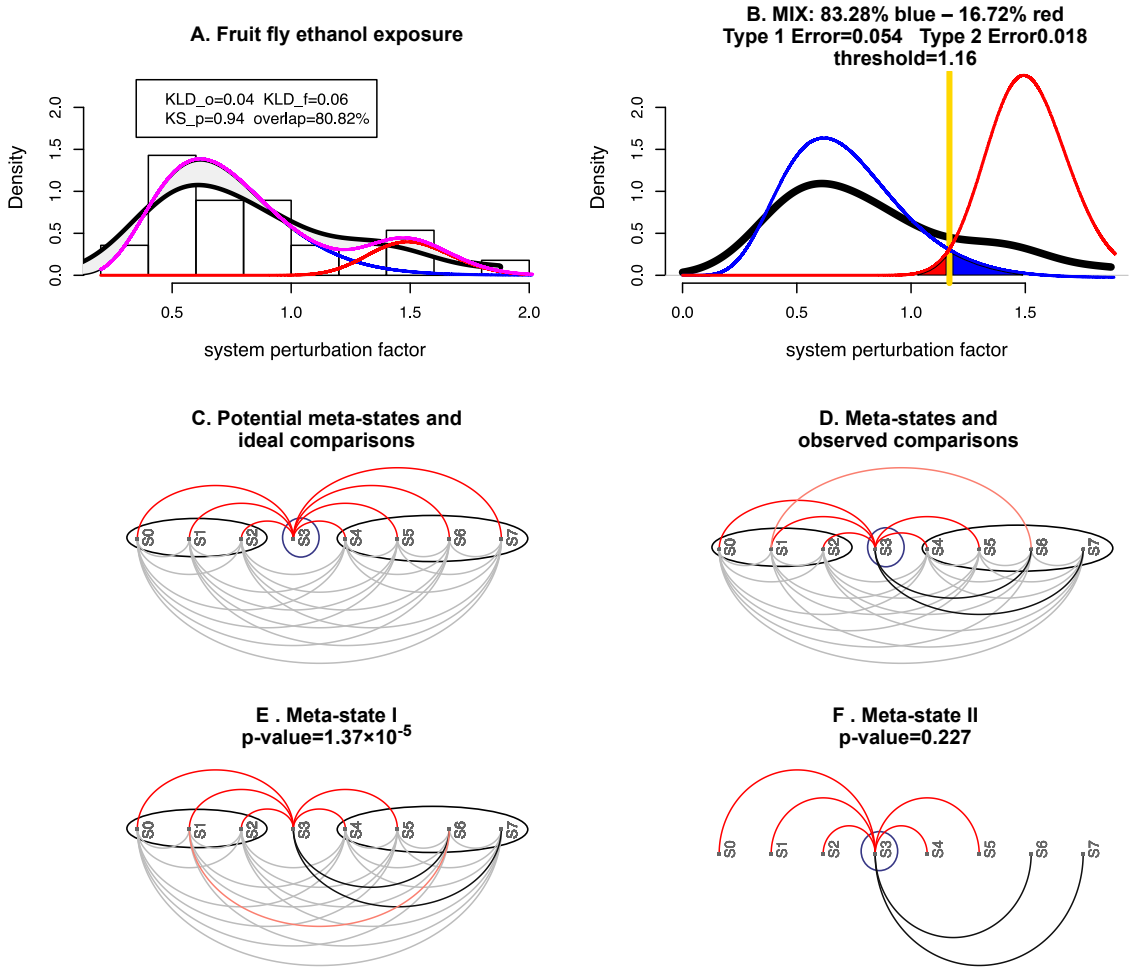

Figure 3.19: The results of QCD on real data of the fruit fly ethanol exposure. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case). Panel E shows the comparisons considered for meta-state I. Panel F shows the comparisons considered of meta-state II.

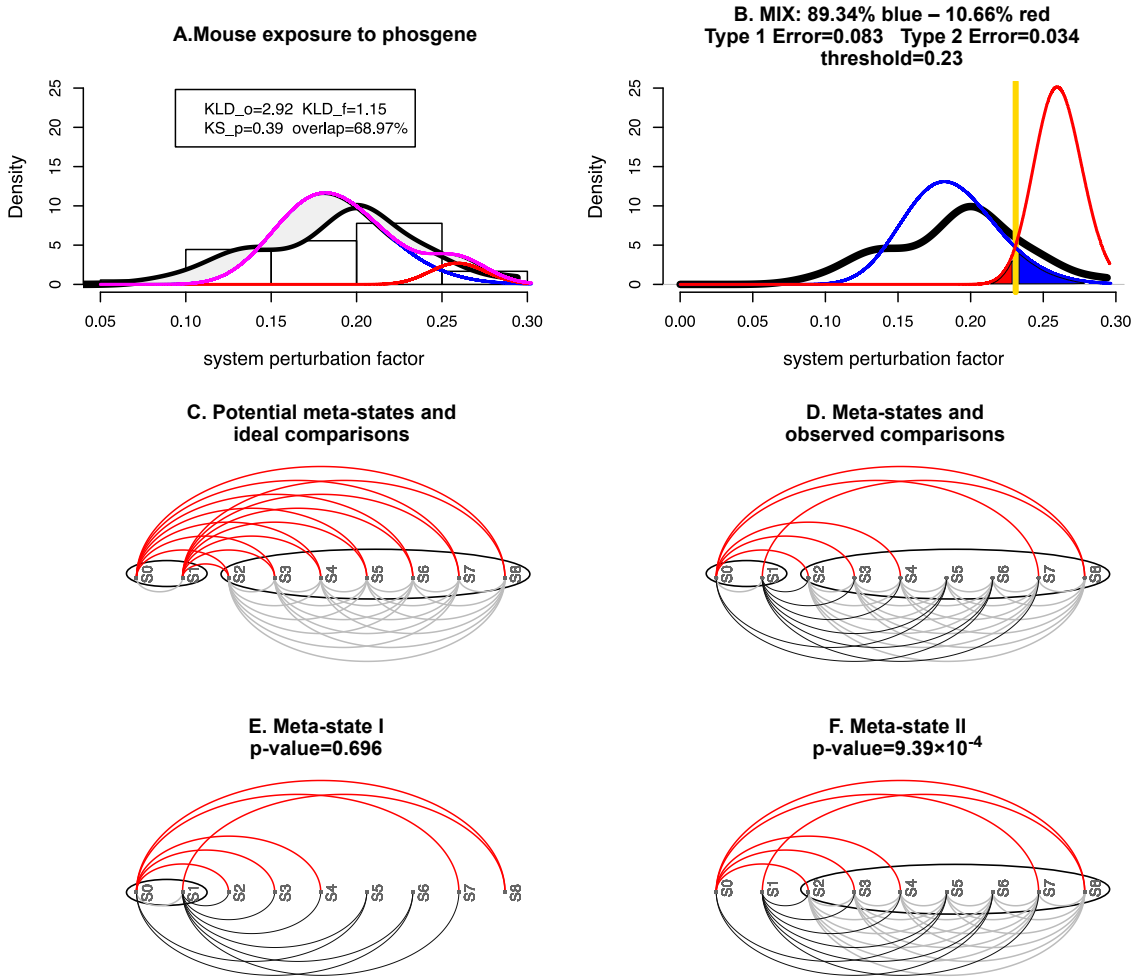

Figure 3.20: The results of QCD on real data of the mouse exposure to carbonyl chloride. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case). Panel E shows the comparisons considered for meta-state I. Panel F shows the comparisons considered of meta-state II.

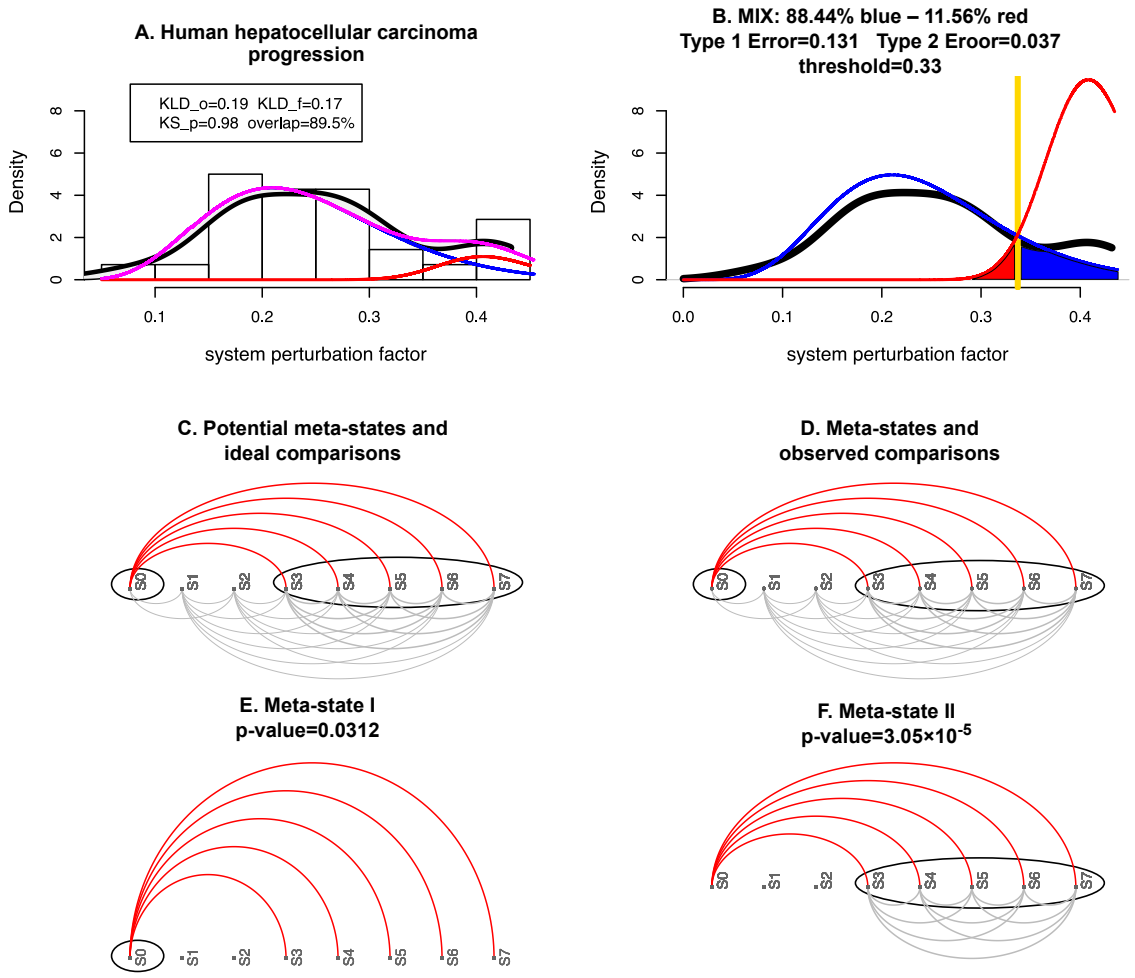

Figure 3.21: The results of QCD on real data for the human hepatitis C virus (HCV) to hepatocellular carcinoma (HCC) progression. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows the gamma mixture model used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The yellow vertical line is the threshold used to separate the small and large perturbation factors. The null hypothesis is that there are no change intervals and therefore there are only small system perturbations (blue distribution). The Type 1 and Type 2 errors are marked by the blue and red areas, respectively. Panel C shows the potential meta-states (black ellipses) together with the ideal comparisons between the time points within these meta-states (red - large perturbation, gray - small perturbation). Panel D shows the same meta-states (black ellipses) together with the observed comparisons (red - large perturbation, gray and black - small perturbation). Black comparisons are between states from different meta-states (these are red in the ideal case). Panel E shows the comparisons considered for meta-state I. Panel F shows the comparisons considered of meta-state II.

3.5.2 Results of followup analysis for HCV progression to HCC

To further investigate the results of our analysis in the case of HCC progression we identified the genes that change (absolute log2 fold change greater than 1) when comparing control to high-grade dysplasia and control to very advanced HCC. In the control versus high-grade dysplasia comparison there are 149 DE genes, while in the control versus very advanced HCC comparison there are 1,355 DE genes, which is almost an order of magnitude higher. This suggests that using the genes that are differentially expressed across the change interval, as opposed to the genes that are different between control and very advanced HCC, offers a more focused analysis. In essence, the comparison across the narrowest change interval targets the genes involved in the initial tumor formation, rather than all genes that change as a consequence of the cancer. The number of common DE genes among the two comparisons is 80, representing 53% of the initial 149 genes.

