DISS. ETH NO. 25633

Exploring multi-modal learning approaches towards precision medicine

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

Matteo Manica

Laurea Magistrale in Ingegneria Matematica, Politecnico di Milano
born on September 11, 1988
citizen of Italy

accepted on the recommendation of

Prof. Dr. Rudolf Aebersold
Dr. María Martínez Rodríguez
Prof. Dr. Manfred Claassen
Prof. Dr. med. Peter J. Wild

2018

to Alessandra, my family, my friends and my colleagues :)

Summary

Recent years have testified unprecedented advances in the field of molecular systems biology. The increasing amount of data produced from disparate sources is giving us the possibility to study biological systems from a wide variety of angles at an incredibly fine scale. In this context keeping up with the pace of data production and fully exploiting the availability of information from multiple modalities is fundamental.

In this work, methods to extract information from different data types are investigated and developed in order to improve complex diseases understanding and develop explainable personalized models for patient stratification. The different data modalities considered ranging from molecular data, such as genomics, transcriptomics, and proteomics to data from the literature, either in form of natural language from publications or structured data from databases. The common denominator to handle these different modalities is machine learning based on graph topologies summarizing molecular interactions that provide a high-level representation of the molecular processes governing cells behavior. The main focus is the study of these molecular interaction networks and their potential applications to personalized medicine.

The thesis is structured around three main pillars: network reconstruction, integration of network information in interpretable machine learning algorithms and the development of personalized models. Network reconstruction is analyzed on two data modalities: on molecular data by implementing state-of-the-art inference models and investigating consensus strategies to ensure robust prediction and on natural language proposing a novel deep learning-based methodology. Integration of network information into machine learning models is tackled by making use of a multiple kernel learning algorithm that exploits *pathway-induced* kernels, a concept introduced in this work. Going towards patient personalized models, networks are also exploited in a dynamic perspective to perform large-scale logical modeling through the implementation of a framework for accelerated attractor analysis. Model personalization is also tackled at genomic level by considering two data modalities: copy number alterations and somatic mutations. These data are used to feed a novel inference algorithm, implemented during this work, to estimate patient-specific phylogenetic trees.

Zusammenfassung

Die letzten Jahre haben beispiellose Fortschritte auf dem Gebiet der molekularen Systembiologie gezeigt. Die zunehmende Menge an Daten aus unterschiedlichen Quellen gibt uns die Möglichkeit, biologische Systeme aus einer Vielzahl von Blickwinkeln und in einem unglaublich feinen Maßstab zu untersuchen. In diesem Zusammenhang ist es von grundlegender Bedeutung, mit der Geschwindigkeit der Datenproduktion Schritt zu halten und die Verfügbarkeit von Informationen aus mehreren Modalitäten vollständig zu nutzen.

In dieser Arbeit werden Methoden zur Extraktion von Informationen aus verschiedenen Datentypen untersucht und entwickelt, um das Verständnis komplexer Krankheiten zu verbessern und erklärbare personalisierte Modelle für die Patientenstratifizierung zu entwickeln. Die verschiedenen Datenmodalitäten reichen von molekularen Daten wie Genomik, Transkriptomik und Proteomik bis hin zu Daten aus der Literatur, entweder in Form natürlicher Sprache aus wissenschaftlichen Publikationen oder strukturierter Daten aus Datenbanken. Der gemeinsame Nenner im Umgang dieser unterschiedlichen Modalitäten ist maschinelles Lernen auf der Grundlage von Graphentopologien, mit welchen molekulare Wechselwirkungen zusammengefasst und eine hochrangige Darstellung der molekularen Prozesse, welche das Verhalten von Zellen steuern, ermöglicht werden. Der Schwerpunkt liegt auf der Erforschung dieser molekularen Interaktionsnetzwerke und ihrer möglichen Anwendungen für die personalisierte Medizin.

Die Arbeit gliedert sich in drei Hauptpfeiler: Netzwerkrekonstruktion, Integration von Netzwerkinformationen in interpretierbare maschinelle Lernalgorithmen und die Entwicklung personalisierter Modelle. Die Netzwerkrekonstruktion wird an zwei Datenmodalitäten analysiert: an molekularen Daten durch Implementierung modernster Inferenzmodelle und Untersuchung von Konsensstrategien, um eine robuste Vorhersage zu gewährleisten, und an natürlicher Sprache, mit einer neuartigen auf Deep Learning basierender Methodik. Die Integration von Netzwerkinformationen in maschinelle Lernmodelle wird durch die Verwendung eines Multiple-Kernel-Lernalgorithmus mit Integration von Signalweg-induzierten Kerneln angegangen, ein Konzept, das in dieser Arbeit eingeführt wurde. Mit Blick auf personalisierte Patientenmodel-

le werden Netzwerke zudem in einer dynamischen Perspektive genutzt: Durch die Implementierung eines Frameworks für eine beschleunigte Attraktoranalyse wird eine umfangreiche logische Modellierung ermöglicht. Die Personalisierung von Modellen wird auch auf genomischer Ebene angegangen, indem zwei Datenmodalitäten berücksichtigt werden: Kopienanzahlveränderungen und somatische Mutationen. Diese Daten werden mit einem neuartigen Inferenzalgorithmus, der während dieser Dissertation implementiert wurde, verarbeitet, um patientenspezifische phylogenetische Bäume zu schätzen.

Acknowledgements

I would like to start by expressing my most sincere gratitude to my IBM advisor Dr. María Rodríguez Martínez, for giving me the opportunity to work in her amazing research group and helping me to grow as a scientist and as a person. Each discussion and meeting we had always pushed me to dive in-depth in all the research areas I decided to explore during my PhD studies. I want also to explicitly thank her for all the freedom she gave me during these years, and for helping me setting some boundaries to my unbounded curiosity. Without her advice in pushing the brake sometime I wouldn't have been able to accomplish my goals with the same efficacy.

I would also like to thank my ETH advisor Prof. Dr. Rudolf Aebersold for his valuable support in all the meetings we had. His incredible experience and immense knowledge of molecular biology helped me a lot to shape my research around the right questions and find the right applications for my, initially vague, algorithmic ideas. His availability to discuss, given his tight schedule, and put me contact with anyone in the lab boosted my progresses and contributed in a major way to generate relevant scientific results during my PhD.

Besides my advisors, special thanks go to the rest of my thesis committee: Prof. Dr. Manfred Claassen and Prof. Dr. med. Peter J. Wild, for their useful comments and suggestions during the committee meetings. Their insightful observations and their tough questions helped me in a consistent way to shape my research.

There are some colleagues I would like to specially thank. Ali, for the projects we worked on together, the late night working sessions and all the funny discussions about late night shows. Sunil, for the hardcore statistical discussions we had when we were sharing the same side of the office and for his illuminating approach to life. Jelena, for her help and collaboration during the work in the PrECISE project and for taking always time to explain to me experimental procedures over and over again. Elina, for her continuos support and the scientific discussions, especially for helping me to understand clinical implications of the algorithms developed. I would like also to thank her for the help and the work done for organizing our committee meetings and a lot of activities for IBM's Systems Biology group. Joris, for being

an incredibly talented guinea pig and for the amazing work he did during his master thesis under my supervision. I'm sure he'll become an amazing scientist and even more amazing programmer (assuming he will learn how to name variables properly). Antonio, for being "le meilleur des collègues possibles" in "les meilleur des mondes possibles" to cite Leibniz through Voltaire. Our breakfast dissertations over linguistic, TV series and his drawings are among the best things I'll remember from this experience. Roland, for the incredible amount of work we did, time we spent, after-midnight dinners we had at Singapore and ideas we discussed during his stay in IBM. His incredible attitude and the the reciprocal understanding while working on our projects is something that I never experienced before and I think is one of the main reasons behind the success of this work. I also have to thank him for his advice and the help he gave him in prioritizing and organize the work. Marianna, for everything she has done during the last years to help me succeed. She helped me to grow on a multitude of levels: as a scientist with our amazing discussions and as a person with all her lessons on how to handle people and properly addressing conflicts. I really consider her much more than a valuable colleague and I don't exaggerate when I say that I see her like my older greek sister. A sincere thanks goes also to past and current members of the Systems Biology Group in IBM, for all the exciting discussions, the great time, the help and their valuable feedback: Stephan, Jonas, Marcel, Amelia, George, Jannis, Cristina, An-phi, Cyril, Özgen, Pierre and Charlotte. I would also like to thank some people from other IBM groups that I had the pleasure to meet and work with. Raul, for the geeky chats about programming and the interesting scientific discussions we had. Matthew, for all suggestions about prog music and the concerts we attended together. Cristiano, for the help and the appreciation he always demonstrated while working together in multiple projects. Peter, for the opportunity to work with his group and the trust he always showed when collaborating on various projects. Costas, my previous manager for his guidance and the precious suggestions on how to move inside IBM and all the exciting opportunities for collaborations he always promoted. Maria, my current manager for her valuable guidance, our technical discussions and her help in making me understand what does working in IBM mean and, above all, for the great work she is doing in directing the Cognitive Heath Care and Life Science department.

A special thank goes to all my family. My grandfather, Eugenio, that unfortunately passed away last year. He was the only medical doctor in my family that was proud of me studying engineering instead of going to medical school. Knowing I studied what he always wanted to, warms my hearth every time I think about it. My mother, Elvira, for the discussions about oncology and her unconditioned support even if I never went to medical school. My brother, Manfredi, one of the smartest people I know and one of the most unexpressed mathematical

talents I've seen. I'm sure he will soon be one the greatest italian golf players, he's already the best as far as I'm concerned. My aunt, Rosilde, for all the political and musical discussions that helped me to keep up with the world outside of the laptop screen. My cousin, Eugenia, for being my actual older sister and one of the most supportive people I had around during my whole life. My father, Mauro, even if we don't spend much time together he is probably one of my greatest supporters and he is extremely proud of my accomplishments.

Last but absolutely not least, a billion of thanks to Alessandra. I will never find a way to express enough gratitude to her and I can't express in a few words what she and her presence meant for me during these years. She has been, is and I know will always be there supporting me and I'm sure I wouldn't have been able to fulfill this goal without her. This thesis is dedicated to her and represents the reward for the sacrifice we did more than two years ago when we had to stop to live together and start this new chapter of our life. Without her any of this would make sense, and I'm grateful every day knowing she is by my side and having the chance to do the same for her.

Contents

Part IV Appendix

Abbreviations

ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
AUC	Area Under the Curve
AUROC	Area under the Receiver Operating Characteristic
BIC	Bayesian Information Criterion
BMC	Boolean network model circuit
BRCA	Breast Cancer
CAC	Colitis-Associated Colon cancer
CAPI	Coherent Accelerator Processor Interface
CLB	Configurable Logic Blocks
CLR	Context Likelihood of Relatedness
CNA	Copy Number Alteration
CNV	Copy Number Variations
COSIFER	Consensus Interaction Network Inference Service
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CRPC	Castration Resistant Prostate Cancer
DPI	Data Processing Inequality
DREAM	Dialogue on Reverse Engineering Assessment and Methods
EGA	European Genome-phenome Archive

EMT	Epithelial-Mesenchymal Transition
EXPANDS	Expanding Ploidy and Allele Frequency on Nested Subpopulations
FA/BRCA	Fanconi Anemia/Breast Cancer
FFPE	Formalin-Fixed Paraffin-Embedded
FPGA	Field-Programmable Gate Array
FunChiSq	Functional χ^2-Test
GENIE3	Gene Network Inference with Ensemble of Trees
GLasso	Graphical LASSO
GRN	Gene Regulatory Network
GUI	Graphical User Interface
HBV	Hepatitis B Virus
HCC	Hepatocellular Carcinoma
HDL	Hardware Description Language
INtERAcT	Interaction Network Inference from Vector Representations of Words
JRF	Joint Random Forest
KOMD	Kernel method for the Optimization of the Margin Distribution
LARS	Least Angle RegreSsion
LFSR	Linear-Feedback Shift Register
LGL	Large Granular Lymphocyte
LUT	Look-Up Table
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MKL	Multiple Kernel Learning

MRMR	Minimum Redundancy Maximum Relevance
MRNET	Minimum Redundancy/Maximum Relevance Networks
MSigDB	Molecular Signatures Database
NCI	National Cancer Institute
NER	Named Entity Recognition
NLP	Natural Language Processing
ODE	Ordinary Differential Equation
PC	Prostate Cancer
PIMKL	Pathway Induced Multiple Kernel Learning
PPI	Protein–Protein Interaction
PRAD	Prostate Adenocarcinoma
PSL	Property Specification Language
ROBDD	Reduced Ordered Binary Decision Diagrams
ROC	Receiver Operating Characteristic
SCHISM	Sublonal Hierarchy Inference from Somatic Mutations
SLSQP	Sequential Least Squares Programming
SNV	Single-nucleotide somatic variant
SUMMA	Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions
T-LGL	T cell type of Large Granular Lymphocyte
TCGA	The Cancer Genome Atlas
TIGRESS	Trustful Inference of Gene Regulation Using Stability Selection
UZH	University Hospital of Zürich

WES	Whole-Exome Sequencing
WOC	Wisdom Of Crowds
WT	Wild Type

Part I.

Introduction and Background

1 Introduction

> *Philosophy [i.e., physics] is written in this grand book — I mean the Universe — which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics, and its characters are triangles, circles, and other geometrical figures, without which it is humanly impossible to understand a single word of it; without these, one is wandering around in a dark labyrinth.*
>
> – Galileo Galilei, *The Assayer*

In recent years molecular biology research has seen a continuos transformation, especially from the late nineties onwards, with the advent of systems biology, the study of biological systems by means of mathematical and computational models has started to be widely adopted. This is testified by the proliferation of consortium efforts that aim to collect large amounts of data in order to be able to identify and measure all the actors and their complex interactions in a molecular system. Some of the most prominent examples of this trend are: 1000 Genomes Project [1], Encode [2], Roadmap Epigenomics Project [3], Blueprint [4] or more recently The 100000 Genomes Project [5].

The ability to measure molecular entities in a high-throughput fashion at multiple omic levels, together with the increasing availability of publicly available datasets has allowed us to improve our understanding of complex cellular processes governing cell behavior. Modeling approaches are believed to be fundamental in the study of complex diseases, such as cancer [6], diabetes [7] or Alzheimer's disease [8]. Cancer research consortium efforts, like TCGA (The Cancer Genome Atlas) [9] or CPTAC (Clinical Proteomic Tumor Analysis Consortium) [10], provide the community with an extremely valuable source of data that has helped to improve the understanding of key mechanisms behind the disease's development.

The knowledge generated as a by-product of these large consortium studies, by the project

itself and from all the subsequent works, represents another priceless source of information and deserves to be considered as an additional data source. These data are divided in two broad categories corresponding to different modalities: natural language, in the form of papers or research reports, and databases (e.g., molecular interaction databases, drug databases, mutation databases, etc.).

In this context of multiple data availability the need for integrative approaches able to combine disparate data sources guaranteeing robustness to noise is becoming fundamental.

The research work presented here aims to leverage current existing methods in heterogeneous data integration and proposes new holistic methodologies, with a focus on network approaches for interpretable patient stratification and precision medicine.

Thesis outline

This cumulative dissertation is structured around a selection of manuscripts that are either in the preparation phase or already submitted to peer-reviewed journals. The common link between all the works is the integration of data modalities and network approaches to answer specific biological questions related to systems biology research with a strong focus on cancer research. This thesis begins with a description of the research fields where the different scientific contributions are made. The scientific contributions are grouped and are presented by including the related manuscripts adapting their pre-print version.

The first chapters report the research conducted on network reconstruction from multimodal data sources. In Chapter 2, the COSIFER (Consensus Interaction Network Inference Service) manuscript is included. COSIFER is a service that implements a selection of state-of-the-art methods for network reconstruction from molecular data and implements various consensus strategies to integrate predictions from different algorithms and/or data types. Chapter 3 includes the INtERAcT (Interaction Network Inference from Vector Representations of Words) manuscript. INtERAcT is a novel approach based on vector representation of words obtained through deep learning that allows an accurate estimation of molecular interactions from topic-specific text corpora.

In the following chapters a novel method and a technical achievement, exploiting interaction network topologies for patient stratification and personalized modeling respectively, are presented. In Chapter 4, the PIMKL (Pathway Induced Multiple Kernel Learning) paper is included. PIMKL is an algorithm based on multiple kernel learning that introduces the con-

cept of *pathway-induced* kernels to classify a phenotype (e.g., stratify patients), allowing us to recover single pathway contributions in the problem considered and ultimately providing, depending on the meaning of the selected pathways, an interpretable classification framework. Chapter 5 includes a manuscript describing the technical work performed in hardware acceleration for logical modeling simulation and attractor analysis exploiting FPGA cards. The computational architecture proposed enables Boolean model simulations to be potentially scaled up to network including thousands of nodes, permitting the simulation of genome-wide patient profiles and the analysis of personalized responses to network perturbation. The last manuscript included, in Chapter 6, presents a pure personalized medicine approach called Chimaera for patient-specific tumor clonal architecture estimation using multiple biopsies.

Concluding remarks are divided into three chapters. A discussion regarding the major results obtained and lessons learned during this research project, with considerations about further extensions of the most promising methodologies, is presented in Chapter 7. Chapter 8 describes, in detail, contributions and copyright notes for each manuscript included. Finally in Chapter 9, the Appendix, the list of figures and tables included in the thesis are reported.

Background

This section presents the current state of the fields that have been the subject of investigation in this thesis work and describes how the research activities carried out during this PhD propose to overcome current limitations and to provide algorithms and services that are useful for the whole systems biology community.

Network reconstruction

Molecular processes are regulated by the coordinated action of multiple factors. Genomics, transcriptomics and proteomics are the three main layers that are orchestrated to determine cellular function. The ways in which these different levels interact with each other and within themselves determine many aspects of cell behavior: from cell proliferation to cell death, from cell cycle to cholesterol metabolism. These interactions can be summarized in a graphical structure, in what is called an interaction network. Representing interactions between macromolecules, like gene sequences, proteins, metabolites can appear as a mere simplification of a complex cellular environment, but provides us with a high-level view of the system that can help us in discovering and understanding molecular processes happening in a living

cell.

Network approaches can also help us to study in a holistic way phenomena beyond a single-cell system, scaling up to arbitrarily elaborate systems, like human diseases [11]. In an effort to collect multi-layer interactions in so called "interactomes" we have seen, in recent years, a large number of databases being published and made accessible to the community: STRING [12], OmniPath [13], Reactome [14, 15], KEGG [16–18] or Pathway Commons [19], to name a few. These resources condense a tremendous amount of knowledge coming from experimental validation through manual curation of literature and text mining approaches.

Besides these efforts of collecting known interactions in structured sources, the need to infer unknown interactions in a more or less automatic way has pushed many researchers into investigating the problem of inferring molecular interactions from different data types. This problem is also known as network reconstruction or network inference, and is based on building a graph of interacting entities based on evidence collected from the data. Data used in network inference can come from different modalities, for example: a set of single/multi-omic measurements, like RNASeq or CNV data, a representation obtained from text sources, via text mining, or structural molecular properties, like nucleotide sequences or amino acid sequences.

In this thesis the focus is on unsupervised methods to infer networks from molecular measurements (Figure 1.1) and from natural language (Figure 1.2).

Inference from molecular data

The advent of high-throughput technologies, such as SWATH-MS or RNASeq, made available large scale, i.e., genome-wide or proteome-wide, measurements of molecular entities providing a high resolution portrait of a cell's internal state. Inferring how genes, transcripts and proteins interact from these datasets is a daunting task that many consortiums and research groups have tackled over the last decade. A tangible proof of this interest is the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project that through its challenges gathered the efforts of a multitude of research groups to infer both relevance and causal networks from *in silico* and cancer cell line datasets [20, 21].

Several methodologies have been proposed over the years, methods based on correlations with FDR (False Discovery Rate) correction [22], on mutual information [23–25], on regression [26, 27] on functional hypothesis testing [28] or on ensemble learning on trees [29, 30]. In this work the focus is on reconstruction of relevance networks, represented as undirected

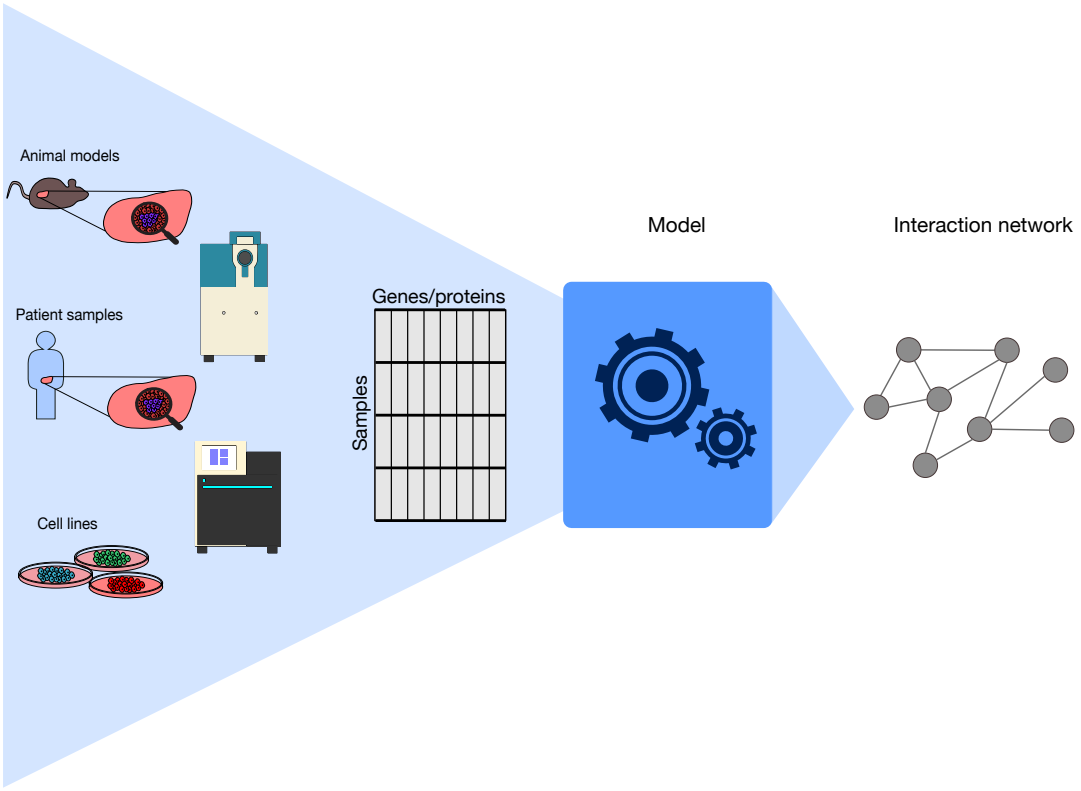


Figure 1.1: **Reconstructing networks from molecular data.** Schematic representation of a network reconstruction problem from molecular data. Samples generated with different experimental designs are measured and used to compile data tables fed to network inference models.

weighted graphs, given their flexibility to summarize relations between entities, and the large availability of suitable data to use in the inference. Indeed any high-throughput omic dataset can be used to infer the topology in a specific condition, as opposed to causal networks where a time series or perturbation experiments are needed to recover the directed topology.

COSIFER [1], presented in Chapter 2, represents an attempt to collect, in a single open access web-service, the most effective algorithms for relevance network inference and to provide the community with a platform that eases the access to these resources, something that is currently missing and might be used advantageously by many researchers. COSIFER also implements multiple consensus strategies to integrate the networks reconstructed using the single methods: Wisdom of Crowds (WOC) as adopted in Marbach *et al.* [20], a modification of the WOC proposed in our work and SUMMA [31] an unsupervised methodology for aggregating weighted contributions from inference algorithms.

[1] `https://sysbio.uk-south.containers.mybluemix.net/cosifer/`, as of November 2018

Inference from publications

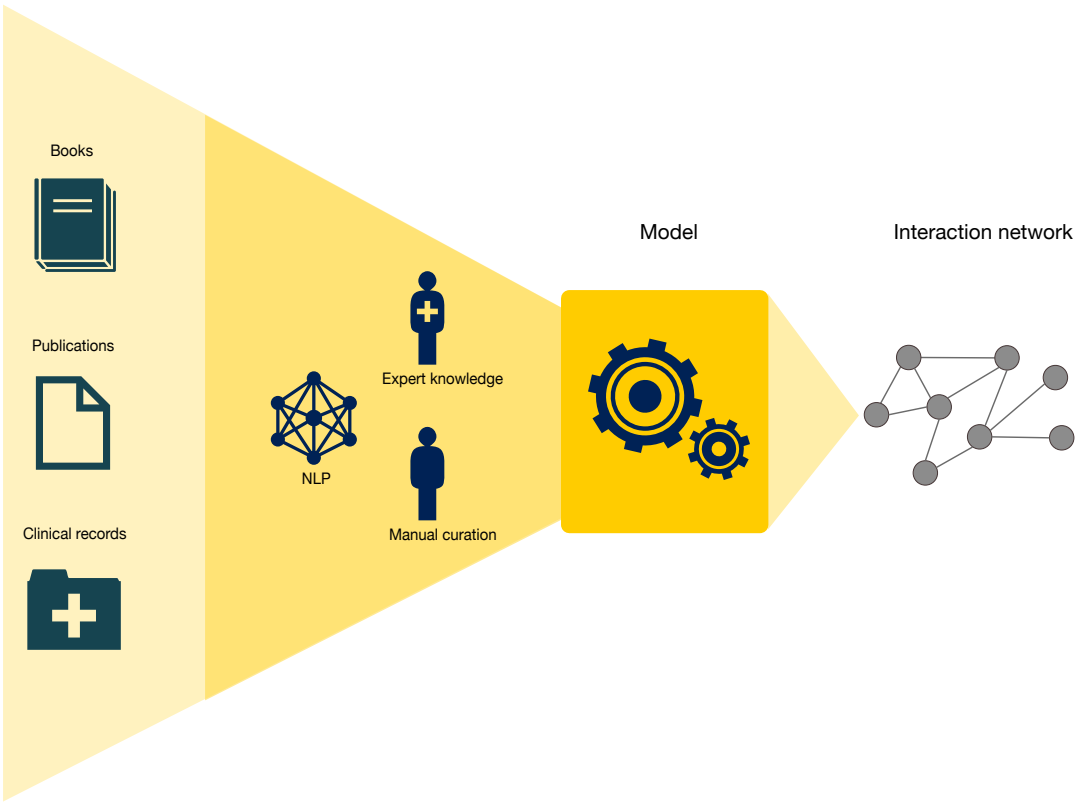


Figure 1.2: **Reconstructing networks from literature.** Schematic representation of a network reconstruction problem from literature. Heterogenous sources of knowledge are available and different approaches can be adopted to feed models for network inference.

Advances in measurement technologies have led to a better understanding of how molecular systems behave and regulate cell functions. Indeed, the development in data acquisition technologies has directly translated into a large amount of newly annotated interactions reported in as many publications We can observe how many papers are indexed on PubMed [1] [32] for topics like cancer (3,743,058) or diabetes (642,967) to have a rough idea of the number of publications produced. A thorough manual curation of all the research produced is clearly unfeasible and this is the main reason behind the proliferation of text mining and natural language processing (NLP) approaches.

In recent years different methodologies to semi-automatically extract interaction from literature have been proposed. Most methodologies fall into three broad categories: machine learning-based, usually characterized by supervision or semi-supervision [33, 34]; terms co-

[1] `https://www.ncbi.nlm.nih.gov/pubmed/`, as of November 2018

occurrence-based, unsupervised but lacking semantic information [35–38]; and rule-based [39]. The main issue with most of the aforementioned approaches is the need for supervision or semi-supervision coming in the form of manual or semi-automatic annotation steps that require expert or domain-specific knowledge. This pitfall partially defies the purpose of skipping manual curation by moving the burden to manual annotation.

INtERAcT [40], see Chapter 3, overcomes this problem by leveraging recent advances in deep learning text analysis [41, 42] and exploits vector representation of words generated on a corpus of interest to score interactions between molecular entities in a completely unsupervised way. The metric proposed in this work can be used on corpora of limited sizes and makes use of word vectors to add contextual information to words representing genes and proteins, exhibiting a high agreement with prior knowledge coming from STRING [43]. To ease the accessibility of the algorithm, INtERAcT has been deployed as an open access web service on IBM Cloud [1].

Interpretable patient stratification

Building interpretable models for patient stratification based on molecular data is an increasingly popular topic nowadays. Developing models with an high level of interpretability together with robustness to noise is a fundamental step towards adoption in a clinical setting. Stable models can also enable complex biomarker identification and, in the case where explainable approach is used, help shed light on molecular processes characterizing a specific phenotype, such as cancer grade or patient survival (see Figure 1.3).

Particularly successful models in this respect are the ones that make use of existing prior knowledge to inform their predictions and drive the selection of most relevant features [44]. An approach usually adopted to represent this prior knowledge is making use of molecular interaction networks to encode relations between the measured entities in the model by means of the underlying graph topology. These methods leverage known interactions between molecular features to create meta-features usually related to a specific set of molecular entities or pathways, thus providing two main advantages. Firstly, we have a direct link between phenotypic traits and well defined mechanisms, considerably increasing our understanding of predicted outcomes. Secondly, we aggregate and transform features by reducing the number of parameters to be estimated. This aspect is extremely valuable in tackling a common issue in patient stratification using molecular measurements: fitting models on datasets characterized

[1] `https://sysbio.uk-south.containers.mybluemix.net/interact/`, as of November 2018

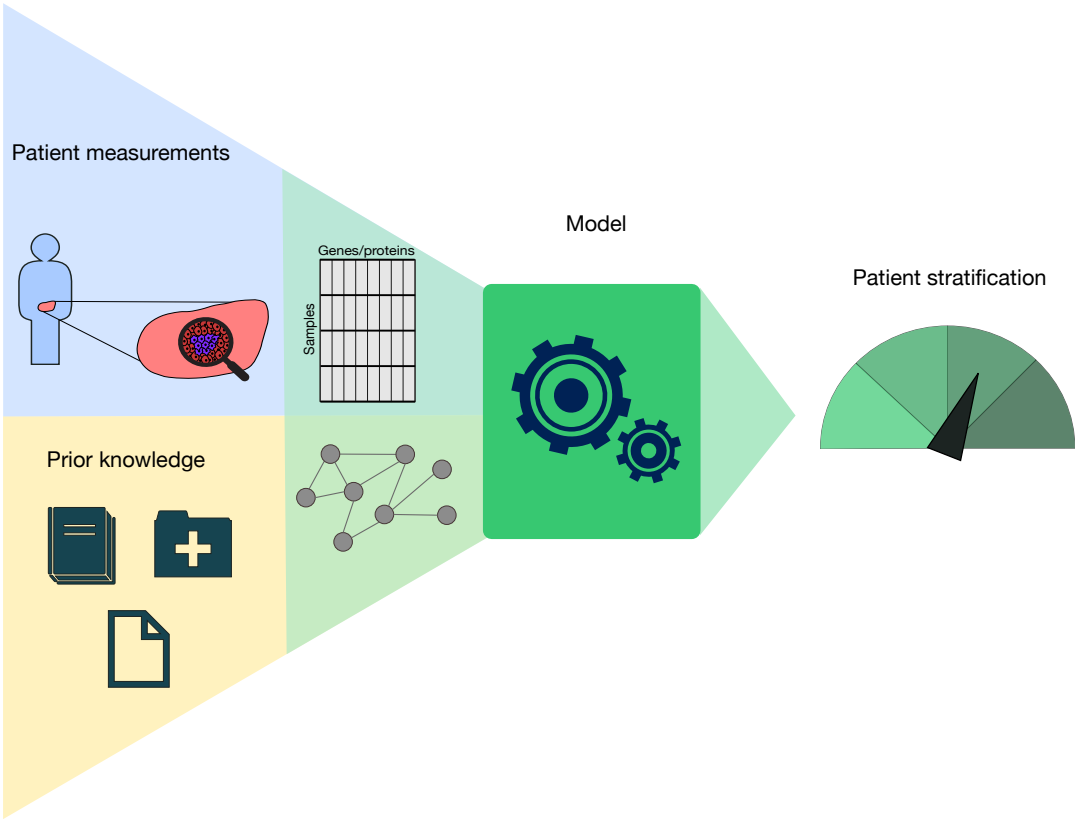


Figure 1.3: **Interpretable patient stratification.** Schematic representation of prior knowledge-informed patient stratification. Predictive models able to integrate knowledge about gene-gene, gene-protein or protein-protein interactions with molecular data to stratify patients into relevant groups.

by few observations, usually in the order of hundreds samples, and a large number of features, in the order of thousands like RNASeq (~17,000-20,000 genes) or SWATH-MS (~3,000-5,000 proteins).

An interesting benchmark of these models was assembled by Cun and Fröhlich [45], where a selection of machine learning algorithms exploiting pathway annotation and molecular interaction information have been compared in terms of predictive performance and molecular signature stability. The benchmark's results highlight how the way prior knowledge is integrated doesn't have a major impact on performance in terms of sensitivity and specificity, but has a great influence on the robustness of the molecular signature estimated. Another aspect to examine for the considered models is the lack of a natural extension to integrate multi-omic data, or more in general multi-modal data.

With PIMKL [46], see Chapter 4, we propose a model able to integrate prior knowledge

on molecular interactions that at the same time is capable of increasing prediction performance and integrating data from multiple modalities. PIMKL achieves this goal by using multiple kernel learning [47], an approach that has been already successfully adopted for drug sensitivity prediction [44] and that is extremely effective in integrating heterogeneous data modalities [48]. In the paper we introduce the concept of *pathway-induced* kernels and we implement a high performance implementation of EasyMKL [49], to handle arbitrarily large numbers of kernels. PIMKL has been made available to the community as an open access service [1].

Modeling for precision medicine

Precision medicine is an emerging field that is attracting the interest of many researchers. The term is used to indicate a therapeutic approach consisting in tailoring treatment to patient subpopulations, or even to a single patient, see Figure 1.4. The idea of precision or personalized medicine is that by collecting information at various omic levels for a patient we are able to determine specific traits that help designing an optimal therapy, that maximizes efficacy and it is cost effective.

In the last few years we have witnessed a trend of investing in this field, an example is the Precision Medicine Initiative, announced in January 2015 by former U.S. president President Barack Obama [50], now named All of Us Research Program [2]. The study's goal is to gather data from more than one million U.S. citizens and make it available for a broad research community, in an effort to better define the variability between individuals and accelerating adoption of personalized medicine approaches in clinical practice. Precision medicine has been proven successful in increasing our understanding of multiple diseases [51]: cystic fibrosis [52–55], melanoma [56–58], head and neck squamous cell carcinoma [59], prostate cancer [60].

Various modeling approaches have been proposed in a precision medicine perspective. In this work the focus is on logical modeling, specifically acceleration of Boolean model simulations, and clonal composition inference from tumor biopsies.

[1] `https://sysbio.uk-south.containers.mybluemix.net/pimkl/`, as of November 2018
[2] `https://allofus.nih.gov/`, as of November 2018

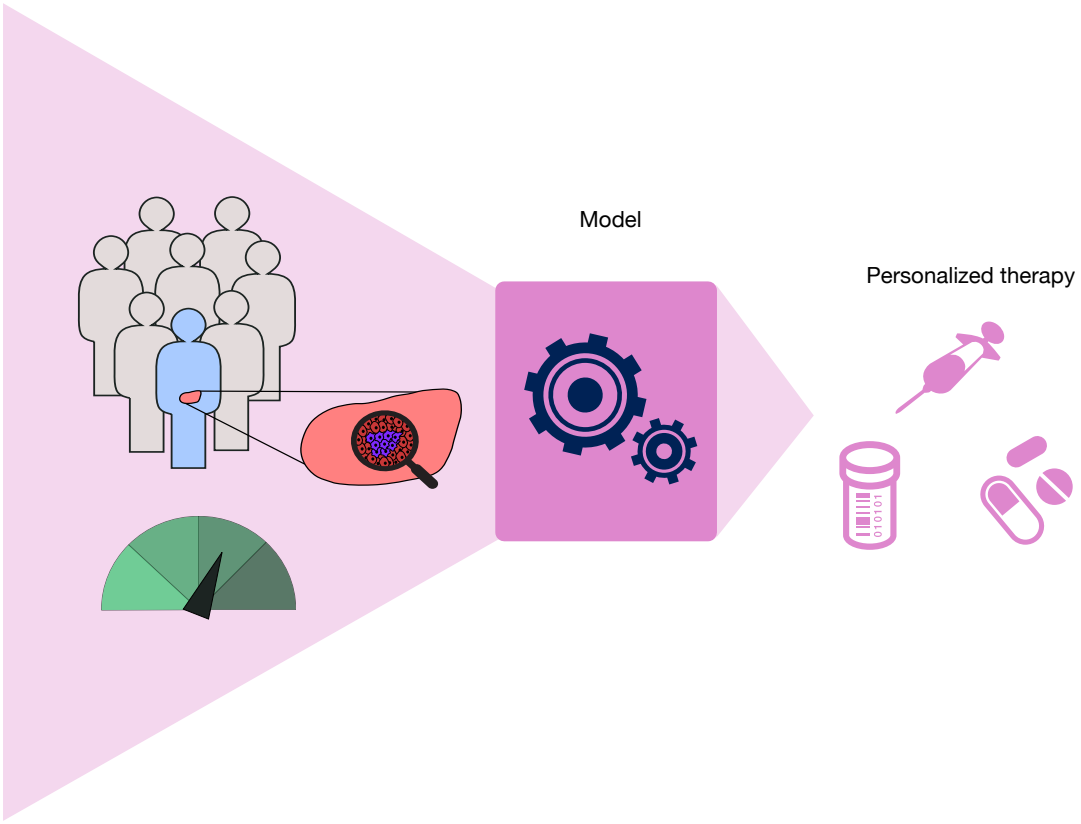


Figure 1.4: **Precision medicine paradigm.** Schematic representation of a precision medicine approach. Characterizing variation between patients using omic data and appropriately identifying the sub-type or sub-population of origin inform models to define a personalized treatment.

Boolean models

Boolean models are used to simulate dynamics in directed interaction networks. They are a good compromise between static network analysis and more complex ODE-based models [61]. In some cases they have also been able to reproduce results in close agreement to more complex ODE-based models [62]. Boolean models are mostly parameter-free, this makes them extremely flexible in providing useful insights about regulatory dynamics [63] and experiment design [64]. In addition to simulation of a biological system Boolean models also enable attractor analysis. By analyzing a model's steady states, it has been shown how we can improve our understanding of specific phenotype traits observed in the system of interest [65, 66].

The logical modeling community is currently active in increasing the size and the range of

the systems analyzed, since this can really increase our understanding of complex biological systems behavior. Unfortunately, simulation and attractor analysis in logic-based models are two problems strictly dependent on the size, in terms of nodes and edges, of the interaction network considered. Most of the currently available software tools are not capable of handling the computational burden imposed by simulating system dynamics in large networks (over hundreds of nodes).

To tackle this problem, in Chapter 5, a Boolean model simulation accelerator based on FPGA is presented. By leveraging the massive parallelism of the problem considered and the architecture of an FPGA card, the computational framework presented exhibits a consistent speedup compared to software simulations and attractor analysis. The results obtained on published models presented show, in principle, how our approach can extend analysis capability up to genome-wide scale solving one of the main issues that is currently present in the field.

Clonality inference from tumor biopsies

Advances in sequencing technologies helped to identify genetic mutations that recur in different tumor types [67]. Despite finding these mutations becoming easier, interpreting them is an open challenge. Deciphering tumor heterogeneity through clonal evolution can help us identify alterations responsible for tumorigenesesis, refractory and proliferative subclones [68, 69]. Information about clonal composition and evolution helps in understanding tumor's potential for metastasis or drug resistance, since they are subclone dependent and play an important role in patient survival [70]. Models to infer a tumor's clonal structure are therefore fundamental in determining a personalized treatment that targets specific subclones and may significantly alter, positively, a patient's disease prognosis. These methods can be divided into two broad categories depending on the data they rely on: single-cell-level mutation data or tumor-level mutation data.

In this work we focus on the second category of methods. The aim is to reconstruct clonal evolution from single-nucleotide somatic variants (SNVs) by deconvolving mutation frequencies from molecular profiles. A commonly used approach is to sequence multiple biopsies from the same tumor across time [71] or across spatial locations [70, 72], and find co-occurring mutations that can be associated with subclones consistently appearing in different biopsies. One of the main limitations of this approach comes into play when dealing with extremely unstable genomes, characterized by the presence of strong copy number alterations (CNA), a

behavior observed in many tumors, like prostate cancer or hepatocellular carcinoma. In these cases mutation frequencies might be wrongly estimated and cause inaccurate estimations of the clonal composition.

In Chapter 6, a manuscript presenting our methodology to perform clonality inference in unstable genomes, Chimaera [73], is included. Chimaera corrects SNVs frequencies by accounting for copy number alterations when determining tumor subclones. Validation on different synthetic datasets to compare our method with other methodologies [74–77] proves its efficacy in estimating tumor clonal composition independently from the genomic instability. Chiamera has also been applied to different tumor types exhibiting its potential to provide patient-specific therapeutic indications. Chimaera is available as an open access service on IBM Cloud [1].

[1] `https://sysbio.uk-south.containers.mybluemix.net/chimaera/`, as of November 2018

References

[1] A. Auton, G. R. Abecasis, D. M. Altshuler *et al.* "A global reference for human genetic variation". *Nature* **526**:7571 (2015), pp. 68–74. arXiv: 15334406.

[2] I. Dunham, A. Kundaje, S. F. Aldred *et al.* "An integrated encyclopedia of DNA elements in the human genome". *Nature* **489**:7414 (2012), pp. 57–74. arXiv: 1111.6189v1.

[3] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman *et al.* "Integrative analysis of 111 reference human epigenomes". *Nature* **518**:7539 (2015), pp. 317–329.

[4] J. M. Fernández, V. de la Torre, D. Richardson *et al.* "The BLUEPRINT Data Analysis Portal". *Cell Systems* **3**:5 (2016), 491–495.e5.

[5] M. Caulfield, J. Davies, M. Dennys *et al. The 100,000 Genomes Project Protocol*. Tech. rep. 2017.

[6] T. C. Archer, E. J. Fertig, S. J. Gosline *et al.* "Systems approaches to cancer biology". *Cancer Research* **76**:23 (2016), pp. 6774–6777.

[7] M. Kussmann, M. J. Morine, J. Hager, B. Sonderegger, and J. Kaput. "Perspective: a systems approach to diabetes research". *Frontiers in Genetics* **4** (2013), p. 205.

[8] J. L. Rollo, N. Banihashemi, F. Vafaee *et al.* "Unraveling the mechanistic complexity of Alzheimer's disease through systems biology". *Alzheimer's and Dementia* **12**:6 (2016), pp. 708–718.

[9] J. N. Weinstein, E. A. Collisson, G. B. Mills *et al.* "The cancer genome atlas pan-cancer analysis project". *Nature Genetics* **45**:10 (2013), pp. 1113–1120. arXiv: arXiv:1403.6652v2.

[10] M. J. Ellis, M. Gillette, S. A. Carr *et al.* "Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium". *Cancer Discovery* **3**:10 (2013), pp. 1108–1112. arXiv: NIHMS150003.

[11] M. Vidal, M. E. Cusick, and A. L. Barabási. "Interactome networks and human disease". *Cell* **144**:6 (2011), pp. 986–998.

[12] D. Szklarczyk, J. H. Morris, H. Cook *et al.* "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible." *Nucleic acids research* **45**:D1 (2017), pp. D362–D368.

[13] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez. "OmniPath: guidelines and gateway for literature-curated signaling pathway resources". *Nature methods* **13**:12 (2016), p. 966.

[14] D. Croft, A. F. Mundo, R. Haw *et al.* "The Reactome pathway knowledgebase". *Nucleic Acids Research* **42**:D1 (2014), pp. 472–477. eprint: /oup/backfile/content_public/journal/nar/42/d1/10.1093/nar/gkt1102/2/gkt1102.pdf.

[15] A. Fabregat, S. Jupe, L. Matthews *et al.* "The Reactome Pathway Knowledgebase". *Nucleic Acids Research* **46**:D1 (2018), pp. D649–D655. eprint: /oup/backfile/content_public/journal/nar/46/d1/10.1093_nar_gkx1132/2/gkx1132.pdf.

[16] J. D. Zhang and S. Wiemann. "KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor". *Bioinformatics* **25**:11 (2009), pp. 1470–1471.

[17] Tenenbaum D. *KEGGREST:Client-side REST access to KEGG*. 2016.

[18] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research* **28**:1 (2000), pp. 27–30.

[19] E. G. Cerami, B. E. Gross, E. Demir *et al.* "Pathway Commons, a web resource for biological pathway data". *Nucleic Acids Research* **39**:SUPPL. 1 (2011).

[20] D. Marbach, J. C. Costello, R. Küffner *et al.* "Wisdom of crowds for robust gene network inference". *Nature Methods* **9**:8 (2012), pp. 796–804. arXiv: arXiv:1511.08814v1.

[21] S. M. Hill, L. M. Heiser, T. Cokelaer *et al.* "Inferring causal molecular networks: Empirical assessment through a community-based effort". *Nature Methods* **13**:4 (2016), pp. 310–322. arXiv: 15334406.

[22] A. J. Butte and I. S. Kohane. "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". *Biocomputing 2000*. World Scientific, 1999, pp. 418–429.

[23] A. A. Margolin, K. Wang, W. K. Lim *et al.* "Reverse engineering cellular networks". *Nature protocols* **1**:2 (2006), pp. 662–671.

[24] J. J. Faith, B. Hayete, J. T. Thaden *et al.* "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles". *PLoS biology* **5**:1 (2007), e8.

[25] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. "Information-theoretic inference of large transcriptional regulatory networks". *EURASIP journal on bioinformatics and systems biology* **2007**:1 (2007), p. 79879.

[26] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". *Biostatistics* **9**:3 (2008), pp. 432–441.

[27] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. "TIGRESS: trustful inference of gene regulation using stability selection". *BMC systems biology* **6**:1 (2012), p. 145.

[28] Y. Zhang and M. Song. "Deciphering interactions in causal networks without parametric assumptions". *arXiv preprint arXiv:1311.2707* (2013).

[29] F. Petralia, W.-M. Song, Z. Tu, and P. Wang. "New method for joint network analysis reveals common and different coexpression patterns among genes and proteins in breast cancer". *Journal of proteome research* **15**:3 (2016), pp. 743–754.

[30] A. Irrthum, L. Wehenkel, P. Geurts *et al.* "Inferring regulatory networks from expression data using tree-based methods". *PloS one* **5**:9 (2010), e12776.

[31] M. E. Ahsen, R. Vogel, and G. Stolovitzky. "Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions" (2018). arXiv: 1802.04684.

[32] N. R. NCBI Resource Coordinators. "Database resources of the National Center for Biotechnology Information [ftp://ftp.ncbi.nih.gov/genomes/Bacteria]". *Nucleic acids research* **44**:D1 (2016), pp. D7–19.

[33] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. "A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature". *PLoS Computational Biology* **6**:7 (2010). (Visited on 07/06/2017).

[34] E. Tjioe, M. W. Berry, and R. Homayouni. "Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization)". *BMC Bioinformatics* **11**:Suppl 6 (2010), S14. (Visited on 07/04/2017).

[35] A. Barbosa-Silva, J.-F. Fontaine, E. R. Donnard *et al.* "PESCADOR, a web-based tool to assist text-mining of

biointeractions extracted from PubMed queries". *BMC Bioinformatics* **12** (2011), p. 435. (Visited on 07/05/2017).

[36] W. W. Fleuren, E. J. Toonen, S. Verhoeven *et al.* "Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining". *BioData Mining* **6** (2013), p. 2. (Visited on 07/04/2017).

[37] K. Raja, S. Subramani, and J. Natarajan. "PPInterFinder—a mining tool for extracting causal relations on human proteins from literature". *Database: The Journal of Biological Databases and Curation* **2013** (2013). (Visited on 07/04/2017).

[38] A. Usie, H. Karathia, I. Teixidó, R. Alves, and F. Solsona. "Biblio-MetReS for user-friendly mining of genes and biological processes in scientific documents". *PeerJ* **2** (2014). (Visited on 07/04/2017).

[39] M. Torii, C. N. Arighi, G. Li *et al.* "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information". *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **12**:1 (2015), pp. 17–29. (Visited on 07/04/2017).

[40] M. Manica, R. Mathis, and M. R. Martínez. "INtERAcT: Interaction Network Inference from Vector Representations of Words" (2018). arXiv: 1801.03011.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". *arXiv:1301.3781 [cs]* (2013). arXiv: 1301.3781. (Visited on 05/01/2017).

[42] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and Their Compositionality". *Proceedings of the 26th International Conference on Neural Information Processing Systems*. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119. (Visited on 04/02/2017).

[43] D. Szklarczyk, A. Franceschini, S. Wyder *et al.* "STRING v10: protein–protein interaction networks, integrated over the tree of life". *Nucleic Acids Research* **43**:Database issue (2015), pp. D447–D452. (Visited on 07/06/2017).

[44] J. C. Costello, L. M. Heiser, E. Georgii *et al.* "A community effort to assess and improve drug sensitivity prediction algorithms". *Nature biotechnology* **32**:12 (2014), p. 1202.

[45] Y. Cun and H. Fröhlich. "Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge". *BMC bioinformatics* (2012).

[46] M. Manica, J. Cadow, R. Mathis, and M. R. Martínez. "PIMKL: Pathway Induced Multiple Kernel Learning" (2018). arXiv: 1803.11274.

[47] M. Gönen and E. Alpaydın. "Multiple kernel learning algorithms". *Journal of machine learning research* **12**:Jul (2011), pp. 2211–2268.

[48] J. Mariette and N. Villa-Vialaneix. "Unsupervised multiple kernel learning for heterogeneous data integration". *Bioinformatics* **34**:2009 (2017).

[49] F. Aiolli and M. Donini. "EasyMKL: A scalable multiple kernel learning algorithm". *Neurocomputing* **169** (2015), pp. 215–224.

[50] F. S. Collins and H. Varmus. "A New Initiative on Precision Medicine". *New England Journal of Medicine* **372**:9 (2015), pp. 793–795. arXiv: arXiv:1011.1669v3.

[51] E. A. Ashley. *Towards precision medicine*. 2016.

[52] J. L. Taylor-Cousar, A. Munck, E. F. McKone *et al.* "Tezacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del". *New England Journal of Medicine* **373**:3 (2017), NEJMoa1709846. arXiv: 15334406.

[53] J. J. Brewington, G. L. McPhail, and J. P. Clancy. *Lumacaftor alone and combined with ivacaftor: Preclinical and clinical trial experience of F508del CFTR correction*. 2016.

[54] M. Brodlie, I. J. Haq, K. Roberts, and J. S. Elborn. *Targeted therapies to improve CFTR function in cystic fibrosis*. 2015.

[55] B. W. Ramsey, J. Davies, N. G. McElvaney *et al.* "A CFTR Potentiator in Patients with Cystic Fibrosis and the G551D Mutation". *New England Journal of Medicine* **365**:18 (2011), pp. 1663–1672. arXiv: NIHMS150003.

[56] J. A. Sosman, K. B. Kim, L. Schuchter *et al.* "Survival in BRAF V600–Mutant Advanced Melanoma Treated with Vemurafenib". *New England Journal of Medicine* **366**:8 (2012), pp. 707–714. arXiv: NIHMS150003.

[57] C. Linnemann, M. M. Van Buuren, L. Bies *et al.* "High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+T cells in human melanoma". *Nature Medicine* **21**:1 (2015), pp. 81–85.

[58] D. Schadendorf, F. S. Hodi, C. Robert *et al.* "Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma". *Journal of Clinical Oncology* **33**:17 (2015), pp. 1889–1894.

[59] M. S. Lawrence, C. Sougnez, L. Lichtenstein *et al.* "Comprehensive genomic characterization of head and neck squamous cell carcinomas". *Nature* **517**:7536 (2015), pp. 576–582. arXiv: NIHMS150003.

[60] A. Abeshouse, J. Ahn, R. Akbani *et al.* "The Molecular Taxonomy of Primary Prostate Cancer". *Cell* **163**:4 (2015), pp. 1011–1025. arXiv: 15334406.

[61] M. K. Morris, J. Saez-Rodriguez, P. K. Sorger, and D. A. Lauffenburger. "Logic-based models for the analysis of cell signaling networks". *Biochemistry* **49**:15 (2010), pp. 3216–3224.

[62] M. L. Wynn, N. Consul, S. D. Merajver, and S. Schnell. "Logic-based models in systems biology: A predictive and parameter-free network analysis method". *Integrative Biology (United Kingdom)* **4**:11 (2012), pp. 1332–1337.

[63] S. Pandey, R. S. Wang, L. Wilson *et al.* "Boolean modeling of transcriptome data reveals novel modes of heterotrimeric G-protein action". *Molecular Systems Biology* **6** (2010), p. 372.

[64] S. Li, S. M. Assmann, and R. Albert. "Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling". *PLoS Biology* **4**:10 (2006). Ed. by J. Chory, pp. 1732–1748.

[65] A. Fauré, A. Naldi, C. Chaouiya, and D. Thieffry. "Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle". *Bioinformatics*. Vol. 22. 14. 2006, pp. 124–131.

[66] D. A. Orlando, C. Y. Lin, A. Bernard *et al.* "Global control of cell-cycle transcription by coupled CDK and network oscillators". *Nature* **453**:7197 (2008), pp. 944–947. arXiv: NIHMS150003.

[67] P. A. Futreal, L. Coin, M. Marshall *et al.* "A census of human cancer genes". *Nature Reviews Cancer* **4**:3 (2004), pp. 177–183.

[68] P. C. Nowell. "The clonal evolution of tumor cell populations". *Science* **194**:4260 (1976), pp. 23–28.

[69] I. J. Fidler and I. R. Hart. "Biological diversity in metastatic neoplasms: Origins and implications". *Science* **217**:4564 (1982), pp. 998–1003.

[70] P. C. Boutros, M. Fraser, N. J. Harding *et al.* "Spatial genomic heterogeneity within localized, multifocal prostate cancer". *Nature Genetics* **47**:7 (2015), pp. 736–745.

[71] J. Wang, H. Khiabanian, D. Rossi *et al.* "Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia". *eLife* **3** (2014).

[72] G. Gundem, P. Van Loo, B. Kremeyer *et al.* "The evolutionary history of lethal metastatic prostate cancer". *Nature* **520**:7547 (2015), pp. 353–357.

[73] M. Manica, P. Chouvarine, R. Mathis *et al.* "Inferring clonal composition from multiple tumor biopsies". *arXiv* (2017). arXiv: 1701.07940.

[74] S. L. Carter *et al.* "Absolute quantification of somatic DNA alterations in human cancer : Nature Biotechnology : Nature Publishing Group". *Nature biotechnology* **30** (2012), pp. 413–421.

[75] N. Andor, J. V. Harness, S. Müller, H. W. Mewes, and C. Petritsch. "Expands: Expanding ploidy and allele frequency on nested subpopulations". *Bioinformatics* **30**:1 (2014), pp. 50–60.

[76] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael. "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data". *Bioinformatics* **31**:12 (2015), pp. i62–i70.

[77] N. Niknafs, V. Beleva-Guthrie, D. Q. Naiman, and R. Karchin. "SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing". *PLoS Computational Biology* **11**:10 (2015). Ed. by Q. Morris, e1004416.

Part II.

Scientific Contributions

2 COSIFER: Consensus Interaction Network Inference Service [1]

Matteo Manica[1,2,*], Charlotte Bunne[1,2,*], Roland Mathis[1,*], Joris Cadow[1], María Rodríguez Martínez[1]

[1] IBM Research Zürich

[2] ETH - Zürich

[*] Shared first authorship

Abstract

Molecular interaction networks regulate every cellular process such as evolution, proliferation and cell death. When these mechanisms fail, normal cell behavior is compromised and different diseases, like cancer, may occur. The complexity of these networks resides in the large number and variability of the molecular interactions involved. Advent of high-throughput technologies such as microarrays and RNA sequencing provided researchers with whole-transcriptome measurements and made available a snapshot of the internal regulatory apparatus of a cell. However, network inference and key regulators detection is a daunting task, that pushed international consortia to intensively work on the development of computational methods. Despite the efforts to compare and develop gene regulatory network inference methods, easy to access inference tools available to everyone are still missing in the research community. COSIFER is a web-based platform providing a service for the inference of molecular networks using a consensus between state-of-the-art methodologies given high-throughput molecular measurements and a list of molecular entities of interest.

COSIFER integrates a set of network inference methods with different theoretical approaches

[1] In preparation. See Chapter 8 for details about contributions and copyright.

as well as robust consensus methodologies. COSIFER is validated by extensively benchmarking it on synthetic data generated from a known network topology exhibiting stability in prediction accuracy. To test the potential of the tool, a study at pathway level of two separate breast cancer cohorts (TCGA-BRCA and METABRIC) is conducted, showing how COSIFER can be used to generate new insights in the mechanisms underlying complex diseases.

2.1 Introduction

Gene regulatory networks govern every kind of cellular decision such as differentiation, proliferation and apoptosis and when these control mechanisms fail, cancer and other diseases may arise [1]. The complexity of these networks originates from the large number of molecules involved and the wide range of interactions occurring between them. High-throughput technologies, like microarrays and RNA sequencing, provide measurements of the transcriptome and enable insights into internal regulatory mechanisms of a cell. However, inferring the topology of these networks and identifying its key regulators is a challenging problem and international consortia have intensively worked on the development of computational methods tackling this problem [2]. The *Dialogue on Reverse Engineering Assessment and Methods* (DREAM) project assessed the performance of over 30 network inference methods on *in silico*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Staphylococcus aureus* microarray data.

Despite the effort of comparison and development of gene regulatory network inference methods, the research community still lacks easy to access inference tools available to everyone. Many method implementations are unavailable to researchers and performing network inference on a larger gene set requires computational resources not readily available. In this work we introduce COSIFER (*Consensus Interaction Network Inference Service*), a web based platform providing a service for inferring relevance networks from uploaded molecular data. Given inferred networks, it is possible to detect central genes, gene communities, analyze network topology and discover further inherent features.

To reconstruct networks with high precision and accuracy, algorithms with various methodological foundation are developed. Basic approaches simply compute the correlation of expression patterns between network components, while other algorithms apply information theoretic approaches, solve regression problems or apply Bayesian heuristics [13]. Being predominantly based on unsupervised learning algorithms, COSIFER integrates a subset of these methods which proved reliable network reconstruction performance and are based on various theoretical concepts (see Table 2.1).

Method	Source
Correlation	
Pearson's correlation coefficient	[3]
Spearman's correlation coefficient	[3]
Mutual Information	
Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE)	[4], [5]
Context Likelihood of Relatedness (CLR)	[6], [5]
Minimum Redundancy/Maximum Relevance Networks (MRNET)	[7], [5]
Regression	
Graphical LASSO (GLasso)	[8]
Trustful Inference of Gene Regulation Using Stability Selection (TIGRESS)	[9]
Other Approaches	
Joint Random Forest (JRF)	[10]
Functional χ^2-Test (FunChiSq)	[11]
Gene Network Inference with Ensemble of Trees (GENIE3)*	[12]

Table 2.1: **Implemented methods.** List of methods implemented in COSIFER web application.

Since every method to infer molecular networks has its advantages as well as limitations and given different conditions such as varying data sources, noise levels and underlying network topologies, methods might complementary outperform others. Thus, combining the results of individual methods into a consensus network might be a promising strategy to obtain robust and accurate results. This approach have been successfully applied in different fields [2, 14]. Marbach *et al.* further showed that a consensus approach between various methods outperforms a consensus between similar algorithms as well as the performance of single methods [2]. We denote as WOC and WOC (hard) consensus strategies inspired to these works. Besides these standard consensus approaches a novel method based on rank aggregation called SUMMA (Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions) is considered [15].

Next to single method implementations, COSIFER applies the introduced consensus algorithms on a set of methods chosen by the user, providing a an inference result robust to different biological as well as experimental setting.

In the following we introduce the web application COSIFER and evaluate the methods' performance as well as the robustness of the consensus approach on different noise models, sample and network sizes. Additionally, COSIFER is applied to breast cancer gene expression data. For this analysis, we consider gene expression profiles in breast tumors from *Molecular Taxonomy of Breast Cancer International Consortium* (METABRIC) as well as the *The Cancer Genome Atlas* (TCGA) [16, 17]. In a subsequent analysis we detect central genes for each hallmark pathway [18] in both cohorts in an effort to define common potential regulators. The detection of central genes can help to cast light on potential disease mechanisms of this cancer type that can be subsequent tested in further experimental studies.

2.2 Results

COSIFER web application integrates basic functionality such as: data upload, method and network entity selection and interactive network visualization using *bokeh* [19], see Figure 2.1.

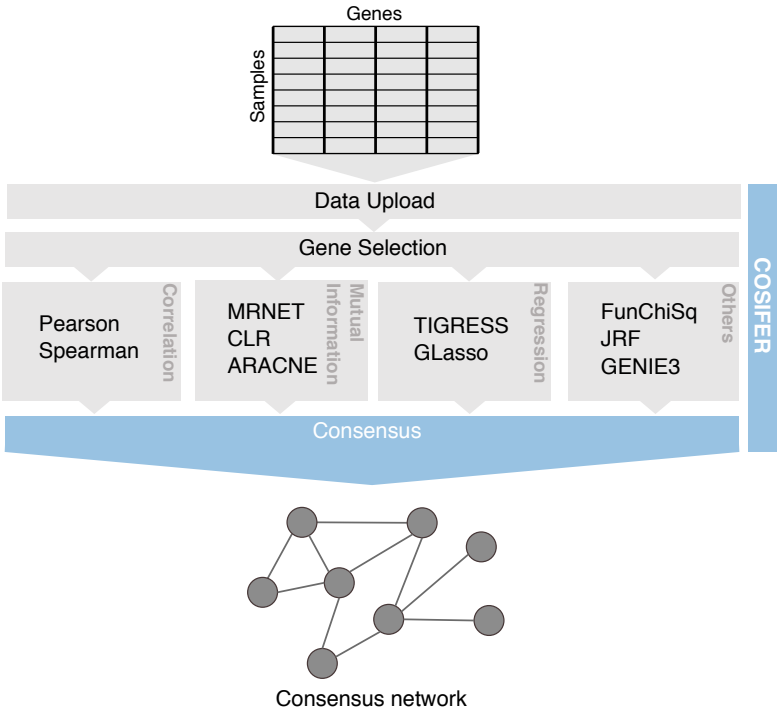


Figure 2.1: **COSIFER workflow.** Once molecular data are uploaded, difference infer methods can be selected. COSIFER integrates single method predictions into a consensus network that can be visualized or downloaded by the user for further analysis.

The user is able to upload a dataset as well as an additional list of gene on which the consensus network inference is performed. If no list containing gene identifiers is given, the network is predicted using the all genes of the uploaded dataset. The consensus network can be visualized in the web application and/or downloaded in a tabular format and used for subsequent network analysis steps.

Applying unsupervised methods to infer gene regulatory networks is convenient because no prior knowledge about the network to be inferred is required and the training step necessary for supervised algorithms can be skipped.

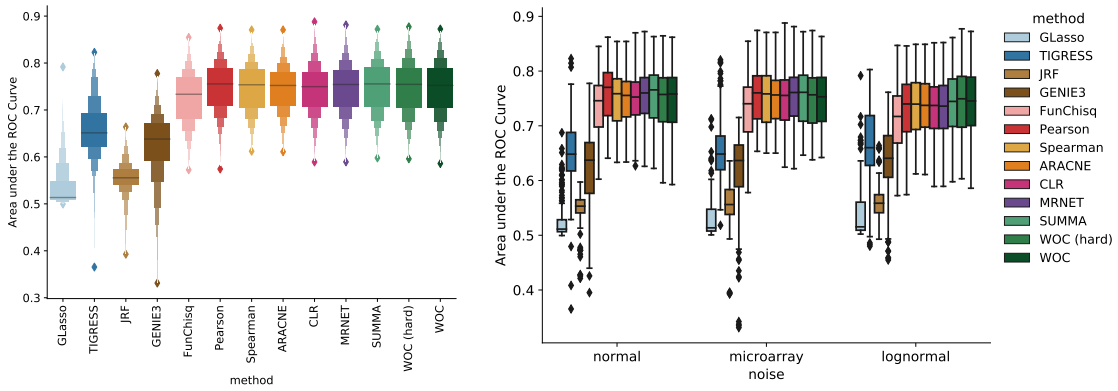
To this moment, the COSIFER web app comprises ten methods and provide as a result a network inferred using SUMMA as a consensus approach on single algorithms outcomes. The selection of this set of methods is based on different criteria: we selected methods for

which previous studies reported superior accuracy [2, 13, 20] as well as methods with varying theoretical foundations to capture a broad variety of approaches. Thus, COSIFER comprises methods based on correlation, including Pearson as well as Spearman correlation, mutual information, regression, tree ensembles and functional χ^2-test based methods. The whole list of methods as well as their original publications are presented in Table 2.1. Details on the theoretical foundations as well as the default parameters used for every method are reported in Section 2.4. The list of methods can be easily extended and further methods can be included into the web application.

2.2.1 Evaluation of COSIFER on *in silico* benchmark datasets

To evaluate the performance of different unsupervised methods as well as their consensus for reverse engineering of molecular networks, COSIFER is applied to *in silico* gene expression data for which the underlying network structure is known. This artificial datasets are generated using an open-source simulator called GeneNetWeaver, a tool intensively used in the DREAM challenge providing researchers benchmark datasets to validate their work [21]. Originating from *S. cerevisiae* the generated gene expression data comprise of a differing number of sample set sizes (100, 500, 1000 and 2000) as well as several network sizes (100, 500, 1000 and 2000). By varying the noise model underlying the simulation, we analyze the robustness of each method to different noise sources (Gaussian noise, lognormal noise as well as microarray noise) [22]. Therefore, each method as well as the consensus approach are analyzed in 48 different settings, repeating each simulation 10 times.

Figure 2.2a shows the result of the evaluation of COSIFER, i.e., the consensus approach, as well as all integrated inference methods. The performance of the methods strongly differs and ranges from an average area under the receiver operating characteristic (ROC) curve (AUC) of 0.53 to 0.75. The best performing methods are based on computing the mutual information matrix or the correlation matrix between the expression profiles of each network component. In terms of variability information theoretic methods performed slightly better than correlation-based methods, a behavior already observed during the DREAM study [2]. The accuracy of the single methods highly varies with a AUC ranging from ~ 0.3 to ~ 0.9. Especially in settings, where the number of samples from which the prediction is made is low, methods such as GLasso, TIGRESS, JRF or GENIE3 perform slightly better than a random guess (see Figure 2.4 center). The size of the network to be inferred has a weak positive influence on best performing methods, while it is more evident in the cases of worst performing ones, where inference with more entities exhibits reduced performance for the majority



(a) **(b)**

Figure 2.2: **Evaluation of COSIFER performance.** AUC values of each network inference method as well as the consensus (SUMMA, WOC (hard), WOC; different shades of green) under different noise models, sample and network sizes (48 settings) over ten simulations. **(a)** Boxen plot (Letter-value plot) of the AUC values. **(b)** Box plots grouped by different noise types.

of them (see Figure 2.4 top). All methods are robust to different sources of noise and thus, the accuracies of the methods' prediction do not vary strongly between *in silico* expression data generated using the Gaussian, lognormal and microarray noise model (see Figure 2.2b Figure 2.4 bottom).

Next to the AUC, the methods differ in the computation time they take for the inference. Despite the quadratic complexity of computing pairwise mutual information as well as correlation between all network entities, methods based on information theoretical principles such as ARACNE, CLR and MRNET as well as Pearson and Spearman correlation based methods are the fastest. FunChiSq, GLasso, TIGRESS, GENIE3 and JRF are slower and for networks exceeding the size of 1000 nodes the latter three methods do not scale properly.

As shown in Figure 2.2, consensus approaches exhibit similar or superior median performance compared to best performing methods. Furthermore the variance in their predictions is slightly lower compared to the mutual information and correlation-based algorithms and is not heavily influenced in its performance by methods, performing poorly on the inference task. Overall SUMMA performs better than WOC (hard) and WOC in terms of median AUC as reported in Figure 2.2.

2.2.2 COSIFER on breast cancer potential regulators detection

This section describes results for COSIFER consensus inference application to datasets from TCGA-BRCA and METABRIC cohorts (summary reported in Table 2.2). COSIFER is used to identify potential regulators that characterize the disease by applying it on each dataset separately and by merging the inference results in a consensus network from both cohorts using SUMMA.

Specifically, the proposed approach, is applied to infer pathway-wise network topologies in both datasets, with the goal of finding common central genes that potentially characterize the disease regardless the cohort considered. The main hypothesis is that a central node in the network can be considered as a proxy for a potential regulator, since highly central nodes represent genes that are strongly connected with multiple neighbors.

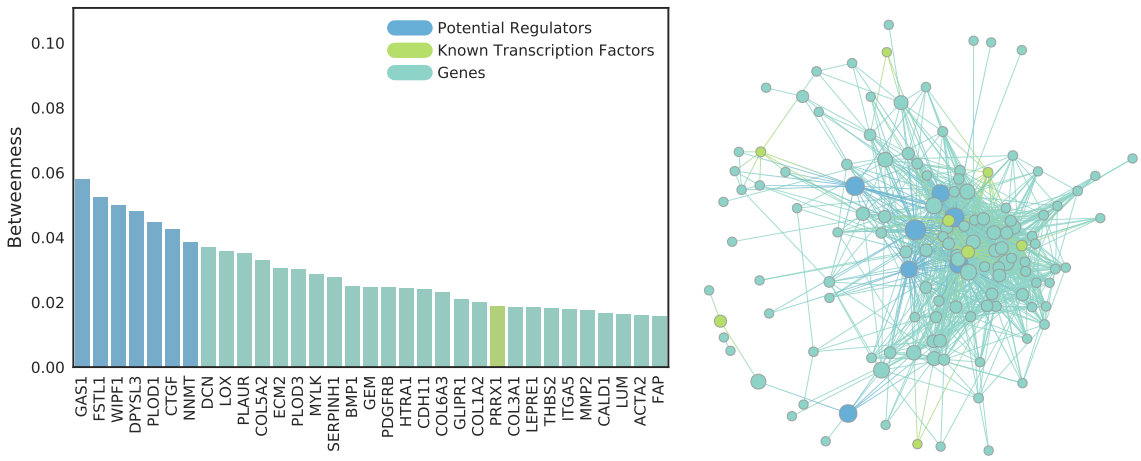
The decision of applying COSIFER at single pathway level is mainly due to two aspects. First, inference in networks of limited size is computationally more efficient (some of the considered methods were not able to produce an output for large number of genes). Secondly, comparing inferred topologies at single pathway level between the datasets enables to find pathways that are consistently reconstructed between the independent cohorts. Similar pathways indicate aspects that are common to both datasets and ideally identify disease-specific features.

To define pathways, 50 *hallmark* gene sets [18] from MSigDB are considered. By analyzing the adjacency similarities (see 2.4.7) between TCGA-BRCA and METABRIC estimated networks for each pathway, it is possible to determine a ranking that defines which gene sets are reconstructed in a comparable fashion independently from the dataset (for the complete list of ranked pathways see 2.5).

	TCGA-BRCA	**METABRIC**
data type	RNASeq	mRNA
number of genes	4303	4135
number of samples	1100	1905

Table 2.2: **Data statistics.** Data considered for breast-specific network reconstruction.

Epithelial-Mesenchymal Transition(EMT) emerges as the most relevant gene set. This is not surprising, since EMT plays an essential role in regeneration of tissue and cell development and its activation has been associated with breast cancer progression and metastatic behavior [23, 24]. In Figure 2.3 betweenness centrality values and a network representation for EMT are shown.



(a) **(b)**

Figure 2.3: **COSIFER inferred consensus network for Epithelial-Mesenchymal Transitiongene set.** This Figure describes the analysis of high confidence regulatory interactions (pruning edges using a threshold $t = 0.9$) for the most stable *hallmark* set, Epithelial-Mesenchymal Transition. The network has been obtained using the consensus network estimated after merging the results from both TCGA-BRCA and METABRIC cohorts. In Panel **(a)** potential regulators, sorted using betweenness measure, are reported. The legend shows the colors associated with the different genes based on their source. The known transcription factors are recovered from TFcheckpoint [25] (green). The genes reported have a centrality betweennes above the 75th percentile of the centrality distribution while the ones highlighted as potential regulators (blue) above the 95th percentile. In Panel **(b)** a graph reporting all the high confidence interactions is shown. Edge width is a function of the intensity, node size depends on their betweenness and the color scheme is the same used in Panel **(a)**.

The network is obtained by merging the topologies reconstructed in single cohorts using COSIFER consensus approach. The highlighted potential regulators are selected by analyzing the pathway-wise distributions of the centrality values and by considering extreme values (see 2.4.8).

The same procedure is repeated for each considered pathway and a list of 239 potential regulators has been compiled.

To validate the set of genes obtained, a pathway enrichment analysis is performed. We use *Enrichr* [26, 27] to test whether regulation-specific processes among GO Biological Processes [28, 29] are significantly enriched by the list of potential regulators found.

The results reported in Table 2.3 show how among the significantly enriched processes (adjusted p-value ≤ 0.05), the majority is composed by regulatory ones. Not surprisingly,

GO Biological Process	Overlap	Adjusted p-value
Negative regulation of apoptotic process (GO:0043066)	17/323	4.19E-04
T cell activation (GO:0042110)	6/35	1.83E-03
Positive regulation of ubiquitin protein ligase activity (GO:1904668)	4/11	2.25E-03
Signal transduction (GO:0007165)	26/861	3.92E-03
Sister chromatid cohesion (GO:0007062)	8/97	4.71E-03
Negative regulation of mitotic cell cycle (GO:0045930)	4/18	7.33E-03
Positive regulation of epithelial to mesenchymal transition (GO:0010718)	5/34	7.33E-03
Cellular response to lipopolysaccharide (GO:0071222)	6/55	7.33E-03
Leukocyte migration (GO:0050900)	10/198	1.41E-02
Positive transcription regulation from RNA polymerase II promoter (GO:0045944)	21/712	1.41E-02
Angiogenesis (GO:0001525)	7/93	1.41E-02
Positive regulation of cell proliferation (GO:0008284)	13/326	1.47E-02
Protein destabilization (GO:0031648)	4/26	2.05E-02
Actin filament bundle assembly (GO:0051017)	4/27	2.21E-02
Positive ubiquitin-protein ligase regulation in mitotic cell cycle (GO:0051437)	6/77	2.34E-02
Protein phosphorylation (GO:0006468)	12/309	2.49E-02
Anaphase-promoting complex-dependent catabolic process (GO:0031145)	6/80	2.54E-02
Negative regulation of cell proliferation (GO:0008285)	11/276	3.12E-02
Mitotic cell cycle (GO:0000278)	6/86	3.35E-02
Negative regulation of neuron apoptotic process (GO:0043524)	5/60	4.11E-02
Positive regulation of DNA binding (GO:0043388)	3/16	4.41E-02

Table 2.3: **Pathway enrichment analysis of the potential regulators.** The set of potential key regulators across all *hallmark* pathways have been extracted. The list is then used for an enrichment analysis to extract GO Biological Processes enriched for the provided list of genes. The second column indicates the number of genes that were overlapping with the pathway and the last column indicates the p-value after correction for multiple testing.

positive regulation of epithelial to mesenchymal transition is present with a high significance. It is interesting to notice also the presence of multiple processes related to proliferation, suggesting that the potential regulators estimated with COSIFER networks analysis might play a fundamental role in breast cancer development and onset.

2.3 Discussion

Gene regulatory network inference is an important task toward understanding biological systems and a variety of methods can be applied to reconstruct them from high-throughput data. It has been observed that no single inference method performs optimally across all datasets. While each method has advantages and limitations, all methods suffer from weaknesses either in terms of robustness or precision across heterogeneous datasets, which are characterized by different error sources, noise levels as well as origin and biological processes covered.

The wisdom of crowds (WOC) is a powerful approach to use the collective knowledge of a community instead of individual knowledge [30] and has demonstrated robustness and

superior performance across species and datasets [2]. In this paper, we have applied this concept and developed a Consensus Interaction Network Inference Service (COSIFER) to address one of the long-standing challenges in molecular and computational biology, which is to uncover and model gene regulatory networks.

COSIFER integrates a wide range of network inference methods that have been selected for their reported superior accuracy as well as their diverse theoretical foundation, so to capture a broader range of the true signal. Furthermore, COSIFER implements different consensus strategies, including the Wisdom of the Crowds [2] (WOC) and the more recently published method SUMMA, which exploits a weighted rank aggregation approach [15].

We have tested COSIFER on a collection of *in silico* benchmark datasets, generated using GeneNetWeaver, an open-source molecular data simulator, for which the underlying interaction network structure was known. We observe that while the size of network slightly influences the performance of the methods, with higher variability observed in worse performing methods, all tested methods show robustness to the different sources of noise investigated, i.e., Gaussian, lognormal and microarray noise models. Overall, all consensus methods returned networks of high accuracy and are robust to weak inference results of less precise methods. Thus, the wisdom of the crowds (WOC and WOC (hard)) and weighted rank aggregation approaches (SUMMA) represent a promising strategy to accurately reconstruct networks of biological systems.

We have also applied COSIFER to reconstruct the network topologies of two different breast cancer datasets, namely the TCGA-BRCA and METABRIC cohorts. The inferred networks were useful to detect potential cancer regulators and central genes whose dysregulation might produce strong perturbations in cellular homeostasis without using any prior knowledge on known transcription factors-targets associations.

We examined inference robustness across both datasets at a pathway level, using for this task a collection of 50 *hallmark* gene sets from MSigDB [18]. We identified the Epithelial-Mesenchymal transition (EMT) gene set as the pathway with the highest reconstruction similarity between cohorts, which reflects the central role of this pathway in breast cancer etiology and a higher content of disease–specific information compared to other pathways.

In conclusion, COSIFER is an useful service to automatically reconstruct molecular networks using a consensus approach. Currently COSIFER includes ten different inference methods exploiting very different theoretical approaches, and three different consensus approaches. We have demonstrated COSIFER performance in a synthetic dataset and in two breast cancer

cohorts. To facilitate its use, we have implemented COSIFER as a web service freely available to the scientific community.

2.4 Methods

2.4.1 Web Application

COSIFER is available as a service on IBM Cloud for the community. Access to the application is free and credentials can be obtained registering on `http://sysbio.uk-south.containers.mybluemix.net/cosifer`. The currently deployed version of the application can estimate a consensus network given an expression matrix (samples on rows and genes on the columns) and a list of gene names. The maximum supported network size at the moment is 250 nodes. The consensus network estimated can be interactively visualized in the web app or downloaded in tabular format within 24 hours from it is creation for further analysis.

2.4.2 Network Inference Methods

This section presents a variety of unsupervised methods whose implementations are provided by COSIFER.

Correlation

Basic correlation based approaches such as Spearman and Pearson correlation are used to infer a relevance network where an interaction between network components is present if their expression levels are significantly correlated. The commonly used Pearson correlation coefficient is a measure of linear correlation and defined as:

$$corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)},$$

where X is the expression level of gene i or j, respectively, $\sigma(\cdot)$ the standard deviation and $cov(\cdot, \cdot)$ the covariance. Spearman's correlation coefficient simply is Pearson's correlation coefficient between the rank values of two variables and thus assesses monotonic relationships.

Both correlation-based approaches are corrected for multiple testing using *Benjamini-Hochberg* procedure [31].

Mutual Information

To infer interactions between network components the concept of mutual information can be applied as well. This measure assesses the mutual dependence between two expression levels X_i and X_j and is defined as:

$$I(X_i, X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} P(x_i, x_j) log \frac{P(x_i, x_j)}{P(x_i)P(x_j)},$$

where $P(x_i, x_j)$ is the joint probability distribution of X_i and X_j, while $P(x_i)$ is the marginal distribution of X_i.

Several algorithms are based on mutual information measure. So is the *Context Likelihood of Relatedness* (CLR) algorithm, which predicts an interaction between network components based on their mutual information but takes the background distribution of $I(X_i, X_j)$ into account [6]. MRNET uses the subset selection algorithm *Minimum Redundancy Maximum Relevance* (MRMR) to infer the regulators of each gene based on the pairwise mutual information between target and regulator genes [7]. The *Algorithm for the Reconstruction of Accurate Cellular Networks* (ARACNE) is another technique based on pairwise mutual information as well as the *Data Processing Inequality* (DPI) [4]. DPI states that:

$$I(X_i, X_k) \leq min\left(I(X_i, X_j), I(X_j, X_k)\right)$$

if a gene X_i interacts with X_j through gene X_k. The weakest edge of a triplet is removed if the interaction triangle violates the DPI beyond a specified tolerance threshold ϵ ($\epsilon = 0.2$). The idea behind using the DPI is that it allows us to remove edges containing redundant information. ARACNE, CLR and MRNET are all implemented in the `R/Bioconductor` package *minet* [5].

Regression

Trustful Inference of Gene Regulation using Stability Selection (TIGRESS) is next to GENIE3 and FunChisq the only method which is able to predict interaction directions of the network [9]. The principle of feature selection is applied to infer the regulators of a gene.

For this, the *Least Angle RegreSsion* (LARS) combined with stability selection is used. If the expression levels of genes are correlated, the performance of LARS is significantly reduced. Thus, stability selection is conducted by applying a randomized LARS many times on randomly perturbed data. Each feature is scored with respect to the number of times it is selected. This score then quantifies the evidence that a target gene is regulated by another gene. We perform 1000 runs of the stability selection to compute the scores (score='area'), and 5 LARS steps, as well as $\alpha = 0.2$ which controls the random re-weighting of each expression array in each stability selection run.

The method *graphical Lasso* (GLasso) estimates the sparse inverse covariance matrix with a ℓ_1-penalty [8]. The inverse covariance matrix represents partial correlations between gene pairs in the associated network. GLasso's objective function for estimating the sparse inverse covariance matrix Θ is thus defined as:

$$\max_{\Theta} \quad log(det(\Theta)) - tr(S\Theta) - \rho\|\Theta\|_1,$$

where S is the sample covariance matrix computed for the dataset. The problem can also be easily extended to the case that uses a penalization matrix P to exploit prior information on the network to be estimated:

$$\max_{\Theta} \quad log(det(\Theta)) - tr(S\Theta) - \|P * \Theta\|_1.$$

The penalization matrix entries correspond to relations between entities and can be set to arbitrary low values for well known interactions and high values for entities that are known to be unrelated. Once the sparse inverse covariance Θ is estimated it is possible to compute partial correlations between all pairs of variables and build a graph defined by:

$$G_{ij} = \begin{cases} -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} & i \neq j \\ 1 & otherwise \end{cases}.$$

Other Approaches

Similarly to TIGRESS, *Gene Network Inference with Ensemble of Trees* (GENIE3) transforms the task into a feature selection problem assuming the expression of each gene to be a function of the noisy expression of the other genes in the network. So each sub-problem is a supervised, non-parametric regression problem solved using regression trees. Two tree-based

ensemble methods based on randomization are used, namely *Random Forests* (RF) and *Extra-Trees* (ET) [12]. For this inference task, RF is used as tree-based ensemble method. The expression of each target gene k is modeled as a function of all other genes. A score $I_{j\to k}$ quantifies the importance of each regulatory event ($j \to k$). The individual rankings obtained from each sub-problem are then aggregated to get a global ranking of all regulatory links. Since the splitting variable at each node of a decision tree is chosen from a randomly sampled data subset and each decision tree of the RF is constructed from a random subset of sample, the result is not deterministic and varies between runs. GENIE3 uses 1000 trees as default parameter for the inference. The square root of the total number of candidate regulators is used as the number of candidate regulators randomly selected at each tree node.

Like GENIE3, *Joint Random Forest* (JRF) is based on RF [10]. Unlike other network inference methods, this approach simultaneously estimates multiple co-expression networks by integrating different data sources such as protein and gene expression data By requiring the class-specific RFs to use the same genes for the splitting rules, information between gene and protein expression data is borrowed. COSIFER applies the inference only to one data type and thus, JRF reduces to GENIE3. Due to differences in the implementation, GENIE3 outperforms JRF, even though they are based on the same theoretical approach. JRF uses 500 trees as default parameter for the inference.

FunChiSq bases its inference on a functional χ^2 test. To run the algorithm the data have to be processed to discretize each variable. This has been implemented as described by Zhang *et al.* [11]. The discretization is achieved using k-means clustering, where k is estimated using a Gaussian mixture model that optimizes the *Bayesian Information Criterion* (BIC). The minimal value k can take is 3, its maximum is 7. Then, a χ^2-test is performed for every pair of genes given the contingency table of size $n_i \times n_j$, where n is the number of estimated clusters of a gene. The χ^2_{ij}-statistic indicates evidence for the existence of a functional interaction from gene i to gene j. The predicted interactions are directed.

For the above methods, no parameter optimization is performed but instead the default parameter preferences are used for subsequent inference if not stated differently.

Consensus Methods

Given that the methods to be integrated into a consensus prediction and so the rank of one specific interaction determined by a method are statistically independent, a consensus approach outperforms individual results. The central limit theorem of probability theory states that the

more predictions are averaged, the distribution of average ranks will approach a normal distribution whose variance shrinks with increasing number of integrated methods [2].

WOC and WOC (hard) have been implemented to combine individual network predictions in an analogous fashion. The weighted adjacency matrices are scaled to obtain a comparable interaction score between methods. Thus, for each method, the weights are scaled between 0 and 1 by applying feature rescaling. Next, the scaled interaction scores are summed and an unweighted rank average over all methods is taken. WOC ignores methods not providing predictions in the rank average while WOC (hard) considers a zero contribution. The score obtained is the interaction score of the resulting consensus network.

Similarly SUMMA, combines scores from weighted adjacency matrices but weights the rank contribution from each method proportionally to its AUROC (Area under the Receiver Operating Characteristic) [15]. Under the assumption of mutual conditional independence between predictions from different methods such weights can be computed from the covariance matrix of the ranked prediction in a completely unsupervised fashion. The score computed by SUMMA is the interaction score of the resulting consensus network.