Followup analysis using the cancer gene census

We downloaded the curated list of cancer genes available in the cancer gene census [53] (accessible at: http://cancer.sanger.ac.uk/census). This list is is presented together with the catalogue of somatic mutations in cancer (COSMIC) [51] (accessible at: http://cancer.sanger.ac.uk/cosmic). We used this list of cancer genes to filter the 80 common genes. The result consists of two genes: CHEK2, a tumor suppressor, and FAT1, which is known to act both as a tumor suppressor as well as an oncogene. These are genes highly relevant to the condition under study considering CHEK2 mutations have been linked to various cancers [160, 37] and it has also been shown to be a mediator of a tumorigenic mechanism specifically in HCC [115]. In addition, FAT1 has been shown to have an oncogenic role in HCC [126, 161], as well as it has been identified as a biomarker in multiple cancers [32, 168]. Fig. 3.22 shows the expression of CHEK2 and FAT1 over the disease progression stages. The expression of both genes increases with disease progression with a

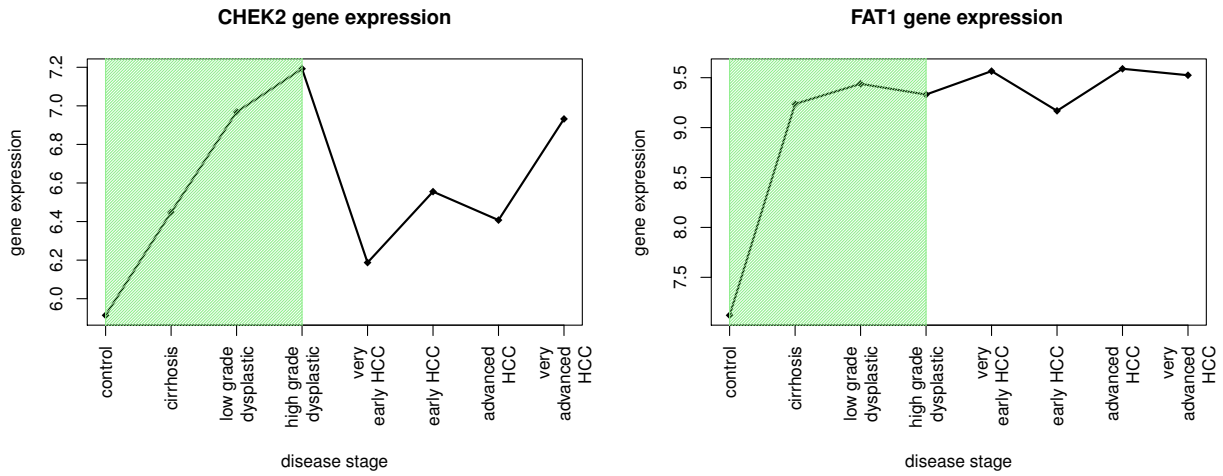

Figure 3.22: The log 2 expression of CHEK2 (left) and FAT1 (right) over 8 stages of disease progression from healthy to advanced hepatocellular carcinoma (HCC). Data from GEO (GSE6764) for the human hepatitis C virus (HCV) to HCC progression. The green shaded area is the change interval detected using this gene expression dataset and the viral carcinogenesis pathway from KEGG (hsa05203). CHEK2 and FAT1 steadily increase throughout the disease stages with a sharp increase in expression during the change interval. CHEK2 and FAT1 could potentially be targeted for down regulation. CHEK2 is a tumor suppressor, and FAT1 is known to act both as a tumor suppressor as well as an oncogene. CHEK2 mutations have been linked to various cancers [160, 37]. FAT1 has been studied in HCC where it has been shown to have an oncogenic role [126, 161] and has been identified as a biomarker in multiple cancers [32, 168].

sharp increase taking place during the change interval, which may be a potential window for treatment.

Followup analysis using KEGG pathways

The "Viral carcinogenesis pathway" from KEGG was used to identify the change interval for the HCV induced HCC progression. To further investigate the results of QCD we used this pathway in a followup analysis. As mentioned above, we compared the control with the high-grade dysplasia stage and the control with the very advanced HCC stage and identified 80 common differentially expressed genes. We also used this pathway to filter the 80 common genes and obtain a "Viral carcinogenesis" gene set, which contains genes from the pathway that change at the onset of the disease. The result consists of two early growth response genes: EGR2 and EGR3. EGR2 has been shown to be an apoptosis promoter

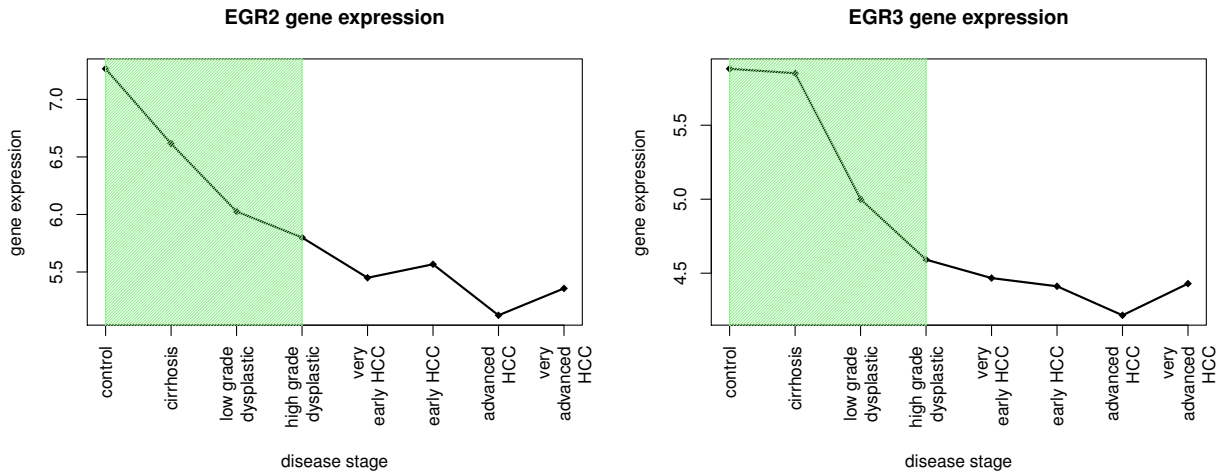

Figure 3.23: The log 2 expression of EGR2 (left) and EGR3 (right) over 8 stages of disease progression from healthy to advanced hepatocellular carcinoma (HCC). Data from GEO (GSE6764) for the human hepatitis C virus (HCV) to HCC progression. The green shaded area is the change interval detected using this gene expression dataset and the viral carcinogenesis pathway from KEGG (hsa05203). EGR2 and EGR3 steadily decrease throughout the disease stages with a sharp decrease in expression during the change interval. EGR2 and EGR3 could potentially be targeted for over expression. EGR2 is known to apoptosis promoter gene [159] that is known to be downregulated by miRNAs in cancer [178, 94]. EGR3 is known to be involved in a number of cancers and the regulation of the immune response [77, 136, 129, 27] and has recently been liked to HCC where it was used to inhibit the growth of tumor cells [187].

gene [159], which is downregulated by miRNAs in cancer [178, 94]. EGR3 has been shown to be involved in a number of cancers and the regulation of the immune response [77, 136, 129, 27] and has recently been liked to HCC where it was used to inhibit the growth of tumor cells [187]. Fig. 3.23 shows the expression of EGR2 and EGR3 over the disease progression stages. The expression of both genes decreases with disease progression with the sharpest decrease taking place during the change interval.

3.5.3 QCD behavior under the null hypothesis

An important question for the proposed method is to demonstrate that the approach does not report qualitative changes in the case of experiments that do not involve any system perturbations (false positive changes).

The hypothesis is that if there is a change interval the system state comparisons will yield a mix of large and small system perturbations. Small system perturbation are expected when comparing system states before the change interval or system states after the change interval. Large system perturbations are expected when comparing system states before the change interval with states after the change interval. Therefore we used a mixture of two gamma distributions, one for the comparisons in which the system is unperturbed (which is also the null hypothesis) and another for comparisons in which the system is perturbed. The mixture model will be initialized with two distributions having the mode the minimum and maximum of the perturbation factors. The mixture model fitting will provide two distributions that best fit the data together with a percentage which estimates how much of the observed data comes from each of these two distributions. If any of the distributions has a percentage of less than 10%, we consider that there is only one distribution and therefore we will not report any significant change.

To investigate the behavior of this approach under the null distribution, when the system is only affected by random noise and small random fluctuations, we used the time-course data from the control samples involved in the perturbation experiments above.

Control data from the fruit fly ethanol exposure experiment

The study by Kong et al. [88] on fruit fly exposure to ethanol contains both condition and control time course data. The experiment spans 3.5 hours (210 min) of recovery after a 30 min ethanol exposure sedating up to 75% of the flies and is sampled at 8 time points. The time-points include one control before exposure, one at 0h right after exposure, and every 30 minutes after that up to 3.5h with a missing data point at 2.5h (150 min) which was not provided in the dataset. Treatment conditions used in this experiment were exposure to humidified air or ethanol vapor (60%) for 30 min, and then recovery for up to 210 minutes [88]. Samples exposed to humidified air are the control samples. Fig. 3.24 presents the first two steps of the QCD method on the control data. A mixture of two gamma distributions is fitted to the perturbation factors computed for all time points comparisons on the control

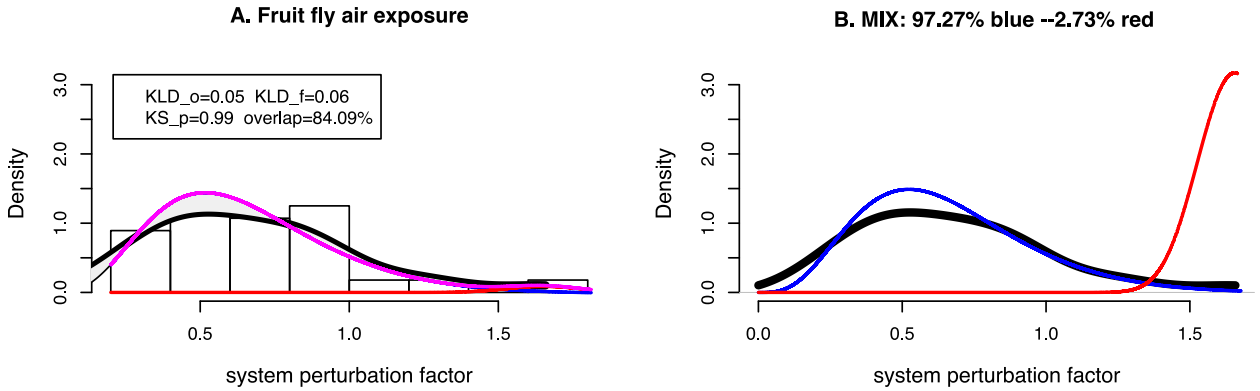

Figure 3.24: The results of QCD on real data of the fruit fly exposure to air. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the observed perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportions. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence, which is computed between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and a sample of the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The red distribution (large perturbation) contributes only 2.73% of the mixture. In other words, the comparisons between system states show mostly small perturbation, which means there is no significant system change.

data. The large perturbations (red distribution) contribute only 2.73% of the mixture. In other words, the comparisons between system states show mostly small perturbation, which means there is no significant system change.

Control data from the mouse phosgene exposure experiment

In the study by Sciuto et al. [143], mice were exposed to 32 mg of phosgene per cubic meter for 20 min and samples were collected from lung tissue at 9 time points: untreated (0), 30 min, 1, 4, 8, 12, 24, 48, 72 hours after exposure. As a control, samples were collected at the same time points from mice exposed to air. Fig. 3.25 presents the first two steps of the QCD method on the control data. A mixture of two gamma distributions is fitted to the perturbation factors computed for all time points comparisons on the control data. Results show that the distribution of large system perturbations contributes less than 10%

of the mixture. In other words, the comparisons between system states show mostly small perturbation, which means there is no significant system change.

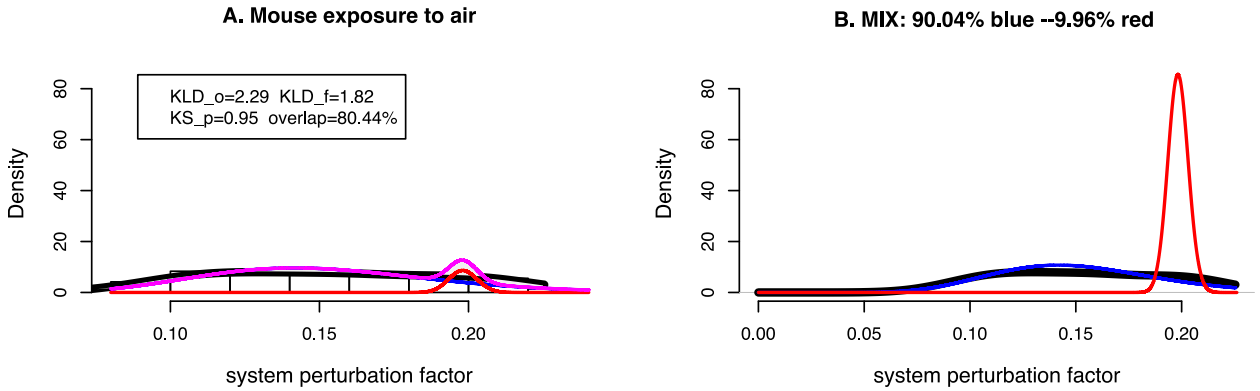

Figure 3.25: The results of QCD on real data of the mouse exposure to air. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportions. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence, which is computed between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and a sample of the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The red distribution (large perturbation) makes up for only 9.96% of the mixture. In other words, the comparisons between system states show mostly small perturbation, which means there is no significant system change.

Behavior on random data

An important requirement is to demonstrate that the approach does not report significant changes in random data. In order to investigate this, we generated 10,000 perturbation factors using random samples from the *E coli* flagellum building data. We select randomly 10 time points out of the 21 available, and we compute a perturbation factor comparing the average of the selected 10 time points with the average of the other 11 time points. This process generates 10,000 random perturbation factors. We fit a mixture of two gamma distributions to these data (see Fig. 3.26). The large perturbations (red distribution) make up for only 2.64% of the mixture. In other words, the comparisons between system states show

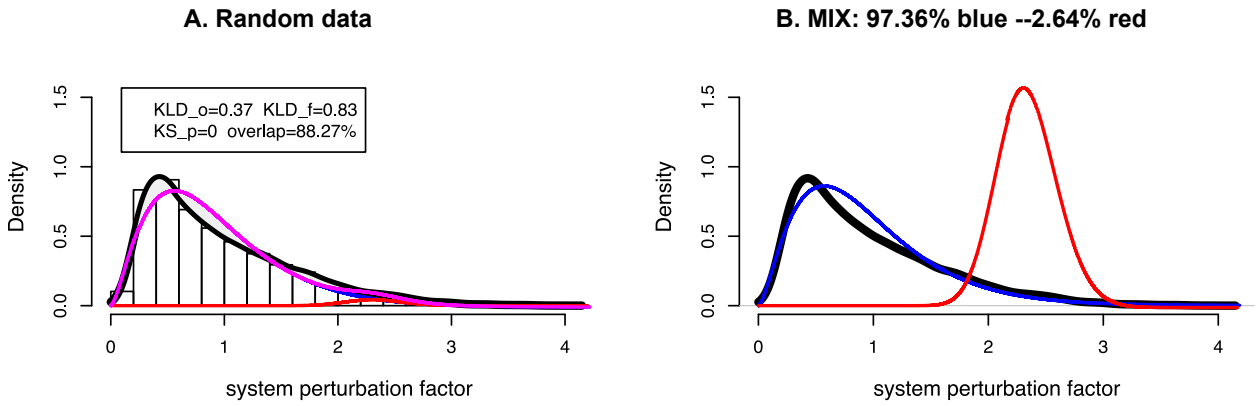

Figure 3.26: The results of QCD on random data generated using the *E. coli* flagellum building synthetic data. Randomly selected groups of 10 (out of a total of 21) system states are compared to the group of remaining 11 system states to generate random perturbation values. Data is generated using 10,000 iterations. Panel A shows the mixture (magenta line) of two gamma distributions (blue and red lines) that is fitted to the perturbation factors (histogram, and density - thick black line). The blue and red density lines are scaled using the mixture proportion. We evaluate the goodness using three statistics: (i) the Kullback-Leibler divergence, which is computed between the density of the observed perturbation and the density of the fitted mixture model (KLD_o-observed first, KLD_f-fitted first); (ii) the p-value of the Kolmogorov-Smirnov test (KS_p) between the observed perturbation and a sample of the fitted mixture model; and (iii) the overlap, which is the ratio between the intersection and union of the areas delimited by the observed (thick black) and fitted (magenta) density lines. Panel B shows how the gamma mixture model is used to separate small (blue line) and large perturbations (red line). The blue and red distributions which compose the mixture model are unscaled in this panel and the mixture proportion is reported. The red distribution (large perturbation) makes up for only 2.64% of the mixture. In other words, the comparisons between system states show mostly small perturbation, which means there is no significant system change.

mostly small perturbation, which means there is no significant system change. Thus, QCD does not report any false positives when the data is random.