2.4.3 *In Silico* Data

The gene expression data have been generated using GeneNetWeaver, a tool for *in silico* benchmark generation and performance profiling of network inference methods [21]. GeneNetWeaver extracts sub-networks from known transcription networks such as those of *S. cerevisiae* and *E. coli*. By applying detailed dynamical models of gene regulatory networks emulating transcription and translation with ordinary and stochastic differential equations, *in silico* data are generated.

In the following, ODE based simulations of *S. cerevisiae* sub-networks of varying network component numbers (100, 500, 1000 and 2000) with differing number of samples (100, 500, 1000 and 2000) each have been adopted. The subsets have been randomly extracted from a network based on well-studied pathways in *S. cerevisiae* containing 4441 nodes with 12873 edges [32]. The ODE simulations include both, transcription and translation processes. GeneNetWeaver implements measurement noise independently from molecular noise and provides different noise models such as Gaussian noise, lognormal noise as well as a noise observed in microarrays [22]. Steady state rather than time series expression data of the wild-type (unperturbed network) have been used to validate network inference. For each dataset, variables have been standardized to mean zero and unit variance.

2.4.4 Breast Cancer Data

COSIFER has been applied to two different breast cancer cohorts: METABRIC and TCGA-BRCA.

In METABRIC case, the proposed approach is used to analyze 1905 samples where Illumina Human v3 microarray (mRNA) measurements have been considered. All primary data of the METABRIC study are provided at the *European Genome-phenome Archive* (EGA) under study accession number EGAS00001001753. The gene expression data of the patients from the original METABRIC publication are freely available [1]. The data were normalized as described by Margolin *et al.* [33].

In TCGA-BRCA case, Illumina HiSeq 2000 RNA Sequencing Version 2 data of breast tumor samples have been obtained from FireBrowse web API [2] using data version *2016_01_28*. The level 3 gene expression data have been processed using two analysis pipelines [34, 35].

For each dataset, variables have been standardized to mean zero and unit variance.

2.4.5 Performance Evaluation

The precision accuracy of each algorithm is measured by comparing the topology of the inferred network with the true network. Since most of the performed methods predict undirected rather than directed interactions, self-interactions and interaction direction are ignored for the accuracy assessment of the methods. Measurement of inference accuracy is based on how many edges are correctly inferred given a threshold and is assessed using *Receiver Operating Characteristic* (ROC) curves.

To draw the ROC curve, the false (FPR) and true positive rate (TPR) are needed. The false positive rate is defined as

$$FPR = \frac{FP}{TN + FP}$$

and the true positive rate as

$$TPR = \frac{TP}{TP + FN}$$

[1] `https://www.synapse.org/#!Synapse:syn1757063`, as of November 2018
[2] `http://firebrowse.org/api-docs/`, as of November 2018

also known as recall or sensitivity, where TP are true positives, TN true negatives, FP are false positives and FN are true negatives. The ROC curve is quantified by the *Area under the Receiver Operating Characteristic* curve (AUROC). The area under a curve is defined as

$$AUC = \frac{1}{2}\sum_{i=1}^{n}(X_i - X_{i-1})(Y_i + Y_{i-1})$$

where X_i is the false positive rate and Y_i the true positive rate for the i-th output in the ranked list of predicted edge weights. While an AUC of 1.0 indicates perfect prediction, an AUC of 0.5 indicates a performance comparable to a random estimator.

2.4.6 Key regulator detection

In order to detect important regulators of the network, the centrality of each network component is computed. A betweenness centrality metric [36] for all the genes contained in the networks has been calculated. Most central genes, with respect to this metric, should represent the main actors in regulatory events given their high-intensity interactions and strong connectivity [37]. The betweenness centrality is defined as

$$\delta_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

where $\sigma_{st}(v)$ is the number of shortest geodesic paths from node s to node t passing through v and σ_{st} the total number of shortest geodesic paths between s and t.

2.4.7 Graph similarity analysis

To quantify the similarity of each pathway inferred from two different cohorts, the weighted adjacency similarity between the two inferred pathway networks is computed. Adjacency similarity is the sum of equal entries in the adjacency matrix, given a vertex ordering determined by the vertex labels. It's a weighted count of the number of edges which have the same source and target labels in both graphs. For undirected weighted graphs, it is defined as

$$S(A_1, A_2) = E - d(A_1, A_2),$$

where

$$d(A_1, A_2) = \sum_{i<j} |A_{ij}^{(1)} - A_{ij}^{(2)}|$$

is the distance between graphs, A_k with $k \in \{1, 2\}$ are the weighted adjacency matrices of the graphs considered, and $E = \sum_{i<j} |A_{ij}^{(1)}| + |A_{ij}^{(2)}|$. The weights are normalized using $S(A_1, A_2)/E$.

2.4.8 Gene Enrichment Analysis

Gene enrichment analysis was performed using *Enrichr* [26, 27] through *gseapy* [1]. The selection of potential regulators tested for enrichment has been performed analyzing the distribution of the betweenness centrality in each gene set. For each pathway, genes with a betweenness centrality value above the 95th percentile, considered as extremely central, have been combined in a list of 239 candidates.

Declarations

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 668858.

Author contributions

See Chapter 8 for details about contributions and copyright.

Competing financial interests

The authors declare no competing financial interest.

[1] `https://github.com/BioNinja/GSEApy`, as of November 2018

Availability of data and materials

Networks inferred and potential regulators produced and presented in this work are available via download link [1]. COSIFER is freely available and can be accessed via `http://sysbio.uk-south.containers.mybluemix.net/cosifer`. A set of anonymous credentials can be created for reviewers.

[1] `https://ibm.box.com/s/o6glkb7lrlgrmlng0dr6ft6eivmdzknz`, as of November 2018

2.5 Supplementary information

Synthetic data evaluation

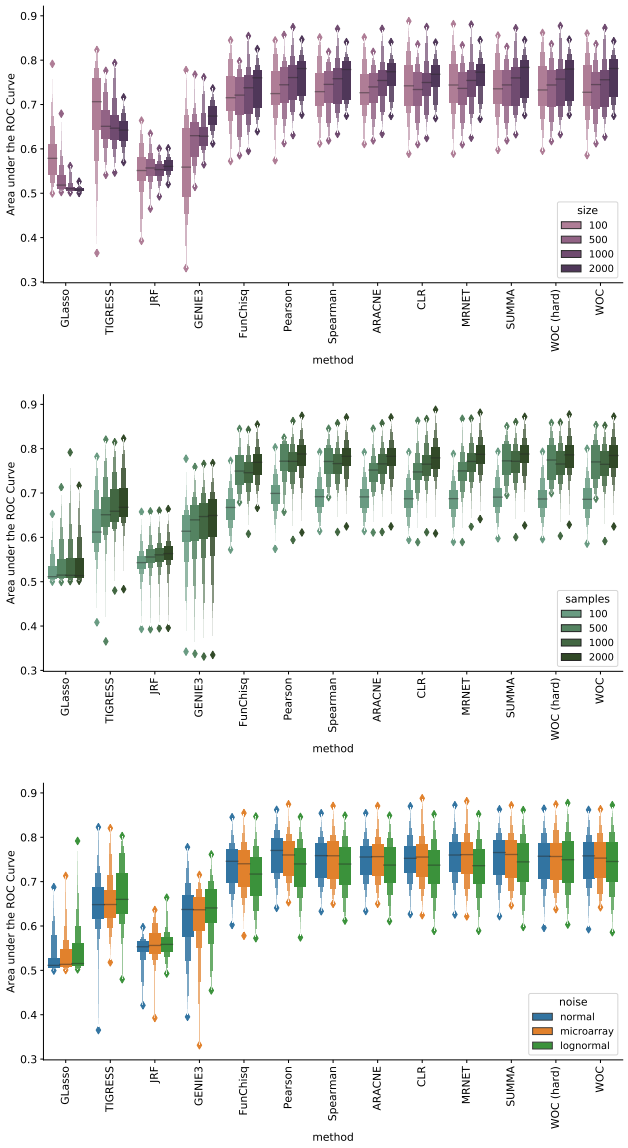


Figure 2.4: **Evaluation of the performance of COSIFER with repsect to simulation parameters.** AUC values of each network inference method as shown in Figure 2.2a, with each subplot showing the data in respect to a certain parameter. **top** Boxen plot (Letter-value plot) of the AUC values with respect to network size. **middle** Boxen plot of the AUC values with respect to sample number. **bottom** Boxen plot of the AUC values with respect to noise type.

Pathway similarity between cohorts

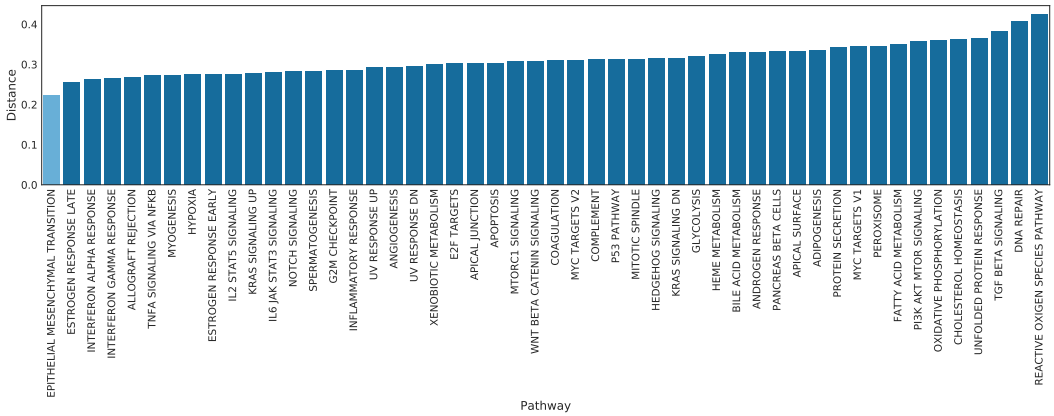


Figure 2.5: **Pathway similarities between cohorts.** Similarity analysis of gene regulatory networks estimated with COSIFER for the hallmark gene sets between the METABRIC and TCGA-BRCA cohorts. Pathways with high similarity between cohorts are expected to contain a higher degree of breast cancer–specific information compared to pathways with low similarity where the cohort effects are influencing the network. The most similar cancer hallmark pathways across cohorts is the pathway Epithelial-Mesenchymal Transition, highlighted in light blue.

References

[1] D. Pe'er and N. Hacohen. "Principles and strategies for developing network models in cancer". *Cell* **144**:6 (2011), pp. 864–873.

[2] D. Marbach, J. C. Costello, R. Küffner *et al.* "Wisdom of crowds for robust gene network inference". *Nature methods* **9**:8 (2012), pp. 796–804.

[3] A. J. Butte and I. S. Kohane. "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements". *Biocomputing 2000*. World Scientific, 1999, pp. 418–429.

[4] A. A. Margolin, K. Wang, W. K. Lim *et al.* "Reverse engineering cellular networks". *Nature protocols* **1**:2 (2006), pp. 662–671.

[5] P. E. Meyer, F. Lafitte, and G. Bontempi. "minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information". *BMC bioinformatics* **9**:1 (2008), p. 461.

[6] J. J. Faith, B. Hayete, J. T. Thaden *et al.* "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles". *PLoS biology* **5**:1 (2007), e8.

[7] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. "Information-theoretic inference of large transcriptional regulatory networks". *EURASIP journal on bioinformatics and systems biology* **2007**:1 (2007), p. 79879.

[8] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". *Biostatistics* **9**:3 (2008), pp. 432–441.

[9] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. "TIGRESS: trustful inference of gene regulation using stability selection". *BMC systems biology* **6**:1 (2012), p. 145.

[10] F. Petralia, W.-M. Song, Z. Tu, and P. Wang. "New method for joint network analysis reveals common and different coexpression patterns among genes and proteins in breast cancer". *Journal of proteome research* **15**:3 (2016), pp. 743–754.

[11] Y. Zhang and M. Song. "Deciphering interactions in causal networks without parametric assumptions". *arXiv preprint arXiv:1311.2707* (2013).

[12] A. Irrthum, L. Wehenkel, P. Geurts *et al.* "Inferring regulatory networks from expression data using tree-based methods". *PloS one* **5**:9 (2010), e12776.

[13] A. S. Iyer, H. U. Osmanbeyoglu, and C. S. Leslie. "Computational methods to dissect gene regulatory networks in cancer". *Current Opinion in Systems Biology* **2** (2017), pp. 115–122.

[14] A. L. Hofmann, J. Behr, J. Singer *et al.* "Detailed simulation of cancer exome sequencing data reveals differences and common limitations of variant callers". *BMC bioinformatics* **18**:1 (2017), p. 8.

[15] M. Eren Ahsen, R. Vogel, and G. Stolovitzky. "Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions". *ArXiv e-prints* (2018). arXiv: `1802.04684 [stat.ML]`.

[16] C. Curtis, S. P. Shah, S.-F. Chin *et al.* "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". *Nature* **486**:7403 (2012), p. 346.

[17] C. G. A. Network *et al.* "Comprehensive molecular portraits of human breast tumours". *Nature* **490**:7418 (2012), p. 61.

[18] A. Liberzon, C. Birger, H. Thorvaldsdóttir *et al.* "The molecular signatures database hallmark gene set collection". *Cell systems* **1**:6 (2015), pp. 417–425.

[19] Bokeh Development Team. *Bokeh: Python library for interactive visualization*. 2018.

[20] S. R. Maetschke, P. B. Madhamshettiwar, M. J. Davis, and M. A. Ragan. "Supervised, semi-supervised and unsupervised inference of gene regulatory networks". *Briefings in bioinformatics* **15**:2 (2013), pp. 195–211.

[21] T. Schaffter, D. Marbach, and D. Floreano. "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods". *Bioinformatics* **27**:16 (2011), pp. 2263–2270.

[22] G. Stolovitzky, A. Kundaje, G. Held *et al.* "Statistical analysis of MPSS measurements: application to the study of LPS-activated macrophage gene expression". *Proceedings of the National Academy of Sciences of the United States of America* **102**:5 (2005), pp. 1402–1407.

[23] J. Felipe Lima, S. Nofech-Mozes, J. Bayani, and J. Bartlett. "EMT in breast carcinoma—A review". *Journal of clinical medicine* **5**:7 (2016), p. 65.

[24] F. Liu, L.-N. Gu, B.-E. Shan, C.-Z. Geng, and M.-X. Sang. "Biomarkers for EMT and MET in breast cancer: An update". *Oncology letters* **12**:6 (2016), pp. 4869–4876.

[25] K. Chawla, S. Tripathi, L. Thommesen, A. Lægreid, and M. Kuiper. "TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors". *Bioinformatics* **29**:19 (2013), pp. 2519–2520.

[26] E. Y. Chen, C. M. Tan, Y. Kou *et al.* "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". *BMC bioinformatics* **14**:1 (2013), p. 128.

[27] M. V. Kuleshov, M. R. Jones, A. D. Rouillard *et al.* "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". *Nucleic acids research* **44**:W1 (2016), W90–W97.

[28] M. Ashburner, C. A. Ball, J. A. Blake *et al.* "Gene Ontology: tool for the unification of biology". *Nature genetics* **25**:1 (2000), p. 25.

[29] G. O. Consortium. "Expansion of the Gene Ontology knowledgebase and resources". *Nucleic acids research* **45**:D1 (2016), pp. D331–D338.

[30] J. Surowiecki. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. New York, NY, US: Doubleday & Co, 2004.

[31] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* **57**:1 (1995), pp. 289–300.

[32] S. Y. Kim, S. Imoto, and S. Miyano. "Inferring gene networks from time series microarray data using dynamic Bayesian networks". *Briefings in bioinformatics* **4**:3 (2003), pp. 228–235.

[33] A. A. Margolin, E. Bilal, E. Huang *et al.* "Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer". *Science translational medicine* **5**:181 (2013), 181re1–181re1.

[34] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey. "RNA-Seq gene expression estimation with read

mapping uncertainty". *Bioinformatics* **26**:4 (2009), pp. 493–500.

[35] K. Wang, D. Singh, Z. Zeng *et al.* "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery". *Nucleic acids research* **38**:18 (2010), e178–e178.

[36] L. C. Freeman. "A set of measures of centrality based on betweenness". *Sociometry* (1977), pp. 35–41.

[37] D. Koschützki and F. Schreiber. "Centrality analysis methods for biological networks and their application to gene regulatory networks". *Gene regulation and systems biology* **2** (2008), GRSB–S702.

3 INtERAcT: Interaction Network Inference from Vector Representations of Words [1]

Matteo Manica[1,2,*], Roland Mathis[1,*], Joris Cadow[1], María Rodríguez Martínez[1]
[1] IBM Research Zürich
[2] ETH - Zürich
[*] Shared first authorship

Abstract

In recent years, the number of biomedical publications has steadfastly grown, resulting in a rich source of untapped new knowledge. Most biomedical facts are however not readily available, but buried in the form of unstructured text, and hence their exploitation requires a time-consuming manual curation of published articles. Here we present INtERAcT, a novel approach to extract interactions from a corpus of biomedical articles related to a broad range of scientific domains in a completely unsupervised way. INtERAcT exploits vector representation of words, computed on a corpus of domain specific knowledge, and implements a new metric that estimates an interaction score between two molecules in the space where the corresponding words are embedded.

An extensive validation of INtERAcT against other commonly adopted similarity metrics varying embedding and algorithm parameters is performed. We evaluate INtERAcT using STRING database as a benchmark and training embeddings on PubMed abstracts and full text papers. We show that our proposed approach gives accurate estimates of the network topology

[1] In revision, pre-print available [1]. See Chapter 8 for details about contributions and copyright.

and it is highly robust to parameter choices. INtERAcT performs especially well in the regime of small corpus sizes. To demonstrate the capabilities of INtERAcT, we reconstruct the molecular pathways of ten different cancer types using a corpus of disease-specific articles for each cancer type. Our metric outperforms currently adopted approaches for similarity computation in the word-space and identifies known molecular interactions in all studied cancer types. Furthermore, our approach does not require text annotation, manual curation or the definition of semantic rules based on expert knowledge, and hence it can be easily and efficiently applied to different scientific domains. To conclude, INtERAcT helps summarize the understanding of a specific disease using the published literature in an automated and completely unsupervised fashion. INtERAcT is freely available as a web service.

3.1 Introduction

As the number of scientific publications continues to grow exponentially, search engines such as PubMed [1] provide an unprecedented amount of information in the form of unstructured written language. With the accelerating growth of available knowledge, particularly in biomedical literature, and the breakdown of disciplinary boundaries, it becomes unfeasible to manually track all new relevant discoveries, even on specialized topics. As an example, recent advances in high-throughput experimental technologies have yielded extensive new knowledge about molecular interactions in the cell; however most of this knowledge is still buried in the form of unstructured textual information only available as written articles.

As of October 2017, PubMed comprises more than 27.8 million references[2] consisting of biomedical literature from MEDLINE, life science journals, and online books. Most references include links to full text content from PubMed Central® (PMC [3]), a free full text archive of biomedical and life sciences journal literature, or publisher web sites. Currently 14.2 million PubMed articles have links to full text articles, 4.2 million of which are freely available. The numbers remain high even when focusing on specific fields such as prostate cancer. For instance, a simple query [4] for prostate cancer related papers on PMC returns 143,321 publications[5].

While a fraction of the information currently available in biomedical publications can be

[1] `https://www.ncbi.nlm.nih.gov/pubmed/`, as of November 2018
[2] The current size of the database can be obtained by typing "1800:2100[dp]" into the search bar.
[3] `https://www.ncbi.nlm.nih.gov/pmc/`, as of November 2018
[4] `https://www.ncbi.nlm.nih.gov/pmc/?term="prostate+cancer"`, as of November 2018
[5] Number obtained as of 12 October 2017

extracted from public databases, the rate at which new research articles are published greatly exceeds the rate at which this information can be currently processed, resulting in an ever wider gap between available knowledge and easily accessible information, e.g., information stored in a database. Clearly the development of novel methodologies that can automatically analyze textual sources, extract facts and knowledge, and produce summarized representations that capture the most relevant information are needed more than ever.

We present here a novel approach to automatically extract knowledge from biomedical publications. Our approach is generic and can be applied to any knowledge domain, but we focus here, as a proof of concept, on the problem of identifying and extracting Protein–Protein Interactions (PPIs) from the biomedical literature related to prostate cancer (PC), a complex disease with multi-factorial etiology.

PC is the second most common cancer type and the fourth leading cause of cancer death in men worldwide [2]. Despite the large number of newly diagnosed cases, the majority of them, in older men, are clinically insignificant, meaning that the life expectancy of the patient is shorter than the time required by the disease to manifest any symptoms [3]. However a small fraction of new cases are aggressive cancers that require intervention. The current prognostic factors are not sufficient to precisely stratify these two groups [4], and thus PC is prone to overdiagnosis and treatment associated with debilitating side effects[5].

While various approaches to automatically extract PPIs information from unstructured text are already available, many of these methods require feature engineering and expert-domain knowledge for good performance, hence preventing full automation. Commonly proposed methodologies exploit machine learning approaches [6, 7], data mining tools [8], co-occurrences [9–12], or rules-based text mining [13].

Recently, word embedding techniques based on deep learning have been proposed as a more advanced approach to process textual information in an unsupervised fashion. Word embedding is a term used to identify a set of methods for language modeling and feature learning, where words in a vocabulary are mapped into vectors in a continuous, high dimensional space, typically of several hundred dimensions [14]. In this representation, words that share a similar context in the corpus are located in close proximity in the word embedding vector space. Besides representing words' distributional characteristics, word vectors can capture the semantic and positional information of a word in a text, providing a richer vector representation than frequency-based approaches. Word vector representations have gained broad recognition thanks to the recent work of Mikolov *et al.* [15, 16], who demonstrated that word embeddings can facilitate very efficient estimations of continuous space word representations from huge

datasets (~1.6 billion words).

Since this seminal work, word embeddings based on neural networks have been adopted to address different tasks of natural language processing. For instance, word embeddings have been used for the task of event trigger identification [17], i.e., to automatically detect words or phrases that typically signify the occurrence of an event. Zhou *et al.* [18] used a combination of features extracted from a word embedding plus syntactic and semantic context features to train a support vector machine classifier for the task of identifying event triggers. Such approaches have been shown to be efficient in identifying the semantic and syntactic information of a word and incorporate it into a predictive model. Word embeddings have also been used as token features, i.e., semantic units of words and characters extracted from a corpus for further processing, to extract complete events represented by their trigger words and associated arguments [19]; to build knowledge-regularized word representation models that incorporate prior knowledge into distributed word representations for semantic relatedness ranking tasks [20]; and to simultaneously analyze the semantic and contextual relationship between words [21]. Finally, alternative deep learning approaches based on autoencoders and a deep neural network have been proposed to extract PPIs, where the features are extracted by a Named Entity Recognition (NER) module coupled to a parser and principal component analysis [22]. While these methodologies have shown the versatility of word embeddings to support text analysis through current natural language processing tools, approaches that can automatically information from unstructured text in a completely unsupervised manner are still missing. To bridge this gap we present our methodology hereby referred as INtERAcT (Interaction Network infErence from vectoR representATions of words).

Our approach can be summarized as follows. We first create a word embedding from a corpus of interest. Next, we cluster the learned word vectors in the embedded word space and find groups of words that convey a close semantic and contextual similarity. Then we develop a novel similarity measure based on the Jensen-Shannon divergence to predict interactions from the embedded word space. As a proof of concept, we focus on proteins and predict PPIs using a biomedical corpus of cancer-related publications.

We benchmark INtERAcT against known similarity metrics for word vectors, such as the Euclidean metric, cosine and correlation distances, using the STRING [1] database [23] as a ground truth. We test the performance of INtERAcT on a wide range of parameters using embeddings built both using article abstracts and full texts. Our method exhibits a strong robustness against parameters choices and good accuracy, especially in the small corpus regime.

[1] `https://string-db.org/`, as of November 2018

To demonstrate the potential of our approach in this regime, INtERAcT is applied to multiple sets of publications related to 10 different cancer types showing superior performance compared to the other considered metrics.

3.2 Results

3.2.1 Applying INtERAcT to prostate cancer publications

Building a word embedding specific for prostate cancer

In the following section we describe the application of INtERAcT to the problem of reconstructing a prostate cancer pathway. Text pre-processing and building of the word embedding follows the methodology described in Section 3.4. Briefly, a text corpus is assembled by downloading the XML version of ~140,000 PubMed Central publications matching the query *"prostate cancer"*. Only abstracts are processed, as we find they provide a concise and cleaner summary of the article's main findings than the article full text (see Figure 3.7), while substantially reducing the computational cost associated with building the embedding.

Rare words and bi-grams occurring less than 50 times in the corpus are removed. The remaining sentences are tokenized, i.e., segmented into linguistic units, and used to build a word embedding. After processing (see Section 3.4.1), our dictionary is composed of ~21,000 single words and common bi-grams, e.g., prostate_cancer, cell_proliferation and gene_expression. Using this dictionary, we build a word embedding using a vector representation of 500 dimensions and a context window of nine words (4 words to the right and 4 to the left of each target word). These parameters have been chosen as those leading to optimal performance after extensive space parameter exploration (see Section 3.4.4 and Figures 3.5 and 3.6).

Applying INtERAcT

The word vectors are clustered into groups conveying similar semantic meaning using K-means with 500 clusters. We next identify the k-nearest neighbors of each protein as described in Section 3.4. The neighborhood size is set to $k = 500$. These parameters are also selected making use of the results obtained in Section 3.4.4.

We use neighbors' cluster assignment distribution of selected words to calculate a pairwise similarity scores based on the Jensen–Shannon divergence (JSD) as shown in Equation 3.5.

This last step is performed on a subset of words, in this example, a list of molecular entities defined using UniProt [24]. We interpret this JSD-based distance metric as the likelihood of a PPI. See Section 3.4 and Fig. 3.1 for details.

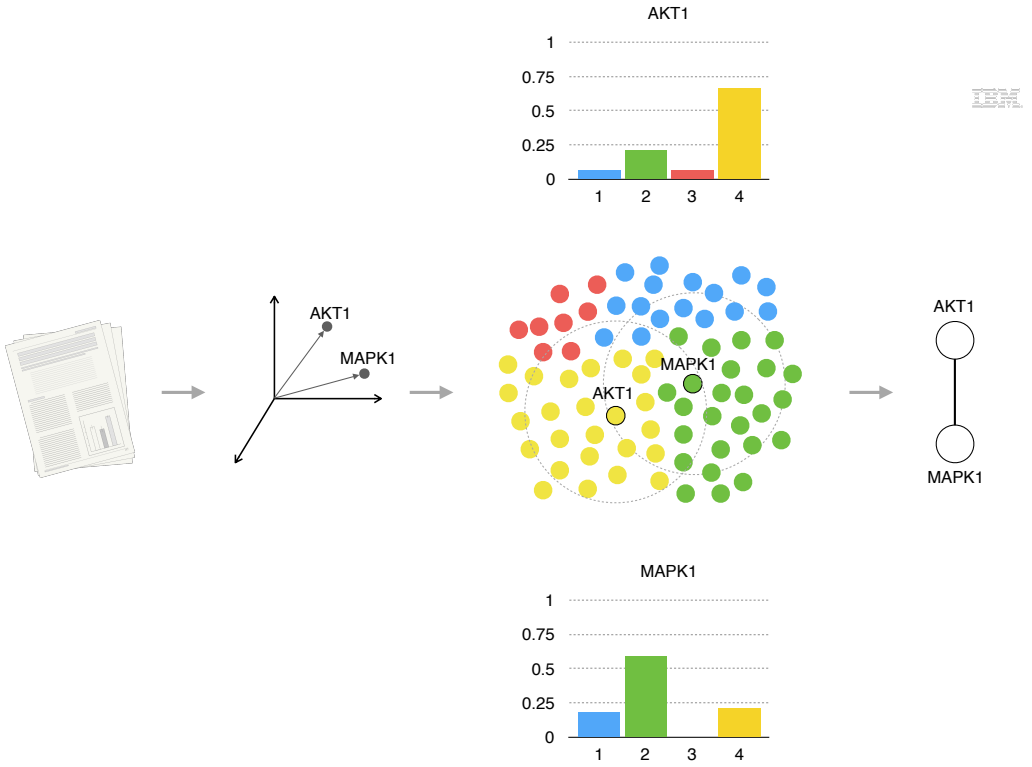


Figure 3.1: **Schematic representation of INtERAcT.** Text is used as input to generate a word embedding. The word vectors are clustered into groups of similar semantic meaning and the distributions of each word's neighbors across clusters are used to compute and predict interactions between molecular entities.

To benchmark the inferred network we focus on the list of molecular entities reported in the prostate cancer pathway as defined by the Kyoto Encyclopedia of Genes and Genomes [1] (KEGG) [25, 26] and apply INtERAcT to the task of reconstructing the connectivity between these entities. Out of the 87 molecular entities that constitute the KEGG pathway, 67 are found in the embedding, and thus can be used as a validation set.

We interrogate INtERAcT and query the interactions between the 67 proteins of our validation set. Fig. 3.2a graphically shows the top-50 inferred interactions in our prostate cancer gene validation set. The full set of interactions with similarity scores can also be found as

[1] `http://www.genome.jp/dbget-bin/www_bget?pathway+map05215`, as of November 2018

a table in the Supplementary Material 3.2. Please, notice that while KEGG provides a well-established reference for function-specific pathways, KEGG merges gene family members in a single node-entity (e.g., AKT1, AKT2 and AKT3 become AKT), and hence a direct comparison between KEGG prostate cancer pathway and INtERAcT inferred results is not possible.

Comparing INtERAcT to STRING

The results from INtERAcT with a prostate cancer specific embedding are compared to other metrics by benchmarking against STRING (version 10.5). For the purpose of benchmarking, we use the combined score provided by STRING, as using a variety of interaction evidences approximates better the true network connectivity than confidence scores predicted using a single method (see Section 3.4.4).

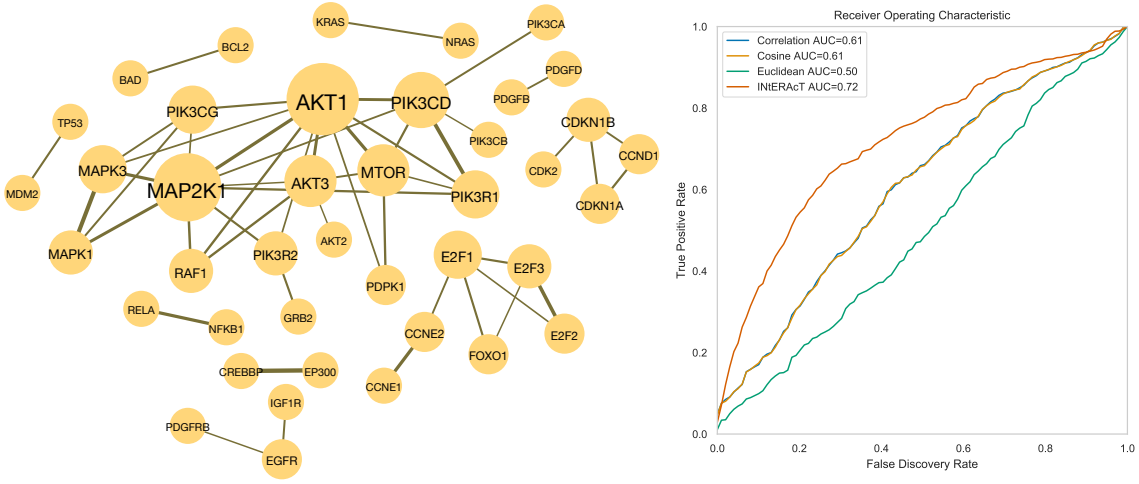
Figure 3.2b reports a summary of our findings. The Receiver Operating Characteristic (ROC) curve for the INtERAcT score (red curve), a cosine distance-based score (orange curve), a correlation-based score (blue curve) and an Euclidean distance-based score (green curve) are comparatively shown. INtERAcT achieves a 0.72 AUC, significantly better than both cosine and correlation scores, which achieve a 0.61 AUC. The Euclidean based distance measure performs closely to a random predictor with an AUC value of 0.50. This poor performance is expected as the Euclidean distance, and more generically, L_k norms, tend to map pairs of points to uniform distances in high dimensional spaces [27].

The curves' trends reinforce the intuition that a neighborhood-aware metric is better able to capture functional associations from unstructured text than methods that limit their analysis to the positions of word vectors in the embedding.

As an additional measure of agreement, we also compute the rank correlation between INtERAcT and STRING scores. To compute the correlation values, all predicted interactions by INtERAcT and STRING have been used without applying any confidence cut-off. INtERAcT and STRING scores used to compute the correlations are provided as additional supplementary tables. The resulting correlation value is positive and very significant ($\rho = 0.39$, $p = 3.2e^{-68}$), and is higher compared to the correlation obtained using cosine, correlation and Euclidean distance-based metrics ($\rho = 0.30$, $p = 4.0e^{-39}$, $\rho = 0.30$, $p = 2.9e^{-39}$, and $\rho = 0.19$, $p = 1.3e^{-16}$ respectively). INtERAcT outperforms again cosine, correlation and Euclidean distance-based metrics.

We note that while the correlation values obtained for INtERAcT and correlation-based or cosine-based scores seem to be relatively close, their difference is highly significant with a p-

value of $p = 0.001$ when the number of interaction scores used to compute the correlations is taken into account (number of interactions = 1825). The significance of the difference of two correlation values can be computed using the Fisher z-transformation [28], which transforms the Spearman correlation values into normally distributed variables whose difference can be evaluated using a standard t-test.



(a) **(b)**

Figure 3.2: **(a) Top 50 prostate cancer protein-protein interactions inferred by INtERAcT**. The prostate cancer gene set has been defined according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) prostate cancer pathway, and includes molecular entities known to be important in prostate cancer onset and development. The interactions and associated scores have been computed using a word embedding trained on ~140000 prostate cancer open-access abstracts from PubMed Central and INtERAcT. Node size is proportional to node degree while edge width is proportional to the intensity of the interaction. **(b) Performance of INtERAcT on a prostate cancer gene validation set compared to other distance measures using STRING as a ground truth.** We use ROC (Receiver Operating Characteristic) curves to quantify the accuracy of the inferred interactions in a set of prostate cancer-related genes. INtERAcT (red curve) significantly outperforms alternative, commonly used metrics on a word embedding such as a cosine distance-based similarity (orange curve), correlation-based similarity (blue curve) and a similarity score based on the Euclidean distance (green curve).

3.2.2 Applying INtERAcT on other cancer pathways

We next focus on investigating the generalization of INtERAcT to other knowledge domains. For this task, we extend our analysis to nine additional cancer types: acute myeloid leukemia,

bladder cancer, chronic myeloid leukemia, colorectal cancer, glioma, small cell lung cancer, non-small cell lung cancer, pancreatic cancer and renal cell carcinoma.

The gene sets for each cancer type are taken from their respective cancer-specific pathway as annotated in KEGG. These cancer types are selected according to two criteria: first, there is a cancer-specific KEGG pathway to define a gene set, and second, we could retrieve at least 10000 cancer-specific publications in PubMed Central. The second criterion is needed in order to obtain a corpus size that guarantees a good reconstruction of the word vectors when building the word embedding. We then defined new query words specific to each cancer type and repeated the procedure described in 3.2.1. The full list of used query words for each cancer type can be found in the Supplementary Material (Section 3.5.5),

In Figure 3.3 we report the average ROCs for the four considered distance metrics: INtERAcT (red curve), cosine (orange curve), correlation (blue curve) and Euclidean (green curve) metrics. In order to obtain a confidence for the curves using the different pathways considered, we built empirical confidence intervals (CIs). The CIs are generated performing an empirical bootstrap on the different pathways. Namely, for each false positive rate level, values are sampled with replacement from the true positive rate values obtained from the different pathways to generate an empirical distribution and build the intervals. The CIs at level 68% are reported (one standard deviation from the mean) in Figure 3.3.

Finally, we compare the similarity of scores predicted by INtERAcT and STRING by computing the Spearman rank correlation between both sets of scores. The values are shown in Table 3.1. For all analyzed pathways, the correlations are positive with a strongly significant p-value (the p-values range from 10^{-08} to 10^{-70}).

Our findings suggest that while having a large corpus is of paramount importance to obtain accurate word vector representations, INtERAcT is still able to find strong correlations with STRING score used as ground-truth even when the size of the corpus is reduced (see Table 3.1). For instance, the highest correlation value 0.55 is found in colorectal cancer, which has the second highest number of publications used to build the embedding. However, prostate cancer only shows a moderate correlation of 0.39, while having the largest number of publications used. We hypothesize that while having a large corpus of publications is beneficial to build a high quality embedding, very active fields of research where a high number of publications are available may also be prone to having a high rate of *noisy publications*. Here noise can take the form of low quality publications that report inconsistent results, or studies based on high-throughput analyses with a high false discovery rate. We also note that in taking STRING as ground truth we are implicitly absorbing its false and true discovery rates into

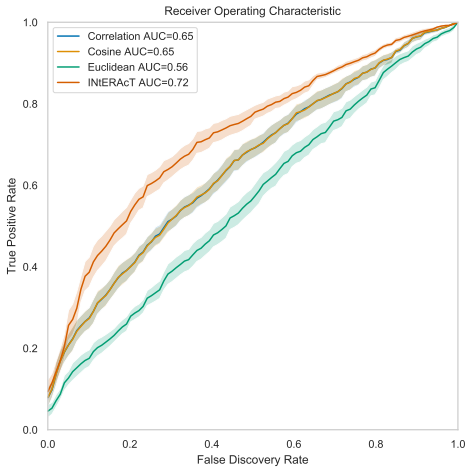


Figure 3.3: **INtERAcT performance compared to other distance measures using STRING as a ground truth.** We use ROC (Receiver Operating Characteristic) curves to quantify the quality and performance of inferred interactions. The curves here reported refer to the inference performed on the KEGG cancer pathways considered in the analysis. Using naive approaches such as a similarity based on the Euclidean distance (green curve) between word vectors led to poor results. Other methods such as cosine-based similarity (orange curve) or correlation-based similarity (green curve) showed an improvement. INtERAcT (red curve) achieved the best performance predicting interactions reported in STRING. The confidence intervals (CIs) at level 68% are reported (one standard deviation from the mean). To generate the empirical distribution we used sampling with replacement at different false positive rates of the true positive rates given by the different pathways. The confidence intervals reported are at level 68% (one standard deviation from the mean)

our error rates. For instance, interactions reported by STRING that might occur in a different context but not in cancer (e.g., mouse interactions not occurring in cancer) will get penalized as false negatives if INtERAcT correctly predicts them as a non-interaction.

Taken all together and within the limitation of not having an unbiased ground truth to evaluate our predictions, INtERAcT shows a good agreement with the information reported by STRING. Our results indicate that our unsupervised approach is able to recapitulate to a large extent the knowledge obtained through manual curation of scientific literature.

3.3 Discussion

We have presented, INtERAcT, a fully unsupervised method to automatically extract context-specific information from a corpus of biomedical publications, without any doubt, the fastest

Pathway	Correlation	p-value	Proteins	Papers
KEGG Acute Myeloid Leukemia	0.401425	1.06e-19	34	34532
KEGG Bladder Cancer	0.436745	1.67e-19	30	35331
KEGG Chronic Myeloid Leukemia	0.386765	1.50e-09	23	14247
KEGG Colorectal Cancer	0.550515	1.75e-70	48	118336
KEGG Glioma	0.359502	1.20e-18	36	64712
KEGG Small Cell Lung Cancer	0.315544	1.68e-08	28	32233
KEGG Non Small Cell Lung Cancer	0.406438	1.27e-17	31	67048
KEGG Pancreatic Cancer	0.372069	6.40e-30	47	62668
KEGG Prostate Cancer	0.392346	3.20e-68	67	132357
KEGG Renal Cell Carcinoma	0.436745	1.67e-19	30	37169

Table 3.1: **INtERAcT-STRING rank-correlation on KEGG's cancer pathways.** The table reports the Spearman correlation and p-values of INtERAcT predictions and STRING-derived scores for different KEGG pathways. The number of proteins in each pathway, as well as the number of papers used to build each embedding is also reported. For all analyzed pathways and cancer types, the correlation is positive and highly significant.

growing source of scientific information. As a proof of concept, we have focused on the problem of extracting protein-protein interactions from cancer-specific corpora, although our approach can be easily applied to other scientific domains and research questions. The strength of INtERAcT lies in being completely unsupervised: no time-consuming manual curation, expert knowledge or labelling of the text is required for knowledge extraction.

We have described the steps to reconstruct a context-specific pathway from prostate cancer publications. When comparing the inferred interactions to STRING, our method outperforms other scores built on commonly used metrics (cosine, correlation and Euclidean metric). On a more extensive validation on multiple cancer pathways, the results remain consistent and we have a significant agreement on the information reported by STRING. We would like to highlight that STRING predicts interactions using a combined score that integrates information from many disparate data sources including genomic proximity, gene fusion events, phylogenetic co–occurrences, homology, co–expression, experimental evidence of interaction, simultaneous annotation in databases and automatic text-mining, where text-mining is done using a combination of co-occurrences and natural language processing based on a rule-based system [29]. Opposed to this, our methodology is a completely unsupervised approach that does not require expert knowledge or rules setting. When focusing on reconstructing a prostate cancer pathway, we achieved a 0.72 AUC score using STRING as benchmark. We notice that the choice of benchmark is likely overpenalising the evaluation of the precision and recall of

our method, as STRING reports many interactions that are not cancer-specific.

We have also performed a systematic exploration of the space of parameters in order to identify those leading to an optimal performance. Remarkably, on the large corpus size regime (~4 millions freely available publications in PubMed), INtERAcT is very robust to changes in parameters at the price of a small decrease in performance compared to the peak performance of other metrics. This is in stark contrast to other similarity metrics that, while achieving a peak performance slightly superior to INtERAcT, show very variable results as a function of parameters. This variability would require a very computational intensive parameter optimization for any new application of those metrics to a different text corpus. On small corpus sizes, ~10,000-100,000, characteristic of disease-specific literature, INtERAcT performance clearly surpasses the other similarity metrics.

To summarize, we expect INtERAcT to be highly relevant for a variety of state-of-the-art text-mining methods. Especially, we are convinced that the proposed methodology can be used to generate hypotheses for detection of biological processes relevant to common and complex diseases and can establish a novel, unsupervised and high-throughput approach to drive drug discovery and advance the frontier of targeted therapies. In order to facilitate the exploitation of INtERAcT, we have developed an open access web service to freely access INtERAcT.

3.4 Methods

In this section we present the elements that constitute INtERAcT and describe the approach adopted to automatically build a network of molecular interactions starting from a domain-specific text corpus.

3.4.1 Text processing

We begin by using a basic and lightweight pipeline for text processing. First, we filter out non-informative words such as extremely common words (e.g., a, the, and other stop-words), rare words (low occurrence in the corpus), non-informative characters like punctuation or isolated numbers and convert text to lower-case. Please, notice that we only remove isolated numbers in order to leave intact and be able differentiate gene names (e.g., AKT1, AKT2 and AKT3). We next identify n-grams, i.e., sequences of words that often appear together

and thus are considered a single entity (e.g., New_York), by summing up the occurrences of words appearing sequentially together in the corpus and setting a threshold of the minimal number of occurrences. The names of a gene, its aliases and corresponding protein are treated as synonyms and mapped to a single name entity using a dictionary obtained from UniProt [1]. Sentences are generated using an English language tokenizer from *nltk* [30] before punctuation is removed. The result of this process is a corpus of sentences that can be used for further analysis.

3.4.2 Word embeddings

Word embeddings are the output of a family of methods that produce, starting from raw text, a real vector representation of each word and phrase contained in the original text corpus. In this work we build the embedding using the Word2Vec implementation proposed by Mikolov et al. [16], a shallow, two-layer neural network based on a skip-gram model. Briefly, the skip-gram model aims to predict the surrounding words, i.e., the context, of a target word given as an input (see Fig. 3.4). In practice, a word's context is defined by considering a window of an arbitrary size to the left and the right of each target word. Each pair target-context word is then fed into the neural network with a single hidden layer of dimension d that is trained to optimize the probability of predicting context words given a target word as input. It has been reported that the quality of the word embedding increases with the dimensionality of the internal layer that produces the vector representation, d, until it reaches a critical point where marginal gain diminishes [15]. Hence this parameter has to be appropriately chosen according to the size of the vocabulary and text corpus.

The word embedding learning process is naturally optimized to capture the contextual associations between words: If two words tend to appear in similar contexts, they will be mapped into similar word vectors. In practice, it has been shown that word embeddings outperform methods based on counting co-occurrences of words on a wide range of lexical semantic tasks and across many parameter settings [31].

[1] `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz`, as of November 2018

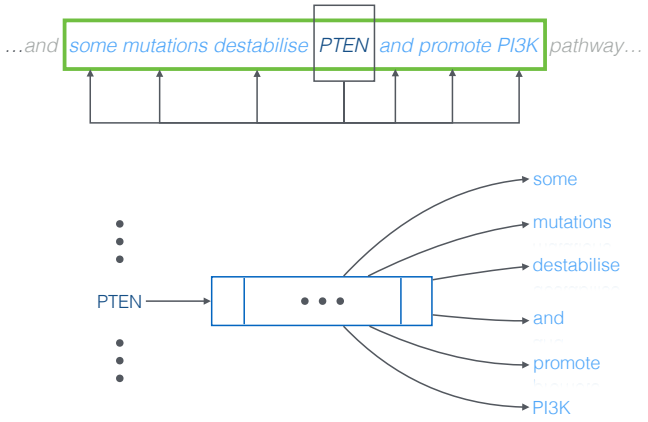


Figure 3.4: **Description of the skip-gram model.** Skip-gram model used in Word2Vec to find an optimal representation to predict the surrounding context of a target word. The example highlights the window around PTEN, a gene implicated in many cancer processes. The target word, PTEN, is linked to each of its neighboring words and the pairs are fed into the network. The learning process optimizes the probability of predicting the contextual words of PTEN.

3.4.3 Extracting interactions from the embedding

Once the embedding is built, our aim is to design a methodology that can predict PPIs based on the distribution of word vectors in the word embedding. We exploit the idea that molecular entities that interact with each other and are involved in similar biological processes are likely to appear in similar word contexts, and thus will be mapped to neighboring positions in the word vector space. It is hence possible to predict functional similarities between molecular entities based on their mapping and proximity in the word embedding.

Our task is therefore to find optimal ways of measuring proximity in the word embedding. A first, obvious approach to define proximity between two word vectors is to use the Euclidean distance and a distance threshold: molecular entities within this threshold can be considered similar and thus predicted to interact. However, the use of the Euclidean metric, and more generically, the use of L_k norms, is problematic as the high dimensionality of the space can make certain regions of the space too sparse. In addition, in high dimensional spaces L_k norms map points to uniform distances from each other, and hence the concepts of proximity, distance or nearest neighbor are not quantitatively meaningful [27].

INtERAcT exploits an alternative approach that does not rely on the direct use of L_k norms, but instead defines similarities between words by looking at the distributional properties of their neighbors. Specifically, we predict PPIs by comparing the neighborhoods of words representing molecular entities. To do so, we first need to cluster the word vectors of the embedding.

Clustering words

We start by defining $\mathcal{W}$ as the set of n words present in the embedding $\mathcal{E} \in \mathbb{R}^{n \times d}$ where d is the embedding dimension, which corresponds to the dimension of the neural network's hidden layer used to build the embedding. We cluster the word vectors in the embedding space using a K-means algorithm with C clusters. The number of clusters is chosen according to the vocabulary size in order to have both a fine grained word representation and a sufficient number of words per cluster. Each word is hence associated with a cluster according to the following mapping:

$$CL : \mathcal{W} \rightarrow \{1, \ldots, C\} \tag{3.1}$$

.

The obtained clusters group together words that are close in the vector representation space and hence tend to appear in similar contexts in the corpus. These clusters can then be used to build fingerprints of each entity in the embedding and to convey the semantic meaning of a word based on the cluster membership of its neighbors.

Finding nearest neighbors

In order to build word fingerprints, our algorithm requires the identification of the nearest neighbors of each target word. An efficient method to retrieve the topological neighbors without having to compute all pairwise distances at each query is k-d trees, a space-partitioning data structure that can be used to organize points in a k-dimensional space [32]. A nearest-neighbor-search can then associate every word in the embedding with a set $\mathcal{N}$ of K nearest neighbors in the embedding:

$$KNN : \mathcal{W} \rightarrow \mathcal{N} \tag{3.2}$$

The optimal number of neighbors depends on the number of clusters C, and it is chosen

as a trade-off between the benefit of having enough cluster assignment variability among the neighbors, while keeping the neighborhood of each word small when compared to the total word count of the embedding. The mapping KNN can be used to efficiently retrieve the shortest paths between two words and identify their nearest words.

Word distribution

We are now able to associate each word in the embedding with a discrete probability distribution that can be computed by analyzing the cluster membership of the nearest neighbors. The number and cluster occupancy of the neighbors can then be interpreted as a discrete probability distribution conveying the semantic meaning of each target word. Furthermore, pair-wise comparisons of these distributions enable us to define similarities between words (see Fig. 3.1).

The pseudo-code used by the described algorithm can be found in the Supplementary Material, Section 3.5.2. The output of the algorithm is a matrix of probability distributions $\mathcal{D} \in \mathbb{R}^{n \times C}$ where each row contains the cluster assignments of each target word.

Computing similarity scores

We can finally compute the functional association between words of interest by computing the similarities between the neighbors' cluster assignments of protein entities in the embedding. We use a score based on the Jensen-Shannon divergence (JSD), defined as follows:

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{3.3}$$

where $M = \frac{1}{2}(P+Q)$ and D_{KL} is the Kullback-Leibler divergence for discrete probability distributions:

$$D_{KL}(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) \tag{3.4}$$

.

The choice of the JSD as a scoring function is motivated by its useful properties. In addition to providing a symmetrized version of the Kullback-Leibler divergence, JSD is a finite value comprised in the interval $[0, \log(2)]$ [33], the lowest bound being reached when two dis-

tributions are identical. Furthermore, the square root of the Jensen–Shannon divergence is a metric [34], and thus JSD is an appropriate function to capture similarities between distributions.

Here we take advantage of the non-negativity of JSD to define the similarity S_{ij} between words i and j as follows:

$$S_{ij} = \exp(-\alpha JSD_{ij} + \beta) \tag{3.5}$$

where $JSD_{ij} = JSD(\mathcal{D}_i||\mathcal{D}_j)$ and α and β are a scaling and an offset parameters respectively. Here, we set the offset parameter $\beta = 0$. Under the transformation defined by Equation 3.5, two identical distributions have a score equal to 1 and substantially different distributions (with a divergence close to the JSD upper bound) have a score ~ 0.0. While larger values of α can bring this theoretical minimal value closer to 0, a very high α will make S_{ij} decay too steeply, shrinking the regime where S_{ij} can effectively rank pairs of words according to their similarity. We found that the choice of $\alpha = 7.5$ and $\beta = 0.0$ is empirically efficient at capturing similarities between words given the theoretical bounds for the JSD (see Supplementary Material Figure 3.8).

Equipped with the similarity score as defined in Equation 3.5, we are now in a position to build a weighted interaction graph where the nodes are the chosen entities (proteins in our case) and the edges are weighted by the similarity value of the nodes they connect.

3.4.4 Benchmarking INtERAcT against STRING

In order to assess the quality of our predictions, we use STRING database [23] as a benchmark. STRING is a comprehensive protein interaction database currently including experimental data from DIP [1] [35], BioGRID [2] [36], IntAct [3] [37], and MINT [4][38], and curated data from BioCyc [5] [39], GO [6] [40], KEGG [7] [25, 26], and Reactome [8][41]. STRING provides a confidence score that integrates information about genomic proximity, gene fusion events,

[1] `http://dip.doe-mbi.ucla.edu/dip/`, as of November 2018
[2] `https://thebiogrid.org/`, as of November 2018
[3] `http://www.ebi.ac.uk/intact/`, as of November 2018
[4] `http://mint.bio.uniroma2.it/`, as of November 2018
[5] `https://biocyc.org/`, as of November 2018
[6] `http://www.geneontology.org/`, as of November 2018
[7] `http://www.kegg.jp/`, as of November 2018
[8] `http://www.reactome.org/`, as of November 2018

phylogenetic co-occurrences, homology, co-expression, experimental evidence of interaction, simultaneous annotation in databases and automatic text-mining [42].

Importantly for the sake of comparing STRING and INtERAcT results, STRING text-mining uses a combination of co-occurrences and natural language processing based on a rule-based system [29]. Throughout this publication, we have used the combined STRING score as the ground truth, as a combination of independent evidences is likely to provide a more accurate and robust network topology.

To quantitatively evaluate the goodness of INtERAcT predictions, we employ the Area Under the receiver operating characteristic Curve (AUC metric [43]) using STRING interactions as a ground truth, and compare our JSD-based score (Equation 3.5) with other similarity scores commonly used in the literature, namely scores based on correlation, cosine and Euclidean distance.

Parametric analysis

The results of the multiple benchmarks presented in the work depend on the ground truth used to compare against, i.e., using the STRING combined score or the text-mining STRING score, as shown in Figure 3.5. Using the STRING text-mined interactions naturally results in higher performance for all four considered similarity metrics, as all of them exploit information available in text corpora. The STRING database suggests three levels of confidence, with lower scores resulting in lower measured performance for metrics considered.

Though INtERAcT extract information from a word embedding, its performance is largely insensitive to the choice of embedding parameters (Figure 3.5). This is in contrast to other distance metrics, for which the performance deteriorates with increasing dimensionality of the word embedding, especially in the Euclidean metric case. This is important as typical word embeddings sizes are usually in the regime where INtERAcT achieves comparable performance to the other metrics, and higher stability (Figure 3.5 bottom left panel and Figure 3.6 right).

Regarding the parameters defining clusters and neighborhoods within the embedding, INtERAcT performs better with larger cluster sizes (Figure 3.6 left), has a slightly better performance for moderately sized neighborhoods (500 neighbors) and, as mentioned above, is largely independent from the dimensionality of the embedding.

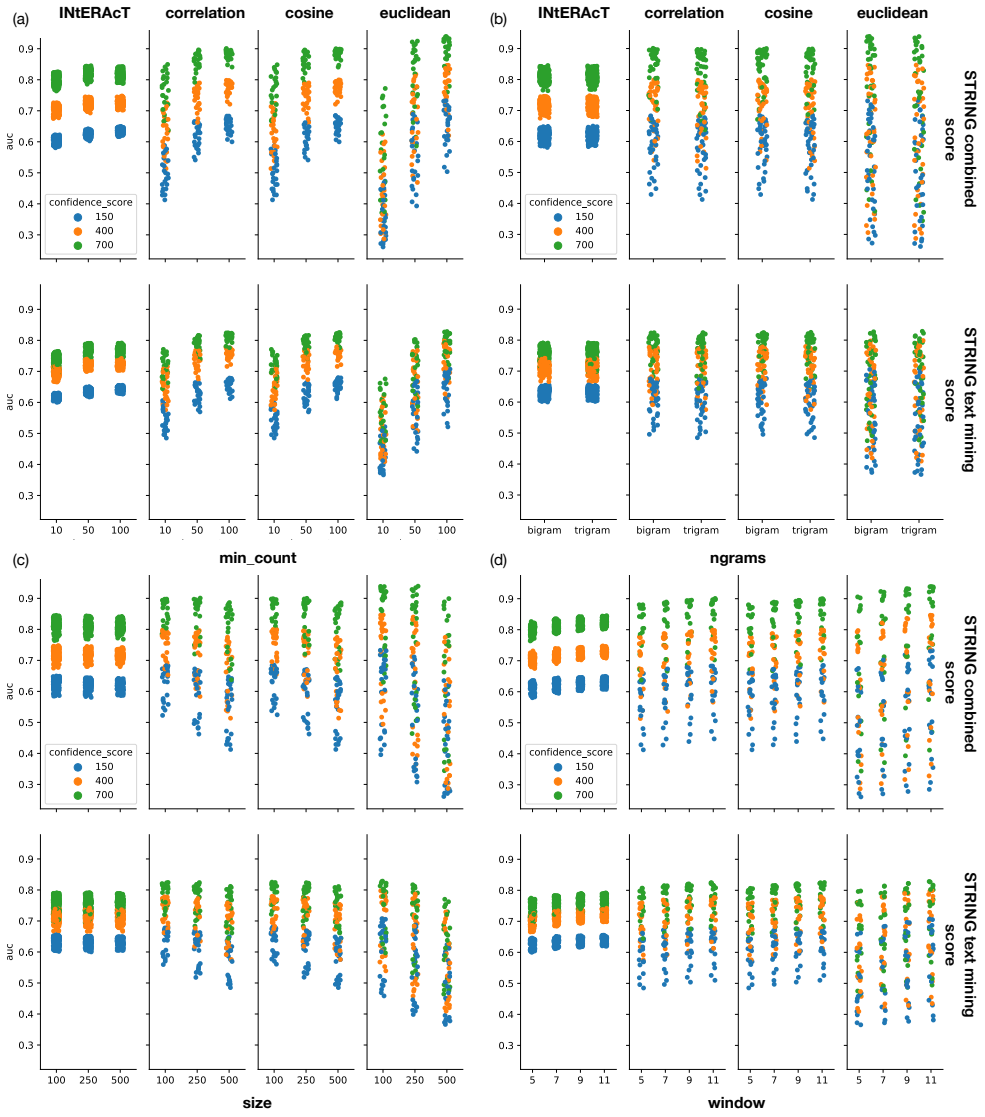


Figure 3.5: **Exploration of the influence of word embedding parameters on AUC for different methods and ground truths.** Many word embeddings with parameters min_-count , n-grams, size and window have been trained. Each embedding has been used for inference with the four studied methods. The inferences have been evaluated using two different ground truths: STRING text-mined interactions and STRING interactions predicted through a combined score integrating diverse computational and experimental evidences. In both cases, ground-truth interactions are color-separated (blue, orange and green) according to the STRING confidence score. Each such combination results in a single data point in the Figures 3.5 and 3.6, except for INtERAcT, which has additional parameters (clusters and neighborhood sizes), making it appear denser. Each of the four panels investigates variation of a single parameter: **(a)** min_count, the minimal allowed occurrence of words in the entire corpus to be included in the embedding; **(b)** n-grams, prior substitution of bi-grams or tri-grams as single tokens. **(c)** size, the dimensionality of the embedding vector. **(d)** window, the size of the window surrounding each target word to be predicted during the learning of the embedding. INtERAcT is largely insensitive to the choice of embedding parameters, with small gains in performance for larger values for min_count and window size.

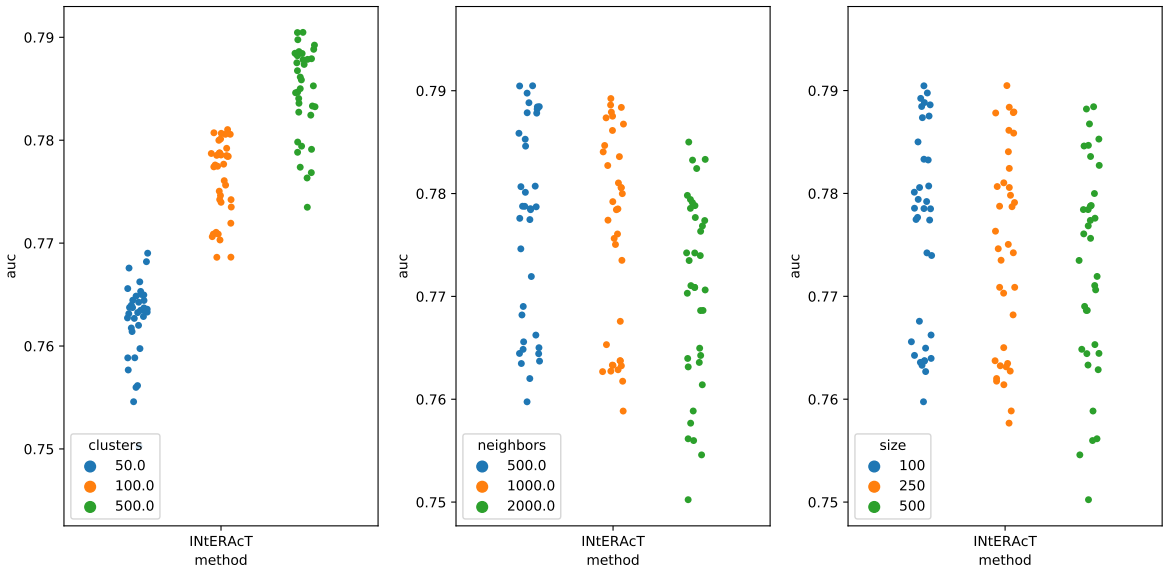


Figure 3.6: **Parametric dependency of INtERAcT using STRING as a ground truth.** AUC for different word embeddings compared to STRING with a confidence score of 700. Word embeddings for both bi-grams and tri-grams, window sizes of 9 and 11 as well as sizes (dimensionality) 100, 250 and 500 with fixed parameter min_count=50 are shown here. Word embeddings were used varying the cluster size and number of neighbors. **Left** Increasing the number of clusters representing similar context in the corpus improves performance of INtERAcT. **Middle** The number of neighbors has a small effect on the performance of INtERAcT. **Right** The size (dimensionality) of the word embedding has no noticeable influence on INtERAcT.

Declarations

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858. We would like to acknowledge Dr. Costas Bekas and Dr. Yves Ineichen for the useful discussions.

Author contributions

See Chapter 8 for details about contributions and copyright.

Competing financial interests

The authors declare no competing financial interest.

Availability of data and materials

INtERAcT is currently available as a service hosted on IBM Cloud at the following address: `https://sysbio.uk-south.containers.mybluemix.net/interact/`. The web service builds a molecular interaction network given word vectors in Word2Vec binary format and a list of molecular entities (example data are made available through a download link in the app). The article abstracts used to generate INtERAcT protein-protein interaction scores can be freely downloaded from PubMed Central. The article collection as well as access to INtERAcT are also available from the corresponding author on request. For this project, STRING interaction scores were downloaded on 19/10/17 (version 10.5). STRING historical data can be downloaded from `https://string-db.org/cgi/access.pl?footer_active_subpage=archive`, or obtained from the corresponding author on request. The networks generated for KEGG cancer pathways and the corresponding word vectors and entities lists are available via download link [1]. All the other data generated or analysed during this study are included in this published article and its supplementary information files.

[1] `https://ibm.box.com/s/njtqm1alrkq1dxm5jtdnqm8vei4uo58m`, as of November 2018

3.5 Supplementary information

3.5.1 Parametric analysis

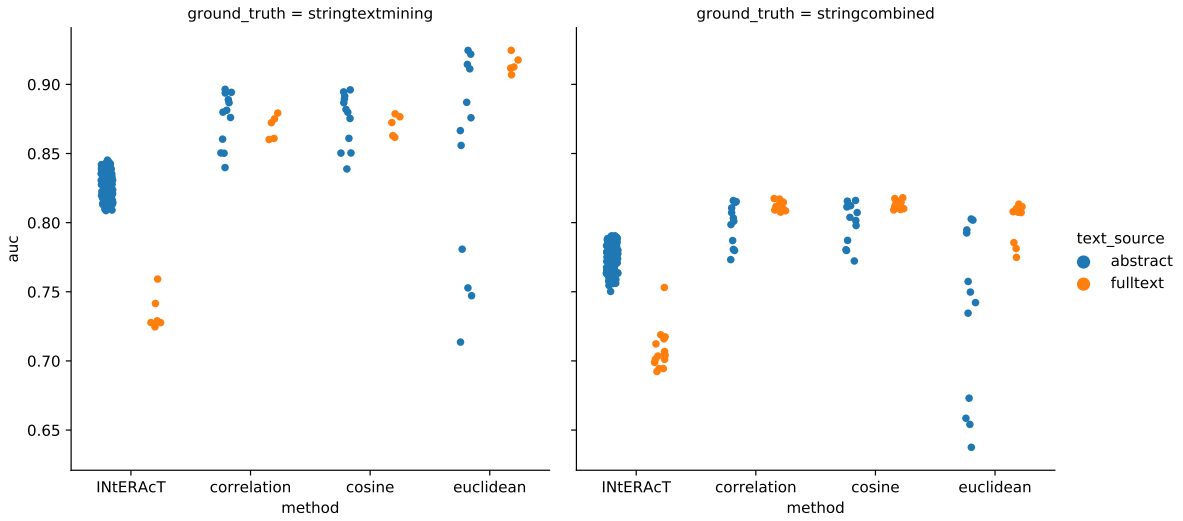


Figure 3.7: **Word embedding source comparison.** Full text versus abstract. AUC for different word embeddings compared to two ground truths obtained by STRING with a confidence score of 700. Word embeddings for both bi-grams and tri-grams, window sizes of 9 and 11 as well as sizes (dimensionality) 100, 250 and 500 with fixed parameter min_count=50 are shown here. **Left** Text-mining ground truth. **Right** Combined ground truth.

3.5.2 Word distributions

Algorithm 1 Word distributions

procedure WORDDISTRIBUTIONS.($\mathcal{W}, CL, KNN$)
  $\mathcal{D} \leftarrow \{\}$ ▷ Define a matrix to store distributions
  for $w \in \mathcal{W}$ **do** ▷ For all the words
    $D \leftarrow []$ ▷ Define a vector to store w neighbors cluster
    $NE \leftarrow KNN(w)$ ▷ Getting K neighbors
    for $ne \in NE$ **do**
      append $CL(ne)$ to D
    $H \leftarrow histogram(D)$
    append row H to $\mathcal{D}$
  return $\mathcal{D}$

3.5.3 Score analysis

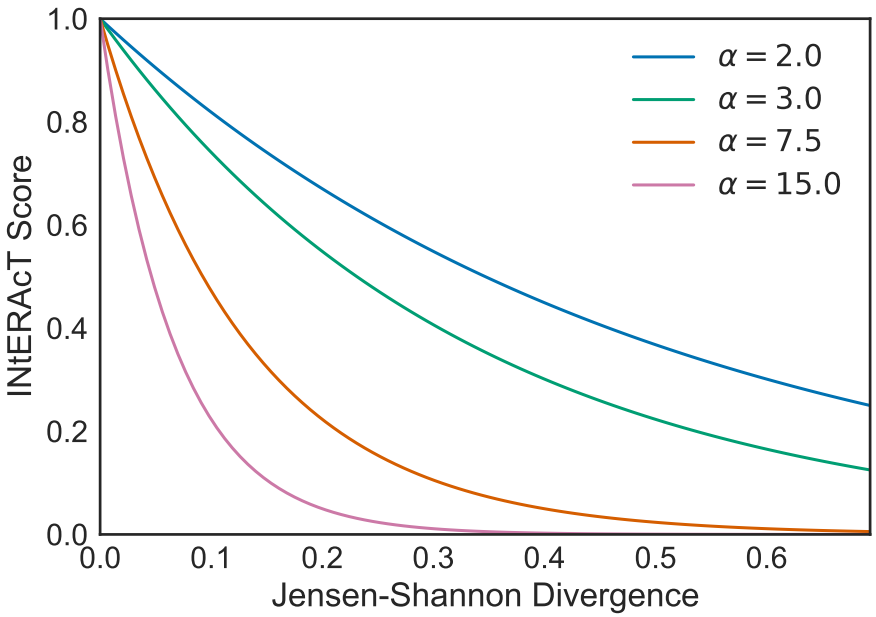


Figure 3.8: **INtERAcT score analysis.** The curves reported describe how the divergence values are mapped into scores by Equation 3.5 setting $\beta = 0.0$ and for different α values. The orange line corresponds to the selected value of $\alpha = 7.5$. Other α values don't map properly the divergence values in a [0,1] interval.

3.5.4 prostate cancer scores

Protein	Protein	Score
CREBBP	EP300	0.78
MAPK1	MAPK3	0.76
PIK3CD	PIK3R1	0.73
E2F2	E2F3	0.71
CCNE1	CCNE2	0.68
AKT1	MAP2K1	0.68
AKT1	MTOR	0.68
NFKB1	RELA	0.68
MAP2K1	MAPK3	0.67
MAP2K1	MAPK1	0.66
AKT1	AKT3	0.66
AKT1	PIK3CD	0.65
MAP2K1	RAF1	0.61
AKT3	RAF1	0.61
AKT1	RAF1	0.60
MTOR	PDPK1	0.60
MAP2K1	PIK3R1	0.60
MTOR	PIK3CD	0.59
CCND1	CDKN1A	0.59
AKT1	PIK3R1	0.59
MAP2K1	PIK3R2	0.59
E2F1	E2F3	0.59
AKT1	PIK3CG	0.58
GRB2	PIK3R2	0.58
E2F1	FOXO1	0.58

Protein	Protein	Score
CDKN1A	CDKN1B	0.58
MDM2	TP53	0.58
BAD	BCL2	0.58
KRAS	NRAS	0.57
MAPK1	PIK3CG	0.57
PIK3CA	PIK3CD	0.57
EGFR	IGF1R	0.57
AKT3	MTOR	0.57
MAPK3	PIK3CG	0.57
CDK2	CDKN1B	0.56
CCNE2	E2F1	0.56
CCND1	CDKN1B	0.56
MAP2K1	PIK3CD	0.56
PDGFB	PDGFD	0.56
AKT1	PDPK1	0.55
AKT1	MAPK3	0.55
AKT1	PIK3R2	0.55
MAP2K1	PIK3CG	0.54
E2F1	E2F2	0.54
MTOR	PIK3R1	0.54
E2F3	FOXO1	0.53
AKT3	MAP2K1	0.53
EGFR	PDGFRB	0.52
AKT2	AKT3	0.52
PIK3CB	PIK3CD	0.52

Table 3.2: **INtERAcT top-50 scores for KEGG prostate cancer pathway.** Top-50 interactions predicted from KEGG prostate cancer pathway using INtERAcT corresponding to the edges of the graph shown in Figure 3.2a.

3.5.5 PubMed Search Queries

	Query on PubMed
KEGG Acute Myeloid Leukemia	"acute myeloid leukemia"
KEGG Bladder Cancer	"bladder cancer"
KEGG Chronic Myeloid Leukemia	"chronic myeloid leukemia"
KEGG Colorectal Cancer	"colorectal cancer"
KEGG Glioma	"glioma"
KEGG Small Cell Lung Cancer	"small cell lung cancer"
KEGG Non Small Cell Lung Cancer	"non small cell lung cancer"
KEGG Pancreatic Cancer	"pancreatic cancer"
KEGG Prostate Cancer	"prostate cancer"
KEGG Renal Cell Carcinoma	"renal cell carcinoma"

Table 3.3: **PubMed Search queries for KEGG's cancer pathways.** In the Table we report the search query that was used for each KEGG cancer pathway. We used the quotation marks to increase specificity.

References

[1] M. Manica, R. Mathis, and M. R. Martínez. "INtERAcT: Interaction Network Inference from Vector Representations of Words" (2018). arXiv: 1801.03011.

[2] J. Ferlay, I. Soerjomataram, R. Dikshit *et al.* "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012". *Int. J. Cancer* **136**:5 (2015), E359–386.

[3] A. R. Zlotta, S. Egawa, D. Pushkar *et al.* "Prevalence of Prostate Cancer on Autopsy: Cross-Sectional Study on Unscreened Caucasian and Asian Men". *Journal of the National Cancer Institute* **105**:14 (2013), pp. 1050–1058.

[4] M. R. Cooperberg, J. M. Broering, and P. R. Carroll. "Risk Assessment for Prostate Cancer Metastasis and Mortality at the Time of Diagnosis". *Journal of the National Cancer Institute* **101**:12 (2009), pp. 878–887.

[5] C. R, C. JM, D. T, and et al. "Screening for prostate cancer: A review of the evidence for the u.s. preventive services task force". *Annals of Internal Medicine* **155**:11 (2011), pp. 762–771.

[6] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. "A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature". *PLoS Computational Biology* **6**:7 (2010). (Visited on 07/06/2017).

[7] E. Tjioe, M. W. Berry, and R. Homayouni. "Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization)". *BMC Bioinformatics* **11**:Suppl 6 (2010), S14. (Visited on 07/04/2017).

[8] S. Mandloi and S. Chakrabarti. "PALM-IST: Pathway Assembly from Literature Mining - an Information Search Tool". *Scientific Reports* **5** (2015). (Visited on 07/04/2017).

[9] A. Barbosa-Silva, J.-F. Fontaine, E. R. Donnard *et al.* "PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries". *BMC Bioinformatics* **12** (2011), p. 435. (Visited on 07/05/2017).

[10] W. W. Fleuren, E. J. Toonen, S. Verhoeven *et al.* "Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining". *BioData Mining* **6** (2013), p. 2. (Visited on 07/04/2017).

[11] K. Raja, S. Subramani, and J. Natarajan. "PPInterFinder—a mining tool for extracting causal relations on human proteins from literature". *Database: The Journal of Biological Databases and Curation* **2013** (2013). (Visited on 07/04/2017).

[12] A. Usie, H. Karathia, I. Teixidó, R. Alves, and F. Solsona. "Biblio-MetReS for user-friendly mining of genes and biological processes in scientific documents". *PeerJ* **2** (2014). (Visited on 07/04/2017).

[13] M. Torii, C. N. Arighi, G. Li *et al.* "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information". *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **12**:1 (2015), pp. 17–29. (Visited on 07/04/2017).

[14] R. Collobert and J. Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning". *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167. (Visited on 07/06/2017).

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". *arXiv:1301.3781 [cs]* (2013). arXiv: 1301.3781. (Visited on 05/01/2017).

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and Their Compositionality". *Proceedings of the 26th International Conference on Neural Information Processing Systems*. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119. (Visited on 04/02/2017).

[17] Y. Nie, W. Rong, Y. Zhang, Y. Ouyang, and Z. Xiong. "Embedding assisted prediction architecture for event trigger identification". *Journal of Bioinformatics and Computational Biology* **13**:03 (2015), p. 1541001. (Visited on 07/05/2017).

[18] D. Zhou, D. Zhong, and Y. He. "Event trigger identification for biomedical events extraction using domain knowledge". en. *Bioinformatics* **30**:11 (2014), pp. 1587–1594. (Visited on 07/05/2017).

[19] C. Li, R. Song, M. Liakata *et al.* "Using word embedding for bio-event extraction". *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015). Stroudsburg, PA: Association for Computational Linguistics*. 2015, pp. 121–126. (Visited on 07/04/2017).

[20] Y. Wang, Z. Liu, and M. Sun. "Incorporating Linguistic Knowledge for Learning Distributed Word Representations". *PLOS ONE* **10**:4 (2015), e0118437. (Visited on 07/05/2017).

[21] Z. Jiang, L. Li, and D. Huang. "An Unsupervised Graph Based Continuous Word Representation Method for Biomedical Text Mining". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13**:4 (2016), pp. 634–642.

[22] Z. Zhao, Z. Yang, H. Lin, J. Wang, and S. Gao. "A protein-protein interaction extraction approach based on deep neural network". *International Journal of Data Mining and Bioinformatics* **15**:2 (2016), pp. 145–164. (Visited on 07/04/2017).

[23] D. Szklarczyk, A. Franceschini, S. Wyder *et al.* "STRING v10: protein–protein interaction networks, integrated over the tree of life". *Nucleic Acids Research* **43**:Database issue (2015), pp. D447–D452. (Visited on 07/06/2017).

[24] T. U. Consortium. "UniProt: the universal protein knowledgebase". *Nucleic Acids Research* **45**:D1 (2017), pp. D158–D169. eprint: /oup/backfile/content_public/journal/nar/45/d1/10.1093_nar_gkw1099/4/gkw1099.pdf.

[25] M. Kanehisa and S. Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Res* **28**:1 (2000), pp. 27–30. (Visited on 12/13/2016).

[26] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. "KEGG as a reference resource for gene and protein annotation". en. *Nucleic Acids Research* **44**:D1 (2016), pp. D457–D462. (Visited on 12/13/2016).

[27] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. "On the Surprising Behavior of Distance Metrics in High Dimensional Spaces". *Proceedings of the 8th International Conference on Database Theory*. ICDT '01. London, UK, UK: Springer-Verlag, 2001, pp. 420–434. (Visited on 03/29/2017).

[28] R. A. Fisher. "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population". *Biometrika* **10**:4 (1915), pp. 507–521.

[29] J. Šarić, L. J. Jensen, R. Ouzounova, I. Rojas, and P. Bork. "Extraction of regulatory gene/protein networks from Medline". *Bioinformatics* **22**:6 (2006), pp. 645–650.

[30] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[31] M. Baroni, G. Dinu, and G. Kruszewski. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, 2014, pp. 238–247.

[32] J. L. Bentley. "Multidimensional Binary Search Trees Used for Associative Searching". *Commun. ACM* **18**:9 (1975), pp. 509–517. (Visited on 05/01/2017).

[33] J. Lin. "Divergence measures based on the Shannon entropy". *IEEE Transactions on Information Theory* **37**:1 (1991), pp. 145–151.

[34] D. M. Endres and J. E. Schindelin. "A new metric for probability distributions". *IEEE Transactions on Information theory* **49**:7 (2003), pp. 1858–1860. (Visited on 05/01/2017).

[35] L. Salwinski, C. S. Miller, A. J. Smith *et al.* "The database of interacting proteins: 2004 update". *Nucleic acids research* **32**:suppl_1 (2004), pp. D449–D451.

[36] A. Chatr-aryamontri, R. Oughtred, L. Boucher *et al.* "The BioGRID interaction database: 2017 update". *Nucleic acids research* **45**:D1 (2017), pp. D369–D379.

[37] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington *et al.* "IntAct: an open source molecular interaction database". *Nucleic Acids Res* **32**:Database issue (2004), pp. D452–D455. (Visited on 12/13/2016).

[38] L. Licata, L. Briganti, D. Peluso *et al.* "MINT, the molecular interaction database: 2012 update". *Nucleic acids research* **40**:D1 (2011), pp. D857–D861.

[39] R. Caspi, H. Foerster, C. A. Fulcher *et al.* "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases". *Nucleic acids research* **36**:suppl_1 (2007), pp. D623–D631.

[40] M. Ashburner, C. A. Ball, J. A. Blake *et al.* "Gene Ontology: tool for the unification of biology". *Nature genetics* **25**:1 (2000), p. 25.

[41] D. Croft, A. F. Mundo, R. Haw *et al.* "The Reactome pathway knowledgebase". *Nucleic Acids Res* **42**:Database issue (2014), pp. D472–D477. (Visited on 12/13/2016).

[42] A. Franceschini, D. Szklarczyk, S. Frankild *et al.* "STRING v9.1: protein-protein interaction networks, with increased coverage and integration". *Nucleic Acids Research* **41**:Database issue (2013), pp. D808–D815.

[43] C. M. Florkowski. "Sensitivity, Specificity, Receiver-Operating Characteristic (ROC) Curves and Likelihood Ratios: Communicating the Performance of Diagnostic Tests". *The Clinical Biochemist Reviews* **29**:Suppl 1 (2008), S83–S87. (Visited on 12/12/2017).

4 PIMKL: Pathway Induced Multiple Kernel Learning[1]

Matteo Manica[1,2,*], Joris Cadow[1,2,*], Roland Mathis[1,*], María Rodríguez Martínez[1]

[1] IBM Research Zürich

[2] ETH - Zürich

[*] Shared first authorship

Abstract

Reliable identification of molecular biomarkers is essential for accurate patient stratification. While state-of-the-art machine learning approaches for sample classification continue to push boundaries in terms of performance, most of these methods are not able to integrate different data types and lack generalization power, limiting their application in a clinical setting. Furthermore, many methods behave as black boxes, and we have very little understanding about the mechanisms that lead to the prediction. While opaqueness concerning machine behavior might not be a problem in deterministic domains, in health care, providing explanations about the molecular factors and phenotypes that are driving the classification is crucial to build trust in the performance of the predictive system.

We propose Pathway Induced Multiple Kernel Learning (PIMKL), a novel methodology to reliably classify samples that can also help gain insights into the molecular mechanisms that underlie the classification. PIMKL exploits prior knowledge in the form of a molecular interaction network and annotated gene sets, by optimizing a mixture of *pathway-induced* kernels using a Multiple Kernel Learning (MKL) algorithm, an approach that has demonstrated excellent performance in different machine learning applications. After optimizing the combi-

[1] In revision, pre-print available [1]. See Chapter 8 for details about contributions and copyright.

nation of kernels for prediction of a specific phenotype, the model provides a stable molecular signature that can be interpreted in the light of the ingested prior knowledge and that can be used in transfer learning tasks.

4.1 Introduction

Designing reliable and interpretable predictive models for patient stratification and biomarker discovery is a daunting challenge in computational biology. A plethora of methods based on molecular data have been proposed throughout the years, many of which exploit prior knowledge about the molecular processes involved in the regulation of the phenotype to be predicted. Prior knowledge is frequently encoded as a molecular interaction network, where nodes represent genes or proteins and edges represent relationships between the connected nodes. Supporting the development of such methods, the number of databases reporting protein-protein interactions has seen an unprecedented growth in recent years, and databases such as STRING [2], OmniPath [3], Reactome [4, 5], IntAct [6], MINT [7], MatrixDB [8], HPRD [9], KEGG [10–12] or Pathway Commons [13], just to name a few, provide an incredibly useful resource for designing models informed about the underlying molecular processes.

Several studies have focused on comparing prior knowledge-based classification methods. For instance, Cun and Fröhlich [14] evaluated 14 machine learning approaches to predict the survival outcome of breast cancer patients. The methods included, among others: average pathway expression [15], classification by significant hub genes [16], pathway activity classification [17], and a series of approaches based on Support Vector Machines (SVMs), such as network-based SVMs [18], recursive feature elimination SVMs [19], and graph diffusion kernels for SVMs [20, 21]. The study concluded that, while none of the evaluated approaches significantly improved classification accuracy, the interpretability of the gene signatures obtained was greatly enhanced by the integration of prior knowledge.

A more recent benchmarking effort was provided by a collaboration between the National Cancer Institute (NCI) and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project [22]. The NCI-DREAM challenge aimed to identify the top-performing methods for predicting therapeutic response in breast cancer cell lines using genomic, proteomic, and epigenomic data profiles. A total of 44 prediction algorithms were scored against an unpublished and hidden gold-standard dataset. Two interesting conclusions emerged from the challenge. First, all top-performing methods modeled nonlinear relationships and incorporated biological pathway information, and second, performance was increased by including

multiple, independent datasets. Interestingly, the top-performing methodology, Bayesian Multitask Multiple Kernel Learning, exploited a multiple kernel learning (MKL) framework [23].

MKL methods aim to model complex and heterogeneous datasets by using a weighted combination of base kernels. While in more traditional kernel methods the parameters of a single kernel are optimized during training, in MKL, the weights of each kernel are tuned during training.

Compared to single-kernel methods, the advantages of MKL are two-fold. First, different kernels can encode various levels of information (e.g., different definitions of similarity or different types of data) endowing the algorithm with the flexibility required to model heterogeneous or multi-modal datasets. Second, after optimizing the combination of kernels, the weights associated with each kernel can provide valuable insights about the sets of features that are most informative for the classification task at hand.

In this paper, we seek to augment the predictive power and interpretability of MKL methods, by supplementing them with the use of prior knowledge. Towards this end, we introduce Pathway Induced Multiple Kernel Learning (PIMKL), a supervised classification algorithm for phenotype prediction from molecular data that jointly exploits the benefits of MKL and prior knowledge ingestion. PIMKL uses an interaction network and a set of annotated gene sets to build a mixture of *pathway-induced* kernels from molecular data, whose mixture is then optimized with an MKL algorithm, see Figure 4.1.

After PIMKL is trained, the weight assigned to each kernel provides information about the importance of the corresponding pathway in the mixture. As a result, a molecular signature characterizing the phenotype of interest is derived.

While there are currently many approaches that take advantage of the known graph structure of a molecular system [20, 24], or use collections of annotated gene sets as prior knowledge to reduce the dimensionality of molecular profiles and enable the analysis of tumor profiles [25, 26], to our knowledge PIMKL is the first methodology that integrates both levels of prior knowledge, molecular networks and collections of pathways, with state-of-the-art machine learning approaches. We demonstrate that the use of MKL enhances classification performance, and the use of prior knowledge ensures that the results are interpretable and shed light on the molecular interactions implicated in the phenotype.

This paper is structured as follows. We first describe PIMKL and validate it by predicting disease-free survival for breast cancer samples from multiple cohorts. We benchmark PIMKL by comparing it with the methods analyzed in [14]. To evaluate its generalization power,

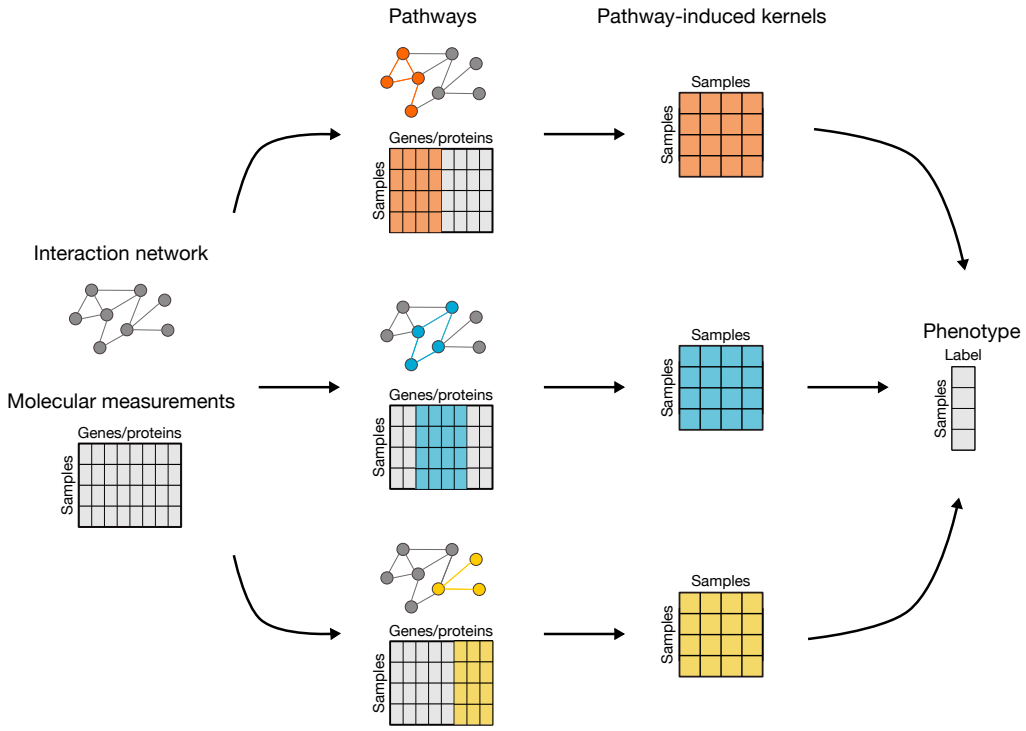


Figure 4.1: **PIMKL concept.** Given a network topology describing molecular interactions, relevant sub-networks can be extracted to generate a mixture of *pathway-induced* kernels. The combination of kernels is then optimized to predict a phenotype of interest. The weights of the mixture provide a measurement of the importance of each pathway, thereby shedding light on the molecular mechanisms that contribute to the phenotype.

we use a PIMKL-generated molecular signature to predict disease-free survival on a different dataset, the METABRIC breast cancer cohort [27]. Finally, we test PIMKL robustness against noise and its capabilities to integrate distinct data types by simultaneously using METABRIC gene expression (mRNA) and copy number alteration (CNA) data for the same classification task.