3.6 A summary of this chapter

We designed and implemented an analytical method capable of detecting qualitative changes in the state of a biological system by monitoring its gene expression levels. This has been conducted with no training on previous examples, with no expert supervision, and with thresholds set using sound statistical criteria. The only hypothesis used here is that a qualitative change will involve enough pathway components to perturb the pathway in a significant way.

To evaluate the proposed method, we used both synthetic and real data. The cases used for validation cover a wide range of biological phenomena and model organisms as presented in section 3.3 (see Table 3.2 for a summary). Identifying a change interval implies recognizing the transition the system goes through from a state of relative equilibrium to another. The states of relative equilibrium the system transitions from are denoted here as meta-states and the transition as the change interval. Notably, in each case study, the system transitions between meta-states that are of great importance if we hypothesize that such transitions are infrequent and that a qualitative change is required for a system to undergo such transitions. We also assessed the statistical significance of the potential meta-states for each of the eight case studies. Results show that out of 16 putative meta-states, 13 are significant at a threshold of 5%.

The proposed method was applied on a wide range of biological phenomena and was able to detect important transitions between system meta-states with high accuracy in the first six case studies having a known change interval: building a motility motor in *E. coli*, spore formation in *B. subtilis* and *S. cerevisiae*, backwards movement triggered by the nose touch in *C. elegans*, and both acute ethanol exposure and metamorphosis in *D. melanogaster*.

We also compared QCD to an existing method developed by Liu et al. [92] for detecting the pre-disease state and network biomarkers on two datasets. These are two case studies where the phenomena are more complex. When analyzing the data for the exposure to the toxic gas phosgene in mice, QCD identified the cellular damage at an earlier time point, when treatment is still effective [141].

When analyzing data for hepatitis C virus infection progression to hepatocellular carcinoma (HCC) in humans, QCD identified the transition from control to high-grade dysplasia. In this case, the existing method identified as the pre-disease state, i.e., the "very early HCC" stage, which can be interpreted as the start of the malignant state. Importantly, the change interval detected by QCD immediately precedes this pre-disease state detected

by the existing method and marks the transition from benign to malignant. Intervention during this interval may prevent this transition and disease progression may be halted.

To summarize, we have evaluated the proposed method QCD on both synthetic (noise free) and real (noisy) data, on a total of eight case studies for six model organisms and one human dataset and the QCD identified the qualitative changes in each case. We have also used both time course data as well as disease stages as system states in our analyses, and QCD performed well for both types of data.

An immediate application for QCD could be to identify when the transition between different disease stages happens for other diseases. However, QCD is a versatile approach that can be applied to systemic states in different contexts (time course, disease progression, drug dose, BMI, age).

The QCD method can also be applied in the study of drug synergies and synthetic lethality where it could identify the time interval when one drug sensitizes the cell and the second drug has maximum efficacy in a time-dependent way. In turn, this could maximize the effect of combination therapies for various diseases. Another important application for the conceptual framework described in this chapter is the prediction of obstetrical disease in early pregnancy, so interventions can mitigate or prevent the "great obstetrical syndromes" that are primarily observed during the third trimester of pregnancy [135]. In future work, we plan to use the QCD method to predict obstetrical disease based on transcriptomics, metabolomics, proteomics, lipidomics, and other data. A system state in the QCD framework can be any of, but not limited to, the following: a developmental stage, the response to a certain therapeutic dose, the stage of a disease, patients who share physiological traits or disease outcome. The analysis of time series expression data using QCD could potentially be used to decide the duration of adjuvant chemotherapy, disease recurrence, etc. However, the most important application of this approach would imply a paradigm shift: one could use a QCD-like approach with the aim of identifying the departure from the healthy state instead of diagnosing the onset of disease.

CHAPTER 4: METABOLIC PATHWAY ANALYSIS

4.1 Challenges in metabolic pathway analysis

A multitude of methods have been developed for gene expression data while the progress in the development of topology-base pathway analyses for metabolite data is lagging behind [111, 118]. Until recently, a big limitation was given by technology, but that is not the case any more [124]. The scarcity of the data is and will be a big limitation until the high-throughput metabolic datasets reach a comparable number to the ones available for gene expression. The complex structure of the metabolic pathways further restricts the development of metabolic pathway analyses (see Fig. 4.1). Due to these limitations, very few metabolic pathway analysis that consider the pathway structure [111, 118], while a larger number of metabolite set analyses are available [99]. Metabolite set analyses consider the metabolites as independent entities, which is not the case. Such analyses are not able to able to identify disease mechanisms involving multiple metabolites. There is a need for pathway structure aware analyses to identify mechanisms.

Considering the wealth of information in pathway databases, there is little consensus among them in terms data structures used to store the data (see Fig. 4.1). In a KEGG signaling pathway nodes are gene products and edges are regulatory signals such as activation or inhibition (see http://www.genome.jp/kegg/document/help_pathway.html) for details). In a KEGG metabolic pathway nodes are biochemical compounds and edges are chemical reactions. Reactions are catalyzed by enzymes which are proteins encoded by genes. Therefore, in a metabolic pathway genes are associated with edges. This makes the structure of the KEGG metabolic pathways very challenging for the development of analysis methods that consider the topology of the pathway. Other pathway databases are available that have a representation more suitable for the analysis of metabolite data in the context of bio-chemical reactions. Such databases are The Reactome Database (https://reactome.org) and The Small Molecule Pathway Database [52, 83] (SMPDB, http://smpdb.ca). Reactome

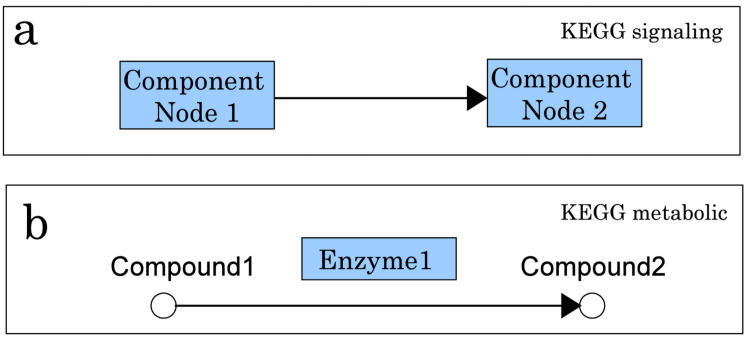

Figure 4.1: Signaling versus metabolic pathway representation. Panel (a) contains the representation for a signaling pathway from the Kyoto Encyclopedia of Genes and Genomes (KEGG). The nodes are genes and the edges represent relations between genes such as activation and repression. Panel (b) contains the representation for a metabolic pathway from KEGG. Nodes are metabolic compounds and edges are bio-chemical reactions catalyzed by enzymes. The differences in the pathway structure makes it challenging to use the multitude of pathway analysis methods developed for signaling pathways for the analysis of metabolic pathways.

has a hierarchical structure for the pathways which introduces a statistical challenge when trying to compute independent p-values for overlapping pathways. Even though an absolute independence between pathways is an ideal that cannot be achieved, they all work within the cell, a hierarchical structure is recommended to be resolved before a pathway analysis is ran on such data. SMPDB will be described in the following as our choice of pathway database for metabolic pathway analysis.

4.2 Pathway analysis using the stoichiometry of the reaction

To help address the challenge of identifying mechanisms in complex metabolic diseases, we propose a novel framework for the analysis of metabolic pathways using the stoichiometry of bio-chemical reactions. We developed the reaction limiting factor analysis for metabolic pathways (RAMP), an approach designed to detect metabolic pathways significantly perturbed when comparing two phenotypes (i.e disease vs. healthy). The method is designed to leverage the information regarding the stoichiometry of bio-chemical reactions stored in pathway databases.

The method takes as input a list of pathways stored as a list of bio-chemical reactions and measurements of reaction components as provided by high-throughput biological experiments for two phenotypes to be compared.

To describe the RAMP analysis method, let us consider the following working example for a single reaction.

$$4A + 3B \rightarrow 2C + 5D \tag{4.1}$$

The measurements of the reaction components (metabolites A, B, C and D) for the working example are presented in Table 4.4.

Input data – metabolite measurements				
Phenotype/Metabolites	**mA**	**mB**	**mC**	**mD**
Healthy (H)	24	15	18	12
Disease (D)	28	36	20	27
Stoichiometry	4	3	2	5
Realization & limiting factor H (min)	$24/4 = 6$	**$15/3 = 5$**		
Realization & limiting factor D (min)	**$28/4 = 7$**	$36/3 = 12$		
Differential realization D/H 4.2	**$7/5 = 1.4$**			

Table 4.4: Example of measurements for the components (metabolite A, metabolite B, metabolite C, metabolite D) of a bio-chemical reaction for two phenotypes: healthy and disease. The first two rows of the table contain the measurements for the left-hand components (metabolite A, metabolite B, metabolite C, metabolite D) of the three bio-chemical reactions for two phenotypes: healthy (H) and disease (D). The next 4 rows display: (1) the stoichiometry of the reaction, (2) the reaction realization and limiting factor for the healthy phenotype, (3) the reaction realization and limiting factor for the disease phenotype, and (4) the reaction differential realization for the disease versus the healthy phenotype

In the first step, we compute **metabolite realization factors** as normalized metabolite measurements using the stoichiometry as normalization factor for each of the two phenotypes. The reactants (left side of the equation) are the components that will be transformed and therefore to evaluate the rate of the reaction it is important to evaluate how much of these components is measured (available) for the reaction. **Realization factors estimate how many times a reaction can take place given the amount of reactants avail-**

able. For the data given in Table 4.4, the realization factors in the healthy healthy state for the reactants are as follows: for metabolite A (mA) is 6 (24/4); and for metabolite (mB) is 5 (15/3). Similarly, the realization factors in the disease state for the reactants are as follows: for metabolite (mA) is 7 (28/4); and for metabolite (mB) is 12 (36/3).

In the second step of the analysis, we compute **reaction limiting factors** as the minimum of the realization factors for computed for the reactants of a reaction for each phenotype. **The limiting factor is the component that limits the reaction to a minimum number of times it can take place.** For the data given in Table 4.4, the limiting factor for healthy is reactant mB with a realization factor ($\mathrm{RF}_{healthy}$) equal to 5. The limiting factor for disease is reactant mA with a realization factor ($\mathrm{RF}_{disease}$) equal to 7.

In the third step of the analysis, **the reaction differential realization** statistic is computed as the ratio between the limiting factors of the reaction in the two phenotypes. The differential realization for our reaction is 1.4 (7/5, $\mathrm{RF}_{disease}/\mathrm{RF}_{healthy}$). **The differential realization is a reaction statistic that evaluates how disrupted is a reaction when comparing two conditions (phenotypes).** When no disruption occurs the reaction statistic (ratio of limiting factors between the two phenotype) would be 1. A differential realization above 1 would signal that the reaction can occur more times in the phenotype under study (usually disease) which means more of the product(s) of the reaction would be available in that phenotype. A differential realization below 1 would signal that the reaction can occur fewer times in the phenotype under study (usually disease) which means less of the product(s) of the reaction would be available in that phenotype. A similar logic can be used when the pathway differential realization (average of the differential realization of the reactions in the pathway) is computed.

4.3 Change propagation for bio-chemical reactions

Reactions in a pathway do not occur independently of one another therefore, to provide a better model for the reaction perturbation would include the interactions between the reactions in a pathway.

Let us consider again a working example, this time, using three reactions. We use multiple reaction to illustrate the different ways a reaction interacts with other reactions.

$$4A + 3B \rightarrow 2C + 5D \tag{4.2}$$

$$4C + 3D \rightarrow 2E + 5F \tag{4.3}$$

$$5A + 4C \rightarrow 2G + 5H \tag{4.4}$$

The measurements of the left-side reaction components for the working example are presented in Table 4.5. The values for metabolites E,F,G and H are 5,6,7,and 8, respectively, and do not change between disease and healthy. The table includes the first three steps of the analysis presented previously in Section 4.2. Specifically, it displays the computation of realization factors for each component (metabolite), identifies the limiting factors and computes the reaction differential realization, $rDR = rRF_{disease}/rRF_{healthy}$, as the ratio between the reaction realization in disease versus healthy for the reaction limiting factor.

The fourth analysis step consists of the computation of the **reaction perturbation** through the propagation of the reaction realization of the upstream reactions using eq 4.5. An upstream reaction is a reaction whose reactants overlap with the products of a given reaction. **The reaction perturbation estimates the disruption produced at the reaction level by the change in phenotype considering the interactions between reactions.**

$$PF(r) = rDR(r) + \sum_{u \in US(r)} \frac{PF(u)}{\#DS(u)} \tag{4.5}$$

where $PF(r)$ is the perturbation factor for reaction r, $rDR(r)$ is the reaction differential realization for the reaction r as the ratio between the reaction realization in disease versus healthy for the reaction limiting factor, $US(r)$ is the set of reactions directly upstream of r, $DS(r)$ is the set of genes directly downstream of r, and # denotes set cardinality.