Our analysis suggest that PIMKL provides an extremely robust approach for the integration of multiple types of data with prior knowledge that can be successfully applied to a wide range of phenotype prediction problems.

4.2 Results

In the following sections, we discuss the application of PIMKL to different breast cancer cohorts. First, in Section 4.2.1, PIMKL is compared to a previous study by Cun and Frölich [14]

where different algorithms for phenotype prediction and gene selection using prior knowledge were compared. Later, in Section 4.2.2, PIMKL is applied to gene expression and copy number data from the METABRIC cohort [27] with two purposes: first, we aim to test whether transfer learning between different studies is possible, and, second, we want to evaluate PIMKL performance in the analysis of multi-omics analysis in the presence of noise or uninformative data.

4.2.1 PIMKL on breast cancer microarray cohorts

PIMKL is tested on microarray gene expression data from six breast cancer cohorts (see Supplementary Table 4.1 for details about the cohorts). The classification task consists in stratifying breast cancer samples according to occurrence of relapse within 5 years. To ensure the fairest possible comparison, we use the same interaction sources as in the study by Cun and Fröhlich, namely a merge between KEGG pathways and Pathway Commons. As access to older releases of KEGG is restricted, the most recent versions from both sources are used. A collection of 50 *hallmark* gene sets from the Molecular Signatures Database (MSigDB) version 5.2 [28] is used to define the sub-graphs considered for *pathway induction*, generating $P = 50$ kernels. The classification performance is evaluated by means of the Area Under the receiver operating characterstic Curve (AUC). We closely follow the same data processing procedures and the cross-validation scheme as proposed in the original study (for details, see Supplementary Algorithm 2).

The results of PIMKL compared to the 14 algorithms considered by Cun and Fröhlich are reported in Figure 4.2. Overall AUC values for the 6 cohorts over the cross-validation rounds for all considered methods are shown in Figure 4.2a. AUC values for the single cohorts can be found in Supplementary Figure 4.7, where PIMKL exhibits the highest median value and consistently outperforms the other methods or is in the top performers group on single cohorts.

Results are consistent when other gene sets are used, even when we use randomized versions of functionally related gene sets (see Supplementary Figure 4.8). These results prove that PIMKL performance does not depend on the specific selection of pathways, and that through the MKL optimization we can identify the informative gene sets in disparate collections of genes. Notice, however, that while choosing random gene sets does not worsen PIMKL performance, interpretation of the molecular signatures, as we will discuss next, is only possible when the sets have a well defined biological function.

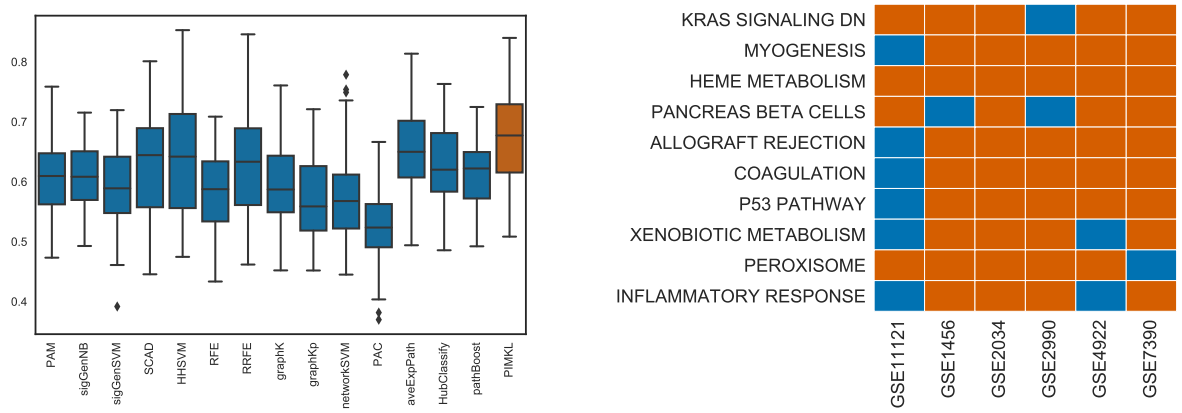
As discussed in Section 4.4, PIMKL generates a molecular signature given by the weighted

contribution of each kernel. Each weight represents the relative importance of each hallmark pathway used for *pathway induction* to explain the phenotype. To evaluate the stability of the signature, the pathway weight distribution over cross-validation rounds is analyzed. Our baseline is the case where all kernels have the same weight: $w_b = \frac{1}{P}$, representing a situation where no pathway contributes more than the others to the phenotype prediction.

To find whether a pathway is significant for the phenotype, the distribution of the kernel weights with median above w_b is tested against the baseline using a one-sample Wilcoxon signed-rank test.

p-values at significance level 0.001 were corrected for multiple testing using Benjamini-Hochberg (for details see Supplementary Figure 4.9). Pathways where the significance was achieved in at least four of six cohorts are reported in Figure 4.2b.

Interestingly, heme metabolism pathway is significant in all cohorts. This pathway is involved in the metabolism of heme and erythroblast differentiation. A possible explanation is that heme metabolism might reflect an active vascularization of the samples, a phenomenon widely observed in cancer progression [29]. A more intriguing hypothesis is a possible association between elevated heme metabolism and cancer progression, as has been reported in



(a) **(b)**

Figure 4.2: **PIMKL cross-validation results. (a)** Box plots for AUC values over all cohorts for the methods considered. PIMKL results are reported in red, while other methods' results are colored in blue. Box plots are obtained from ten (repeats of) mean AUC values over 10-fold cross-validation splits, see algorithm 2. **(b)** Heat map showing significant pathways selected by PIMKL across the different cohorts considered in the study. Significant pathways are highlighted in red, while non-significant are colored in blue.

non-small-cell lung cancer cells and xenograft tumors [30]. It is also interesting to look at the pathways that are significant in at least five cohorts: KRAS signaling, myogenesis, allograft rejection, coagulation, P53 pathway and peroxisome. All of these pathways are associated with breast cancer. For instance, activation of KRAS signaling has been reported to promote the mesenchymal features of basal-type breast cancer [31, 32]. Myogenesis, or the process of formation of muscular tissue, is commonly disrupted in cancer [33]. Allograft rejection might reflect an immune-mediated tumor rejection signature following administration of immunotherapeutic agents [34]. Several studies have suggested a role for blood coagulation proteins in tumor progression [35–37]. P53 is the most commonly mutated protein in cancer [38, 39]. Finally, peroxisomes are small, membrane-enclosed organelles that contain enzymes involved in a variety of metabolic reactions, including several aspects of energy metabolism. Altered peroxisome metabolism has been linked to various diseases, including cancer [40, 41].

Finally, Figure 4.3 reports the correlation of PIMKL molecular signatures estimated across multiple cohorts and highlights their stability across different studies, suggesting that a cohort-independent disease free survival signature for breast cancer has been learned.

4.2.2 PIMKL on METABRIC cohort

To test PIMKL applicability to multi-modal datasets, we use our methodology to predict disease free survival in METABRIC breast cancer cohort, consisting of 1890 samples profiled with Illumina Human v3 microarray data (mRNA) and Affymetrix SNP 6.0 copy number data (CNA), see Supplementary Table 4.2 for details.

In order to validate the generalization power of PIMKL-generated molecular signatures, we first focused on the analysis of METABRIC microarray data. Our hypothesis here is that the underlying molecular mechanisms associated with disease free survival are the same in different cohorts and, as such, knowledge learned in one cohort can be transferred to another one.

After computing the *pathway-induced* kernels with the same procedure adopted in Section 4.2.1, a set of pathway weights was defined using the median of the weights obtained in the six previously analyzed cohorts. Figure 4.4 shows the results obtained by training a KOMD classifier using the weights transferred from the six independent cohorts and by learning METABRIC-specific pathway weights (for details see Section 4.4 and Supplementary Algorithm 3). It is evident that both molecular signatures perform very similarly. Indeed, the

two signatures are highly correlated (Pearson correlation $\rho = 0.72$, p-value $= 3.34 \cdot 10^{-9}$, Figure 4.10). It is important to notice that the variance of the prediction results is also consistently reduced, probably due to the newer microarray technology used by the METABRIC study.

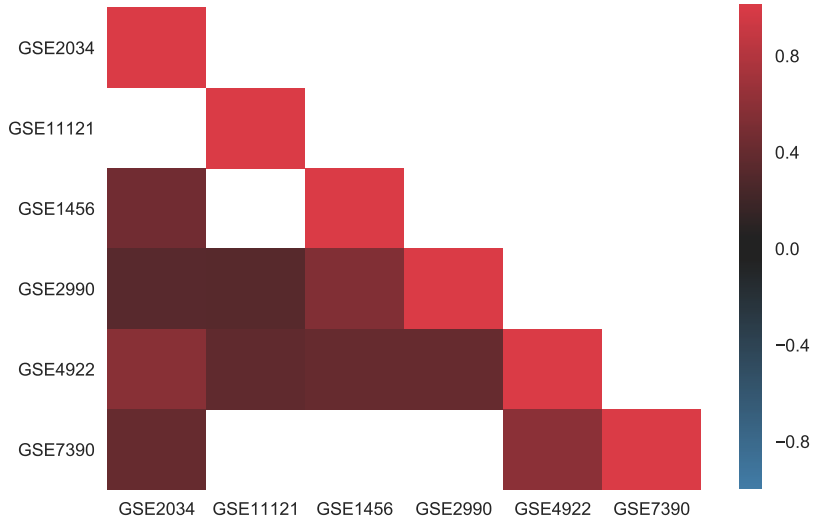


Figure 4.3: **Correlation in molecular signatures.** Heat map reporting the correlation of the molecular signature estimated across multiple cohorts. Correlation values are reported in the lower triangular part of the heat map (since it is symmetric) on blue to red scale, white squares indicate non significant correlations. All cohorts exhibit a positive correlation, significant in most cases, proving the stability of the molecular signature obtained with PIMKL.

To test PIMKL capability to integrate multi-omic data, both the mRNA and CNA data from the METABRIC cohort are jointly utilized in the same predictive task. A set of additional kernels are generated using the copy number data and then used in two ways: first, the CNA kernels are independently optimized with PIMKL, and second, a mixture of CNA and mRNA kernels are jointly optimized.

From Figure 4.5, it is evident that the CNA data are not as predictive as mRNA regarding disease free survival. However, it is interesting to notice that PIMKL is able to discard noisy kernels, associated with CNA data, to achieve similar levels of performance when using the more informative mRNA data and when using a mixture of CNA and mRNA data. This suggests that the application of the proposed algorithm is feasible even when no prior knowledge about the information content of each single omic type is available.

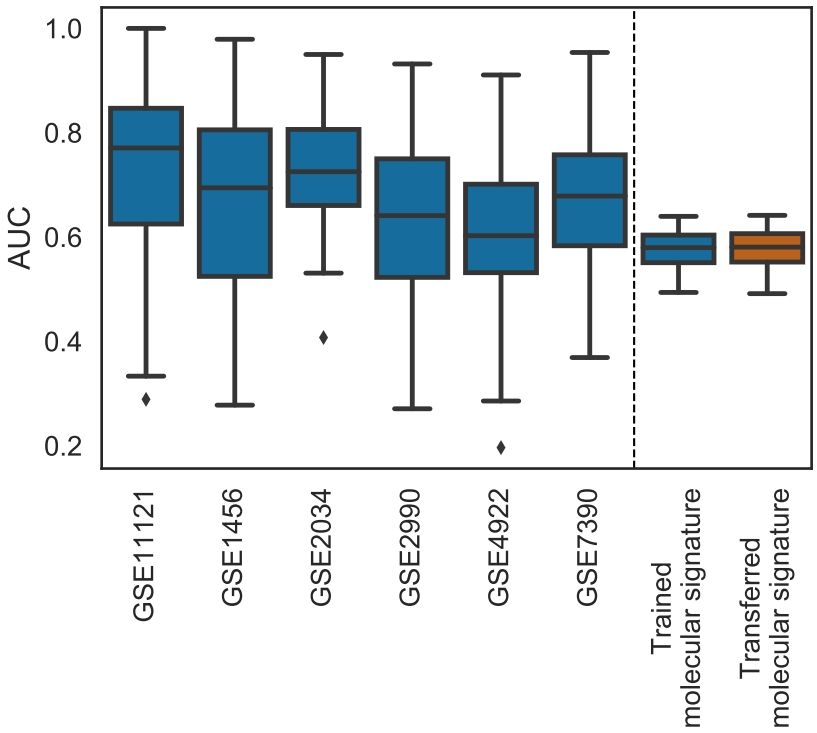


Figure 4.4: **PIMKL performance on METABRIC.** Box plots of the performance of PIMKL over the six cohorts used to benchmark the method (left of the dashed vertical line) and its application on METABRIC for disease free survival prediction (right of the dashed vertical line). Optimized weights at training by EasyMKL (blue); provided weights from taking the pathway-wise median weights of the six signatures obtained during benchmarking (red).

4.3 Discussion

We have presented here PIMKL (Pathway Induced Multiple Kernel Learning), a novel, effective and interpretable machine learning methodology for phenotype prediction using multi-modal molecular data. PIMKL is based on a multiple kernel learning (MKL) framework, a kernel-based method that has demonstrated excellent capabilities to integrate multi-omic datasets [22]. In addition, PIMKL also exploits prior knowledge in the form of molecular interaction networks and sets of annotated pathways with known biological functions to build a mixture of *pathway-induced* kernels. The main novelty introduced in this work is the definition of multiple interaction-aware kernel functions, which enables us to encode information about the molecular prior knowledge related to a phenotype, and facilitates the interpretation of the results in terms of known biological functions or specific molecular interactions. We achieve this by using such kernels to map samples into the space of network edges, i.e.,

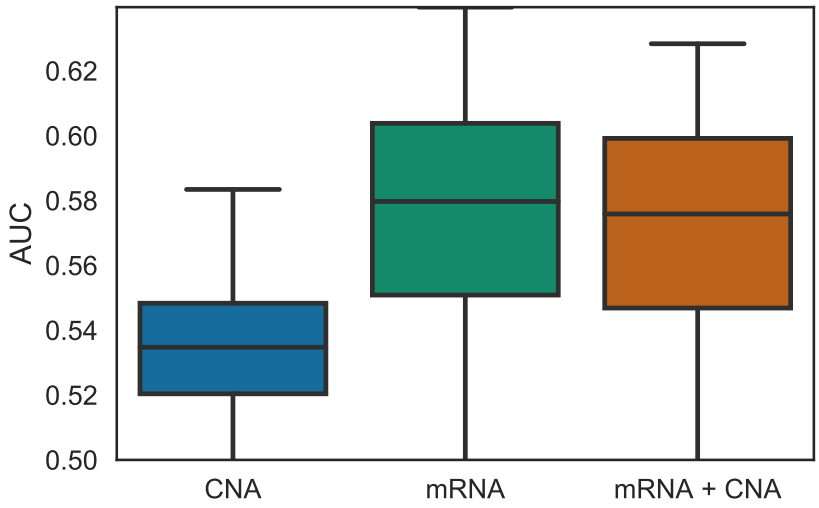


Figure 4.5: **PIMKL performance on METABRIC multi-omics.** Box plots for AUC values obtained applying PIMKL on different data types and their integration. CNA only results are reported in blue, mRNA ones in green and their integration in red.

molecular interactions, recovering a direct biological interpretation. The kernel weights are later optimized to classify a phenotype or a clinical variable of interest.

In this work, PIMKL has been extensively tested in the context of predicting disease-free survival from breast cancer samples. We have demonstrated that the resulting weighted combination of kernels can be interpreted as a phenotypic molecular signature and provides insights into the underlying molecular mechanisms.

As a benchmark, a well-studied set of cohorts previously analyzed using a range of stratification methods has been adopted [14]. The quality and the stability of the obtained signatures has been thoroughly investigated, and we have shown that PIMKL outperforms other methods and finds stable molecular signatures across different breast cancer cohorts. We also investigated the generalization power of the found signatures by testing them on unseen mRNA breast cancer data from the METABRIC cohort and the associated disease-free survival data. The obtained results have confirmed that the algorithm can be used to effectively gain insights into disease progression and that this knowledge can be transferred to other cohorts without loss of performance.

Furthermore, PIMKL can be seamlessly applied to integrate data from different omic layers. Its intrinsic capability to discard noisy molecular features has been demonstrated by applying it on METABRIC, where it was possible to integrate multiple types of data with varying predictive power. Even when non-informative data were mixed with informative data, PIMKL

was able to discard uninformative kernels and achieve similar levels of performance. Evidently, PIMKL is not restricted to breast cancer, the specific omic data types or the sources of prior information used in this work. Its application is open to other disease types using any available combination of data together with any suitable prior network and sets of genes.

Besides being capable of using different types of prior knowledge, the proposed approach is also highly flexible with regard to the number and nature of the selected kernels. Indeed, PIMKL has been developed by making use of an efficient implementation of EasyMKL [42], an extremely scalable MKL algorithm with constant memory complexity in the number of kernels. This efficiency can potentially allow the user to define smaller pathways, leading to a more fine-grained characterization and understanding of the molecular mechanisms involved in disease progression with limited performance drawbacks.

Finally, possible extensions of PIMKL, such as optimizing the kernel mixture using semi-supervised or unsupervised multiple kernel learning methodologies [43], may help discovery phenotype-independent pathway signatures and will be explored in the future. To summarize, PIMKL provides a flexible and scalable method to translate prior knowledge and molecular data into actionable insights in a clinical setting. In order to ease the usage of PIMKL, the algorithm has been made available as an open access web service.

4.4 Methods

PIMKL is a methodology for phenotype prediction from multi-omic measurements (e.g., mRNA, CNA, etc.) based on the optimization of a mixture of *pathway-induced* kernels. Such kernels are generated by exploiting prior knowledge in a dual fashion. First, prior knowledge is injected in PIMKL in the form of a molecular interaction network, and second, as a set of annotated gene sets or pathways.

A key aspect of PIMKL is *pathway induction*, a method to generate similarity functions using the topological properties of an interaction network. In practice, we use pathway gene sets with well-defined biological functions to define sub-networks from which we generate *pathway-induced* kernels. This mixture of *pathway-induced* kernels is then optimized to classify a phenotype of interest, and in doing so, each pathway is assigned a weight representing its importance to explain the phenotype. The established link between kernels and pathways enables PIMKL to identify which molecular mechanisms are important for the prediction of the considered phenotype, as shown in Figure 4.1.

4.4.1 Pathway Induction

PIMKL encodes information from the topology of each pathway's sub-network. The approach of integrating pathway information into *interaction-aware* kernel similarity functions is here termed *pathway induction*. Specifically, we design kernel functions by utilizing a positive semidefinite (PSD) matrix that encodes the topological properties of a graph. Given any PSD matrix M, a valid kernel can be *induced* through the following weighted inner product [44]:

$$k(x, y) = x^T M y$$

Hence, in order to have a *pathway-induced* kernel, we only need to define M such as it encodes the graph topological information of the pathway.

In this work, an encoding based on the symmetric normalized Laplacian matrix is adopted. Pathways are defined as weighted undirected graphs $\mathcal{P} = (V, E, W)$, with $N_v = |V|$ nodes, $N_e = |E|$ edges and a diagonal weight matrix $W \in \mathbb{R}^{N_e \times N_e}$, representing respectively the molecular entities (e.g., genes, proteins, etc.), their interactions and the weights associated with them.

For any pair of samples $x, y \in \mathbb{R}^{N_v}$, we define a *pathway-induced* kernel using the following similarity function:

$$\begin{cases} k_{\mathcal{L}}(x, y) = x^T \mathcal{L} y = x^T \mathcal{S}\mathcal{S}^T y = \Pi(x)^T \Pi(y) \\ \mathcal{S} = D^{-\frac{1}{2}} S W^{\frac{1}{2}} \end{cases}$$

where $\mathcal{L} \in \mathbb{R}^{N_v \times N_v}$, $D \in \mathbb{R}^{N_v \times N_v}$ and $S \in \mathbb{R}^{N_v \times N_e}$ are respectively the normalized Laplacian, the diagonal degree matrix and an ordered incidence matrix for graph $\mathcal{P}$ associated with a pathway (see Supplementary 4.5.1 for a detailed explanation about the formulation and the design of the kernel function).

The normalized Laplacian can be interpreted as a discrete Laplace operator representing a diffusion process over graph nodes. A *pathway induction* process based on it introduces a mapping Π from the original space of the N_v molecular measurements to an N_e-dimensional feature space, where each pathway interaction is a dimension, and the value along the edge is the discrete diffusion potential between the nodes that the edge connects. A schematic illustration of the mapping introduced using *pathway induction* can be seen in Figure 4.6.

4.4.2 Pathway Induced Multiple Kernel Learning

PIMKL makes use of the concept of *pathway induction*, defined in 4.4.1, to implement a multiple kernel learning classifier.

Consider a network that recapitulates a comprehensive set of known molecular interactions represented by a graph $\mathcal{G} = (V, E, W)$ with $N_v = |V|$ nodes, $N_e = |E|$ edges and a set of molecular measurements $X \in \mathbb{R}^{N \times N_v}$ with associated labels for a relevant phenotype y.

Given a selection of pathways P (e.g., gene sets from ontologies or inferred via community detection), it is possible to extract for each pathway $p \in P$, a corresponding sub-graph $\mathcal{P}^p = (V^p, E^p, W^p) \subset \mathcal{G}$ with $N_v^p = |V^p|$ nodes, $N_e^p = |E^p|$ edges and a sub-selection of measurements corresponding to the genes contained in the pathway $X^p \in \mathbb{R}^{N \times N_v^p}$.

For every pathway, a Gram matrix K^p can be used to represent the *pathway-induced* kernel, where K^p is computed for each pair of samples i and j as follows:

$$K_{ij}^p = k_{\mathcal{L}^p}(x_i, x_j)$$

In the above equation, $x_i, x_j \in \mathbb{R}^{N_v^p}$ and $\mathcal{L}^p$ is the normalized Laplacian for $\mathcal{P}^p \; \forall p \in P$.

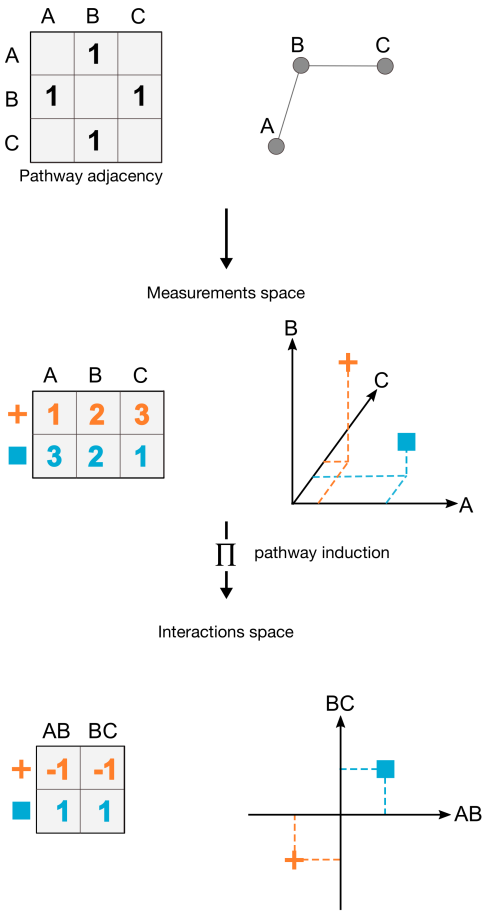


Figure 4.6: **Pathway induction.** Given a pathway adjacency matrix, it is possible to map sample measurements from their original space, the space of the nodes, to the space of the interactions between the molecular entities. The example above shows how the mapping using *pathway induction* transforms the considered samples.

For the problem of finding the optimal mixture of kernels over the different *pathway-induced* kernels, any supervised MKL algorithm can be used. In this work, a custom version of EasyMKL [42] has been implemented as it achieves high performance at a low computational cost. EasyMKL is based on the Kernel method for the Optimization of the Margin Distribution (KOMD) [45] and focuses on optimizing a linear combination of kernels:

$$K = \sum_{p=1}^{P} w_p K^p, \quad w_p \geq 0$$

In PIMKL, the weights obtained are divided by their sum, as we are interested in evaluating the relative contribution of each kernel. This normalization does not affect the quality of the kernel mixture, which is invariant under positive scalar multiplication. In addition, to account for differences in sub-graph sizes we force the kernel matrices to have equal trace, ensuring comparable Gram matrices between different pathways.

It is important to note that PIMKL formulation enables a seamless integration of multi-omics data. Kernels from different data types can be easily generated and added to the mixture. The same applies to multi-modal data integration: kernels generated from other data modalities associated with a specific sample (e.g., histopathology images or clinical records) can be added to the mixture and weighted accordingly to their contribution in the classification problem.

Declarations

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 668858. We thank Yupeng Cun for providing results [14] for recreation of Figures 4.7 and 4.2a.

Author contributions

See Chapter 8 for details about contributions and copyright.

Competing financial interests

The authors declare no competing financial interest.

Availability of data and materials

Processed data and materials used to produce the results presented in this work are available via download link [1]. PIMKL as a service is available on IBM Cloud at the following link `https://sysbio.uk-south.containers.mybluemix.net/pimkl/`. A set of anonymous credentials can be created for reviewers.

4.5 Supplementary information

4.5.1 Pathway induction

Similarity functions can be designed by making use of a PSD matrix to *induce* a weighted inner product:

$$k(x, y) = x^T M y \quad \forall x, y \in \mathbb{R}^N$$

represents a valid kernel if matrix $M \in \mathbb{R}^{N \times N}$ is PSD [44], indeed this ensures the existence of a matrix U:

$$M = U^T U$$
$$\phi(x) = Ux$$

where ϕ is a mapping describing a transformation in the feature space.

By making use of a PSD matrix encoding the topological properties of a graph representing a pathway, it is possible to design *interaction-aware* kernels.

Let's consider an undirected graph representing a pathway:

$$\mathcal{P} = (V, E)$$

with $N_v = |V|$ nodes and $N_e = |E|$ edges representing the genes/proteins and their interactions respectively. Such a graph is defined by a symmetric adjacency matrix $A \in \{0, 1\}^{N_v \times N_v}$:

$$A_{ij} = 1 \quad \forall (i, j) \in E$$

[1] `https://ibm.box.com/s/ac2ilhyn7xjj27r0xiwtom4crccuobst`, as of November 2018

and its diagonal degree matrix $D \in \mathbb{R}^{N_v \times N_v}$:

$$D_{jj} = \sum_i A_{ij}$$

For such a graph, the Laplacian matrix $L \in \mathbb{R}^{N_v \times N_v}$ is computed using the following:

$$L = D - A$$

The Laplacian is a PSD matrix and therefore represents a suitable candidate for *induction* of a weighted inner product based on a pathway topology. This can be shown by defining an ordered incidence matrix $S \in \mathbb{R}^{N_v \times N_e}$ for $\mathcal{P}$ that by construction satisfies the relation $L = SS^T$. As in [46], after introducing an index set $\mathcal{E}$ for the edges E, S can be defined as:

$$S_{ne} = \begin{cases} 1 & \text{if } n = i \wedge i \leq j \\ -1 & \text{if } n = j \\ 0 & \text{otherwise} \end{cases}$$

where $e \in \mathcal{E}$ corresponds to edge $(i, j) \in E$ and $n \in V$

Moreover, the Laplacian can be interpreted as a discrete Laplace operator. Indicating with $X \in \mathbb{R}^{N \times N_v}$ a set of N samples, a discrete diffusion process over graph nodes is described as:

$$LX^T = SS^T X^T \tag{4.1}$$

where the term $S^T X^T$ computes the discrete diffusion potential (a difference) along the edges and Equation 4.1 describes how the flow of this potential is effected, aggregating incoming and outgoing flows on the nodes.

Decomposing the Laplacian using an ordered incidence matrix shows how samples X are mapped from the original space with measurements for N_v molecular entities into an N_e-dimensional feature space, where each interaction from the pathway is a dimension and the value along the edge is the discrete diffusion potential between respective node's measurements.

The inner product in this space is the resulting similarity function defined as:

$$k_L(x, y) = x^T L y = x^T S S^T y \quad \forall x, y \in \mathbb{R}^{N_v}$$

Similar considerations can be applied to weighted graphs with non-negative weights. Given a weighted undirected graph $\mathcal{P} = (V, E, W)$, indicating by $W \in \mathbb{R}^{N_e \times N_e}$ its diagonal weights matrix, the Laplacian L is defined as:

$$L = SWS^T$$

$$L_{ij} = \begin{cases} d_i - W_e & \text{if } i = j \\ -W_e & \text{otherwise} \end{cases}$$

where $e \in \mathcal{E}$ corresponds to edge $(i, j) \in E$

To ensure an equal contribution from all the nodes in the considered pathway, the degree-normalized version of the Laplacian $\mathcal{L}$ can be adopted:

$$\mathcal{L} = D^{-\frac{1}{2}} SWS^T D^{-\frac{1}{2}}$$

$$\mathcal{L}_{ij} = \begin{cases} 1 - \dfrac{W_e}{d_i} & \text{if } i = j \text{ and } d_i \neq 0 \\ -\dfrac{W_e}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

where $e \in \mathcal{E}$ corresponds to edge $(i, j) \in E$

This pathway encoding directly leads to the definition of *pathway induction* used in this work. Given any two samples measurement $x, y \in \mathbb{R}^{N_v}$:

$$\begin{aligned} k_{\mathcal{L}}(x, y) =& x^T \mathcal{L} y = \\ =& x^T D^{-\frac{1}{2}} SWS^T D^{-\frac{1}{2}} y = x^T (D^{-\frac{1}{2}} SW^{\frac{1}{2}})(W^{\frac{1}{2}} S^T D^{-\frac{1}{2}}) y = \\ =& x^T \mathcal{S}\mathcal{S}^T y = \Pi(x)^T \Pi(y) \end{aligned}$$

with:

$$\Pi(x) = \begin{cases} \sqrt{W_e} \dfrac{x_i}{\sqrt{d_i}} & \text{if } i = j \text{ and } d_i \neq 0 \\ \sqrt{W_e} (\dfrac{x_i}{\sqrt{d_i}} - \dfrac{x_j}{\sqrt{d_j}}) & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}$$

where $e \in \mathcal{E}$ corresponds to edge $(i, j) \in E$

A similar concept was proposed [47] at the complete network level. The normalized Laplacian was used as a regularizer to constrain the optimization problem when training an SVM. In PIMKL, we arrive at a similar formulation of the problem by introducing a feature mapping instead of using the Laplacian as a regularizer. We define a kernel function which allows easy application to any kernelized method and any further kernel transformation (e.g., polynomial, Gaussian, etc.). The decomposition of $\mathcal{L}$ can be derived from the graph but is implicit, and can be easily extended to the multiple kernel learning case, allowing us to work at pathway/sub-network level.

It should be noted that in PIMKL the individual *pathway-induced* kernels are set to equal trace (equal average self similarity of the samples) to learn fair relative weights independent of the sub-network size.

4.5.2 Breast cancer microarray cohorts

GEOid [48]	**Patients**	**dmfs/rfs $\leq$ 5 years**	**dmfs/rfs $>$ 5 years**
GSE2034 [49]	286	93	183
GSE1456 [50]	159	34	119
GSE2990 [51]	187	42	116
GSE4922 [52]	249	69	159
GSE7390 [53]	198	56	135
GSE11121 [54]	200	28	154

Table 4.1: **Breast cancer benchmark cohorts.** Brief description of sample counts in the different classes for the cohorts considered in [14] (all Affymetrix Human Genome U133A Array). In GSE4922 and GSE11121 metastasis free survival (dmfs) is considered, in other cohorts relapse free survival (rfs).

Data types	**Patients**	**Recurred/Progressed**	**DiseaseFree**
Illumina Human v3 microarray (mRNA) Affymetrix SNP 6.0 copy number (CNA)	1890	647	1333

Table 4.2: **Breast cancer METABRIC cohort.** Brief description of sample counts in the different classes for the considered data types in the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) cohort [27].

Algorithm 2 PIMKL Cross-validation on [14]. Cross-validation analysis of PIMKL for each of the breast cancer cohorts as suggested in [14] (with internal optimization of parameters). Given as input: X gene expression measurements with related clinical labels y, a set of P pathways and a set of hyper-parameters Λ = {0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0} for EasyMKL.

for $i \leftarrow 1, 10$ **do**
  for $(X_{train}, y_{train}), (X_{validation}, y_{validation}) \leftarrow$ stratified 10-fold split X, y **do**
    learn feature-wise normalization on X_{train} and apply to $X_{validation}$
    for $(X^{\lambda}_{train}, y^{\lambda}_{train}), (X^{\lambda}_{test}, y^{\lambda}_{test}) \leftarrow$ stratified 3-fold split of X_{train}, y_{train} **do**
      for $\lambda \in \Lambda$ **do**
        train PIMKL(λ) on $\{k_{\mathcal{L}^p}(X^{\lambda}_{train}, X^{\lambda}_{train}) : p \in P\}$ and y^{λ}_{train}
        record prediction accuracy on $\sum_{p=1}^{P} k_{\mathcal{L}^p}(X^{\lambda}_{train}, X^{\lambda}_{test})$
    $\lambda^* \leftarrow$ argmax(mean prediction accuracy over cross-validation)
    PIMKL(λ^*) on $\{k_{\mathcal{L}^p}(X_{train}, X_{train}) : p \in P\}$ and y_{train}
    report kernel weights $\boldsymbol{w}$
    report area under the curve for prediction on $\sum_{p=1}^{P} w_p k_{\mathcal{L}^p}(X_{train}, X_{validation})$
  report mean area under the curve over 10-fold splits ▷ for figure 4.7 and 4.2a

Algorithm 3 PIMKL Cross-validation on METABRIC. Cross-Validation on METABRIC single omics or multi-omics. Given as input: X molecular measurements comprised of a selection of data types T (CNA, mRNA or both) with related clinical labels y, a set of P pathways with a respective pathway for each data type in T and λ = 0.2 for EasyMKL.

for 100 folds with 20 samples per class in (X_{train}, y_{train}) **do**
  for $type$ in T **do**
    learn feature-wise normalization on $X_{type,train}$ and apply to $X_{type,validation}$
  train PIMKL(λ) on $\{k_{\mathcal{L}^p}(X_{train}, X_{train}) : p \in P\}$ and y_{train}
  report kernel weights $\boldsymbol{w}$
  report area under the curve for prediction on $\sum_{p=1}^{P} w_p k_{\mathcal{L}^p}(X_{train}, X_{validation})$

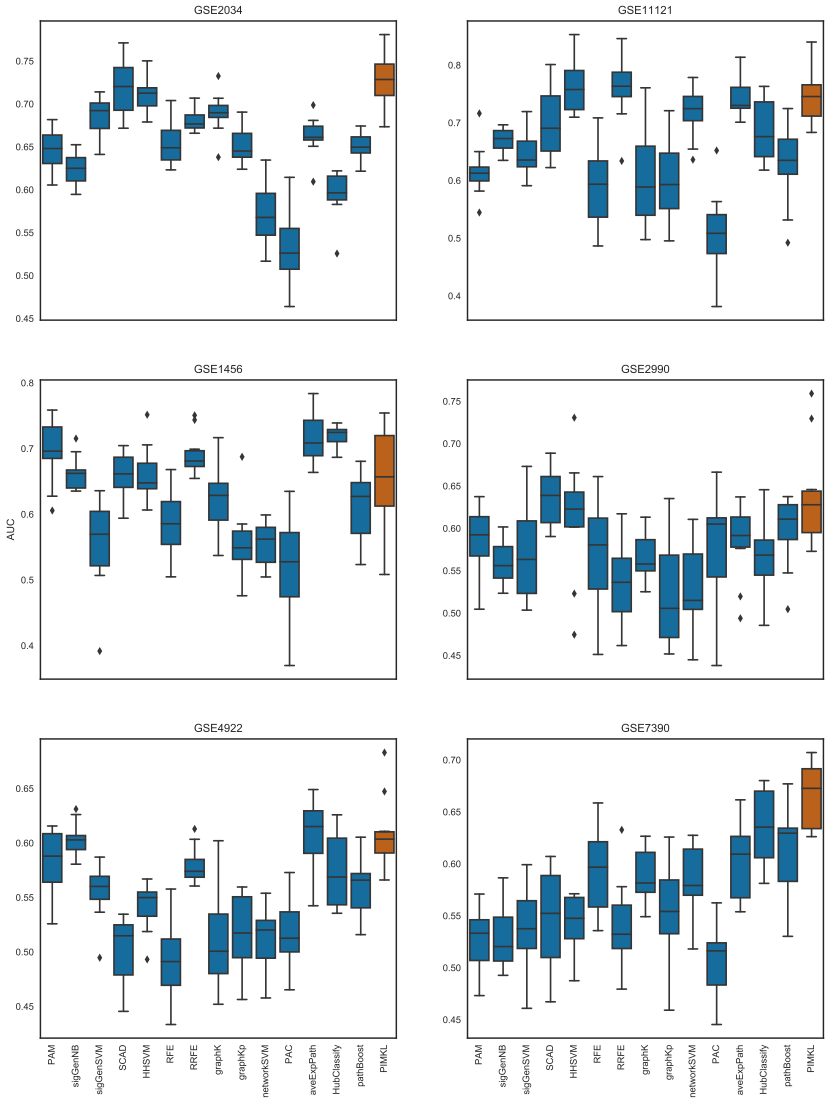


Figure 4.7: **PIMKL cross-validation AUC.** Box plots of the AUC values for the methods analyzed in [14] (blue) and PIMKL (red). PIMKL clearly outperforms other methods in four out of six datasets. For GSE1456 is performing close to other methods average while for GSE11121 is in the top group. Results are presented as in [14], where each box is drawn from ten (repeats of) mean AUC values over 10-fold cross-validation splits, see algorithm 2.

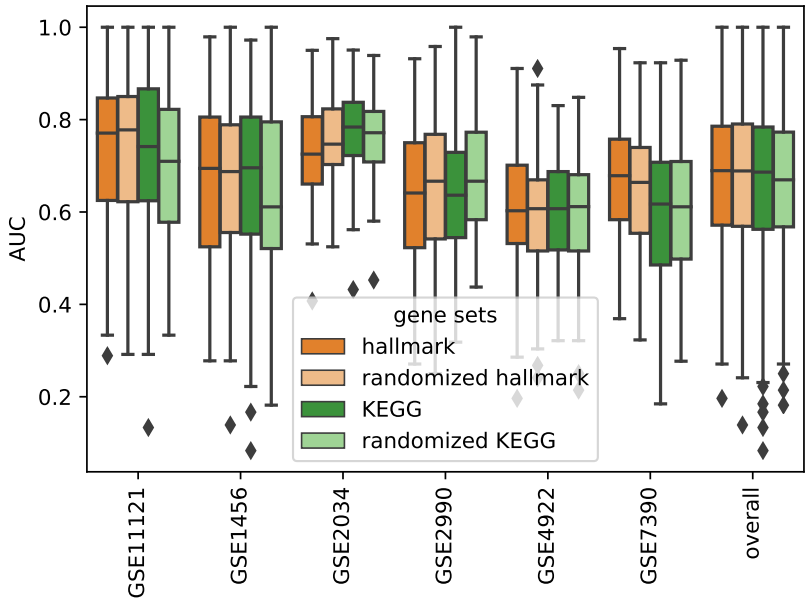


Figure 4.8: **PIMKL cross-validation AUC for different gene sets.** Box plots of all 100 AUC values (overall 600) for pathway induced MKL obtained by algorithm 2 with different gene sets to define the pathways given the same aforementioned interactions. In addition to the 50 previously introduced hallmark gene sets, results for 186 KEGG gene sets from the Molecular Signatures Database (MSigDB) version 5.2 [28] and also respective randomized gene sets are reported. For randomization, the same number of gene sets is created, each set with random size between 50 and 250 genes by sampling from the union of all gene sets. The quartiles are comparable within each cohort proving the stability of the methods towards gene sets selection.

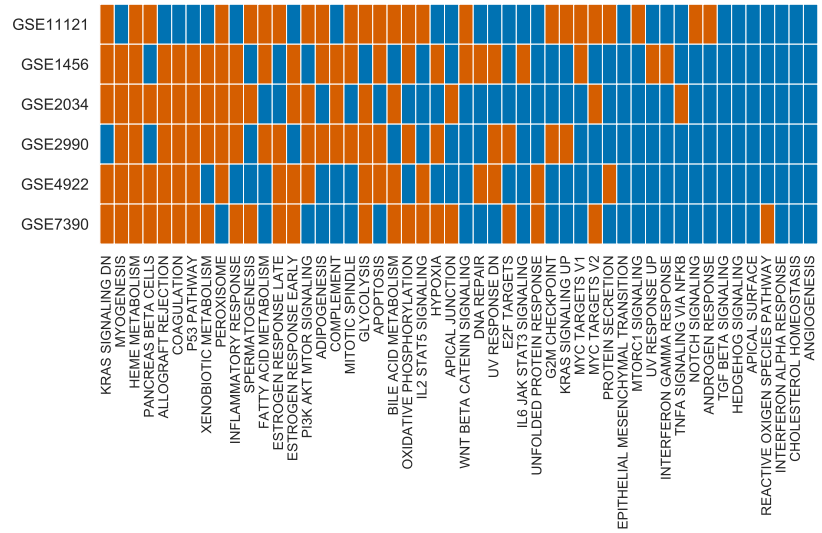


Figure 4.9: **PIMKL cross-validation weights.** Significance of weights over 100 cross-validation folds for the 50 hallmark pathways are reported. Significant pathways are colored in red, while non-significant in blue.