In the fifth analysis step, we evaluate the **perturbation at the pathway (reaction set) level**. We consider that the reactions in a pathway interact with one another. We define an interaction between two reactions $r1$ and $r2$ if any of the products of reaction $r1$ (right side components) is a reactant (left side component) of reaction $r2$. With this definition, we compute a incidence-type matrix computed for the reactions in a pathway. The reaction matrix for the three reactions described by equations 4.2, 4.3, 4.4 is computed in Table 4.6. The reaction matrix, together with equation 4.5 is used to compute the reaction perturbation

Input data – metabolite measurements - 3 reactions impact analysis				
Phenotype/Metabolites	**mA**	**mB**	**mC**	**mD**
Healthy (H)	24	15	18	12
Disease (D)	28	36	32	27
Stoichiometry for 4.2	4	3		
Identify limiting factor H 4.2	24/4 = 6	**15/3 = 5**		
Identify limiting factor D 4.2	**28/4 = 7**	36/3 = 12		
Differential realization D/H 4.2	**7/5 = 1.4**			
Stoichiometry for 4.3			4	3
Identify limiting factor H 4.3			18/4 =4.5	**12/3 = 4**
Identify limiting factor D 4.3			**32/4=8**	27/3 = 9
Differential realization D/H 4.3	**8/4 = 2**			
Stoichiometry for 4.4	5		4	
Identify limiting factor H 4.4	24/5 = 4.8		**18/4 = 4.5**	
Identify limiting factor D 4.4	**28/5 = 5.6**		32/4 = 8	
Differential realization D/H 4.4	**5.6/4.5 = 1.24**			

Table 4.5: Example of the perturbation analysis for a set of three reactions. The first two rows of the table contain the measurements for the left-hand components (metabolite A, metabolite B, metabolite C, metabolite D) of the three bio-chemical reactions for two phenotypes: healthy (H) and disease (D). The next 12 rows in groups of 4 display: (1) the stoichiometry of the reaction, (2) the reaction realization and limiting factor for the healthy phenotype, (3) the reaction realization and limiting factor for the disease phenotype, and (4) the reaction differential realization for the disease versus the healthy phenotype.

Pathway reaction matrix - 3 reactions impact analysis				
Reaction	**4.2** perturbs	**4.3** perturbs	**4.4** perturbs	Reaction perturbation
4.2 is perturbed by	0	0	0	1.4
4.3 is perturbed by	1	0	0	2 + 1.4/2 = 2.7
4.4 is perturbed by	1	0	0	1.24 + 1.4/2 = 1.94
Number of down stream reactions	2	0	0	
Reaction differential realization	1.4	2	1.24	
Pathway perturbation				(1.4 + 2.7 + 1.94)/3 = 2.01

Table 4.6: Example of the incidence-type matrix computed for the reactions in a pathway in order to compute a pathway-level perturbation value. The 3 top rows and columns represent reactions. The values for the cells are either 0 or 1, where a value of 0 represents no interaction between reactions and a value of 1 represents an interaction. The columns represent a "perturbs" relation, where reaction the column header perturbs the reactions on the rows where there is a value of 1. The rows represent a "is perturbed by" relation where the row header is perturbed by the reactions on the columns where there is a value of 1. The first row following the reaction matrix contains the reaction differential realization for the three reactions. Then, in the "Number of down stream reactions" row, for each reaction we compute the number of downstream reactions as the sum of the column values. Using these values, we observe that the first reactions perturbs the other two, while the second reaction only perturbs the last one, and the last one does not perturb any other reactions. The last column, titled "Reaction perturbation" contains the calculation of the reaction perturbation for each reaction (row) using the reaction matrix and the reaction realization. The last row, titled "Pathway perturbation" contains the computation of the perturbation computed at the pathway level where the pathway is the set of the three reactions.

factor, that estimates the perturbation produced by the change in phenotype at the reaction level. Finally, we evaluate the perturbation produced by the change in phenotype at the pathway level using the mean of the perturbations of the reactions in the pathway (eq. 4.6). The pathway perturbation for the example pathway comprised of three reactions described (eq. 4.2, 4.3, 4.4) is computed in the last row of Table 4.6.

$$PF(P) = \frac{\sum_{r \in P} PF(r)}{\#P} \tag{4.6}$$

where $PF(P)$ is the perturbation factor for pathway P, $PF(r)$ is the perturbation factor for reaction r, and # denotes set cardinality, in this case the number of reactions in pathway P.

The statistic at the pathway level estimates the how the change in phenotype disrupts the pathway. However, this statistic, weather high or low can happen just by chance. In the last two steps of the analysis a resampling approach is employed to compute an empirical distribution of this statistic, which is then used to compute a p-value that evaluates how likely it is that the observed value occurred just by chance. This p-values computed for the pathway perturbation, together with a predefined threshold, is then used to identify significant pathways.

All together, the workflow of the analysis consists of the following steps:

1. Compute **metabolite realization factors**, as normalized measures using metabolite values measured for each phenotype and the stoichiometry as normalization factor

2. Identify **limiting factors** as the minimum of the realization factors for each reaction for each of the two phenotypes;

3. Compute the **reaction differential realization** statistic as the ratio between the limiting factors of the reaction in the two phenotypes;

4. Compute the **reaction perturbation factor** by propagating the reaction differential realization statistic between reactions that share products and reactants using an impact analysis at the pathway level;

5. Compute the **pathway perturbation factor** for each pathway in the input list as the average of reaction perturbation factors;

6. Compute the **significance of the pathway perturbation** using a permutation approach to generate the distribution of the pathway perturbation under the null hypothesis and compute an empirical p-value using this distribution;

7. Select **significantly impacted pathways** using a predefined significance threshold.

4.4 Evaluation on simulated data

To evaluate the novel pathway analysis method we will use controlled simulated data for both the pathways (bio-chemical reactions) and the metabolites. The goal is to assess if the novel method is able to identify relevant changes at the pathway level. We will consider two pathways and a set of metabolite data that includes measurements for all metabolites in these two pathwys. We will also compare the results of RAMP with the classical over-representation approach, for which we will use the hypergeometric test.

In order to test our method we use the pathway and data created in Section 4.3. We have pathway 1 with three reactions (eq. 4.2, 4.3, 4.4) and 8 metabolites measured (A, B, C, D, E, F, G and H, see Table 4.7, left). We also generate another pathway, pathway 2, with the same structure (equations) but different metabolites (A1, B1, C1, D1, E1, F1, G1, and H1, see Table 4.7, right). The measurements for these metabolites are as follows: A1, B1, C1, D1 do not change between disease and healthy and have the values 5,6,7,and 8, respectively (same values as E,F,G and H in the first dataset), while E1,F1,G1 and H1 have the same values as A, B, C and D in the first dataset. We created these case-studies where we have pathway 1 with significant changes, and pathway 2 with non significant changes. Also, for both pathways, we have the same number (4) of metabolites that change between phenotypes. These case-studies allow us to evaluate the behavior of the new method and compare it with the over-representation approach.

For these data, the expected results of the over representation to be the same and to be non significant, even though there are metabolites changing in the pathway and for one of the pathways metabolites that perturb all pathway reactions change. This result is due to the fact that in both pathways the same number of metabolites change (4/8), and they change in the same proportion if we consider all pathways 8/16. RAMP considers the stoichiometry of the reaction as well as how the metabolites that change are distributed in the pathway, based on the structure of the bio-chemical reactions. We expect RAMP to identify pathway 1 as changing significantly and pathway 2 as changing non-significantly.

Data for pathway1		
Metabolite	**Healthy values**	**Disease values**
A	16	48
B	15	36
C	8	32
D	12	27
E	5	5
F	6	6
G	7	7
H	8	8

Data for pathway 2		
Metabolite	**Healthy values**	**Disease values**
A1	5	5
B1	6	6
C1	7	7
D1	8	8
E1	16	48
F1	15	36
G1	8	32
H1	12	27

Table 4.7: Simulated datasets for RAMP evaluation. The left-side table contains the values measured in healthy and disease phenotypes for the metabolites involved in the reactions from pathway1. The right-side table contains the values measured in healthy and disease phenotypes for the metabolites involved in the reactions from pathway2. In pathway1 metabolites A, B, C and D change between phenotypes, while in pathway2 metabolites E1, F1, G1 and H1, change between phenotypes. Values that change between phenotypes are highlighted in pink in the table.

With this set-up, we ran both the hypergeometric test and the RAMP method using pathways 1 and 2 and the dataset presented in Table 4.7. For the RAMP method, so far we have shown how to compute the pathway-level statistic called perturbation factor, in order to assess the significance of this statistic we employ a resampling approach called bootstrapping in order to build the distribution of the statistic under the null hypothesis where we have random changes. We randomly permute the labels of the input measured metabolite data for a number of times ($n = 10,000$) and compute the pathway perturbation using these values. Based on this distribution an empirical p-value is computed as the number of times the observed statistic is more extreme (higher) than the values in the null distribution. For the hypergeometric test, we use the total number of metabolites (16) and the total number of changed metabolites (8), and for each pathway, there is a total of 8 metabolites and 4 changed metabolites.

The results for RAMP show pathway 1 as significant at 5% (p-value FDR 0.0246), while the hypergeometric test did not identify any significant pathways (see Table 4.8).

Pathway/p-value	RAMP		hypergeometric test	
	p-value	FDR	p-value	FDR
Pathway1	**0.0127**	**0.0254**	0.6903	0.6903
Pathway2	1	1	0.6903	0.6903

Table 4.8: Results for RAMP and the hypergeometric test evaluation on simulated data. The table contains the list of pathways and corresponding p-values computed using the RAMP method (left 2 columns) and hypergeometric test (right 2 columns). RAMP identifies pathway 1 as significant at a 5% significance threshold with a FDR corrected p-value of 0.0254, highlighted in pink (corresponding row p-value 0.0127 is displayed in bold). The hypergeometric test does not identify any pathway as significant reporting p-values of 0.6903.

These evaluation results highlight the importance of taking into consideration the important information provided by the pathway structure.

The next step would be to evaluate the novel pathway analysis method for practical purposes. The goal is to assess if the novel method is able to identify relevant changes at the pathway level in practice, using real data. For that, we will use data from biological experiments and pathways from public biological pathway databases. To select the data, we examine a pathway database repository and a metabolite data repository.

4.5 Data source for metabolic data: case study HMDB

The Human Metabolome Database [176, 175, 174] (HMDB, http://www.hmdb.ca) is a database that contains a wealth of information about metabolites that can be found in the human body. It has a search functionality implemented that allows easy access to the data specific to a metabolite through simple queries. The data for all metabolites can be downloaded in the XML machine readable format. Parsed data can then be queried for different statistics. It contains data for 3,295 compounds detected for healthy individuals and 1,780 compounds for individuals that have a specific condition/disease (http://www.hmdb.ca/statistics).

HMDB can be queried for various statistics. For instance, the data stored in this database is available for different biofluids including blood, tears, urine, and saliva. To produce a dataset, we were interested to see how many biosamples are available for each

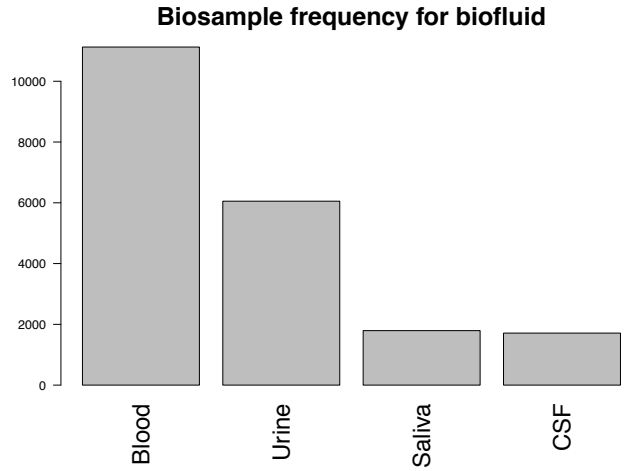

Figure 4.2: The top 4 biofluids available in HMDB based on the biosample frequency for each biofluid. CSF is the abbreviation for cerebrospinal fluid.

biofluid. Fig. 4.2 shows the number of samples for the top 4 biofluids out of a total of 16. The top one is blood with 11,126 biosamples, followed by urine with 6,053 biosamples.

Once we identified the tissue/biofluid, the next step in selecting a dataset is choosing a condition that we want to study. For analysis purposes, a dataset that has more data is a

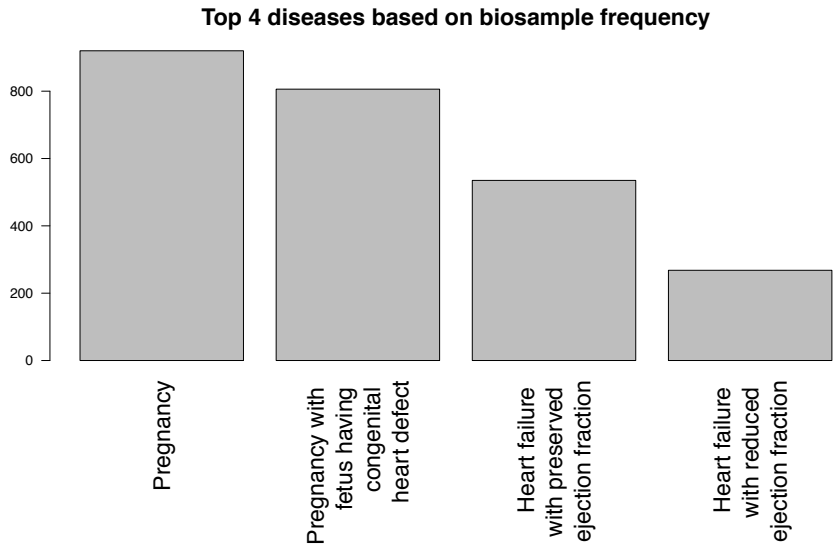

Figure 4.3: The top 4 diseases available in HMDB based on the biosample frequency for blood biosamples.

better dataset. Therefore we would like to know how many biosamples are available for each condition for the tissue previously selected (blood). There are biospecimens available for this tissue for 472 diseases. Fig. 4.3 shows how many biosamples are available for each disease. Pregnancy is the top disease with 920 biosamples for blood and therefore our condition of choice for the evaluation of our analysis.

4.6 Metabolic pathway database: case study SMPDB

The Small Molecule Pathway Database [52, 83] (SMPDB, http://smpdb.ca) is a database designed to store data in machine readable format for more than 40,000 pathways found in humans and other model organisms (*E. coli*, yeast, mouse). It has a search functionality implemented that allows easy access to a specific pathway through simple queries. The data for all pathways can be downloaded in various machine readable formats (BioPAX, SBGN, SBML, PWML). It also contains a visual representation of the pathways that can also be exported. Pathway data can be parsed in different data structures and then queried for different statistics. In the latest version (v2.75) SMPDB contains 55,700 compounds in 57,402 reactions (http://smpdb.ca/stats). There are regular releases that update the database. Table 4.9 shows the evolution of the number of pathways and the number of metabolites from one version to another of the SMPDB.

	SMPDB v1.0	**SMPDB v2.0**	**SMPDB v2.5**	**SMPDB v2.75**
Pathway no.	351	618	61345	48690
Metabolites no.	772	1493	70469	55700

Table 4.9: The change in the number of pathways and metabolites available in the Small Molecule Pathway Database (SMPDB, http://smpdb.ca) is presented for four releases (http://smpdb.ca/stats).

SMPBD can be interrogated for a multitude of purposes. For instance, we were interested to see what is the distribution of the number of reactions per pathway. This is a useful information when developing an analysis tool. It can provide the scale of the data

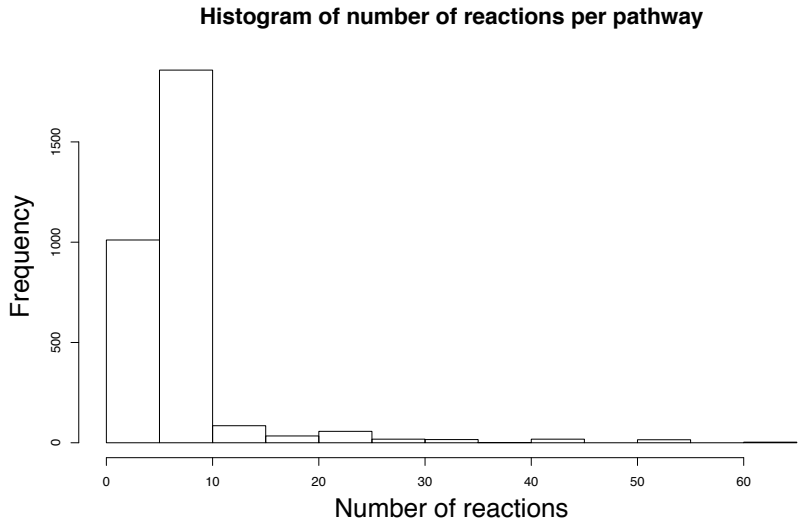

Figure 4.4: Histogram of the number of reaction per pathway for the SMPDB database. Most of the pathways have between 5 and 10 reactions.

to be analyzed. In this case the distribution of the number of reaction is between 1 and 63, with most of the pathways having less than 10 reactions (see Fig 4.4).

4.7 Evaluation on experiment data

To evaluate the proposed method on real data, we will consider pathways from the Small Molecule Pathway Database (SMPDB, http://smpdb.ca, v2.0, release February 19, 2015) and a set of metabolite data form the Human Metabolome Database (HMDB, http://www.hmdb.ca, v3.6) that includes measurements for 577 metabolites in blood samples collected from pregnant and non-pregnant women. We will again compare the results of RAMP with the classical over-representation approach, for which we will use the hypergeometric test.

The experiment data and the pathways are stored in different databases (HMDB, SMPDB) and the metabolites have specific identifiers in each database. To make the connection between the two databases we used the metabolite (small molecule) identifiers used in a third database called the Chemical Entities of Biological Interest (ChEBI, https://www.ebi.ac.uk/chebi/).

For RAMP analysis we used as input metabolite levels in pregnancy and non-pregnancy for 577 metabolites and 60,912 pathways. The experiment data was processed into the format: metabolite ChEBIID, value in non-pregnancy, value in pregnancy. The values were averaged across all samples (Biospecimen: blood) available in HMDB for the condition. Only 3,116 had a least one measured metabolite among the pathway components. Pathways were downloaded in BioPAX Level 3 [33] (machine readable format for data exchange) and processed using custom R (v3.4.2) scripts into a list-type data structure that contains for each pathway the list of reaction and for each reaction the left-side reactants and right-side products with the stoichiometric coefficients. After processing the input data, we run the RAMP analysis to compute pathway perturbation factors and using a bootstrapping (random resampling) approach we compute empirical p-values for the pathway perturbation factors. RAMP reports these values as the results (see Table 4.10). Notably, after an false discovery rate (FDR) correction for multiple comparisons all p-value are 1, which is due to the high number of pathways that are analyzed. A solution may be to pre-select a subset of the pathways in order to do this analysis.

	Pathway Name	**p-value**
1	Folate Metabolism	**0.004**
2	Histidine Metabolism	**0.004**
3	Betaine Metabolism	0.154
4	Ethanol Degradation	0.27
5	Selenoamino Acid Metabolism	0.926
6	Homocysteine Degradation	0.934
7	Glucose-Alanine Cycle	0.942
8	Catecholamine Biosynthesis	0.945
9	Primary Hyperoxaluria Type I	0.953
10	Aromatic L-Aminoacid Decarboxylase Deficiency	0.954

Table 4.10: Results for RAMP evaluation on pregnancy versus non-pregnancy data. The table contains the top 10 pathways and corresponding p-values (computed using 1,000 permutations) as resulted from applying the RAMP method on pregnancy data from the Human Metabolome Database (HMDB, http://www.hmdb.ca) and pathways from the Small Molecule Pathway Database (SMPDB, http://smpdb.ca). RAMP identifies the "Folate Metabolism" pathway as the top perturbed pathway with a very low p-value. Folate metabolism has been shown to be an important metabolic process in pregnancy.

For the hypergeometric analysis we use the 577 metabolites and the 3,116 pathways. We compute the number of metabolites that fall on these pathways, and that number is 74. We compute the number of metabolites that change at least 50% between pregnancy and non-pregnancy, and that number is 28. For each pathway, we compute the number of measured metabolites on the pathway and the number of metabolites that change at least 50% for that pathway. With these numbers we compute a hypergeometric test and report the p-value (see Table 4.11). The p-values reported by the hypergeometric test show that the results could occur by random chance since these p-values are on the higher end with a minimum reported p-value of 23.41%. If we correct the p-values using the FDR correction we have again all corrected p-values equal to 1, due to the high number of pathways.

The results show that RAMP was able to identify at the top of the ranked list and with a low p-value (uncorrected for multiple comparison) pathways relevant to pregnancy. The top ranked pathway is the "Folate metabolism" pathway. Folate, known as vitamin B9, is essential in pregnancy for normal development, prevents the risk of birth defects, and protects the against future complex disease of the child [105, 158]. In addition to the random nature of the results, given by the high p-values, the results of the hypergeometric test show

	Pathway Name	**p-value**
1	Phenylalanine metabolism	0.2341
2	Alanine metabolism	0.3831
3	Cysteine Metabolism	0.3831
4	L-alanine metabolism	0.3831
5	Lactic Acidemia	0.3831
6	Phenylalanine and Tyrosine Metabolism	0.3831
7	Phenylketonuria	0.3831
8	Phosphatidylethanolamine Biosynthesis PE(14:0/16:0)	0.3831
9	Phosphatidylethanolamine Biosynthesis PE(14:0/18:1(11Z))	0.3831
10	Phosphatidylethanolamine Biosynthesis PE(14:0/18:1(9Z))	0.3831

Table 4.11: Results for the hypergeometric test on pregnancy versus non-pregnancy data. The table contains the top 10 pathways and corresponding p-values as resulted from applying the the hypergeometric test on pregnancy data from the Human Metabolome Database (HMDB, http://www.hmdb.ca) and pathways from the Small Molecule Pathway Database (SMPDB, http://smpdb.ca).

on top of the list pathways that involve the metabolism of compounds that are not specific to pregnancy, such as phenylalanine, which is found in the artificial sweetener aspartame.

4.8 A summary of this chapter

In this chapter, we present RAMP a new stoichiometry-driven reaction perturbation inference analysis for metabolic pathways. The method takes as input metabolite data for two phenotypes and metabolic pathways given as sets of bio-chemical reactions. The output is a list of ranked metabolic pathways together with a p-values which estimate if the pathways was perturbed just by random chance or the change in phenotype produce a significant pathway change. The RAMP algorithm uses the stoichiometry for the input measured metabolite level to compute a reaction rate-type of statistic for each metabolite and identify the metabolite with the minimum statistic as the reaction limiting factor in each of the phenotypes. The ratio of the limiting factor is computed as the reaction differential realization that is propagated from one reaction to another reaction when some of the products of one reaction are the reactants of the other reaction. The final reaction statistic that includes values propagated from other, upstream, reactions is the reaction perturbation factor. The average of the reactions perturbation factor is computed for each pathway as the pathway perturbation. A permutation approach is further used to compute a significance level for the perturbation of the pathways. The list of pathways ranked by the significance values (p-values, lower on top) is reported as the result of the RAMP method.

RAMP was evaluated using both simulated and experiment data and in both cases outperformed the classical hypergeometric test. These results are promising and the applicability of the method for identifying mechanisms of complex metabolic diseases warrants attention. The validation performed in this chapter is limited to one simulation case and one experiment data case, as well as only one set of pathways. For a thorough evaluation multiple datasets are needed for both simulation and experiment data, which is part of future work and will be reiterated in the next chapter.

CHAPTER 5: CONCLUSION

5.1 A summary of contributions

In Chapter 1, we detailed the concepts related to systems biology and pathway analysis, specifically topology-based pathway analysis, which takes advantage of the existing knowledge related to the interactions between genes.

In Chapter 2, we present a survey of 22 pathway analysis methods is presented comparing and contrasting the input, mathematical model, and output of the surveyed methods. We also present information on 12 additional methods surveyed by a follow-up book chapter [118]. The lack of benchmarking datasets is a real challenge for developing and evaluating novel methods leading to large number of methods being developed without thorough evaluation. The scarcity of high-throughput metabolite data is a major challenge in developing metabolic pathway analysis methods, as shown by the low number of metabolic pathway analysis methods available (4/34, see [118]).

In Chapter 3, we present a paradigm shift from treating the disease to maintaining the healthy state. We present a qualitative change detection method, QCD, able to identify when a system transitions between qualitative states (e.g. fly metamorphosis). QCD was evaluated on 8 datasets for 7 model organisms, where it accurately identified the respective change intervals. We also compared QCD with an existing method that identifies the pre-disease state. On the two datasets that the comparison was performed QCD identified an earlier change, thus allowing for earlier intervention.

In Chapter 4, we present a novel method for data analysis in the context of metabolic data comparing two phenotypes and metabolic pathways given as sets of bio-chemical reactions that represent various biological processes. The proposed method uses the change in metabolite concentrations and the stoichiometry of the bio-chemical reactions together with an impact analysis approach to evaluate the disruption the phenotype change produces at the pathway level. Our hypothesis is that identifying the sets of bio-chemical reactions

that are significantly perturbed in a specific disease versus a healthy control would provide novel insights into the disease mechanisms. On simulated data, our approach identified as significant the known highly perturbed metabolic pathways and performed better than the classical over-representation approach, which could not distinguish between pathways that have the same number of reactions and the same number of metabolites that changed. Notably RAMP identifies as significant the pathways in which the metabolites that influence the most reactions are changed. On data comparing pregnant versus non-pregnant samples, RAMP identified on top of the list the "Folate metabolism" pathway, which is closely related to pregnancy. This method can be used to further the research into metabolic disease mechanisms.

5.2 Future research directions

Novel benchmarks can be developed for the evaluation of pathway analysis methods and some work has been done in this direction in recent years [155, 76, 6]. Methods to unify the input and output of pathway analysis methods as well as careful selection of benchmark datasets would be the first steps in that direction. Work has been done in creating benchmarks for the pathway analysis methods, software packages that run several methods for pathway analysis [137] and target pathway benchmarks have been created [155].

Future work we consider for pathway analysis evaluation includes:

- creating new benchmarks for pathway analysis using mouse knock-out datasets, where a specific gene is targeted and the affected pathways should be the one that contain the knock-out gene;

- creating a general unified standard format for the input and output of pathway analyses method, maybe independent of the type of data, in order to better evaluate the contribution of new methods.

Our proposed qualitative change detection method currently works only with gene expression data and signaling pathways. Future work we propose for the change detection

analysis will involve adapting QCD to work with RAMP and then change intervals can be identified using metabolic data for diseases or the change detector can be enhanced even further to identify change intervals using multiple types of data. Specific items we consider for for future work related to the change detection method includes:

- expanding the QCD analysis to drug treatment data where we identify the window for co-treatment in diseases that become resistant to therapies;
- expanding the QCD analysis for multiple types of data and explore another source of information for the system, other pathway databases, small protein-protein interaction (PPI) networks.

Future work we consider for the development of the metabolic pathway analyses will primarily involve more validation cases for the RAMP method and it includes:

- evaluation of the RAMP method on more simulated datasets where specific elements are changed such as number of metabolites per reaction, number of reaction per pathway, number of total changed metabolites, the number of changed metabolites per pathway, the position of the changed metabolites on the pathway;
- identifying more experiment datasets that will provide a better coverage of the large number of metabolic pathways;
- designing a method to pre-select pathways, maybe based on the measured metabolites coverage, or select pathways based on a threshold set on the number of reactions;
- refining the RAMP method to better quantify the propagation of the reaction realization, at this time it disregards the values for the products (right side of the reaction).

Future work will also involve the design of an analysis method that would take advantage of multiple types of existing omics data, will incorporate information from multiple

pathway databases, using not only interactions among pathway components but among pathways. In addition, this method will also incorporate information about the dynamics of the condition under study (see Fig. 5.1).