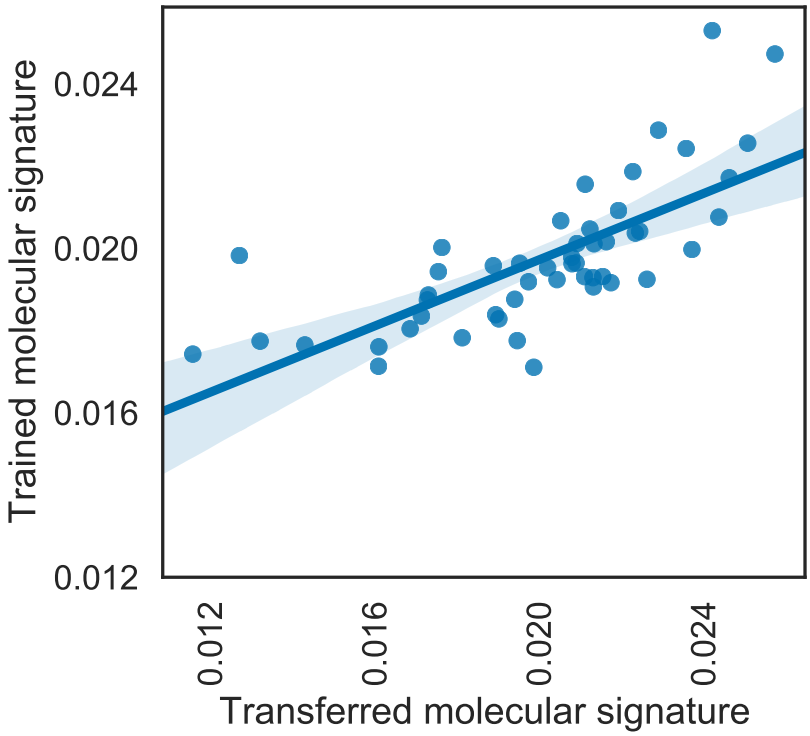


Figure 4.10: **Regression between trained and transferred signature.** Regression of the pathway weights of the signature obtained from directly training on METABRIC (median over 100 cross-validation folds) against the transferred signature obtained from training on six independent cohorts (each median over 100 cross-validation folds) indicating high correlation of the two signatures.

References

[1] M. Manica, J. Cadow, R. Mathis, and M. R. Martínez. "PIMKL: Pathway Induced Multiple Kernel Learning" (2018). arXiv: 1803.11274.

[2] D. Szklarczyk, J. H. Morris, H. Cook *et al.* "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible." *Nucleic acids research* **45**:D1 (2017), pp. D362–D368.

[3] D. Türei, T. Korcsmáros, and J. Saez-Rodriguez. "OmniPath: guidelines and gateway for literature-curated signaling pathway resources". *Nature methods* **13**:12 (2016), p. 966.

[4] D. Croft, A. F. Mundo, R. Haw *et al.* "The Reactome pathway knowledgebase". *Nucleic Acids Research* **42**:D1 (2014), pp. 472–477. eprint: /oup/backfile/content_public/journal/nar/42/d1/10.1093/nar/gkt1102/2/gkt1102.pdf.

[5] A. Fabregat, S. Jupe, L. Matthews *et al.* "The Reactome Pathway Knowledgebase". *Nucleic Acids Research* **46**:D1 (2018), pp. D649–D655. eprint: /oup/backfile/content_public/journal/nar/46/d1/10.1093_nar_gkx1132/2/gkx1132.pdf.

[6] S. Kerrien, B. Aranda, L. Breuza *et al.* "The IntAct molecular interaction database in 2012". *Nucleic acids research* **40**:D1 (2011), pp. D841–D846.

[7] L. Licata, L. Briganti, D. Peluso *et al.* "MINT, the molecular interaction database: 2012 update". *Nucleic acids research* **40**:D1 (2011), pp. D857–D861.

[8] E. Chautard, L. Ballut, N. Thierry-Mieg, and S. Ricard-Blum. "MatrixDB, a database focused on extracellular protein–protein and protein–carbohydrate interactions". *Bioinformatics* **25**:5 (2009), pp. 690–691.

[9] T. Keshava Prasad, R. Goel, K. Kandasamy *et al.* "Human protein reference database—2009 update". *Nucleic acids research* **37**:suppl_1 (2008), pp. D767–D772.

[10] J. D. Zhang and S. Wiemann. "KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor". *Bioinformatics* **25**:11 (2009), pp. 1470–1471.

[11] Tenenbaum D. *KEGGREST:Client-side REST access to KEGG*. 2016.

[12] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic acids research* **28**:1 (2000), pp. 27–30.

[13] E. G. Cerami, B. E. Gross, E. Demir *et al.* "Pathway Commons, a web resource for biological pathway data". *Nucleic Acids Research* **39**:SUPPL. 1 (2011).

[14] Y. Cun and H. Fröhlich. "Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge". *BMC bioinformatics* (2012).

[15] Z. Guo, T. Zhang, X. Li *et al.* "Towards precise classification of cancers based on robust gene functional expression profiles". *BMC Bioinformatics* **6** (2005).

[16] I. W. Taylor, R. Linding, D. Warde-Farley *et al.* "Dynamic modularity in protein interaction networks predicts breast cancer outcome". *Nature biotechnology* **27**:2 (2009), p. 199.

[17] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee. "Inferring pathway activity toward precise disease classification". *PLoS computational biology* **4**:11 (2008), e1000217.

[18] Y. Zhu, X. Shen, and W. Pan. "Network-based support vector machine for classification of microarray samples". *BMC Bioinformatics 2009 10:1* **10**:1 (2009), S21.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. "Gene selection for cancer classification using support vector machines". *Machine learning* **46**:1-3 (2002), pp. 389–422.

[20] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. P. Vert. "Classification of microarray data using gene networks". *BMC Bioinformatics* **8** (2007).

[21] C. Gao, X. Dang, Y. Chen, and D. Wilkins. "Graph ranking for exploratory gene data analysis". *BMC bioinformatics*. Vol. 10. BioMed Central. 2009, S19.

[22] J. C. Costello, L. M. Heiser, E. Georgii *et al.* "A community effort to assess and improve drug sensitivity prediction algorithms". *Nature biotechnology* **32**:12 (2014), p. 1202.

[23] M. Gönen and E. Alpaydın. "Multiple kernel learning algorithms". *Journal of machine learning research* **12**:Jul (2011), pp. 2211–2268.

[24] L. Jacob, P. Neuvial, and S. Dudoit. "More power via graph-structured tests for differential expression of gene networks". *Ann. Appl. Stat.* **6**:2 (2012), pp. 561–600.

[25] A. Livshits, A. Git, G. Fuks, C. Caldas, and E. Domany. "Pathway-based personalized analysis of breast cancer expression data". *Molecular Oncology* **9**:7 (2015), pp. 1471–1483.

[26] Y.-H. Chang, C.-M. Chen, H.-Y. Chen, and P.-C. Yang. "Pathway-based gene signatures predicting clinical outcome of lung adenocarcinoma". *Scientific Reports* **5** (2015), p. 10979.

[27] C. Curtis, S. P. Shah, S.-F. Chin *et al.* "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". *Nature* **486** (2012).

[28] A. Liberzon, C. Birger, H. Thorvaldsdóttir *et al.* "The Molecular Signatures Database Hallmark Gene Set Collection". *Cell Systems* **1**:6 (2015), pp. 417–425.

[29] F. Hillen and A. W. Griffioen. "Tumour vascularization: sprouting angiogenesis and beyond". *Cancer Metastasis Rev.* **26**:3-4 (2007), pp. 489–502.

[30] J. Hooda, M. Alam, and L. Zhang. "Evaluating the association of heme and heme metabolites with lung cancer bioenergetics and progression". *Metabolomics* **5**:3 (2015), p. 1000150.

[31] R. K. Kim, Y. Suh, K. C. Yoo *et al.* "Activation of KRAS promotes the mesenchymal features of basal-type breast cancer". *Exp. Mol. Med.* **47** (2015), e137.

[32] A. K. Najumudeen, A. Jaiswal, B. Lectez *et al.* "Cancer stem cell drugs target K-ras signaling in a stemness context". *Oncogene* **35**:40 (2016), pp. 5248–5262.

[33] K. A. Hogan, D. S. Cho, P. C. Arneson *et al.* "Tumor-derived cytokines impair myogenesis and alter the skeletal muscle immune microenvironment". *Cytokine* (2017).

[34] D. Bedognetti, W. Hendrickx, F. M. Marincola, and L. D. Miller. "Prognostic and predictive immune gene signatures in breast cancer". *Current Opinion in Oncology* **27**:6 (2015), pp. 433–444.

[35] L. Lima and R. Monteiro. "Activation of blood coagulation in cancer: implications for tumour progression". *Bioscience Reports* **33**:5 (2013), pp. 701–710.

[36] M. Belting, J. Ahamed, and W. Ruf. "Signaling of the tissue factor coagulation pathway in angiogenesis and cancer". *Arterioscler. Thromb. Vasc. Biol.* **25**:8 (2005), pp. 1545–1550.

[37] A. FALANGA, M. MARCHETTI, and A. VIGNOLI. "Coagulation and cancer: biological and clinical aspects". *Journal of Thrombosis and Haemostasis* **11**:2 (2013), pp. 223–233.

[38] A. Vazquez, E. E. Bond, A. J. Levine, and G. L. Bond. "The genetics of the p53 pathway, apoptosis and cancer therapy". *Nature Reviews Drug Discovery* **7**:12 (2008), pp. 979–987.

[39] A. Mandinova and S. W. Lee. "The p53 Pathway as a Target in Cancer Therapeutics: Obstacles and Promise". *Science Translational Medicine* **3**:64 (2011), 64rv1–64rv1.

[40] H. K. Delille, N. A. Bonekamp, and M. Schrader. "Peroxisomes and disease - an overview". *Int J Biomed Sci* **2**:4 (2006), pp. 308–314.

[41] M. Fransen, M. Nordgren, B. Wang, and O. Apanasets. "Role of peroxisomes in ROS/RNS-metabolism: Implications for human disease". *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1822**:9 (2012), pp. 1363–1373.

[42] F. Aiolli and M. Donini. "EasyMKL: A scalable multiple kernel learning algorithm". *Neurocomputing* **169** (2015), pp. 215–224.

[43] J. Mariette and N. Villa-Vialaneix. "Unsupervised multiple kernel learning for heterogeneous data integration". *Bioinformatics* **34**:2009 (2017).

[44] C. M.Bishop. "Pattern Recognition and Machine Learning". *Springer* (2006), p. 738.

[45] F. Aiolli, G. Da San Martino, and A. Sperduti. "A kernel method for the optimization of the margin distribution". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **5163 LNCS**:PART 1 (2008), pp. 305–314.

[46] W. N. Anderson and T. D. Morley. "Eigenvalues of the Laplacian of a Graph". *Linear and Multilinear Algebra* **18**:2 (1985), pp. 141–145.

[47] L. Chen, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang. "Identifying cancer biomarkers by network-constrained support vector machines". *BMC Systems Biology* **5** (2011).

[48] T. Barrett, S. E. Wilhite, P. Ledoux *et al.* "NCBI GEO: Archive for functional genomics data sets - Update". *Nucleic Acids Research* **41**:D1 (2013).

[49] Y. Wang, J. G. Klijn, Y. Zhang *et al.* "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer". *Lancet* **365**:9460 (2005), pp. 671–679.

[50] Y. Pawitan, J. Bjöhle, L. Amler *et al.* "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts". *Breast Cancer Research* **7**:6 (2005), R953.

[51] C. Sotiriou, P. Wirapati, S. Loi *et al.* "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis". *Journal of the National Cancer Institute* **98**:4 (2006), pp. 262–272.

[52] A. V. Ivshina, J. George, O. Senko *et al.* "Genetic Reclassification of Histologic Grade Delineates New Clinical Subtypes of Breast Cancer". *Cancer Research* **66**:21 (2006), pp. 10292–10301.

[53] C. Desmedt, F. Piette, S. Loi *et al.* "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series." *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**:11 (2007), pp. 3207–14.

[54] M. Schmidt, D. Böhm, C. Von Törne *et al.* "The humoral immune system has a key prognostic impact in node-negative breast cancer". *Cancer Research* **68**:13 (2008), pp. 5405–5413.

5 Accelerated analysis of Boolean gene regulatory networks via reconfigurable hardware[1]

Matteo Manica[1,2,*], Raphael Polig[1,*], Mitra Purandare[1,*], Roland Mathis[1], Christoph Hagleitner[1], María Rodríguez Martínez[1]

[1] IBM Research Zürich

[2] ETH - Zürich

[*] Shared first authorship

Abstract

Boolean models are a powerful abstraction for qualitative modeling and analysis of gene regulatory networks dynamics. With the development of advanced high-throughput technologies, the availability of experimental data about the molecular interactions within the cell has reached unprecedented volumes and accuracy, and hence modern Boolean models are increasing in size and complexity. Unfortunately, current software simulation tools have not scaled at the same speed and hence cannot handle properly complex Boolean models of large size.

Field Programmable Gate Arrays (FPGAs) are powerful reconfigurable integrated circuits that can offer massive performance improvements. Due to their highly parallel nature, FPGAs are extremely well suited to simulate complex molecular networks. We present here a new simulation framework for Boolean models, which first converts the model to Verilog, a standardized hardware description language, and then connects it to an execution core that runs on an FPGA coherently attached to a POWER8 processor. We report an order of magnitude

[1] Submitted. See Chapter 8 for details about contributions and copyright.

speed up over a multi-threaded software simulation tool running on the same processor on a set of state-of-the-art Boolean models. An analysis of a T-cell large granular lymphocyte (T-LGL) network is performed to show how the implemented framework can help to obtain biological insights with its consistent performance improvements. In addition, we show that our framework allows to perform attractor detection at an unprecedented speed, exhibiting a speedup ranging from one to three orders of magnitude compared to a software solution.

5.1 Introduction

Genes do not work in isolation, but exert their function in complex and tightly connected gene regulatory networks (GRNs) [1]. At the very basis, understanding complex diseases amounts to unravelling normal and dysregulated behavior of GRNs. However due to their complexity and the lack of quantitative knowledge about most kinetic parameters governing molecular interactions, an exact analysis of GRNs, usually based on ordinary differential equations (ODEs), is in most cases not possible.

Boolean models [2, 3] are an attractive alternative approach for the study of GRNs that are consistently used in the systems biology community [4–12]. Boolean models provide a qualitative description of a GRN, where chemical species concentrations or activities are represented using a finite set of discrete values. In a Boolean model, a node corresponds to a species (e.g., gene, protein, etc.) and an edge represents an interaction between species. In its simplest form, a gene can be ON (1) or OFF (0), and its interactions with other genes are defined by means of a Boolean function of its parent nodes in the GRN. Time is represented by discrete steps after which the Boolean functions are evaluated following an update scheme and assigned to their corresponding genes [13]. Various update schemes can be adopted. In the *synchronous* scheme [14], all genes are simultaneously evaluated and updated, resulting in a fully deterministic and computationally tractable system, although often biologically unrealistic. Conversely, the *asynchronous* scheme [15] randomly chooses a gene and updates it to its next value. The *asynchronous* scheme provides a stochastic, and hence more realistic, description of a GRN, although at the price of greatly increasing the computational complexity and running time of the model. In addition, as the model is stochastic, it has to be run multiple times in order to resolve the mean dynamical behavior. Although a Boolean model cannot provide the level of detailed information that an experimentally well-characterized ordinary differential equation (ODE) system can achieve, it can produce a qualitative description of the most salient features of a dynamical system. For instance, Boolean models can be very

useful to identify steady states, cyclic states or attractors – cycles of states A such that no trajectory entering in A can leave A. These model predictions can provide valuable insights about observed phenotypes and the molecular processes underlying them [16–18].

One of the main goal of Boolean network analysis is the determination of attractors (steady states and/or cycles), which can provide valuable information about the underlying mechanisms associated with complex diseases, such as cancer [19]. However, the problem is known to be characterized by a high computational complexity, mostly depending on the number of network nodes. Furthermore, the number and the size of the attractors of a system is known to change dramatically with the update scheme [20]. Some types of attractors, such as self-loops and simple loops, are common to both update schemes and hence can be computed using the less expensive synchronous update scheme. However, in the most general case, the characterization of the attractors landscape of a model requires using asynchronous updates, resulting in high complexity in the number of states conforming the attractor, as well as lengthy transitory states leading to an attractor [13].

Generally speaking, the computational problem of finding all the attractors in a Boolean model is extremely hard. Even the simpler problem of finding the steady states in a Boolean model is NP-hard [21, 22], indicating that it is not possible to efficiently, i.e., in polynomial time, find all attractors in the analysed system. However, we have shown in a previous work that it is possible to provide scalable solutions that are fast enough [23]. Namely, we proposed a hardware accelerated simulation framework based on the use of FPGAs for synchronous and asynchronous simulation of Boolean models. Due to the highly parallel nature and ever-increasing capacity of FPGAs, our approach scaled efficiently, showing a significant speedup compared with BoolNet [24], a popular R package for the construction and analysis of Boolean networks.

In this paper, we extend our FPGA simulator to perform attractor detection. The accelerator is seamlessly integrated with a POWER8 processor, greatly increasing the usability of the proposed framework. We demonstrate the performance of our accelerator using six state-of-the-art Boolean models from literature, including models for T-cell large granular lymphocyte leukemia [25], castration resistant prostate cancer [7], signaling pathways involved in cancer [8], colon cancer [5], Fanconi anemia and breast cancer [26], and the MAPK pathway [9]. Firstly, we compare runtime performance of our framework with multi-threaded implementations of two commonly used software tools: BoolNet and BooleanNet [27], running on a POWER8 processor; and with an existing accelerator proposed by Miskov-Zivanov *et al.* [28]. Our framework demonstrates an order of magnitude speedup over BoolNet, which

already runs significantly faster than BooleanNet; and exhibits better performance compared with the existing accelerator. We also include an analysis of the dynamic behavior of the T cell type of large granular lymphocyte (T-LGL) model [25]. Secondly, we measure performance of our accelerator in attractors detection. By considering BoolNet as a baseline, we observe a speedup ranging from one to three orders of magnitude.

This paper is organized as follows. An overview of existing hardware accelerated solutions and basics of GRN and Boolean models are reported next. Results and discussion are provided in Sections 5.2 and 5.3, respectively. Details about the framework and its implementation are described in Section 5.4.

5.1.1 Comparison with existing literature

Boolean GRNs can be studied with simulators, such as BooleanNet [27] and BoolNet [24]. However, simulation of Boolean models on conventional computers, especially, asynchronous simulation, usually results in prohibitively long execution times due to the intrinsic disparity between the sequential steps executed by a microprocessor program and the highly parallel nature of information flow within biochemical networks [29]. We observe similarly long running times on multi-threaded simulations running on a POWER8 processor.

Regarding the computation of attractors, common methods start with randomly selected initial states and perform exhaustive searches of the state space of a network. However, the time complexity of these methods grows exponentially with the number of nodes in the network, and hence, techniques to reduce the complexity of the state space have been proposed. For instance, the entire network state space can be appropriately broken down into selected subspaces that can be exhaustively searched [30]. However, this approach is not scalable and it is currently limited to networks containing up to 150 nodes. Network reduction techniques that conserve the fixed points and complex attractors of general asynchronous Boolean models have been developed [31]. A systematic removal of state transitions, to render the state transition graph acyclic, transforms all attractors into fixed points that can be enumerated with little effort [20]. Finally, a mathematical model of a pruned portion of the state space, followed by a randomized traversal method to extract the steady states in the remaining state space, has also been proposed to increase speed and scalability [32].

When approximate solutions are not desirable, symbolic approaches can be efficient as they do not perform explicit traversal of the state space. Reduced ordered binary decision diagrams (ROBDDs) use directed acyclic graphs to represent large Boolean functions in a space-efficient

manner, and are computationally suitable for complex Boolean operations (e.g., logical AND, OR, etc.) and set operations (e.g., union, intersection, etc.) [13, 33]. Still, BDDs have generally unpredictable memory requirements. Satisfiability solvers, usually more scalable than BDDs, are also popular in attractor computation [34, 35]. But with increasing number of genes and length of unwinding these approaches become inefficient. Analysis of Networks through TEmporal-LOgic sPEcifications (Antelope) uses model checkers, a collection of techniques for automatically verifying properties of discrete systems, for analyzing and constructing Boolean GRNs [36]. Unlike simulators, model checkers can prove properties of a set of infinitely many paths. In addition, they can handle new, unforeseen properties by simply adding temporal-logic formulas, while simulators require the incorporation of such properties in their program code. Despite these properties, one common disadvantage of symbolic approaches in comparison with explicit approaches is that the attractors are available at the very end of the computation which can take a prohibitively long time.

Explicit approaches are not scalable but can present results as and when available. A practical solution is to accelerate them using highly parallel Field programmable Gate Arrays (FPGAs). Hardware accelerated biological network simulators have been proposed in the past [29, 37], where reprogrammable FPGA hardware has been applied to efficiently simulate the stochastic behavior of biological systems. These early works demonstrated the suitability of FPGA technology for the simulation of variants of the Gillespie algorithm, achieving a performance 20 times faster than a competing general purpose CPU. Here we apply FPGAs to the simulation and analysis of Boolean networks. An FPGA-based accelerator framework for Boolean models has been demonstrated by Miskov-Zivanov *et al.* [28, 38]. This framework performs asynchronous updates and does not perform attractor analysis. To the best of our knowledge, the framework is not fully integrated with the host system limiting its accessibility by the user software. Buttons are used to manually start and stop the simulation on the FPGA. The state of the network is displayed using 7-segment LED displays. This prohibits any further analysis of computed results. The framework we propose in this paper is seamlessly integrated with a POWER8 processor greatly increasing its usability and integration with other software tools.

5.1.2 Simulating biological networks

The central dogma of biology explains the transfer of information between genes (DNA), transcripts, and proteins. Genes are used as templates to create mRNAs through a process called *transcription*. In turn, mRNAs get *translated* to proteins. Some proteins act as *transcription*

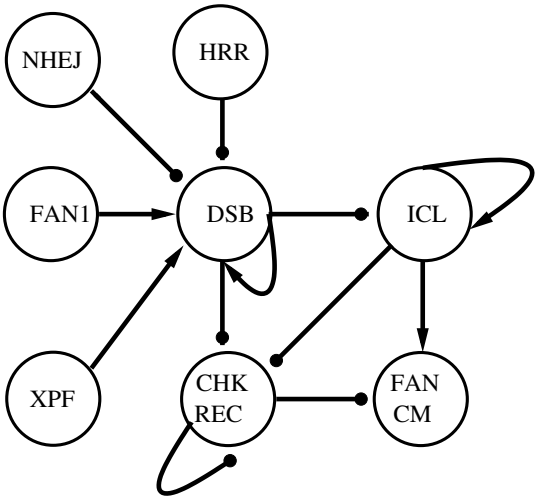


Figure 5.1: **Partial Network of FA-BRCA pathway.** Sub-network representing part of the Boolean model for Fanconi Anemia/Breast Cancer (FA/BRCA) pathway proposed by Rodríguez *et al.*.

factors and can up-regulate or down-regulate the expression of other genes. Collecting evidence from molecular data, either from high-throughput technologies or from literature, can help us to reverse engineer biological systems.

We can formally divide models for gene regulatory networks in two main categories: *static* and *dynamic*. Static models represent a network as a graph (set of nodes connected by edges), which is analysed using graph-theoretic approaches. Figure 5.1 depicts a static network from the FA-BRCA pathway [26]. Each circular node is a gene. Edges indicate the direction of the regulatory interaction between genes and its type, namely, activation or inhibition. In Figure 5.1, the gene ICL is suppressed by the gene DSB (DSB $\multimap$ ICL), whereas the gene XPF activates the gene DSB (XPF$\rightarrow$ DSB).

Dynamic models instead, describe how gene expression values change over time. These approaches can be further divided in *continuous/quantitative* and *discrete/qualitative* [39]. Continuos or quantitative models represent concentrations of constituents as continuous functions of time and their dynamics is represented using ordinary differential equations or stochastic equations. The downside of quantitative models is that they require an accurate knowledge of the kinetic parameters which are not easily obtainable (e.g., growth rate, decay rate, etc.). Though these models can represent detailed interactions between constituents, their use is limited.

In qualitative or discrete models, each constituent is assumed to take discrete values and the interactions between constituents are modeled by discrete functions. It is common knowledge that discretizing continuous-time data results in loss of information. However, these models can be a suitable choice when only the gene network topology and partial qualitative information is known due to the complexity of the network. Popular quantitative models are: Boolean networks [2, 3], multi-valued models [40, 41], and Petri nets [42]. In this work we focus on Boolean models.

5.1.3 Boolean models

A Boolean model of a GRN including n genes assigns a Boolean variable g_i to each gene i. If a gene i is active or expressed, the corresponding variable $g_i = 1$. If a gene i is inactive, the corresponding variable $g_i = 0$. Figure 5.2 depicts a general Boolean model for a set of genes $g_1, g_2, \cdots g_n$. The next state of a gene i is determined by a Boolean function dependent on the current values of its neighboring genes with incoming connections in g_i. The Boolean function consists of logical operators (AND, OR, and NOT).

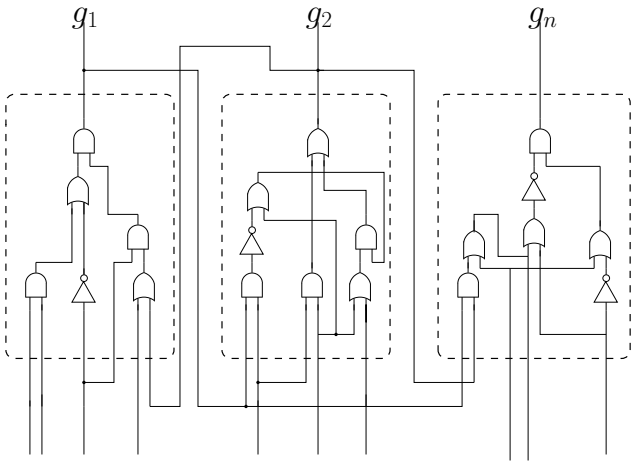


Figure 5.2: **A general Boolean model.** Genes in the model are connected using different types of Boolean gates describing the action of a logical operator. As in a circuit the next value for each node is dependent on the values of all the incoming connections carrying the current value of neighboring nodes.

Table 5.1 illustrates how regulatory functions of some genes in Figure 5.1 can be expressed as Boolean functions. For example, the second row indicates that the genes ICL and CHKREC regulate gene FANCM. The corresponding Boolean function is ICL and not CHKREC, i.e.,

Gene Name	Regulatory Function
ICL	ICL and not DSB
FANCM	ICL and not CHKREC
DSB	(DSB or FAN1 or XPF) and not (NHEJ or HRR)
CHKREC	not ICL and not DSB and not CHKREC

Table 5.1: **Updated rules for some nodes of the network from Figure 5.1.** Given a static network with information about activation and inhibition we can build updated rules for every node. An update rule or Boolean function computing the state of a node must only depend on the values of neighbors connected with incoming edges.

the gene FANCM is active only when ICL is active and CHKREC is inactive.

Update order

In Boolean models, variables are updated only at discrete/fixed times. The state of a Boolean model at a time t is a set of gene values. Depending on the order in which next state values are assigned to gene variables, various update schedule/schemes are possible. All genes simultaneously get the next values in a synchronous scheme. Such a synchronous update scheme is largely unrealistic as biological processes are known to be characterized by different reaction rates. Asynchronous schemes take in to account time diversity and different reaction rates of biological systems by updating variables in a non-synchronous order. These schemes can be stochastic, in this case each variable is updated with a certain probability. There are several flavors of stochastic asynchronous updates. One *general* variant involves updating one randomly selected node/variable at each time step. Another variant, known as, *random order asynchronous update* involves generating a random of permutation of variables each time and updating the variables in that order. A combined strategy [28] has been also proposed, where variables are ranked and grouped according to the rank. Groups are then updated in the order of the rank following a random asynchronous update of all variables in the group. This results in a *ranked asynchronous* order.

Attractors

Representing continuous-time data as two discrete values 0 and 1 results in a loss of precise quantitative information. Despite the drawbacks caused by this discretization, Boolean models are still extremely useful for analyzing long term behavior, namely, cycles of states arising

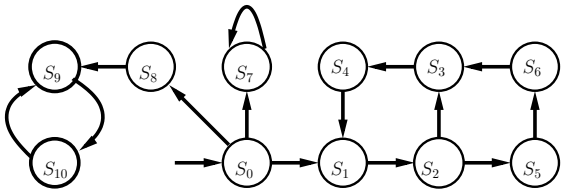


Figure 5.3: **Cycles types.** Schematic depiction of different types of cycles.

from the possible feedback loops in the network. Figure 5.3 illustrates different types of cycles.

A set A of states is an attractor if A is strongly connected (all the states in A are reachable from each other) and no transitions starting in A can leave A. An attractor A is a simple attractor if every state has exactly one successor. The cycles S_9, S_{10}, S_9 and S_7, S_7 are simple attractors. The cycle $S_1, S_2, S_5, S_6, S_3, S_4, S_1$ forms a non-simple, also defined as complex, cycle.

Different update schemes may result in different types of attractors [13]. Simple attractors of a Boolean model can be reached by using a synchronous update strategy, while complex attractors can only be reached by applying asynchronous updates. In the following we focus on analyzing runtime performance of Boolean model simulations using asynchronous updates and simple attractor detection by using a synchronous update scheme.

5.2 Results

The performance of our accelerator framework is evaluated on six published models with various number of nodes and complexity. T-LGL, a Boolean model proposed in [25] for T-cell large granular lymphocyte leukemia (here we consider a simplified version of the model from the set of examples that are provided with BooleanNet). CRPC, a model by Hu et. al [7] including relevant pathways for castration resistant prostate cancer. Fumia, a model integrating the main signaling pathways in cancer [8]. CAC, a model for the development of colitis-associated colon cancer integrating the extracellular environment and intracellular signaling pathways [5]. FA-BRCA, a Boolean model describing Fanconi Anemia/Breast Cancer (FA/BRCA) pathway [26]. MAPK, comprehensive model of MAPK pathway [9].

5.2.1 Asynchronous simulation

Runtime analysis

Software simulations of BoolNet and BooleanNet constitute the baseline and are performed on the same POWER8-based server node that hosts the FPGA accelerator. The server has 20 physical cores running at 2.29 GHz and a total of 512 GB DDR3 RAM. The simulations are run using the BooleanNet simulation package for Python and BoolNet for R. The benchmarks processed all simulation jobs with 20 worker threads simultaneously to fully utilize the server node.

Our framework uses the Xilinx Kintex UltraScale KU060 FPGA and the target frequency is 250 MHz. The measurements include the time for transferring the parameters to the FPGA and transferring the results from the FPGA to the main memory. Only one software thread has been used to perform the memory management and control for the FPGA.

We ran our accelerator for asynchronous simulations on the following models: T-LGL, CRPC and Fumia. Table 5.2 summarizes the results for each case study. All possible input combinations are generated as individual simulation jobs (number of simulations in Table 5.2). Each job is then simulated by BooleanNet, BoolNet, and the FPGA. Only for the CRPC model the number of simulations has been limited due to the long runtime. Each simulation job has been simulated in asynchronous mode on the models for 100 time steps and repeated 100 times.

Model	Number of inputs	Number of outputs	Number of simulations	runtime BooleanNet	runtime BoolNet	runtime FPGA	Time per sim. BooleanNet	Time per sim. BoolNet	Time per sim. FPGA
T-LGL	4	47	16	90.1s	0.88s	0.12s	5.6s	0.043s	0.007s
CRPC	22	69	64	580.5s	3.51s	0.23s	9.1s	0.043s	0.003s
Fumia	6	92	64	7895.7s	3.52s	0.30s	123.4s	0.044s	0.004s

Table 5.2: **Asynchronous simulation benchmark.** Summary of the execution times for evaluated models: T-LGL, CRPC and Fumia. Results for 100 time steps and 100 repetitions in asynchronous mode are reported.

Compared to BooleanNet and BoolNet, the FPGA accelerator exhibits a speedup of 750.8x and 7.3x respectively for the T-LGL model. For the CRPC model, it takes a prohibitively long time to generate all inputs in case of software simulations and hence, the number of simulations is limited to 64. While BooleanNet apparently struggles to simulate the CRPC model, BoolNet runtime is dominated by the number of simulations. In this case the speedup obtained is 2523.9x compared to BooleanNet and 15.2x compared to BoolNet. The FPGA

accelerator demonstrates a speedup of 26,319x and 11.7x times respectively over BooleanNet and BoolNet in Fumia case.

Comparison with Miskov-Zivanov *et al.* [28]. A runtime of 0.019s for the T-LGL model for 200 repeats and 15 time steps using an FPGA accelerator has been previously reported [28]. These experiments have been conducted on a standalone FPGA board at 50 MHz. Adjusting this number for 100 time steps, 100 iterations, and a frequency of 250 MHz, such a simulation would take 0.012s. This is 68% slower than the presented architecture and has roots in the generation of the random update order. The runtime of the architecture is non-deterministic due to its reliance on the random order generated by the linear-feedback shift register (LFSR).

Dynamic behavior analysis of T-LGL

Analysis of the dynamic behavior of the T-LGL leukemia network has identified a fundamental role for the Apoptosis node [43]. When this node is ON a single steady state where programmed cell death is normally occurring is found. When it is instead stabilized at OFF two additional fixed points for which the disease is active appear. This criterion is used to group steady state behavior into the T-LGL leukemia class (disease) and into the apoptosis attractor class (normal). We now present our simulation results related to the nodes Apoptosis and BID, one of the Fas-induced apoptosis pathway elements.

For a given set of initial states, asynchronous simulations necessitate repeating simulation runs several times each time updating the nodes in a random order. We compute the number of times a node is 1/0 at specific time points in all the repetitions. If a simulation is repeated few times, the activation frequencies do not converge as depicted in Figure 5.4. The frequency curve for apoptosis for 10 or 100 repetitions is not smooth. A similar unstable curve can be observed for BID as well. As the number of repetitions are increased, the curves start to look smoother.

The order of magnitude speedup achieved by our hardware accelerator framework enables large number of repetitions of the simulations for the given initial states. The larger the number of repetitions of a simulation, the more accurate and smoother are the estimates of the evolution of node states frequencies over the time steps. This aspect plays a major role when analyzing the dynamics of the system, since node activation frequencies convergence can be considered as an indicator of the presence of an attractor. Figure 5.4 shows how increasing the number of repetitions for a given initial state changes the activation frequency estimates in T-LGL. The curves reported for the different repetitions numbers tend to converge after 1000

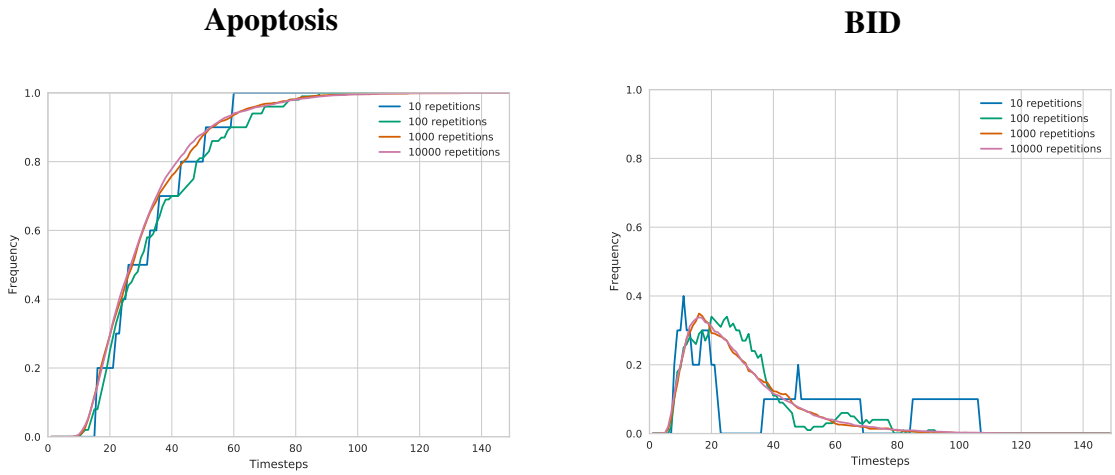


Figure 5.4: **Frequency at different number of repetitions.** Line plots for different number of repetitions of a simulation with a fixed initial state in T-LGL are reported. In the two panels the activation frequencies for Apoptosis (left) and BID (right) are shown. The increased number of repetitions smooth the curves resulting in more accurate frequencies estimates that consistently capture the system dynamic.

repetitions. The curves for the Apoptosis node illustrate how estimates change their evolution over the time steps: fewer repetitions underestimate the steepness. This exhibits how higher number of repetitions capture the dynamics of the system in a more consistent fashion. For the node BID, we observe that the dynamics captured with higher number of repetitions tends to converge towards a curve with a maximum that is consistently estimated only when the simulation is repeated over 1000 times. This example shows that an efficient simulator helps to better capture the dynamic of a biological system by enabling high number of repetitions with low runtime requirements.

5.2.2 Attractor analysis

BooleanNet has been excluded from the benchmark for attractor analysis given its poor performance. BoolNet does not perform exhaustive attractor analysis if the number of nodes in the model is greater than 29. All models considered in the benchmark exceed this limit. Hence, the method for finding the attractors has been set to *chosen* so that the attractor analysis is limited to selected initial states.

Attractor search is performed by running multiple synchronous simulations using different start states. The time to generate the start states is excluded from timing measurements. The

Model	#nodes	Tool	#Start states	Time	Time per state	#attractors
T-LGL	51	BoolNet	2^{12}	0.12 s	32 us	3
		FPGA	2^{28}	6.46 s	0.024 us	5
CRPC	91	BoolNet	2^{16}	13.43 s	205 us	1
		FPGA	2^{20}	1.64 s	1.56 us	135
Fumia	98	BoolNet	2^{16}	15.62 s	238 us	4
		FPGA	2^{20}	0.59 s	0.56 us	26
CAC	70	BoolNet	2^{16}	20.17 s	307 us	4
		FPGA	2^{25}	1.59 s	0.047 us	6
FA-BRCA	28	BoolNet	2^{28}	214.8 s	0.8 us	1
		FPGA	2^{28}	4.31 s	0.016 us	1
MAPK	53	BoolNet	2^{16}	12.12 s	184 us	4
		FPGA	2^{24}	13.01 s	0.7 us	10

Table 5.3: **Attractor search benchmark.** Summary of the evaluated models and results for the synchronous attractor search comparing BoolNet and our framework.

number of start states has been selected such that the runtime for a specific model is sufficiently long (>10 s) to avoid side effects for short runs.

Table 5.3 summarizes the evaluated models and the measured runtime for both BoolNet and our accelerator framework. The number of start states is different for the FPGA and BoolNet runs to adjust for different performance characteristics. When we used the same number of start states for both, it often happened that either BoolNet ran for too long or the accelerator was too fast resulting in a short runtime.

As the overall runtime is dependent on the number of start states that have been used for the attractor search, Table 5.3 includes a runtime per state column to make the tools comparable. The speedup factors of the FPGA framework range from 50x to 6,531x over BoolNet. An observation can be made on the FA-BRCA model which is the smallest model in terms of number of nodes. For BoolNet the runtime per state is significantly better compared to the other models. This is probably due to the fact that in this regime it can perform an exhaustive search of all states.

Performance projections

As we run BoolNet only on a single core, a single accelerator core has been used on the FPGA for a fair comparison. The accelerator consumes only 2 % of the overall resources available on the FPGA. This allows the accelerator core to be replicated at least 20 times on a single FPGA card, resulting in a further speedup of 20x. The server system allows to plug in an additional coherent accelerator processor interface-based (CAPI-based) accelerator to further increase performance and adjust for a multi-threaded software implementation. When utilizing all cores of the POWER8 processor the performance of the software should increase linearly and be 20 times faster as well.

Comparison with boolSim [44]

We also ran boolSim, a symbolic ROBDD-based tool for attractor analysis on our models. Synchronous attractor computation for T-LGL (51 nodes) on an Intel Xeon processor running at 3.5GHz was performed in ~80 seconds and 71 attractors were reported.boolSim took 10 seconds to finish the analysis for MAPK (53 nodes). However, we observed that boolSim is unable to finish the attractor analysis in a reasonable time as the number of nodes increases. Some of the runs of boolSim had to be killed after running for a long time. For the Fumia model (98 nodes), boolSim kept running for >5317 minutes (approximately 4 days). The run time for both CRPC (91 nodes) and CAC (70 nodes) is >8352 minutes (approximately 6 days). We chose to stop boolSim after running for such a long time. As is observed in all symbolic approaches, no response/feedback has been presented to the user during this time. Our approach, though exhaustive, presents the results faster with the help of FPGAs. We ran into problems using the state-of-the-art software tool *geneFAtt* [45]. The source code is publicly available but seems to be incomplete. The compilation failed not only on the POWER system but also on an x86 system.

5.2.3 Further improvements

FPGA utilization

The simulation core is rather small leaving the FPGA resources under utilized. Each model requires around 7,200 to 7,400 look-up tables (LUTs) which is about 2% of the overall resources available on the KU060 FPGA. The BlockRAM requirements are higher due to the collector

Model	LUTs	BRAM	LUTs (%)	BRAM (%)
T-LGL	7227	43	2.1	4.0
CRPC	7345	75	2.2	6.9
Fumia	7405	75	2.2	6.9
PSL	54945	281	16.5	26.0

Table 5.4: **FPGA resources requirements.** Required FPGA resources for the core per model and the property specification language (PSL).

module. The entire core requires 43 BlockRAM instances for the T-LGL model. Each of the larger models requires 75 instances which is about 7% of all BlockRAMs on this FPGA chip. The POWER Service Layer and the interconnecting modules require far more logic resources and BlockRAMs. These are necessary to connect the accelerator to the host system. Table 5.4 summarizes the required resources.

Performance enhancements

As the core requires little resources on the FPGA, multiple instances can be used to further reduce the processing times. This will more efficiently utilize the available bandwidth towards the processor. More results can be sent back concurrently. Since results are sent back only after all simulation repetitions are complete, a single core requires a high bandwidth. The experiments indicate a utilization of less than 1% of the available bandwidth of CAPI. Recent research has demonstrated the use of network-attached FPGAs to accelerate applications [46]. Boolean network simulations can leverage such an architecture by distributing the simulations across multiple FPGAs. This will allow to scale the models even further without sacrificing performance.

5.3 Discussion

In this work we have presented our FPGA-based framework for simulation of Boolean models and computation of attractors. We show that our accelerator can be used to more efficiently simulate network dynamics with asynchronous updates compared to existing software solutions. The proposed framework exhibits an order of magnitude speedup over existing multi-threaded software tools. We also leverage the speedup offered by our accelerator to perform a

massive number of repetitive asynchronous simulations of the T-LGL model. Our framework successfully computes simple attractors of large and complex Boolean models, exhibiting one to three orders of magnitude speedup over existing software solutions.