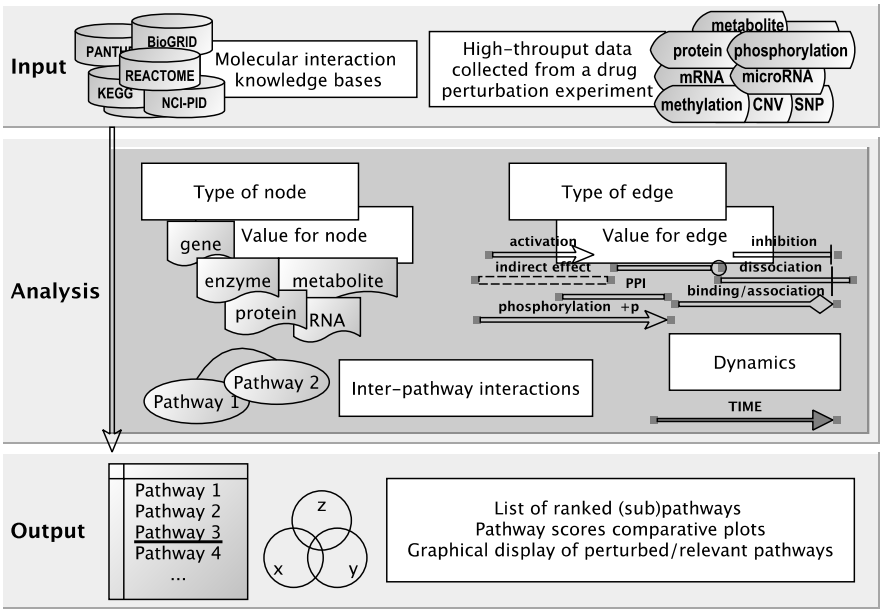

Figure 5.1: The road ahead. An analysis method that would take advantage of existing data will incorporate information from multiple pathway databases, using not only interactions among pathway components but among pathways. Such method will also use multiple datatypes as well as information about the dynamics of the condition.

REFERENCES

[1] Advaita Corporation, "Pathway Analysis with iPathwayGuide," http://www.advaitabio.com/ipathwayguide.html, 2014.

[2] S. Ahsan and S. Draghici, "Identifying significantly impacted pathways with iPathwayGuide," in *Current Protocols in Bioinformatics*. John Wiley and Sons, Inc., 2016.

[3] A. Akhtar, E. Fuchs, T. Mitchison, R. J. Shaw, D. St Johnston, A. Strasser, S. Taylor, C. Walczak, and M. Zerial, "A decade of molecular cell biology: achievements and challenges," *Nature Reviews Molecular Cell Biology*, vol. 12, no. 10, p. 669, 2011.

[4] U. Alon, *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2006.

[5] ——, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 450–461, 2007.

[6] A. Amadoz, M. R. Hidalgo, C. Cubuk, J. Carbonell-Caballero, and J. Dopazo, "A comparison of mechanistic signaling pathway activity analysis methods," *Briefings in Bioinformatics*, p. bby040, 2018. [Online]. Available: http://dx.doi.org/10.1093/bib/bby040

[7] I. Androulakis, E. Yang, and R. Almon, "Analysis of time-series gene expression data: methods, challenges, and opportunities," *Annual Review of Biomedical Engineering*, vol. 9, pp. 205–228, 2007.

[8] S. Ansari, C. Voichiţa, M. Donato, R. Tagett, and S. Drăghici, "A novel pathway analysis approach based on the unexplained disregulation of genes," *Proceedings of the IEEE*, vol. 105, no. 3, pp. 482–495, 2017. [Online]. Available: http://dx.doi.org/10.1109/JPROC.2016.2531000

[9] D. K. Arrell and A. Terzic, "Network systems biology for drug discovery," *Clinical Pharmacology & Therapeutics*, vol. 88, no. 1, pp. 120–125, 2010.

[10] G. D. Bader, I. Donaldson, C. Wolting, F. B. Ouellette, T. Pawson, and C. W. Hogue, "BIND–The Biomolecular Interaction Network Database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 242–245, 2001.

[11] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2493–2503, 2004.

[12] Z. Bar-Joseph, A. Gitter, and I. Simon, "Studying and modeling dynamic biological processes using time-series gene expression data," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 552–564, 2012.

[13] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets–update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, 2013.

[14] A. R. Bassett, C. Tibbit, C. P. Ponting, and J.-L. Liu, "Highly efficient targeted mutagenesis of Drosophila with the CRISPR/Cas9 system," *Cell reports*, vol. 4, no. 1, pp. 220–228, 2013.

[15] R. B. Beckstead, G. Lam, and C. S. Thummel, "The genomic response to 20-hydroxyecdysone at the onset of Drosophila metamorphosis," *Genome Biology*, vol. 6, no. 12, p. R99, 2005.

[16] 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*, vol. 27, no. 15, pp. 2127–2133, 2011.

[17] Y. Ben-Shaul, H. Bergman, and H. Soreq, "Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression," *Bioinformatics*, vol. 21, no. 7, pp. 1129–1137, 2005.

[18] K. J. Beumer, J. K. Trautman, A. Bozas, J.-L. Liu, J. Rutter, J. G. Gall, and D. Carroll, "Efficient gene targeting in Drosophila by direct embryo injection with zinc-finger nucleases," *Proceedings of the National Academy of Sciences*, vol. 105, no. 50, pp. 19 821–19 826, 2008.

[19] J. J. Biagi, M. J. Raphael, W. J. Mackillop, W. Kong, W. D. King, and C. M. Booth, "Association between time to initiation of adjuvant chemotherapy and survival in colorectal cancer," *JAMA*, vol. 305, no. 22, p. 2335, 2011.

[20] K. Bingol, "Recent Advances in Targeted and Untargeted Metabolomics by NMR and MS/NMR Methods," *High-throughput*, vol. 7, no. 2, p. 9, 2018.

[21] BioCarta, "BioCarta - Charting Pathways of Life," http://www.biocarta.com.

[22] BioPAX, "The Biological Pathway Exchange (BioPAX)," http://www.biopax.org, 2002.

[23] B. Bokanizad, R. Tagett, S. Ansari, B. H. Helmi, and S. Drăghici, "SPATIAL: A System-level PAThway Impact AnaLysis approach," *Nucleic Acids Research*, vol. 44, no. 11, pp. 5034–5044, 2016.

[24] M. Borzio, S. Fargion, F. Borzio, A. L. Fracanzani, A. M. Croce, T. Stroffolini, S. Oldani, R. Cotichini, and M. Roncalli, "Impact of large regenerative, low grade and high grade dysplastic nodules in hepatocellular carcinoma development," *Journal of Hepatology*, vol. 39, no. 2, pp. 208–214, 2003.

[25] E. Calura, P. Martini, G. Sales, L. Beltrame, G. Chiorino, M. D'Incalci, S. Marchini, and C. Romualdi, "Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles," *Nucleic Acids Research*, vol. 42, no. 11, p. e96, 2014.

[26] M. Chalfie, J. E. Sulston, J. G. White, E. Southgate, N. J. Thomson, and S. Brenner, "The neural circuit for touch sensitivity in *Caenorhabditis elegans*," *The Journal of Neuroscience*, vol. 5, no. 4, pp. 956–964, 1985.

[27] H. Cheng, S. Hao, Y. Liu, Y. Pang, S. Ma, F. Dong, J. Xu, G. Zheng, S. Li, W. Yuan, and T. Cheng, "Leukemic marrow infiltration reveals a novel role for Egr3 as a potent inhibitor of normal hematopoietic stem cell proliferation," *Blood*, vol. 126, no. 11, pp. 1302–1313, 2015.

[28] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. Brown, and I. Herskowitz, "The transcriptional program of sporulation in budding yeast." *Science*, vol. 282, no. 5393, pp. 699–705, 1998.

[29] H.-Y. Chuang, M. Hofree, and T. Ideker, "A Decade of Systems Biology," *Annual Review of Cell and Developmental Biology*, vol. 26, pp. 721–744, 2010.

[30] F. R. Chung, *Spectral graph theory*. American Mathematical Society, 1997, no. 92.

[31] X. Cui, D. Ji, D. A. Fisher, Y. Wu, D. M. Briner, and E. J. Weinstein, "Targeted integration in rat and mouse embryos with zinc-finger nucleases," *Nature Biotechnology*, vol. 29, no. 1, pp. 64–67, 2011.

[32] C. De Bock, A. Ardjmand, T. Molloy, S. Bone, D. Johnstone, D. Campbell, K. Shipman, T. Yeadon, J. Holst, M. Spanevello, G. Nelmes, R. Catchpoole, L. Lincz, A. Boyd, G. Burns, and R. Thorne, "The Fat1 cadherin is overexpressed and an independent prognostic factor for survival in paired diagnosis–relapse samples of precursor B-cell acute lymphoblastic leukemia," *Leukemia*, vol. 26, no. 5, pp. 918–926, 2012.

[33] 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*, vol. 28, no. 9, pp. 935–942, 2010.

[34] Z. Dezső, Y. Nikolsky, T. Nikolskaya, J. Miller, D. Cherba, C. Webb, and A. Bugrim, "Identifying disease-specific genes based on their topological significance in protein networks," *BMC Systems Biology*, vol. 3, no. 36, 2009.

[35] D. Diaz, M. Donato, T. Nguyen, and S. Draghici, "MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping," in *Pacific*

Symposium on Biocomputing. Pacific Symposium on Biocomputing, vol. 22. New Jersey: World Scientific, 2016, pp. 390–401.

[36] D. Diaz and S. Draghici, *mirIntegrator: Integrating miRNAs into signaling pathways*, 2015.

[37] X. Dong, L. Wang, K. Taniguchi, X. Wang, J. M. Cunningham, S. K. McDonnell, C. Qian, A. F. Marks, S. L. Slager, B. J. Peterson, D. I. Smith, J. C. Cheville, M. L. Blute, S. J. Jacobsen, D. J. Schaid, D. J. Tindall, S. N. Thibodeau, and W. Liu, "Mutations in CHEK2 associated with prostate cancer risk," *The American Journal of Human Genetics*, vol. 72, no. 2, pp. 270–280, 2003.

[38] 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*, vol. 17, no. 10, pp. 1537–1545, 2007.

[39] S. Drăghici, P. Khatri, and C. Voichiţa, "Pathway-Express software url," http://vortex.cs.wayne.edu/projects.htm, 2013, accessed: May 15.

[40] 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*, vol. 17, no. 10, pp. 1537–1545, 2007.

[41] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[42] B. Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[43] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.

[44] S. Efroni, C. F. Schaefer, and K. H. Buetow, "Identification of Key Processes Underlying Cancer Phenotypes Using Biologic Pathway Analysis," *PLoS One*, vol. 2, no. 5, p. e425, 2007.

[45] P. Eichenberger, M. Fujita, S. T. Jensen, E. M. Conlon, D. Z. Rudner, S. T. Wang, C. Ferguson, K. Haga, T. Sato, J. S. Liu, and R. Losick, "The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis." *PLoS Biology*, vol. 2, no. 10, p. e328, 2004.

[46] Z. Fang, W. Tian, and H. Ji, "A network-based gene-weighting approach for pathway analysis," *Cell Research*, vol. 22, no. 3, pp. 565–580, 2011.

[47] ——, "GANPA software url," http://cran.r-project.org/web/packages/GANPA/index.html, 2013, accessed: May 15.

[48] F. Farfán, J. Ma, M. A. Sartor, G. Michailidis, and H. V. Jagadish, "THINK Back: knowledge-based interpretation of high throughput data," *BMC Bioinformatics*, vol. 13, no. Suppl 2, p. S4, 2012.

[49] ——, "THINK-Back-DS software url standalone," http://eecs.umich.edu/db/think/files/density_analysis_1.0.zip, 2013, accessed: May 15.

[50] ——, "THINK-Back-DS software url web-based," http://eecs.umich.edu/db/think/software.html, 2013, accessed: May 15.

[51] S. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C. Cole, M. Jia, C. Kok, H. Boutselakis, T. De, Z. Sondka, L. Ponting, R. Stefancsik, B. Harsha, J. Tate, E. Dawson, S. Thompson, H. Jubb, and P. Campbell, "COSMIC: high-resolution cancer genetics using the catalogue of somatic mutations in cancer," *Current Protocols in Human Genetics*, pp. 10–11, 2016.

[52] A. Frolkis, C. Knox, E. Lim, T. Jewison, V. Law, D. D. Hau, P. Liu, B. Gautam, S. Ly, A. C. Guo, J. Xia, Y. Liang, S. Shrivastava, and D. S. Wishart, "SMPDB: the small molecule pathway database," *Nucleic Acids Research*, vol. 38, no. suppl_1, pp. D480–D487, 2009.

[53] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[54] S. Gao and X. Wang, "TAPPA: topological analysis of pathway phenotype association," *Bioinformatics*, vol. 23, no. 22, pp. 3100–3102, 2007.

[55] L. Geistlinger, G. Csaba, R. Küffner, N. Mulder, and R. Zimmer, "From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems," *Bioinformatics*, vol. 27, no. 13, pp. i366–i373, 2011.

[56] S. Ghosh, Y. Matsuoka, Y. Asai, K.-Y. Hsin, and H. Kitano, "Software for systems biology: from tools to integrated platforms," *Nature Reviews Genetics*, vol. 12, no. 12, pp. 821–832, 2011.

[57] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z.-Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison, H. O. Smith, and J. C. Venter, "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science*, vol. 329, no. 5987, pp. 52–56, 2010. [Online]. Available: http://science.sciencemag.org/content/329/5987/52

[58] E. Glaab, "EnrichNet software url," http://www.enrichnet.org, 2013, accessed: May 15.

[59] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, and A. Valencia, "EnrichNet: network-based gene set enrichment analysis," *Bioinformatics*, vol. 28, no. 18, pp. i451–i457, 2012.

[60] E. Glaab, A. Baudot, N. Krasnogor, and A. Valencia, "TopoGSA: network topological gene set analysis," *Bioinformatics*, vol. 26, no. 9, pp. 1271–1272, 2010.

[61] ——, "TopoGSA software url," http://www.infobiotics.net/topogsa, 2013, accessed: May 15.

[62] S. Greenblum, S. Efroni, C. Schaefer, and K. Buetow, "The PathOlogist: an automated tool for pathway-centric analysis," *BMC Bioinformatics*, vol. 12, no. 1, p. 133, 2011.

[63] S. I. Greenblum, S. Efroni, C. F. Schaefer, and K. H. Buetow, "PathOlogist software url," ftp://ftp1.nci.nih.gov/pub/pathologist/, 2013, accessed: May 15.

[64] Z. Gu, "CePa software url standalone," http://cran.r-project.org/web/packages/CePa/index.html, 2013, accessed: May 15.

[65] ——, "CePa software url web-based," http://mcube.nju.edu.cn/cgi-bin/cepa/main.pl, 2013, accessed: May 15.

[66] Z. Gu, J. Liu, K. Cao, J. Zhang, and J. Wang, "Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes," *BMC Systems Biology*, vol. 6, no. 1, p. 56, 2012.

[67] W. A. Haynes, R. Higdon, L. Stanberry, D. Collins, and E. Kolker, "Differential expression analysis for pathways," *PLoS Computational Biology*, vol. 9, no. 3, p. e1002967, 2013.

[68] M. Heinonen, O. Guipaud, F. Milliat, V. Buard, B. Micheau, G. Tarlet, M. Benderitter, F. Zehraoui, and F. d'Alché Buc, "Detecting time periods of differential gene expression using Gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction," *Bioinformatics*, vol. 31, no. 5, pp. 728–735, 2014.

[69] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler, "IntAct: an open source molecular interaction database," *Nucleic Acids Research*, vol. 32, no. Suppl 1, pp. D452–D455, 2004.

[70] 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*, vol. 37, no. 1, pp. 1–13, 2009.

[71] J.-H. Hung, "PWEA software url," http://zlab.bu.edu/PWEA/, 2013, accessed: May 15.

[72] J.-H. Hung, T. W. Whitfield, T.-H. Yang, Z. Hu, Z. Weng, and C. DeLisi, "Identification of functional modules that correlate with phenotypic difference: the influence of network topology," *Genome Biology*, vol. 11, no. 2, p. R23, 2010.

[73] M. A.-H. Ibrahim, S. Jassim, M. A. Cawthorne, and K. Langlands, "A Topology-Based Score for Pathway Enrichment," *Journal of Computational Biology*, vol. 19, no. 5, pp. 563–573, 2012.

[74] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular Systems Biology*, vol. 8, no. 1, p. 565, 2012.

[75] I. Ihnatova and E. Budinska, "ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data," *BMC bioinformatics*, vol. 16, no. 1, p. 350, 2015.

[76] I. Ihnatova, V. Popovici, and E. Budinska, "A critical comparison of topology-based pathway analysis methods," *PloS One*, vol. 13, no. 1, p. e0191154, 2018.

[77] A. Inoue, Y. Omoto, Y. Yamaguchi, R. Kiyama, and S. I. Hayashi, "Transcription factor EGR3 is involved in the estrogen-signaling pathway in breast cancer cells," *Journal of Molecular Endocrinology*, vol. 32, no. 3, pp. 649–661, 2004.

[78] S. Isci, "BPA software url," http://bumil.boun.edu.tr/bpa, 2013, accessed: May 15.

[79] S. Isci, C. Ozturk, J. Jones, and H. H. Otu, "Pathway analysis of high-throughput biological data within a Bayesian network framework," *Bioinformatics*, vol. 27, no. 12, pp. 1667–1674, 2011.

[80] C. R. Jack Jr, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski, "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade," *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, 2010.

[81] L. Jacob, P. Neuvial, and S. Dudoit, "Gains in power from structured two-sample tests of means on graphs," *Arxiv preprint arXiv:1009.5173*, 2010.

[82] ——, "DEGraph software url," http://www.bioconductor.org/packages/2.12/bioc/html/DEGraph.html, 2013, accessed: May 15.

[83] T. Jewison, Y. Su, F. M. Disfany, Y. Liang, C. Knox, A. Maciejewski, J. Poelzer, J. Huynh, Y. Zhou, D. Arndt, Y. Djoumbou, Y. Liu, L. Deng, A. C. Guo, B. Han, A. Pon, M. Wilson, S. Rafatnia, P. Liu, and D. S. Wishart, "SMPDB 2.0: big improvements to the Small Molecule Pathway Database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D478–D484, 2013.

[84] 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*, vol. 33, no. Database issue, pp. D428–432, 2005.

[85] S. Kalir and U. Alon, "Using a quantitative blueprint to reprogram the dynamics of the flagella gene network," *Cell*, vol. 117, no. 6, pp. 713–720, 2004.

[86] P. Khatri, B. Done, A. Rao, A. Done, and S. Drăghici, "A semantic analysis of the annotations of the human genome," *Bioinformatics*, vol. 21, no. 16, pp. 3416–3421, 2005.

[87] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLOS Computational Biology*, vol. 8, no. 2, p. e1002375, 2012.

[88] E. C. Kong, L. Allouche, P. A. Chapot, K. Vranizan, M. S. Moore, U. Heberlein, and F. W. Wolf, "Ethanol-regulated genes that contribute to ethanol sensitivity and rapid tolerance in Drosophila," *Alcoholism: Clinical and Experimental Research*, vol. 34, no. 2, pp. 302–316, 2009.

[89] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[90] S. L. Lauritzen, *Graphical models*. Oxford University Press, 1996, vol. 17.

[91] J. Lee, K. Jo, S. Lee, J. Kang, and S. Kim, "Prioritizing biological pathways by recognizing context in time-series gene expression data," *BMC Bioinformatics*, vol. 17, no. 17, p. 477, 2016.

[92] R. Liu, M. Li, Z.-P. Liu, J. Wu, L. Chen, and K. Aihara, "Identifying critical transitions and their leading biomolecular networks in complex diseases," *Scientific Reports*, vol. 2, 2012, article number: 813.

[93] R. Liu, X. Wang, K. Aihara, and L. Chen, "Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers," *Medicinal Research Reviews*, vol. 34, no. 3, pp. 455–478, 2014.

[94] X. Liu, H. Shi, B. Liu, J. Li, Y. Liu, and B. Yu, "miR-330-3p controls cell proliferation by targeting early growth response 2 in non-small-cell lung cancer," *Acta biochimica et biophysica Sinica*, vol. 47, no. 6, pp. 431–440, 2015.

[95] W. Luo, M. S. Friedman, K. D. Hankenson, and P. J. Woolf, "Time series gene expression profiling and temporal regulatory pathway analysis of BMP6 induced osteoblast differentiation and mineralization," *BMC Systems Biology*, vol. 5, no. 1, p. 82, 2011.

[96] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf, "GAGE: generally applicable gene set enrichment for pathway analysis," *BMC Bioinformatics*, vol. 10, no. 1, p. 161, 2009.

[97] S. Ma, T. Jiang, and R. Jiang, "Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data," *Bioinformatics*, vol. 31, no. 4, pp. 563–571, 2014.

[98] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemotherapy Reports*, vol. 50, no. 3, pp. 163–170, 1966.

[99] A. Marco-Ramell, M. Palau-Rodriguez, A. Alay, S. Tulipani, M. Urpi-Sarda, A. Sanchez-Pla, and C. Andres-Lacueva, "Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data," *BMC Bioinformatics*, vol. 19, no. 1, p. 1, 2018.

[100] P. Martini, G. Sales, M. S. Massa, M. Chiogna, and C. Romualdi, "Along signal paths: an empirical gene set approach exploiting pathway topology," *Nucleic Acids Research*, vol. 41, no. 1, pp. e19–e19, 2013.

[101] M. S. Massa, M. Chiogna, and C. Romualdi, "Gene set analysis exploiting the topology of a pathway," *BMC Systems Biology*, vol. 4, no. 1, p. 121, 2010.

[102] S. Massa and G. Sales, "TopologyGSA software url," http://cran.r-project.org/web/packages/topologyGSA/index.html, 2013, accessed: May 15.

[103] J. Mata, R. Lyne, G. Burns, and J. Bähler, "The transcriptional program of meiosis and sporulation in fission yeast," *Nature Genetics*, vol. 32, no. 1, pp. 143–147, 2002.

[104] R. A. McLean, W. L. Sanders, and W. W. Stroup, "A unified approach to mixed linear models," *The American Statistician*, vol. 45, no. 1, pp. 54–64, 1991.

[105] J. McPartlin, D. Weir, A. Halligan, M. Darling, and J. Scott, "Accelerated folate breakdown in pregnancy," *The Lancet*, vol. 341, no. 8838, pp. 148–149, 1993.

[106] H.-W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman, "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, vol. 27, no. 1, pp. 44–48, 1999.

[107] H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano, and P. D. Thomas, "The PANTHER database of protein families, subfamilies, functions and pathways," *Nucleic Acids Research*, vol. 33, no. Suppl 1, pp. D284–D288, 2005.

[108] J. Mieczkowski, K. Swiatek-Machado, and B. Kaminska, "Identification of Pathway Deregulation–Gene Expression Based Analysis of Consistent Signal Transduction," *PLoS ONE*, vol. 7, no. 7, p. e41541, 2012.

[109] ——, "ACST software url," http://www.plosone.org/article/info:doi/10.1371/journal.pone.0041541#s4, 2013, accessed: May 15.

[110] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[111] 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*, vol. 4, p. 278, 2013.

[112] D. Nam and S.-Y. Kim, "Gene-set approach for expression pattern analysis," *Briefings in Bioinformatics*, vol. 9, no. 3, pp. 189–197, 2008.

[113] R. E. Neapolitan, *Learning bayesian networks*. Prentice Hall, 2004.

[114] A. M. Neiman, "Sporulation in the budding yeast Saccharomyces cerevisiae," *Genetics*, vol. 189, no. 3, pp. 737–765, 2011.

[115] O. Neumann, M. Kesselmeier, R. Geffers, R. Pellegrino, B. Radlwimmer, K. Hoffmann, V. Ehemann, P. Schemmer, P. Schirmacher, J. Lorenzo Bermejo, and T. Longerich, "Methylome analysis and integrative profiling of human HCCs identify novel protumorigenic factors," *Hepatology*, vol. 56, no. 5, pp. 1817–1827, 2012.

[116] T. Nguyen, D. Diaz, R. Tagett, and S. Draghici, "Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data," *Nature Scientific Reports*, vol. 6, p. 29251, 2016.

[117] T. Nguyen and S. Draghici, *BLMA: A package for bi-level meta-analysis*, Bioconductor, 2017, r package.

[118] T. Nguyen, C. Mitrea, and S. Draghici, "Network-based approaches for pathway level analysis," *Current Protocols in Bioinformatics*, vol. 61, no. 1, pp. 8–25, 2018.

[119] T. Nguyen, R. Tagett, M. Donato, C. Mitrea, and S. Draghici, "A novel bi-level meta-analysis approach-applied to biological pathway analysis," *Bioinformatics*, vol. 32, no. 3, pp. 409–416, 2016.

[120] P. Nurse and J. Hayles, "The cell in an era of systems biology," *Cell*, vol. 144, no. 6, pp. 850–854, 2011.

[121] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[122] K.-H. Pan, C.-J. Lih, and S. N. Cohen, "Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 25, pp. 8961–8965, 2005.

[123] R. Pandey, R. K. Guru, and D. W. Mount, "Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data," *Bioinformatics*, vol. 20, no. 13, pp. 2156–2158, Sep 2004.

[124] G. J. Patti, O. Yanes, and G. Siuzdak, "Innovation: Metabolomics: the apogee of the omics trilogy," *Nature Reviews Molecular Cell Biology*, vol. 13, no. 4, p. 263, 2012.

[125] N. P. Pavletich and C. O. Pabo, "Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A," *Science*, vol. 252, no. 5007, pp. 809–817, 1991.

[126] B. Pereira, S.-F. Chin, O. M. Rueda, H.-K. M. Vollan, E. Provenzano, H. A. Bardwell, M. Pugh, L. Jones, R. Russell, S.-J. Sammut *et al.*, "The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes," *Nature Communications*, vol. 7, p. 11479, 2016.

[127] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. Gandhi, K. Chandrika, N. Deshpande, S. Suresh, B. Rashmi, K. Shanker, N. Padma, V. Niranjan, H. Harsha, N. Talreja, B. Vrushabendra, M. Ramya, A. Yatish, M. Joy, H. Shivashankar, M. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey, "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Research*, vol. 32, no. Suppl 1, pp. D497–D501, 2004.

[128] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo, "WikiPathways: pathway editing for the people," *PLoS Biology*, vol. 6, no. 7, p. e184, 2008.

[129] R. Pio, Z. Jia, V. T. Baron, and D. Mercola, "Early growth response 3 (Egr3) is highly over-expressed in non-relapsing prostate cancer but not in relapsing prostate cancer," *PLoS One*, vol. 8, no. 1, p. e54096, 2013.

[130] S. S. P. M. K. A. A. D. Purvesh Khatri and S. Drăghici, "Recent additions and improvements to the Onto-Tools," *Nucleic Acids Research*, vol. 33, pp. W762–W765, 2005, suppl. S.

[131] J. Rahnenführer, F. S. Domingues, J. Maydt, and T. Lengauer, "Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, 2004.

[132] T. Reuters, "MetaCore software url," http://www.genego.com/metacore.php, 2013, accessed: May 15.

[133] A. D. Rhim, E. T. Mirek, N. M. Aiello, A. Maitra, J. M. Bailey, F. McAllister, M. Reichert, G. L. Beatty, A. K. Rustgi, R. H. Vonderheide *et al.*, "EMT and dissemination precede pancreatic tumor formation," *Cell*, vol. 148, no. 1, pp. 349–361, 2012.

[134] U. Roessner, C. Wagner, J. Kopka, R. N. Trethewey, and L. Willmitzer, "Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry," *The Plant Journal*, vol. 23, no. 1, pp. 131–142, 2000.

[135] R. Romero, "Prenatal medicine: The child is the father of the man*," *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 22, no. 8, pp. 636–639, 2009.

[136] M. Safford, S. Collins, M. A. Lutz, A. Allen, C.-T. Huang, J. Kowalski, A. Blackford, M. R. Horton, C. Drake, R. H. Schwartz, and J. D. Powell, "Egr-2 and Egr-3 are negative regulators of T cell activation," *Nature Immunology*, vol. 6, no. 5, pp. 472–480, 2005.

[137] G. Sales, E. Calura, D. Cavalieri, and C. Romualdi, "graphite–a Bioconductor package to convert pathway topology to gene network," *BMC Bioinformatics*, vol. 13, no. 1, p. 20, 2012.

[138] M. A. Sartor, G. D. Leikauf, and M. Medvedovic, "LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data," *Bioinformatics*, vol. 25, no. 2, pp. 211–217, 2009.