The results reported show that our solution enables analysis of Boolean models with unmatched performance. The low utilization of the FPGA observed in the models analyzed, shows that there is enormous room for improvement in terms of speed. A straightforward way to achieve additional speedup is to synthesize multiple instances of the Boolean model on the FPGA. In addition, simulations can also be distributed across multiple FPGAs if further speedup is necessary. Gaining speed helps us to get rid of current existing limitations for Boolean models in terms of number of nodes and model complexity. Being able to simulate and analyze larger and more complex Boolean networks, up to thousands of nodes, allows us to consider a more comprehensive description of a biological system and to fully exploit the potential of high-throughput molecular data.

Besides performance improvement, our framework can be easily extended to use other update strategies, such as random ranked updates. This will increase its ability to explore the state space, hence improving attractor detection. Additionally, another intriguing extension consists in implementing complex attractor computation on FPGAs to enable fast analysis of the reachable states of a Boolean model.

The integration of the accelerator with a POWER8 processor via CAPI greatly simplifies its usage. We believe that this is a fundamental feature in making our framework a valuable tool for the whole scientific community, offering the possibility to seamlessly integrate it in software applications.

5.4 Methods

This section describes our accelerator framework detailing its architecture and system integration.

Host processor and FPGA integration

The host is an IBM POWER8-based server system with the ability to coherently connect an FPGA via the coherent accelerator processor interface (CAPI). This enables the FPGA to act as part of a software process and access virtual memory locations just as a regular processor

core. Also, it allows the FPGA to access all of the system's main memory that has been allocated by the software process owning the accelerator. This solution proposed allows a seamless integration of the FPGA and the host processor. Fig. 5.5 provides an overview of the overall system architecture.

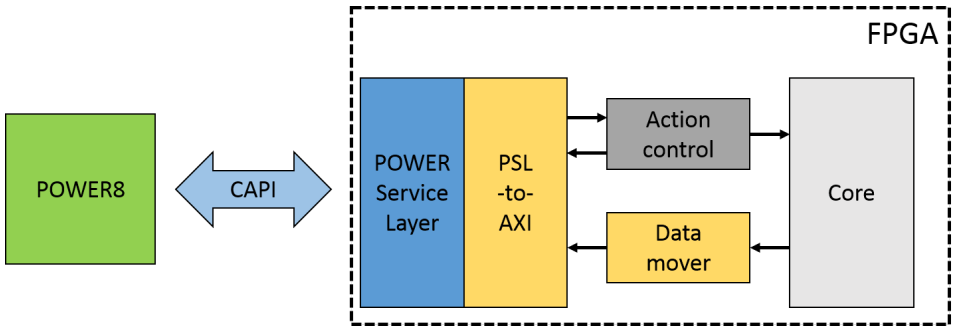


Figure 5.5: **System architecture overview.** Overall system architecture with the FPGA top-level. Communication between the FPGA card and the POWER8 processor is performed through CAPI.

Input arguments

The end-user is required to provide a Boolean model definition, a list of states to analyze, a number of time steps to simulate and a flag indicating whether to perform attractor analysis. In case of asynchronous updates the number of simulation repetitions has to be provided.

Hardware acceleration process in details

Once the arguments are received, our framework performs the following steps. First, the host converts the Boolean model into a hardware description language (HDL) model (Verilog). Second, the host creates a bit stream and configure the FPGA card for the computation. Afterwards, simulation parameters are transferred from the host to the FPGA. The simulation is then started on the FPGA using either a synchronous or asynchronous update strategies. Optionally, if enabled, the FPGA card checks for attractors. As soon as the results are processed the FPGA reports them to the host. Once the results are in the host they can be either displayed via a graphical user interface (GUI) or written on disk for further analysis.

FPGAs are essentially semiconductor devices that are based around a matrix of configurable logic blocks (CLBs) connected via programmable interconnects. FPGAs can be reprogrammed to desired application or functionality requirements after manufacturing. FPGA

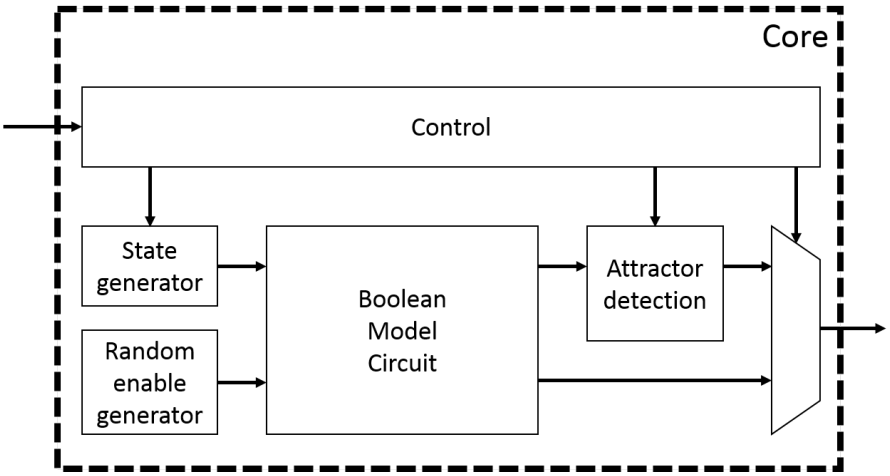


Figure 5.6: **Execution core scheme.** Here are included the top-level modules used in the execution core to implement: synchronous and asynchronous simulations as well as attractor detection.

vendors provide tools that accept a model written in an HDL to create a bit stream and configure the logic blocks and interconnects on the FPGA using this bit stream. The first two steps on the host take care of this.

Execution core

We have put two types of logic modules on the FPGA, namely, communicating modules (to-and-from host) and core modules. The core module contains the Boolean model and is responsible for simulation and analysis. Fig. 5.6 illustrates the top-level of the core with its main components.

The Boolean model is embedded in the Boolean network model circuit (BMC). In addition, the execution core all the necessary components to perform simulations of the Boolean model and further analyze the results. The core receives a *start* signal together with a set of arguments listed before. The implemented computational core is capable of performing synchronous or asynchronous simulations of the Boolean model and can detect simple attractors. The random enable generator takes care of selecting the node update order accordingly.

For asynchronous mode, a simulation is run for a given number of time steps and the simulations can be repeated for a specified number of times. The core captures the states reached in the multiple simulation iterations. It computes the fraction of simulations that had a certain gene node activated (1/ON) at a given time step.

The execution core is also capable of performing an exhaustive search for attractors in the Boolean model. The attractor detection unit collects the states that have been visited during a simulation. If a state is already in the list of visited states then the simulation stops. The states of the attractor are stored in a local memory. The core then moves on to the next initial state supplied to it.

Reporting simulation results

Due to the deterministic nature of synchronous updates, a state of a Boolean model has only one successor state. For a given input and its two simulation repetitions, the value of a particular output at a particular time step remains the same. This being the case, it is feasible to report all the states reached during simulation to the host/software.

This is not the case for asynchronous updates. For a fixed input and two simulation repetitions, the generated sequences of random permutations/updates is potentially different. Different sequences of update orders most likely result in different outputs. Hence, outputs at the same time step can be different for different simulation iterations. We present the results of such simulations in a meaningful manner. We record how often a particular node is active at a given time step. We then divide the value by the number of simulation repetitions. This gives us the percentage of simulations in which a node was active at a given time step. We perform this calculation for every time step. The collector module in the core is responsible for keeping track of the number of times an output has become active at each time step.

Declarations

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 668858.

Author contributions

See Chapter 8 for details about contributions and copyright.

Competing financial interests

The authors declare no competing financial interest.

Availability of data and materials

Results, processed data and materials presented in this work can be provided upon request.

References

[1] L. Koch. "A global view of regulatory networks". en. *Nature Reviews Genetics* **17**:5 (2016), pp. 252–252.

[2] L. Glass and S. A. Kauffman. "The Logical Analysis of Continuous, Non-linear Biochemical Control Networks". *Glass* - (1973), pp. 103–129.

[3] L. Glass and S. Kaufmann. "The logical analysis of continuous non-linear biochemical control networks". *Theoretical Biology* **39** (1973), pp. 103–129.

[4] I. N. Melas, A. D. Chairakaki, E. I. Chatzopoulou *et al.* "Modeling of signaling pathways in chondrocytes based on phosphoproteomic and cytokine release data". *Osteoarthritis and Cartilage* **22**:3 (2014), pp. 509–518.

[5] J. Lu, H. Zeng, Z. Liang *et al.* "Network modelling reveals the mechanism underlying colitis-associated colon cancer and identifies novel combinatorial anti-cancer targets". *Scientific reports* **5** (2015), p. 14739.

[6] H. Chen, G. Wang, R. Simha, C. Du, and C. Zeng. "Boolean models of biological processes explain cascade-like behavior". *Scientific reports* **6** (2016), p. 20067.

[7] Y. Hu, Y. Gu, H. Wang, Y. Huang, and Y. M. Zou. "Integrated network model provides new insights into castration-resistant prostate cancer". *Scientific Reports* **5**:April (2015), pp. 1–12.

[8] H. F. Fumiã and M. L. Martins. "Boolean Network Model for Cancer Pathways: Predicting Carcinogenesis and Targeted Therapy Outcomes". *PLoS ONE* **8**:7 (2013), p. 11.

[9] L. Grieco, L. Calzone, I. Bernard-Pierrot *et al.* "Integrative Modelling of the Influence of MAPK Network on Cancer Cell Fate Decision". *PLoS Computational Biology* **9**:10 (2013), pp. 1–15.

[10] D. P. Cohen, L. Martignetti, S. Robine *et al.* "Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration". *PLoS Computational Biology* **11**:11 (2015), e1004571.

[11] J. Saez-Rodriguez, L. Simeoni, J. A. Lindquist *et al.* "A logical model provides insights into T cell receptor signaling". *PLoS Computational Biology* **3**:8 (2007), pp. 1580–1590.

[12] J. Dorier, I. Crespo, A. Niknejad *et al.* "Boolean regulatory network reconstruction using literature based knowledge with a genetic algorithm optimization method". *BMC Bioinformatics* **17**:1 (2016), p. 410.

[13] A. Garg, A. Di Cara, I. Xenarios, L. Mendoza, and G. De Micheli. "Synchronous versus asynchronous modeling of gene regulatory networks". *Bioinformatics* **24**:17 (2008), pp. 1917–1925.

[14] S. Kauffman. "Homeostasis and differentiation in random genetic control networks". *Nature* **224**:5215 (1969), pp. 177–178.

[15] R. Thomas. "Regulatory networks seen as asynchronous automata: A logical description". *Journal of Theoretical Biology* **153**:1 (1991), pp. 1–23.

[16] A. Fauré, A. Naldi, C. Chaouiya, and D. Thieffry. "Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle". *Bioinformatics*. Vol. 22. 14. 2006, pp. 124–131.

[17] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. "The yeast cell-cycle network is robustly designed". *Proceedings of the National Academy of Sciences* **101**:14 (2004), pp. 4781–4786. arXiv: 0310010 [q-bio].

[18] S. Huang. "Gene expression profiling, genetic networks, and cellular states: An integrating concept for tumorigenesis and drug discovery". *Journal of Molecular Medicine* **77**:6 (1999), pp. 469–480.

[19] D. A. Orlando, C. Y. Lin, A. Bernard *et al.* "Global control of cell-cycle transcription by coupled CDK and network oscillators". *Nature* **453**:7197 (2008), pp. 944–947. arXiv: NIHMS150003.

[20] T. Skodawessely and K. Klemm. "Finding attractors in asynchronous Boolean dynamics". *Advances in Complex Systems* **14**:03 (2010), pp. 439–449. arXiv: 1008.3851.

[21] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. "A System for Identifying Genetic Networks from Gene Expression Patterns Produced by Gene Disruptions and Overexpressions". *Genome Informatics* **9** (1998), pp. 151–160.

[22] S. Q. Zhang, M. Hayashida, T. Akutsu, W. K. Ching, and M. K. Ng. "Algorithms for finding small attractors in boolean networks". *Eurasip Journal on Bioinformatics and Systems Biology* **2007**:1 (2007), p. 20180.

[23] M. Purandare, R. Polig, and C. Hagleitner. "Accelerated analysis of Boolean gene regulatory networks". *2017 27th International Conference on Field Programmable Logic and Applications, FPL 2017*. 2017, pp. 1–6.

[24] C. Müssel, M. Hopfensitz, and H. A. Kestler. "BoolNet-an R package for generation, reconstruction and analysis of Boolean networks". *Bioinformatics* **26**:10 (2010), pp. 1378–1380. arXiv: /arxiv . org / abs / 1604 . 02208 [http:].

[25] R. Zhang, M. V. Shah, J. Yang *et al.* "Network model of survival signaling in large granular lymphocyte leukemia". *Proceedings of the National Academy of Sciences* **105**:42 (2008), pp. 16308–16313. arXiv: arXiv:1408.1149.

[26] A. Rodríguez, D. Sosa, L. Torres *et al.* "A Boolean network model of the FA/BRCA pathway". *Bioinformatics* **28**:6 (2012), pp. 858–866.

[27] I. Albert, J. Thakar, S. Li, R. Zhang, and R. Albert. "Boolean network simulations for life scientists". *Source Code for Biology and Medicine* **3**:1 (2008), p. 16.

[28] N. Miskov-Zivanov, A. Bresticker, D. Krishnaswamy *et al.* "Emulation of biological networks in reconfigurable hardware". *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. BCB '11. ACM, 2011, pp. 536–540.

[29] L. Salwinski and D. Eisenberg. "In silico simulation of biological network dynamics". *Nature Biotechnology* **22**:8 (2004), pp. 1017–1019.

[30] N. Berntenis and M. Ebeling. "Detection of attractors of large Boolean networks via exhaustive enumeration of appropriate subspaces of the state space". *BMC Bioinformatics* **14**:1 (2013), p. 361.

[31] A. Saadatpour, R. Albert, and T. C. Reluga. "A Reduction Method for Boolean Network Models Proven to Conserve Attractors". *SIAM Journal on Applied Dynamical Systems* **12**:4 (2013), pp. 1997–2011.

[32] F. Ay, F. Xu, and T. Kahveci. "Scalable Steady State Analysis of Boolean Biological Regulatory Networks". *PLOS ONE* **4**:12 (2009), pp. 1–9.

[33] D. Zheng, G. Yang, X. Li *et al.* "An Efficient Algorithm for Computing Attractors of Synchronous And Asynchronous Boolean Networks". *PLOS ONE* **8**:4 (2013), pp. 1–7.

[34] E. Dubrova and M. Teslenko. "A SAT-based algorithm for finding attractors in synchronous boolean networks".

IEEE/ACM Transactions on Computational Biology and Bioinformatics **8**:5 (2011), pp. 1393–1399. arXiv: 0901.4448.

[35] W. Guo, G. Yang, W. Wu, L. He, and M. Sun. "A parallel attractor finding algorithm based on boolean satisfiability for genetic regulatory networks". *PLoS ONE* **9**:4 (2014), pp. 1–10.

[36] G. Arellano, J. Argil, E. Azpeitia *et al.* ""Antelope": A hybrid-logic model checker for branching-time Boolean GRN analysis". *BMC Bioinformatics* **12**:1 (2011), p. 490.

[37] J. F. Keane, C. Bradley, and C. Ebeling. "A compiled accelerator for biological cell signaling simulations". *Proceeding of the 2004 ACMSIGDA 12th international symposium on Field programmable gate arrays FPGA 04*. FPGA '04 1. ACM, 2004, p. 233.

[38] N. Miskov-Zivanov, A. Bresticker, D. Krishnaswamy *et al.* "Regulatory network analysis acceleration with reconfigurable hardware". *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol. 2011. 2011, pp. 149–152.

[39] H. de Jong. "Modeling and simulation of genetic regulatory systems: a literature review". *Computational Biology* **9**:1 (2002), pp. 67–103.

[40] D. R. Thomas R. *Biological feedback*. CRC Press, 1990.

[41] S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1992, pp. 61–100. arXiv: 05218657199780521865715.

[42] B. B. Aldridge, J. Saez-Rodriguez, J. L. Muhlich, P. K. Sorger, and D. A. Lauffenburger. "Fuzzy Logic Analysis of Kinase Pathway Crosstalk in TNF/EGF/Insulin-Induced Signaling". *PLoS Computational Biology* **5**:4 (2009), pp. 1–13.

[43] A. Saadatpour, R. S. Wang, A. Liao *et al.* "Dynamical and structural analysis of a t cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia". *PLoS Computational Biology* **7**:11 (2011), e1002267.

[44] A. Garg, A. Di Cara, I. Xenarios, L. Mendoza, and G. De Micheli. "Synchronous versus asynchronous modeling of gene regulatory networks". *Bioinformatics* **24**:17 (2008), pp. 1917–1925. eprint: /oup/backfile/content_public/journal/bioinformatics/24/17/10.1093/bioinformatics/btn336/2/btn336.pdf.

[45] D. Zheng, G. Yang, X. Li *et al.* "An Efficient Algorithm for Computing Attractors of Synchronous And Asynchronous Boolean Networks". *PLoS ONE* **8**:4 (2013). Ed. by I. P. Androulakis, pp. 1–7.

[46] J. Weerasinghe, R. Polig, F. Abel, and C. Hagleitner. "Network-Attached FPGAs for Data Center Applications". *IEEE International Conference on Field-Programmable Technology (FPT '16)*. IEEE. 2016.

6 Inferring clonal composition from multiple tumor biopsies[1]

Matteo Manica[1,2,*], Hyunjae Ryan Kim[3,*], Philippe Chouvarine[3,*], Roland Mathis[1,*], Laura De Vargas Roditi[5], Bence Szalai[4], Ulrich Wagner[5], Kathrin Oehl[5], Karim Saba[5], Angshumoy Roy[3], Donald W. Parsons[3], Julio Saez-Rodriguez[4], Peter J. Wild[6], María Rodríguez Martínez[1], Pavel Sumazin[3]

[1] IBM Research Zürich

[2] ETH - Zürich

[3] Baylor College of Medicine

[4] RWTH Aachen University

[5] University Hospital Zurich

[6] University Hospital Frankfurt

[*] Shared first authorship

Abstract

Knowledge about tumor clonal evolution can help interpret the function of genetic alterations by pointing out initiating events and mutations that contribute to the selective advantage of proliferative, metastatic, and drug-resistant tumor subclones. Clonal evolution can be reconstructed from estimates of the relative abundance (frequency) of subclone-specific alterations in tumor biopsies, which, in turn, informs on the cellular composition of each tumor subclone. However, estimating these frequencies is complicated by the high genetic instability that characterizes many cancers. Models for genetic instability suggest that copy number alterations (CNAs) can dramatically alter mutation-frequency estimates and thus affect efforts

[1] In preparation, previous version pre-print available [1]. See Chapter 8 for details about contributions and copyright.

to reconstruct tumor phylogenies.

Our analysis suggests that a detailed accounting of CNAs is required for accurate mutation frequency estimates, and that such accounting is impossible for many cancer types using molecular profiling of one biopsy per tumor. Instead, we propose an optimization algorithm, Chimaera, to account for the effects of CNAs using profiles of multiple biopsies per tumor. Analyses of simulated data and profiles of a prostate cancer patient and an hepatocellular carcinoma cohort suggest that Chimaera estimates are consistently more accurate than previously-proposed methods, resulting in improved phylogeny reconstructions, and the discovery of recurrent initiating mutations and key tumorigenesis events.

6.1 Introduction

Pan-cancer tumor profiling has identified recurrent alterations that are associated with tumor etiology at the loci of thousands of genes but the interpretation of genetic alterations remains a major challenge [2–4]. Knowledge about the clonal evolution of tumors can point to genetic alterations that both contribute to tumorigenesis, indicate prognostically-relevant intra-tumoral variability, and point to refractory tumor subclones [5–7]. Specifically, clonal evolution, depicted as a phylogenetic tree in Figure 6.1a, can help identify alterations that play a role in tumor initiation as well as those that confer a selective advantage to altered tumor cells. Moreover, information about its subclonal composition is important for predicting the cancer's potential for drug resistance and metastasis, which vary across tumor subclones [8] and are the key determinants of patient survival. Consequently, tumor-subclone characterization is essential for designing personalized therapies that target all tumor subclones and may hold the key to predicting tumor progression, drug sensitivity, and patient outcome.

Current methods that rely on DNA-profiling to reconstruct clonal evolution of tumors can be classified into two categories: methods that primarily rely on single-cell profiles [9–12] and those that computationally resolve mixtures of subclones from molecular profiles of tumors, i.e., profiles of pools of cells that originate from a common malignant lesion [7, 13–15]. Single-cell DNA sequencing can produce more definitive estimates of the proportion of tumor cells that contain each genetic alteration (alteration frequencies) and more complete profiles of tumor subclones, including information about the co-occurrence of alterations within each subclone. Its primary disadvantage is operational given the low availability of high-quality tumor samples that permit single-cell isolation and profiling, and accuracy and cost associated with parallel sequencing DNA from a multitude of cells per tumor. Alternatively, single-cell

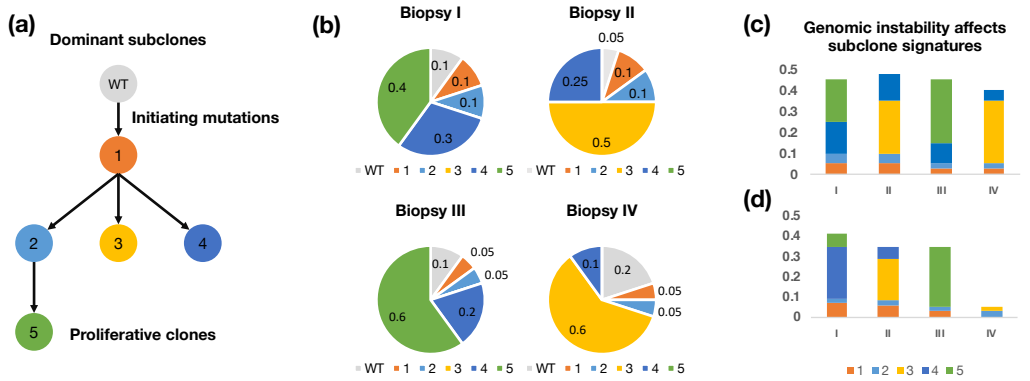


Figure 6.1: **Footprint of clonal evolution across tumor biopsies. (a)** Tumor phylogeny composed of five dominant tumor subclones and wildtype (WT) cells, with no somatic mutations, that make up the cellular composition of four tumor biopsies **(b)**. Subclones 3 and 5 were more proliferative, i.e., the proportion of these subclones (cellularity) in containing biopsies is greatest. **(c)** Failure to account for genetic instability can skew cellularity estimates because fractions of reads (mutated-read fractions) presenting each mutation in WES depend on the copy numbers of the alleles in both mutated and non-mutated cells. Consequently, in genetically-stable tumors, biopsies from **(b)** will have mutated-read fractions that differ from those of **(d)** genetically unstable tumors with the same cellularities.

RNA sequencing or protein profiling can be used to define tumor subclones, but these do not directly point to key driving genetic alterations. Moreover, the accuracy of single-cell mutation profiling is an issue due to limited material availability in single cells [16], and this is not likely to improve as future sequencing technologies focus on profiling formalin-fixed paraffin-embedded (FFPE) tumor samples [17, 18]. Focusing on single-nucleotide somatic variants (SNVs; or simply mutations), we seek to reconstruct clonal evolution from mutation profiles of genetically-unstable cancers. This entails deconvolving mutation frequencies, alteration-subclone associations, and copy number alterations (CNAs) from molecular profiles, including whole-exome sequencing (WES) assays, that produce average estimates across cellular ensembles (see Figure 6.1b). One approach to improve the accuracy of such deconvolutions is to profile multiple biopsies from the same tumor across time points [19] or across regions [8, 20]. This approach relies on the assumptions that genetic alterations that are specific to the same tumor subclone are expected to co-occur with the same frequency across biopsies and that the clonal composition across time or heterogeneous regions varies; i.e., multiple sampling will allow for the aggregation and deconvolution of the frequencies of most mutations with improved power. It's important to note that mutations that underwent convergent evolution [21] will not be aggregated with other mutations form the same tumor subclone because of differing frequency estimates across biopsies. A central challenge for aggregating and esti-

mating mutation frequencies in tumors with unstable genomes is accounting for the influence of CNAs in mutated-read fractions, i.e., the frequencies of observed alternative alleles in the profiling assay. Specifically, CNAs can alter contributions from reference alleles in mutation-free cells, as well as both alternative and reference alleles in mutated cells (see Figure 6.1c). In turn, errors in mutation-frequency estimates can prevent accurate phylogeny reconstructions (see Figure 6.2).

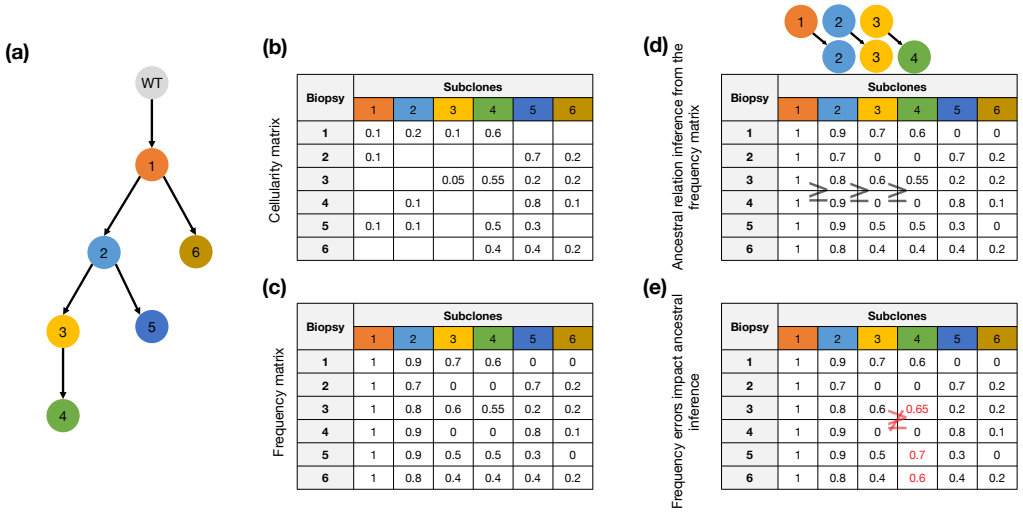


Figure 6.2: **Impact of mutation frequencies on the inference of the ancestral relations.** Small variations in mutation frequency estimates can impact the inference of ancestral relations. **(a)** Simulated tumor phylogeny, **(b)** subclone cellularities, and **(c)** frequencies of subclonal mutations across biopsies. **(d)** Ancestral relations between subclones can be inferred from comparisons of their frequency vectors: Subclone 4 frequencies are greater than those of subclone 3 across all biopsies, but **(e)** errors in frequency estimates (red) can violate this relationship and complicate tumor-phylogeny reconstruction efforts.

Here, we introduce a model for the effects of CNAs on mutated-read fractions in WES. We use this model as a basis for simulations with CNA distributions that are compatible with observations from The Cancer Genome Atlas (TCGA) primary breast cancer (BRCA) and hepatocellular carcinoma (HCC) samples [22, 23]. Data were simulated using synthetically generated phylogenies followed by the duplication or loss of sequencing reads according to simulated effects of copy number variations (CNVs).

Several methods are available in the literature to estimate mutation frequencies and clonal compositions. ABSOLUTE [24] infers tumor purity and malignant cell ploidy directly from the analysis of somatic DNA alterations, by fitting estimates of copy-ratio of both homologous chromosomes with a Gaussian mixture model, with components centred at the discrete

concentration-ratios implied by an initial frequentist estimation. AncesTree [14] provides a combinatorial characterization of the clonal evolution of a tumor by assuming than in an error-free data mutations can be described by a perfect phylogeny matrix, which is found using integer linear programming; the problem is extended to real data using a probabilistic model for errors. EXPANDS [25] clusters mutations based on their cell-frequency probability distributions; clusters are next extended by members with similar distributions, and pruned based on statistical confidence by comparing the cluster maxima and peaks observed outside the core region. PhyloWGS [26] reconstructs phylogenies based on a model for simple somatic mutations in addition to a correction for CNAs, all based on a single biopsy per tumor. SCHISM [15] takes as input mutation cellularity estimations and mutation clustering inferred by other methods, and uses a generalized likelihood ratio to infer lineage precedence and lineage divergence. A genetic algorithms is then used to build phylogenetic trees. Attempts to estimate the frequencies and cellularities of mutations and subclones using ABSOLUTE, AncesTree, EXPANDS, PhyloWGS, and SCHISM revealed variable success rates, with some methods showing consistently poor accuracy. EXPANDS and PhyloWGS, which were designed for phylogeny reconstruction using profiles of one biopsy per tumor, and ABSOLUTE, which is best known and most effective for estimating tumor purity, had consistently poor accuracy in our simulations. While SCHISM and AncesTree, which do not explicitly account for the full range of observed CNAs in tumors, were less accurate on simulations with CNAs. Like PhyloWGS, we concluded that explicit accounting for CNAs is required in order to approximate mutation frequencies accurately. However, more than one biopsy per tumor are required for accurately approximating mutation frequencies and CNAs at mutated loci.

To address this challenge and improve mutation-frequency and CNA estimations from WES of tumors with genetic instability, we developed Chimaera: clonality inference from mutations across biopsies. In our methodology we define the clonality problem as that of associating mutations with subclones and inferring ancestral relations between subclones. The goal of the resulting set-theoretic formulation, for each tumor, is to aggregate co-occurring mutations across biopsies, estimate the frequency of each aggregate in every biopsy, and identify partial orders across aggregates that are consistent across biopsies. When viewed in this way, each tumor subclone can be associated with a frequency vector that describes the proportion of cells containing its mutations in each biopsy. Establishing ancestral order between two subclones then depends on (probabilistic) comparisons between their corresponding mutation frequencies. Hence, Chimaera relies on multiple biopsies for the same tumor to, first, approximate CNAs and mutation frequencies; then, identify mutations with similar approximate frequencies and associate them with subclones; and, finally, to estimate the true frequencies of

these mutations and the associated subclones. As is the case for estimates made by SCHISM, ABSOLUTE and other methods, Chimaera is not able to produce frequency estimates for all mutations, but compared to existing methods is able to process and determine true frequencies for more variants, exhibiting more power in identifying potentially tumor initiating mutations and disease drivers. Finally, to demonstrate that Chimaera is able to reconstruct subclones from tumor profiles we produced Chimaera-inferred subclones and resulting phylogeny from profiles of ten biopsies taken from a castration-resistant prostate cancer (CRPC) tumor and a set of profiles extracted from five different tumor areas from a cohort of hepatocellular carcinoma patients [27].

6.2 Results

We describe our efforts to evaluate method accuracy on simulated data and to reconstruct phylogenies from real tumors: ten biopsies from a CRPC patient and multi-area biopsies from an HCC cohort.

6.2.1 Accuracy of mutation-frequency estimation based on simulated data

We compared the accuracy of EXPANDS, ABSOLUTE, SCHISM, AncesTree, and Chimaera on simulated data, as described in Section 6.4. Phylogeny reconstruction success and clonality-inference accuracy by EXPANDS and ABSOLUTE were the lowest. EXPANDS relies on single biopsies, and when evaluated on phylogenies that were composed of as few as three tumor subclones, EXPANDS-reconstructed phylogenies from profiles of same-tumor biopsies (both simulated and collected from the clinic including the CRPC reported on here) had few common ancestral inferences and performance was poor in every tested simulated instance. In contrast, SCHISM-reconstructed phylogenies from synthetic constructions with three tumor subclones were accurate in 100% of the tested instances. ABSOLUTE can process profiles of multiple biopsies per tumor and has good accuracy for inferring tumor purity in our synthetic data. However, when using default parameters, errors in ABSOLUTE frequency-inferences were more than double those of SCHISM. Parameter optimization through human intervention consistently improved its accuracy, but it remained less accurate than SCHISM. Moreover, the degree of human intervention that this required was not compatible with large-scale benchmarking. Consequently, we focused on accuracy comparisons between inferences by

SCHISM, AncesTree, and Chimaera (as given in Figure 6.3), and excluded EXPANDS and ABSOLUTE from further analyses. AncesTree accepts no external input when estimating mutation frequencies, but SCHISM and can be guided by externally-inferred mutation clusters. SCHISM's implementation includes its own selected clustering methods, and these were also used to compare accuracy. We clustered mutations with tclust [28] based on five iterative optimization methods to determine the appropriate number of subclones: an elbow method based on intra-cluster sum of square (ElbowSSE), a method based on intra-cluster entropy (Entropy), a method based on the standard deviation index (SDIndex), and two approaches based on gaussian mixture models (GMD, Mclust) [29–32].

We compared the accuracy of methods and pipelines on 2000 simulated assays, including both simulated assays with and without modelled genetic instability (varying mutation copy numbers). The accuracy of SCHISM estimates was better on average than that of AncesTree, but it was relatively sensitive to clustering optimization methods, with SDIndex outperforming other methods, including those included in SCHISM's implementation. Comparatively, Chimaera estimates were less dependent on clustering methods and significantly outperformed estimates by SCHISM with SDIndex ($p < 1e-16$ by U test). We note that many mutations were eliminated from the evaluation by both the SCHISM and Chimaera pipelines with tclust-based clustering algorithms. In total, only $\sim 60\%$ of mutations were assigned frequencies by both methods. For this reason we also run Chimaera using hdbscan [33] as a clustering algorithm, since it showed a less strict criterion for outlier detection. While this combination exhibited slightly lower performance compared to other Chimaera runs (see Figure 6.3a), it showed an increasing power in the percentage of mutations used to estimate the frequencies (see Figure 6.3b), thus making it more appealing compared to other settings. Combining SCHISM with hdbscan was not considered a valid option given the poor error estimates observed using stricter clustering algorithms (see Figure 6.3a). Inference accuracy, for both SCHISM and Chimaera, was anti-correlated with the level of genetic instability, which followed truncated normal distributions with varying means and variances (see Figure 6.3c, see Section 6.4 for data generation). To better understand mutation-level behaviour, as opposed to the genome-level comparisons in Figure 6.3c, we rescued individual mutations from each simulation and compared accuracy, mutation by mutation, as a function of their simulated copy numbers (see Figure 6.3d).

The result suggests similar Chimaera accuracy across copy numbers, testifying the efficacy of accounting for copy number alterations. While Chimaera assigned frequencies to all clustered mutations, SCHISM did not successfully estimate mutation frequencies for some sim-

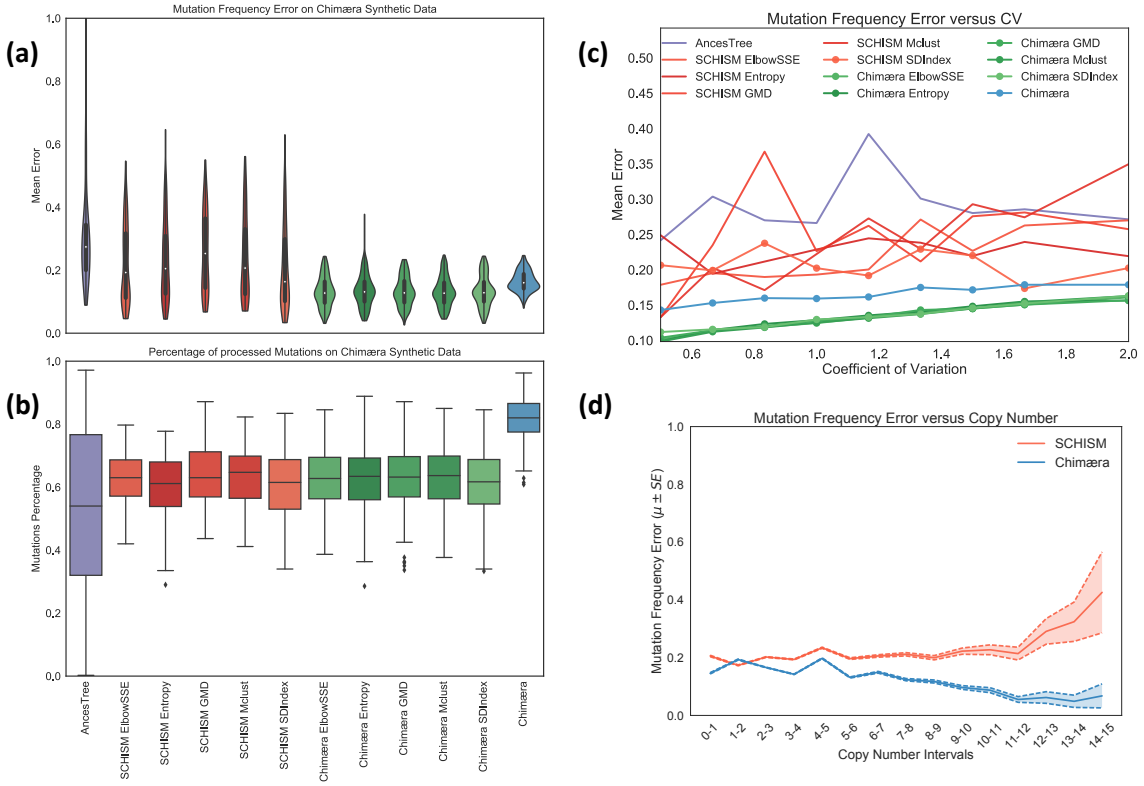


Figure 6.3: **Accuracy on simulated data.** **(a)** Accuracy of mutation-frequency estimates by AncesTree (purple), SCHISM (red) and Chimaera (green and blue) on simulated WES data from genomes with varying mutation copy numbers; SCHISM and Chimaera were evaluated using multiple clustering methods with SDIndex (SCHISM) and ElbowSSE (Chimaera) producing top accuracy, respectively, in blue are reported estimates for Chimaera using hdbscan. **(c)** Percentage of mutations processed applying the three different algorithms. It is evident how Chimaera using hdbscan outperforms clearly other methods, being able of considering over 80% of the mutations considered. **(b)** Accuracy was inversely correlated with genetic instability, which was measured here as the coefficient of variation of the distributions used to simulate CNVs in each simulated WES profile; SCHISM with SDIndex clustering outperformed AncesTree inferences. **(d)** Evaluated independently, mutation copy numbers had relatively little effect on Chimaera accuracy. We report results for Chimera using hdbscan and SCHISM with SDIndex (a representative that resembles results with other clustering methods). Standard errors are reported. Mean Error is the mean of the L^1 distances between true and estimated mutation frequencies after normalizing for the number of biopsies.

ulated genomes. Accuracy comparisons in Figure 6.3 were made using only those mutations that had assigned frequencies by all methods. In its totality, our analysis suggests that, at least under our model, mutation frequency estimation is more challenging for genomes with high copy-number variability. Chimaera shows high inference accuracy for simulated genomes where all mutations had consistently low or consistently high copy numbers was relatively

high. This is in part due to Chimaera's iterative process, where success in mutation clustering is followed by an optimization process that can correct for consistently high or consistently low mutation copy numbers.

6.2.2 Phylogeny inference in CRPC

To test our ability to infer mutation frequencies and ancestral relations between subclones using clinical data, we profiled ten biopsies of a single castrate-resistant prostate cancer (CRPC) tumor (see Figure 6.4). CRPCs are high-risk prostate tumors that are known to have high genomic instability [34]. Each of these biopsies was profiled and analyzed as described in Section 6.4, producing a total of 356 mutations that were used as input to SCHISM, ABSOLUTE and Chimaera. SCHISM did not produce frequency estimates for any of the mutations. ABSOLUTE, following repeated parameter optimization steps, produced frequency estimates for 21 mutations, resulting in four predicted subclones; mutations in three of these subclones had high frequencies in at least one biopsy, but ancestral relations between these subclones could not be inferred.

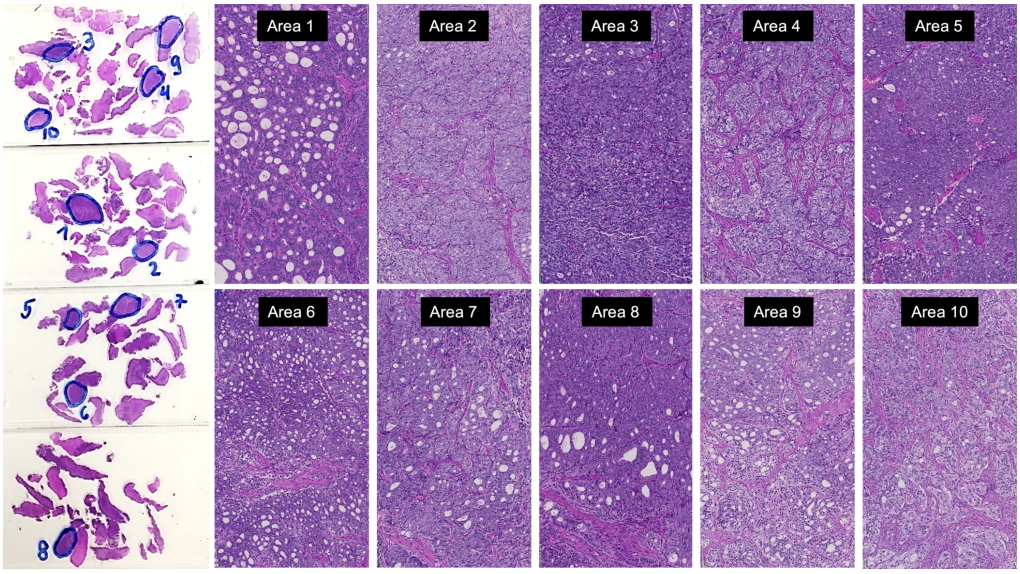


Figure 6.4: **Profiled CRPC regions.** Overview of four hematoxylin-eosin stained histology slides with 10 profiled areas (left); and zoomed-in versions (right) that portray the histological heterogeneity of this tumor. These CRPC regions were profiled by deep WES.

Chimaera combined with hdbscan was able to infer frequencies for 275 mutations that were clustered into four subclones. Using the frequency estimates provided by Chimaera we were

able to build a consistent phylogenetic tree supported by most of the biopsies (8/10), see Figure 6.5. Chimaera inferred potentially initiating mutations that targeted 29 genes, and identified a chain of three subclones that followed in a later stage in tumor evolution. Interestingly, the analysis identified genes that were targeted by multiple mutations, with these mutations inferred from multiple subclones at different stages of the evolution (TBC1D22A and TMEM131).

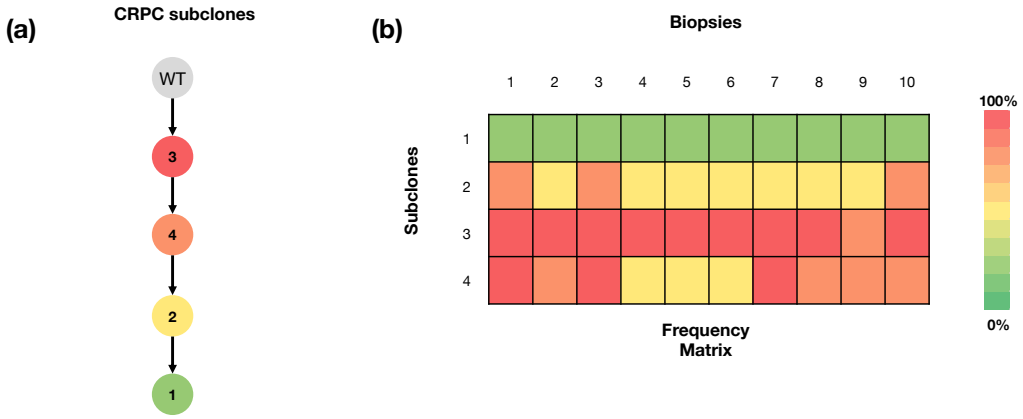


Figure 6.5: **Reconstructed CRPC phylogeny.** **(a)** Chimaera inferred four CRPC subclones that implied a chain phylogeny, clones are colored after the average frequency in the inferred mutations from red (high) to green (low). **(b)** Schematic representation of mutation frequencies across biopsies in each subclone.