[139] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow, "PID: the pathway interaction database," *Nucleic Acids Research*, vol. 37, no. Suppl 1, pp. D674–D679, 2009.

[140] M. Schlageter, L. M. Terracciano, S. D'Angelo, and P. Sorrentino, "Histopathology of hepatocellular carcinoma," *World Journal of Gastroenterology*, vol. 20, no. 43, pp. 15 955–15 964, 2014.

[141] A. M. Sciuto and H. H. Hurt, "Therapeutic treatments of phosgene-induced lung injury," *Inhalation Toxicology*, vol. 16, no. 8, pp. 565–580, 2004.

[142] A. M. Sciuto, R. B. Lee, J. S. Forster, M. B. Cascio, D. L. Clapp, and T. S. Moran, "Temporal changes in respiratory dynamics in mice exposed to phosgene," *Inhalation Toxicology*, vol. 14, no. 5, pp. 487–501, 2002.

[143] A. M. Sciuto, C. S. Phillips, L. D. Orzolek, A. I. Hege, T. S. Moran, and J. F. Dillman, "Genomic analysis of murine pulmonary tissue following carbonyl chloride inhalation," *Chemical Research in Toxicology*, vol. 18, no. 11, pp. 1654–1660, 2005.

[144] K. Sha, S.-H. Choi, J. Im, G. G. Lee, F. Loeffler, and J. H. Park, "Regulation of ethanol-related behavior and ethanol metabolism by the corazonin neurons and corazonin receptor in Drosophila melanogaster," *PloS One*, vol. 9, no. 1, p. e87062, 2014.

[145] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[146] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of Escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[147] A. Shojaie, "NetGSA software url," http://www.biostat.washington.edu/~ashojaie/software/netGSA_1.0.tar.gz, 2013, accessed: May 15.

[148] A. Shojaie and G. Michailidis, "Analysis of Gene Sets Based on the Underlying Regulatory Network," *Journal of Computational Biology*, vol. 16, no. 3, pp. 407–426, 2009.

[149] ——, "Network Enrichment Analysis in Complex Experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 9, no. 1, 2010.

[150] P. Spirtes, "Directed cyclic graphical representations of feedback models," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 491–498.

[151] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, no. Suppl 1, pp. D535–D539, 2006.

[152] O. Stegle, K. Denby, D. L. Wild, Z. Ghahramani, and K. M. Borgwardt, "A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series," *Journal of Computational Biology*, vol. 17, no. 3, pp. 355–367, 2010.

[153] 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*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[154] K. N. Sugahara, T. Teesalu, P. P. Karmali, V. R. Kotamraju, L. Agemy, D. R. Greenwald, and E. Ruoslahti, "Coadministration of a tumor-penetrating peptide enhances the efficacy of cancer drugs," *Science*, vol. 328, no. 5981, pp. 1031–1035, 2010.

[155] A. L. Tarca, S. Drăghici, G. Bhatti, and R. Romero, "Down-weighting overlapping genes improves gene set analysis," *BMC Bioinformatics*, vol. 13, no. 1, p. 136, 2012.

[156] 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*, vol. 25, no. 1, pp. 75–82, 2009.

[157] A. L. Tarca, P. Khatri, and S. Drăghici, "SPIA software url," http://bioconductor.org/packages/release/bioc/html/SPIA.html, 2013, accessed: May 15.

[158] J. R. Thompson, P. F. Gerald, M. L. Willoughby, and B. K. Armstrong, "Maternal folate supplementation in pregnancy and protection against acute lymphoblastic leukaemia in childhood: a case-control study," *The Lancet*, vol. 358, no. 9297, pp. 1935–1940, 2001.

[159] M. Unoki and Y. Nakamura, "EGR2 induces apoptosis in various cancer cell lines by direct transactivation of BNIP3L and BAK," *Oncogene*, vol. 22, no. 14, pp. 2172–2185, 2003.

[160] P. Vahteristo, J. Bartkova, H. Eerola, K. Syrjäkoski, S. Ojala, O. Kilpivaara, A. Tamminen, J. Kononen, K. Aittomäki, P. Heikkilä, K. Holli, C. Blomqvist, J. Bartek, O.-P. Kallioniemi, and H. Nevanlinna, "A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer," *The American Journal of Human Genetics*, vol. 71, no. 2, pp. 432–438, 2002.

[161] D. Valletta, B. Czech, T. Spruss, K. Ikenberg, P. Wild, A. Hartmann, T. S. Weiss, P. J. Oefner, M. Müller, A.-K. Bosserhoff, and C. Hellerbrand, "Regulation and function of the atypical cadherin FAT1 in hepatocellular carcinoma," *Carcinogenesis*, vol. 35, no. 6, pp. 1407–1415, 2014.

[162] C. J. Vaske and S. C. Benz, "PARADIGM software url standalone," http://sbenz.github.com/Paradigm, 2013, accessed: May 15.

[163] ——, "PARADIGM software url web-based," https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/, 2013, accessed: May 15.

[164] 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*, vol. 26, no. 12, pp. i237–i245, 2010.

[165] 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*, vol. 1. Boca Raton, FL, USA: IEEE, 12-15 Dec. 2012, pp. 126–131.

[166] C. Voichiţa and S. Drăghici, *ROntoTools: R Onto-Tools suite*, 2013, r package.

[167] C. Voichita, S. Ansari, and S. Draghici, *ROntoTools: R Onto-Tools suite*, 2016, R package version 2.0.0. [Online]. Available: http://www.bioconductor.org

[168] L. Wang, S. Lyu, S. Wang, H. Shen, F. Niu, X. Liu, J. Liu, and Y. Niu, "Loss of FAT1 during the progression from DCIS to IDC and predict poor clinical outcome in breast cancer," *Experimental and Molecular Pathology*, vol. 100, no. 1, pp. 177–183, 2016.

[169] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[170] S. Watanabe, K. Okita, T. Harada, T. Kodama, Y. Numa, T. Takemoto, and T. Takahashi, "Morphologic studies of the liver cell dysplasia," *Cancer*, vol. 51, no. 12, pp. 2197–2205, 1983.

[171] R. C. Waters, P. W. O'Toole, and K. A. Ryan, "The FliK protein and flagellar hook-length control," *Protein Science*, vol. 16, no. 5, pp. 769–780, 2007.

[172] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[173] A. Wise and Z. Bar-Joseph, "SMARTS: reconstructing disease response networks from multiple individuals using time series gene expression data," *Bioinformatics*, vol. 31, no. 8, pp. 1250–1257, 2014.

[174] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, R. VÃązquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii,

S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, and A. Scalbert, "Hmdb 4.0: the human metabolome database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D608–D617, 2017.

[175] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert, "Hmdb 3.0âĂŤthe human metabolome database in 2013," *Nucleic Acids Research*, vol. 41, no. D1, pp. D801–D807, 2012.

[176] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser, "HMDB: the human metabolome database," *Nucleic Acids Research*, vol. 35, no. suppl_1, pp. D521–D526, 2007.

[177] W. C. Wood, D. R. Budman, A. H. Korzun, M. R. Cooper, J. Younger, R. D. Hart, A. Moore, J. A. Ellerton, L. Norton, C. R. Ferree, A. C. Ballow, E. Frei, and I. C. Henderson, "Dose and dose intensity of adjuvant chemotherapy for stage II, node-positive breast carcinoma," *New England Journal of Medicine*, vol. 330, no. 18, pp. 1253–1259, 1994.

[178] Q. Wu, H. Jin, Z. Yang, G. Luo, Y. Lu, K. Li, G. Ren, T. Su, Y. Pan, B. Feng, Z. Xue, X. Wang, and D. Fan, "miR-150 promotes gastric cancer proliferation by negatively regulating the pro-apoptotic gene EGR2," *Biochemical and biophysical research communications*, vol. 392, no. 3, pp. 340–345, 2010.

[179] E. Wurmbach, Y.-b. Chen, G. Khitrov, W. Zhang, S. Roayaie, M. Schwartz, I. Fiel, S. Thung, V. Mazzaferro, J. Bruix, E. Bottinger, S. Friedman, S. Waxman, and J. M. Llovet, "Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma," *Hepatology*, vol. 45, no. 4, pp. 938–947, 2007.

[180] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic Acids Research*, vol. 28, no. 1, pp. 289–291, 2000.

[181] J. Xia, "MetPA software url," http://metpa.metabolomics.ca, 2013, accessed: May 15.

[182] J. Xia, I. V. Sinelnikov, B. Han, and D. S. Wishart, "MetaboAnalyst 3.0—making metabolomics more meaningful," *Nucleic Acids Research*, vol. 43, no. W1, pp. W251–W257, 2015.

[183] J. Xia and D. S. Wishart, "MetPA: a web-based metabolomics tool for pathway analysis and visualization," *Bioinformatics*, vol. 26, no. 18, pp. 2342–2344, 2010.

[184] ——, "Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst," *Nature Protocols*, vol. 6, no. 6, p. 743, 2011.

[185] Z. Yin, M. Gupta, T. Weninger, and J. Han, "A unified framework for link recommendation using random walks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE, 2010, pp. 152–159.

[186] M. Zampieri, K. Sekar, N. Zamboni, and U. Sauer, "Frontiers of high-throughput metabolomics," *Current Opinion in Chemical Biology*, vol. 36, pp. 15–23, 2017.

[187] S. Zhang, C. Xia, C. Xu, J. Liu, H. Zhu, Y. Yang, F. Xu, J. Zhao, Y. Chang, and Q. Zhao, "Early growth response 3 inhibits growth of hepatocellular carcinoma cells via upregulation of fas ligand," *International Journal of Oncology*, vol. 50, no. 3, pp. 805–814, 2017.

[188] Y. Zhao, M.-H. Chen, B. Pei, D. Rowe, D.-G. Shin, W. Xie, F. Yu, and L. Kuo, "A Bayesian Approach to Pathway Analysis by Integrating Gene–Gene Functional

Directions and Microarray Data," *Statistics in Biosciences*, vol. 4, no. 1, pp. 105–131, 2012.

ABSTRACT

QUALITATIVE CHANGE DETECTION APPROACH FOR PREVENTIVE THERAPIES

by

CRISTINA FLORENTINA MITREA

December 2018

Advisor: Dr. Sorin Drăghici

Major: Computer Science

Degree: Doctor of Philosophy

Currently, most diseases are diagnosed only after disease-associated changes have occurred. In this PhD dissertation, we propose a paradigm shift from treating the disease to maintaining the healthy state. The proposed approach is able to identify when systemic qualitative changes in biological systems happen, thus opening the possibility of therapeutic interventions before the occurrence of symptoms. The change detection method exploits knowledge from biological networks and longitudinal data using a system impact analysis approach. This approach is validated on eight datasets, for seven different model organisms and eight biological phenomena. On these data, our proposed method performs well, consistently identifying the qualitative change in each dataset. Most importantly, the method accurately detected the transition from the control stage (benign) to the early stage of hepatocellular carcinoma on an eight-stage disease dataset. Knowing when a transition (qualitative change) from healthy to disease occurs may help preserve the healthy state.

We also propose a novel analysis approach for metabolic pathway analysis that uses an impact analysis approach and leverages the stoichiometry of bio-chemical reactions to identify which pathways are significantly disrupted by the change in metabolite levels in disease samples versus healthy controls. Our approach outperforms the over-representation approach when evaluated on simulated data. We applied our proposed method to biological experiment data that compares samples from pregnant women to non-pregnant control samples. Our

method was able to identify biologically relevant results on real high-throughput data better than the classical approach.

In summary, we developed two novel methods for the analysis of high-throughput biological data, gene expression and metabolite concentration, respectively. The proposed methods can be adapted to work together in order to capture relevant complementary information stored in time-course datasets for gene expression or metabolite levels that may available for complex diseases in order to identify when a qualitative change happens, before the physiological onset of the disease.

AUTOBIOGRAPHICAL STATEMENT

CRISTINA FLORENTINA MITREA

Education

- PhD, Computer Science with concentration in Bioinformatics, Wayne State University, Detroit, MI, USA, Dec. 2018
- MSc, Computer Science Wayne State University, Detroit, MI, USA, Dec. 2012
- BSc, Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania, Jul. 2005

Publications

1. Teslow EA, Bao B, Dyson G, Legendre C, **Mitrea C**, Sakr W, Carpten JD, Powell I, Bollig-Fischer A. Exogenous IL-6 induces mRNA splice variant MBD2_v2 to promote stemness in TP53 wild-type, African American PCa cells. *Molecular Oncology* 2018;12(7):1138-1152.

2. **Mitrea C**, Wijesinghe P, Dyson G, Kruger A, Ruden DM, Drăghici S, Bollig-Fischer A. Integrating 5hmC and gene expression data to infer regulatory mechanisms. *Bioinformatics*. 2018;34(9):1441-1447.

3. Bao B, **Mitrea C**, Wijesinghe P, Marchetti L, Girsch E, Farr RL, Boerner JL, Mohammad R, Dyson G, Terlecky SR, Bollig-Fischer A. Treating triple negative breast cancer cells with erlotinib plus a select antioxidant overcomes drug resistance by targeting cancer cell heterogeneity. *Scientific Reports*. 2017;7:44125.

4. Shafi A, **Mitrea C**, Nguyen T, Drăghici S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in Bioinformatics*. 2017.

5. Nguyen T, **Mitrea C**, Tagett R, Drăghici S. DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions-applied to biological pathway analysis. *Proceedings of the IEEE*. 2017;105(3):496-515.

6. Nguyen T, Tagett R, Donato M, **Mitrea C**, Drăghici S. A novel bi-level meta-analysis approach: applied to biological pathway analysis. *Bioinformatics*. 2016;32(3):409-16.

7. Bollig-Fischer A, Marchetti L, **Mitrea C**, Wu J, Kruger A, Manca V, Drăghici S. Modeling time-dependent transcription effects of HER2 oncogene and discovery of a role for E2F2 in breast cancer cell-matrix adhesion. *Bioinformatics*. 2014;30(21):3036-43.

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

9. Tin Nguyen T, **Mitrea C**, and Drăghici S. Network-Based Approaches for Pathway Level Analysis. *Current Protocols in Bioinformatics*, 61(1): 8.25.1-24, 2018. (book chapter)