6.2.3 Phylogeny inference in HCC

HCCs are high-risk liver tumors that are known to have high genetic instability [23]. To further validate Chimaera, we studied the profiles of nine hepatitis B virus-positive (HBV-positive) HCC patients, with each tumor profiled in five areas [27]. In total, we obtained mutated-read fractions and CNV estimates for $1,424$ mutation candidates in nine tumors and 43 tumor samples, while seven tumors were profiled in five areas each, profiles from only four areas of tumors HCC5647 and HCC8716 passed quality control (see Section 6.4 part describing CRPC profiling for details about the quality control).

Chimaera inferred frequencies estimates for 60% (858/1424) of all mutations, reconstructing phylogenetic trees for each tumor sample and predicting initiating clones and clones that are associated with a proliferative advantage; see representative trees for three patients in Figure 6.6.

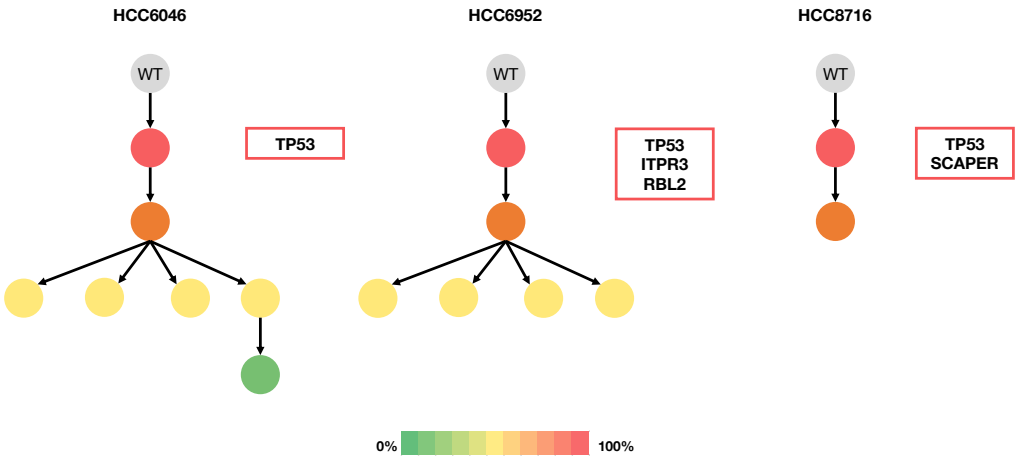


Figure 6.6: **Representative phylogenetic trees for three HCC patients.** A selection of representative trees reconstructed using Chimaera clonality inference. HCC6046 and HCC6952 exihibit a similar structure where, after a chain of two clones a branching event takes place. In HCC9716 Chimaaera identified two clones organized in chain. In the red boxes we report mutated genes, included in WNT-signaling pathway, assigned to each patient's root clone.

Interestingly, 78% (7/9) of the tumors included predicted initiating mutations in WNT-signaling pathway genes ($p < 0.05$ after FDR correction). An examination of 102 TCGA-profiled HBV-positive HCCs suggested that 74% (75/102) of samples carried mutations in WNT-signaling pathway genes, and the majority of these samples (76%) had WNT-signaling pathway mutations with mutated-read fractions above 25%, thus corresponding to mutations that are potentially present in the majority of cells. To test whether WNT-signaling pathway genes were enriched for mutations, and particularly mutations with mutated-read fractions above 25%, we calculated the proportion of tumors with such mutations in each of 186 KEGG [35, 36] pathways in MSigDB [37] and performed a permutation testing (see Section 6.4 for details). The top ten pathways by p-value and mutated-sample fraction is given in Table 6.1, and highlight a significant enrichment of WNT-signaling supporting the hypothesis of the involvement of this pathway in HBV-positive HCC initiation.

6.3 Discussion

We sought to develop a methodology to improve the accuracy of tumor phylogeny reconstruction from tumor WES data by improving mutation-frequency estimates when multiple profiles of the same tumor are available. Mutation-frequency estimates are particularly challenging in the face of high genetic instability, which is characteristic to many tumor types, includ-

Enriched KEGG pathways	Genes	Patients with mutations	p-value	Excluding WNT-signaling genes	
KEGG WNT SIGNALING PATHWAY	151	57%	0.001	**Frequency**	**p-value**
KEGG PROSTATE CANCER	89	54%	0.001	23%	>0.1
KEGG COLORECTAL CANCER	62	52%	0.001	12%	>0.1
KEGG ENDOMETRIAL CANCER	52	51%	0.001	17%	>0.1
KEGG BASAL CELL CARCINOMA	55	50%	0.001	9%	>0.1
KEGG CALCIUM SIGNALING PATHWAY	178	60%	0.002	56%	0.002
KEGG ECM RECEPTOR INTERACTION	84	44%	0.003	44%	0.004
KEGG PATHWAYS IN CANCER	328	68%	0.016	51%	>0.1
KEGG MAPK SIGNALING PATHWAY	267	64%	0.021	58%	0.020
KEGG FOCAL ADHESION	201	58%	0.023	52%	0.042

Table 6.1: **Pathway enrichment analysis of the potential regulators.** The ten most enriched pathways for mutations with mutated-read fractions greater than 25% (high-frequency mutations) in TCGA-profiled virus-positive HCCs. Pathways were sorted by p-values followed by the proportion of patients with a high-frequency mutations in at least one pathway gene. p-values were estimated using permutation testing based on all expressed genes in 186 KEGG pathways; here, for each pathway and given the number of pathway gene, each permutation test selected that number of genes uniformly at random and calculated the fraction of patients with a mutation in one of these genes. The same test was conducted after excluding WNT-signaling genes to establish independence from WNT-pathway signaling.

ing the high-risk prostate cancer tumor whose profiling was reported on here. Our proposed method, Chimaera, is suitable for analyzing tumor profiles across multiple time points and across multiple tumor regions and can be easily modified to process both types of data, but here we focused the discussion on the latter. In addition, while we focused on single nucleotide variants, our methods can be extended to consider other types of genetic alterations.

We outlined the challenges involved in estimating mutation frequencies from WES of genomes with high genetic instability—where the copy numbers of mutations can widely vary. We showed that, even for EXPANDS, which uses the copy number of mutations to infer clonal composition, the accuracy of mutation frequency estimates and cellularities using single biopsies was very poor. Our own investigation suggests that the task is often impossible on simulated assays with varying mutation copy numbers. Consequently, we elected to rely on multiple biopsies per tumor to improve mutation-frequency estimation. We showed that even when profiles of multiple biopsies are available, methods that do not explicitly account for the full range of copy number variability produce inconsistent results with often poor accuracy.

Chimaera is able to improve mutation-frequency estimates by harnessing added information from multiple profiles and by directly accounting for the influence of CNVs on observations from WES. In synthetic data, Chimaera's performance was the most consistent, and the greatest both in terms of accuracy and percentage of mutations assigned to subclones. Interestingly, while Chimaera was able to estimate mutation frequencies with relatively high accuracy even

for significantly high and low copy numbers, its performance declined for unstable genome with a high degree of copy number variability. By profiling and analyzing a CRPC tumor, Chimaera demonstrated the ability to estimate clonal composition of a real unstable cancer, permitting the inference of a phylogenetic tree that describes the evolution of the disease. We also applied Chimaera on a recently published HCC cohort, where the reconstructed phylogeny estimates allowed the discovery of recurrent initiating mutations. Specifically, we observed a significant enrichment of the WNT-signaling pathway in the mutations assigned to root subclones. This finding is in agreement with the known role of WNT in cellular proliferation and tumor initiation [38] and it has been corroborated by analyzing HCC profiles from TCGA.

In conclusion, Chimaera provides an effective way to handle clonal composition inference in unstable cancers, thus enabling reconstruction of phylogenetic trees that elucidate disease progression in a patient-specific fashion. To facilitate its application to the community, Chimaera has been made available as an open-access web service.

6.4 Methods

In the following we report all the methodologies adopted for the analysis presented in Section 6.2.

6.4.1 Clonality reconstruction problem

Let $M = \{m : m \in \mathbb{N}, 1 \leq m \leq n\}$ denote the set of n mutations identified across a set of profiled biopsies S. The mutation burden in any given cell is given as a subset of M, $\gamma \subseteq M$, or as an element of the power set over M, $P(M)$; i.e., $\gamma \in P(M)$ is a specific mutation ensemble that characterizes a tumor subclone. We denote the cellularity of γ and its corresponding subclone in biopsy $s \in S$ as ρ^s_γ, and the frequency of a mutation $m \in \gamma$ in biopsy s as $\phi^s_m = \sum_{\{\gamma : \gamma \in P(M), m \in \gamma\}} \rho^s_\gamma$. Consequently, $\sum_{\gamma \in P(M)} \rho^s_\gamma = 1$ and the assignment $A = \{\rho^s_\gamma : \gamma \in P(M),\ s \in S\}$ produces a solution to the formulated clonality reconstruction problem.

6.4.2 Relation between copy number and mutation frequencies

As defined above, for a mutation m in biopsy $s \in S$, ϕ_m^s denotes the frequency of cells in s with mutation m. The total copy number C^s of the allele targeted by the mutation can be estimated from sequencing data. C^s is composed by: the copy numbers of the allele in cells that lack mutation m, δ^s, the copy number of the wildtype allele in m-mutated cells, δ_w^s and the copy number of the mutated allele in m-mutated cells, δ_m^s (see Figure 6.7).

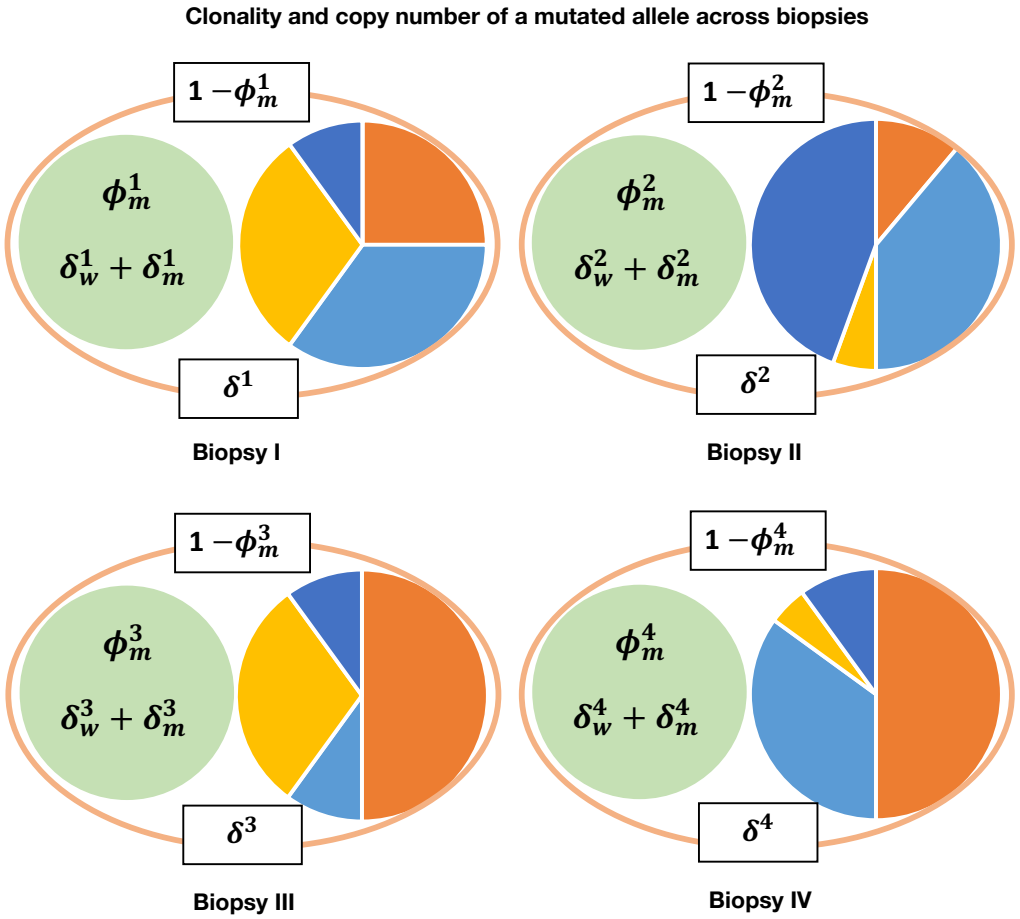


Figure 6.7: **Mutation-centric model for CNV effects.** Our mutation-centric model for the effects of CNVs on mutated-read fractions in WES. In each biopsy s, the mutated-read fraction is a function of the true mutation frequency ϕ_m^s, the copy number of the allele in all profiled cells—tumor and WT—that lack this mutation, δ^s, and the copy number of the wildtype and the mutated allele in tumor cells with the mutation, δ_w^s, δ_m^s.

Notice that if no copy number event has occurred at the locus m : $\delta^s = 2$, $\delta_w^s = 1$ and $\delta_m^s = 1$. Adopting the infinite-sites assumption, we denote the mutated-read fraction, the fraction of reads reflecting the mutated versus wildtype allele, in sample s as f_m^s. Then, we

can formulate the following system of equations:

$$C^s = \delta^s(1 - \phi_m^s) + (\delta_w^s + \delta_m^s)\phi_m^s \tag{6.1}$$

$$f_m^s = \frac{\phi_m^s \delta_m^s}{C^s} \tag{6.2}$$

where Equation 6.1 provides a weighted sum of the copy number contribution from each allele type, and Equation 6.2 gives the ratio of the number of reads coming from the mutated allele to the total number of reads.

6.4.3 Chimaera

Chimaera proceeds in three steps. First, mutation frequencies are approximated from sequencing and CNV data in each biopsy; then, mutations with similar approximated frequency vectors (where each vector component gives the mutation frequency in each biopsy) are clustered together to form subclones; and finally, mutation frequencies and CNVs for these alleles are refined using an optimization process. The optimization assumes that all clustered mutations that are associated with the same subclone have the same frequency in each tumor biopsy and that δ_m^s, the average copy number of the m-mutated allele, is the same across all biopsies from the same tumor.

A first approximation

We first approximate the true frequency of the mutation ϕ_m^s by accounting for tumor purity, i.e., the fraction of tumor cells in the biopsy, and assuming that the allele's average copy number in tumor cells, whether mutated or not, is fixed. Let p^s be the purity of biopsy s, then Equation 6.2 can be rewritten as follows:

$$f_m^s = \frac{\phi_m^s \delta_m^s p^s}{2(1 - p^s) + C^s p^s} \tag{6.3}$$

The experimentally observed copy number, C_{obs}^s, depends on the purity of the sample and the copy number of the sample tumor cells, C^s, as follows:

$$C_{obs}^s = 2(1 - p^s) + C^s p^s \tag{6.4}$$

where C^s_{obs} can be estimated using additional biochemical assays, genetic sequencing, or through computational analysis of WES data [39], and the normal cells are assumed to have been corrected for germline copy number variants associated biases. We assume that the mutated allele's average copy number in tumor cells in each biopsy is constant, i.e., $\delta^s_m = \frac{C^s}{2}$. Under this approximation, we can use Equation 6.3 and Equation 6.4 to eliminate C^s and obtain a first approximation for the mutation frequency:

$$\widetilde{\phi}^s_m = \min\left(\frac{2f^s_m C^s_{obs}}{C^s_{obs} - 2(1-p^s)}, 1\right) \tag{6.5}$$

This constraint will be later removed in the optimization process that follows, but is necessary at this stage to obtain an initial approximation of the mutation frequencies that takes into account copy number variations.

Subclone reconstruction

The approximate mutation frequency vectors, Equation 6.5, are next clustered to identify candidate groups of mutations that form subclones. We considered clustering algorithms with robust treatment of the outliers in order to ensure a good clustering stability and quality. Specifically, we used hdbscan [33], a density-based hierarchical clustering method that aims at maximizing the stability of the obtained clustered against noise and requires minimal parameter selection. The number of clusters is determined automatically based on the minimal number of mutations that has to be considered to constitute a cluster. We also use tclust [28], a non-hierarchical robust clustering that trims outliers based on a probabilistic model. The number of clusters is selected by optimizing intra-cluster entropy or the sum of square errors (SSE), and using a variety of optimization methods including the Elbow method, SD index and gaussian mixture models-based approaches [29–32]. The clustering based on hdbscan, using a distance based on the L^1-norm, exhibits better performance on the generated synthetic data compared to others, especially when considering the number of mutations processed. Furthermore, it has the advantage of avoiding imposing a prior distribution on the mutation frequencies. Once the clusters are found, Chimaera assumes that each cluster represents a subclone and uses the mutation assignments to infer subclone frequencies and copy number estimates for each mutated allele in the final optimization step.

Frequency and copy number inference

Focusing on subclone $\gamma \in P(M)$, Equation 6.3 describes a relationship between the frequencies and copy numbers of mutations in γ:

$$\phi_m^s \delta_m^s = f_m^s \frac{C_{obs}^s}{p^s} \equiv \mathcal{B}_{sm}, \ \forall m \in \gamma, \forall s \in S. \tag{6.6}$$

where, $\mathcal{B}_{sm}$ is the entry of a matrix $\mathcal{B} \in \mathbb{R}^{|S|,|\gamma|}$ corresponding to mutation m and biopsy s. $\mathcal{B}$ is fully determined from analysis of sequencing assays, including purity, observed copy numbers, and observed mutated-read fractions for each mutation. Unfortunately, the central term of Equation 6.6, a multiplication of frequencies and copy numbers, cannot be analytically decoupled. However, mutations from the same subclone occur in cells with shared evolutionary history, and thus are expected to show similar mutation frequencies, i.e., $\phi_{m_i}^s = \phi_{m_j}^s \equiv \phi^s \, \forall m_i, m_j \in \gamma$. Notice that the same mutations may have different frequencies in a different biopsy, as the subclones identified in different biopsies are not constraint to descend from the same ancestral parent. Further, we assume that the copy number of each mutation m is constant across biopsies, i.e., $\delta_m^{s_i} = \delta_m^{s_j} \equiv \delta_m \in [0, CN] \ \forall s_i, s_j \in S$, where CN is a fixed upper bound for the copy number; $CN = 15$ in our simulations and WES analysis. While we expect that this assumption will introduce some errors to the approximation of δ_m^s, it will have limited effects on the selection of optimal mutation frequencies because the variability of copy number averages for the mutated allele across biopsy is expected to be low. We also note that we have not assumed stable genomes in our simulated data, i.e, the generated data displays variable copy numbers for the same mutated allele across biopsies, in order to have an accurate estimate of the committed error. Making use of these assumptions, the optimization problem for each subclone $\gamma \in P(M)$, based on Equation 6.6, can be formulated as:

$$\begin{aligned} \text{minimize} \quad & \|\vec{\phi^s} \otimes \vec{\delta_m} - \mathcal{B}\|_2 \\ \text{subject to} \quad & 0 \leq \delta_m \leq CN, \, \forall m \in \gamma \\ & 0 \leq \phi^s \leq 1, \, \forall s \in S. \end{aligned} \tag{6.7}$$

where $\vec{\phi^s}$ is the mutation frequency vector across biopsies for all mutations in γ; $\vec{\delta_m}$ is the copy-number vector for each mutation in γ; $\mathcal{B}$ as defined in Equation 6.6; and $\vec{\phi^s} \otimes \vec{\delta_m}$ denotes the outer product of vectors $\vec{\phi^s} \in \mathbb{R}^{|S|}$ and $\vec{\delta_m} \in \mathbb{R}^{|\gamma|}$. We used sequential least squares programming (SLSQP) optimization [40] to find an optimal solutions of Equation 6.7, where multiple runs with random initialization are used to avoid being trapped in local optima.

6.4.4 Simulation of WES data

WES simulations are based on phylogenies and associated cellularity matrices that describe ancestral relations between 6 to 12 subclones (see Figures 6.8a and 6.8b).

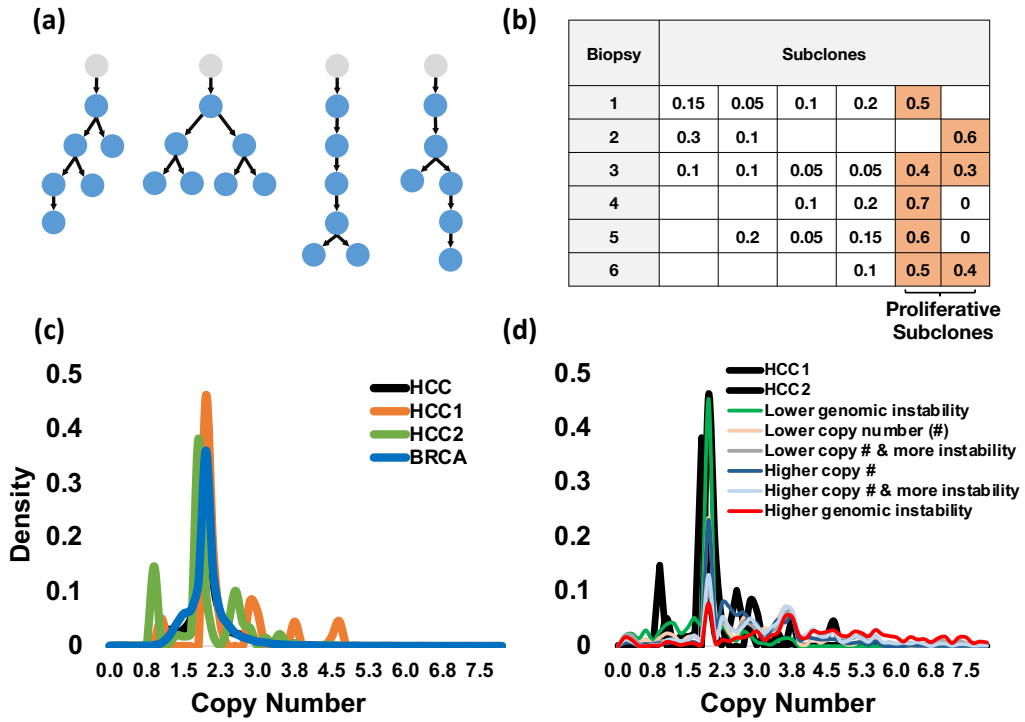


Figure 6.8: **Synthetic data generation.** Our synthetic data generation and a comparison of simulated CNV distributions to those that were observed in tumors. **(a)** Representative phylogenies and **(b)** a representative cellularity matrix. **(c)** Density plots of average copy numbers across profiles of TCGA hepatocellular carcinoma (HCC) and breast (BRCA) tumors. HCC1 and HCC2 show genome-wide CNV distributions in each of two HCC tumors, while HCC and BRCA distributions are taken across genes and tumors. **(d)** Simulated CNVs ranged from 0 to 15x.

Each subclone is associated with 20 to 50 somatic mutations, and each somatic mutation is associated with a trio of copy numbers—δ^s, δ^s_w, and δ^s_m—that are sampled from truncated normal distributions with means $\mu \in \{1, 2, 3\}$, where $\mu = 1$ corresponds to no copy number changes in tumor cells, and standard deviation $\sigma \in \{0, 1, 2, 3\}$; $\sigma = 0$ is used only when $\mu = 1$. The resulting copy numbers model a range of genetic instability conditions that is in line with observed copy number changes in hepatocellular carcinoma (HCC) and breast cancer (BRCA) tumors from TCGA (see Figures 6.8c and 6.8d). We assume no linkage between simulated CNVs of any mutation. In addition, we add up to 10% of wildtype reads for all simulated mutations to account for the potential inclusion of non-tumor cells in the assay (WT

subclone in Figure 6.1a). Total coverage for each allele, i.e., the number of reads covering both wild-type and the mutated genetic position, is taken by sampling mutation coverage values from real tumor biopsies profiled in a CRPC patient. Finally, once simulated reads are produced for both mutated and wild-type alleles, noise is added to simulate duplication or loss of up to 5% of the observations according to a uniform distribution. Each simulation has been repeated to produce six biopsies per tumor using a distinct cellularity vector for each biopsy (as depicted in Figures 6.8a and 6.8b). The availability of six biopsies per tumor increases the likelihood that mutations can be aggregated and subclone mutation frequencies can be compared to infer ancestral relations. We note that while our CRPC tumor has been profiled at ten regions, setting a six-biopsy minimum will exclude the profiling of many tumor types using our method; this represents a compromise between clinical feasibility and power to infer mutation frequencies and phylogenies.

6.4.5 Profiling and analysis of ten CRPC biopsies

To test our ability to infer mutation frequencies and ancestral relations between subclones based on clinical profiles, we profiled ten castration resistant prostate cancer (CRPC) tumor biopsies (see Figure 6.4). The specimen was collected at the Department of Pathology and Molecular Pathology, University Hospital Zürich, Switzerland as previously described [41] with the approval of Cantonal scientific ethics committee Zürich, approval number KEK-ZH-No. 2014-0007, and with informed consent by the patient. Tumor regions were selected for heterogeneous histological presentation by an experienced uropathologist. DNA from peripheral blood and formalin-fixed paraffin-embedded (FFPE) punches (ten cylinders with diameter of $0.6\,\mathrm{mm}$) was isolated with the Maxwell 16 LEV Blood DNA kit (Promega, AS1290) and Maxwell 16 FFPE Tissue LEV DNA Purification Kit (Promega AS1130), respectively, according to manufacturer's recommendations; $300\,\mu\mathrm{L}$ of blood collected in a BD Vacutainer K2 (EDTA $18.0\,\mathrm{mg}$) tube was added to $30\,\mu\mathrm{L}$ of Proteinase K solution (final concentration $2\,\mathrm{mg\,mL^{-1}}$) and subsequently mixed with $300\,\mu\mathrm{L}$ lysis buffer, vortexed and incubated for 20 minutes at $56\,^{\circ}\mathrm{C}$. FFPE cylinders were deparaffinised with xylene, washed twice with ethanol, dried 10 minutes at $37\,^{\circ}\mathrm{C}$ and re-suspended in $200\,\mu\mathrm{L}$ incubation buffer containing $2\,\mathrm{mg\,mL^{-1}}$ Proteinase K. Samples were incubated overnight at $70\,^{\circ}\mathrm{C}$ and mixed with $400\,\mu\mathrm{L}$ lysis buffer. Lysates from both, blood and FFPE tissues, were transferred to well 1 of the supplied cartridge of the corresponding kit and DNA was automatically purified and eluted in $30\,\mu\mathrm{L}$ Tris-buffer, pH 8.0 by the Maxwell instrument. Each biopsy was profiled using Agilent SureSelect Whole Exome Enrichment, v6 (58 Mbp) and 2x75 bp paired-end reads were

used for optimal performance on a HiSeq 4000 (Illumina). Mutation calling was followed by protocols established by TCGA and ExAC [42, 43]. Reads were aligned to hg19 using BWA 39, and variants were called with GenomeAnalysisTK, MuTect 40, Picard MarkDuplicates, and additional post-processing utilities from GATK including BaseRecalibrator. FastQ files were deposited in EBI's ENA project PRJEB19193. Predicted mutations were annotated with estimated read fractions and estimated CNVs by VarScan using default parameters and after setting the maximum amplification to 15x [39]. Mutations that were present in fewer than three biopsies or supported by fewer than three reads were discarded. A total of 356 mutations were used as input for inference methods.

6.4.6 Enrichment analysis of WNT-signaling in HCC

Pathway enrichment analysis for mutated genes present in initiating subclones estimated with Chimaera was performed using Enrichr [44] and Panther [45]. In order to test WNT-signaling pathway genes enrichment for high frequency mutations from HBV-positive HCC patients included in TCGA we consider KEGG [35, 36] pathways reported in MSigDB [37]. A permutation testing was used to estimate p-values, where for each pathway, random same-size gene sets were generated using KEGG pathway genes, and the mutated-sample fraction taken to generate a null distribution. WNT-signaling was our top pathway for enrichment of mutations with mutated-read fractions above 25% or for any mutated-read fraction. To correct for the shadow effect [46], where pathways that overlap a pathway that is mutated in many samples are also significant, we recalculated enrichment significance for each pathway after excluding WNT-signaling pathway genes, and note that MAPK-signaling and two other pathways were still enriched (see Table 6.1).

Declarations

Acknowledgments

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 668858.

Author contributions

See Chapter 8 for details about contributions and copyright.

Competing financial interests

The authors declare no competing financial interest.

Availability of data and materials

Sequencing data is deposited in ENA project PRJEB19193. Results, processed data and materials presented in this work can be provided upon request. Chimaera as a service is available on IBM Cloud at the following link `https://sysbio.uk-south.containers.mybluemix.net/chimaera/`. A set of anonymous credentials can be created for reviewers.

References

[1] M. Manica, P. Chouvarine, R. Mathis *et al.* "Inferring clonal composition from multiple tumor biopsies". *arXiv* (2017). arXiv: 1701.07940.

[2] P. A. Futreal, L. Coin, M. Marshall *et al.* "A census of human cancer genes". *Nature Reviews Cancer* **4**:3 (2004), pp. 177–183.

[3] M. E. Higgins, M. Claremont, J. E. Major, C. Sander, and A. E. Lash. "CancerGenes: A gene selection resource for cancer genome projects". *Nucleic Acids Research* **35**:SUPPL. 1 (2007), pp. D721–726.

[4] L. Ding *et al.* "Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics". *Cell* **173** (2018), pp. 305–320.

[5] P. C. Nowell. "The clonal evolution of tumor cell populations". *Science* **194**:4260 (1976), pp. 23–28.

[6] I. J. Fidler and I. R. Hart. "Biological diversity in metastatic neoplasms: Origins and implications". *Science* **217**:4564 (1982), pp. 998–1003.

[7] S. M. G. Espiritu, L. Y. Liu, Y. Rubanova *et al. The Evolutionary Landscape of Localized Prostate Cancers Drives Clinical Aggression*. Vol. 173. 4. Cell, 2018, 1003–1013.e15.

[8] P. C. Boutros, M. Fraser, N. J. Harding *et al.* "Spatial genomic heterogeneity within localized, multifocal prostate cancer". *Nature Genetics* **47**:7 (2015), pp. 736–745.

[9] Y. Wang, J. Waters, M. L. Leung *et al.* "Clonal evolution in breast cancer revealed by single nucleus genome sequencing". *Nature* **512**:7513 (2014), pp. 155–160. arXiv: NIHMS150003.

[10] H. Suzuki, K. Aoki, K. Chiba *et al.* "Mutational landscape and clonal architecture in grade II and III gliomas". *Nature Genetics* **47**:5 (2015), pp. 458–468.

[11] K. M. Mann, J. Y. Newberg, M. A. Black *et al. Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq*. Vol. 34. 9. Nature biotechnology, 2016, pp. 962–972. arXiv: NIHMS150003.

[12] R. Gao, A. Davis, T. O. McDonald *et al.* "Punctuated copy number evolution and clonal stasis in triple-negative breast cancer". *Nature Genetics* **48**:10 (2016), pp. 1119–1130. arXiv: 15334406.

[13] N. Andor, T. A. Graham, M. Jansen *et al.* "Pan-cancer analysis of the extent and consequences of intratumor heterogeneity". *Nature Medicine* **22**:1 (2016), pp. 105–113. arXiv: 15334406.

[14] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael. "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data". *Bioinformatics* **31**:12 (2015), pp. i62–i70.

[15] N. Niknafs, V. Beleva-Guthrie, D. Q. Naiman, and R. Karchin. "SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing". *PLoS Computational Biology* **11**:10 (2015). Ed. by Q. Morris, e1004416.

[16] W. K. Chu, P. Edge, H. S. Lee *et al.* "Ultraaccurate genome sequencing and haplotyping of single human cells". *Proceedings of the National Academy of Sciences* **201707609** (2017), p. 201707609. arXiv: arXiv:1303.3997.

[17] G. Getz and K. Ardlie. *Mutation Analysis in Frozen and FFPE Tumor Samples*. Crystal City, 2012.

[18] M. Cieslik, R. Chugh, Y. M. Wu *et al.* "The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing". *Genome Research* **25**:9 (2015), pp. 1372–1381.

[19] J. Wang, H. Khiabanian, D. Rossi *et al.* "Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia". *eLife* **3** (2014).

[20] G. Gundem, P. Van Loo, B. Kremeyer *et al.* "The evolutionary history of lethal metastatic prostate cancer". *Nature* **520**:7547 (2015), pp. 353–357.

[21] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel. *Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors*. Vol. 27. 11. Genome Research, 2017, pp. 1885–1894.

[22] C. G. A. Network *et al.* "Comprehensive molecular portraits of human breast tumours". *Nature* **490**:7418 (2012), p. 61.

[23] A. Ally, M. Balasundaram, R. Carlsen *et al.* "Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma". *Cell* **169**:7 (2017), 1327–1341.e23.

[24] S. L. Carter *et al.* "Absolute quantification of somatic DNA alterations in human cancer : Nature Biotechnology : Nature Publishing Group". *Nature biotechnology* **30** (2012), pp. 413–421.

[25] N. Andor, J. V. Harness, S. Müller, H. W. Mewes, and C. Petritsch. "Expands: Expanding ploidy and allele frequency on nested subpopulations". *Bioinformatics* **30**:1 (2014), pp. 50–60.

[26] A. G. Deshwar, S. Vembu, C. K. Yung *et al.* "PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors". *Genome Biology* **16**:1 (2015), p. 35. arXiv: 1406.7250.

[27] D. C. Lin, A. Mayakonda, H. Q. Dinh *et al.* "Genomic and epigenomic heterogeneity of hepatocellular carcinoma". *Cancer Research* **77**:9 (2017), pp. 2255–2265.

[28] H. Fritz, L. A. García-Escudero, and A. Mayo-Iscar. "tclust: An R Package for a Trimming Approach to Cluster Analysis". *Journal of Statistical Software* **47**:12 (2012), pp. 1–26.

[29] W. J. Krzanowski and Y. T. Lai. "A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering". *Biometrics* **44**:1 (1988), p. 23.

[30] C. Legány, S. Juhász, and A. Babos. "Cluster validity measurement techniques". *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases* (2006), pp. 388–393.

[31] G. Celeux and G. Govaert. "Gaussian parsimonious clustering models.pdf". *Pattern recognition* **28** (1995), pp. 781–793.

[32] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R journal* **8**:1 (2016), pp. 289–317.

[33] R. J. G. B. Campello, D. Moulavi, and J. Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". Springer, Berlin, Heidelberg, 2013, pp. 160–172.

[34] D. Robinson, E. M. Van Allen, Y.-M. Wu *et al.* "Integrative clinical genomics of advanced prostate cancer." *Cell* **161**:5 (2015), pp. 1215–1228.

[35] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. "KEGG as a reference resource for gene and protein annotation". en. *Nucleic Acids Research* **44**:D1 (2016), pp. D457–D462. (Visited on 12/13/2016).

[36] M. Kanehisa and S. Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Res* **28**:1 (2000), pp. 27–30. (Visited on 12/13/2016).

[37] A. Liberzon, A. Subramanian, R. Pinchback *et al.* "Molecular signatures database (MSigDB) 3.0." *Bioinformatics (Oxford, England)* **27**:12 (2011), pp. 1739–40.

[38] J. N. Anastas and R. T. Moon. "WNT signalling pathways as therapeutic targets in cancer". *Nature Reviews Cancer* **13**:1 (2013), pp. 11–26. arXiv: 9809069v1 [arXiv:gr-qc].

[39] D. C. Koboldt, Q. Zhang, D. E. Larson *et al.* "VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing". *Genome Research* **22**:3 (2012), pp. 568–576.

[40] D. Sheppard, R. Terrell, and G. Henkelman. "Optimization methods for finding minimum energy paths". *Journal of Chemical Physics* **128**:13 (2008), p. 6.

[41] A. Mortezavi, T. Hermanns, H. H. Seifert *et al.* "KPNA2 expression is an independent adverse predictor of biochemical recurrence after radical prostatectomy". *Clinical Cancer Research* **17**:5 (2011), pp. 1111–1121.

[42] A. Abeshouse, J. Ahn, R. Akbani *et al.* "The Molecular Taxonomy of Primary Prostate Cancer". *Cell* **163**:4 (2015), pp. 1011–1025. arXiv: 15334406.

[43] M. Lek, K. J. Karczewski, E. V. Minikel *et al.* "Analysis of protein-coding genetic variation in 60,706 humans". *Nature* **536**:7616 (2016), pp. 285–291. arXiv: 030338.

[44] M. V. Kuleshov, M. R. Jones, A. D. Rouillard *et al.* "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update." *Nucleic acids research* **44**:W1 (2016), W90–7.

[45] H. Mi, X. Huang, A. Muruganujan *et al.* "PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements". *Nucleic Acids Research* **45**:D1 (2017), pp. D183–D189.

[46] A. Roy *et al.* "Integration of Whole Transcriptome Sequencing into the Genomic Analysis of Pediatric Solid Tumors: Early Experience and Challenges". *Journal of Molecular Diagnostics* **16** (2014), pp. 754–755.

Part III.

Concluding Remarks

7 Discussion and Outlook

> *On the back part of the step, toward the right, I saw a small iridescent sphere of almost unbearable brilliance. At first I thought it was revolving; then I realised that this movement was an illusion created by the dizzying world it bounded. The Aleph's diameter was probably little more than an inch, but all space was there, actual and undiminished. Each thing (a mirror's face, let us say) was infinite things, since I distinctly saw it from every angle of the universe. I saw the teeming sea; I saw daybreak and nightfall; I saw the multitudes of America; I saw a silvery cobweb in the center of a black pyramid; I saw a splintered labyrinth (it was London); I saw, close up, unending eyes watching themselves in me as in a mirror; I saw all the mirrors on earth and none of them reflected me; I saw in a backyard of Soler Street the same tiles that thirty years before I'd seen in the entrance of a house in Fray Bentos; I saw bunches of grapes, snow, tobacco, lodes of metal, steam; I saw convex equatorial deserts and each one of their grains of sand...*
>
> – Jorge Luis Borges, *El Aleph*

This thesis presents a series of methods and algorithms developed with the aim of accelerating the adoption of precision medicine approaches and increasing our ability to better understand complex diseases using multiple data modalities. Starting form the reconstruction of relevance networks and their application for interpretable patient stratification it arrives at the implementation of personalized or sub-population specific models. The research conducted can be seen as an effort to design novel methodologies that constitute a systems biology toolbox integrating state-of-the-art methods from multiple domains, in order to provide the scientific community with a set of algorithms ready to be applied to a wide variety of experimental data to answer relevant biological questions.

As a starting point, molecular interaction network inference has been analyzed using two

data modalities: high-throughput omic measurements and natural language from scientific publications. In Chapter 2, COSIFER [1] scalable cloud service for the inference of molecular networks that creates a consensus network by integrating the output of different inference algorithms, is presented. The COSIFER idea came from the need to access a reliable and robust framework for building disease-specific interactomes from omic data. The ability to infer context-specific interaction networks is a fundamental step towards a comprehensive understanding of the biological system considered. The methodologies considered and the idea of merging predictions using a voting scheme have been inspired by the seminal work performed in two DREAM challenges concerning network inference from molecular data [1, 2]. In the COSIFER manuscript the framework results are validated on a large set of synthetic data with different noise types, sample sizes and network sizes, generated using GeneNetWeaver [3]. Performance of the consensus methods, in agreement with what has been observed in the DREAM challenges, has been proven to be more stable as they surpass or match the top-performing inference algorithms. Furthermore, COSIFER has been applied on network inference using breast cancer gene expression data coming from two independent cohorts, TCGA-BRCA [4] and METABRIC [5]. By comparing the similarity of the networks inferred from a selection of different pathways [6], it has been possible to show which mechanisms are consistently reconstructed from the independent patient measurements, highlighting disease-specific relevant pathways, like Epithelial-Mesenchymal transition, a known pathway for breast cancer development and metastatic behavior [7, 8]. By integrating the data from the two cohorts we have been able to identify a set of potential breast cancer regulators using node betweenness centrality [9] as a proxy for regulatory behavior. The research conducted while developing COSIFER also led to a publication [10] where an inference method based on partial correlation was implemented on memristive devices showing for the first time the potential of mixed-precision in-memory computing. The hardware implementation has been used to estimate an autophagy specific network from cancerous and normal RNASeq measurements from TCGA, exhibiting an altered regulation of genes responsible for autophagosomes formation.

In parallel to the efforts of inferring interaction networks from molecular data, INtERAcT [11] [2] has been developed, see Chapter 3. INtERAcT represents a novel approach to infer relevance networks from publications in an unsupervised way. The main motivation behind the implementation of INtERAcT resides in the need of a method able to keep up with the continuously increasing amount of knowledge coming from scientific articles published. In

[1] `https://sysbio.uk-south.containers.mybluemix.net/cosifer/`, as of November 2018
[2] `https://sysbio.uk-south.containers.mybluemix.net/interact/`, as of November 2018

such a context, rule-based methods [12] as well as semi-supervised or supervised methods [13] are not easily applicable, given the impossibility of reliably annotating an ever-growing text corpus. INtERAcT builds upon recent research in language modeling based on learning vector representation of words without supervision [14, 15]. By leveraging these approaches, no manual curation or expert knowledge in the text processing phase is required, since the vector representation can be obtained without any prior assumption. INtERAcT scores interactions making use of a new metric to compare word vectors. After clustering the words in the embedding space, the metric is defined by using the Jensen-Shannon divergence [16] to compare word's neighbors distribution over the clusters. We have shown how INtERAcT is effectively able to reconstruct cancer-specific networks in significant agreement with interactions reported in STRING [17] using KEGG [18, 19] to define tumor-specific pathways. Specifically, it is extremely effective for variable corpus sizes and exhibits strong robustness towards variations of the hyper-parameters used to train word vectors compared to others metrics usually adopted to compute similarities in word embeddings.

Networks reconstructed using INtERAcT and COSIFER are extremely useful to produce disease-specific interaction networks and summarize evidence contained in two disparate data domains. Moreover, by making use of prediction aggregation methods, like consensus algorithms adopted in COSIFER [20], we can easily integrate the information from the multiple modalities. Besides these clear advantages, some limitations have to be considered. Both methods estimate relevance networks in the form of weighted undirected graphs. While this representation, by using network analysis approaches, can provide us with a lot of useful insights it is still not able to capture the complex dynamic of a gene regulatory network or a signaling pathway. Inferring causal interactions would allow us to enhance the graph representation learned, but this comes at the price of requiring either temporal series data or data from perturbation experiments. Moreover, most causality inference methods come with performance drawbacks that force us to limit the size of the system analyzed, e.g., single signaling pathways. Inspired by current trends in structural representation learning based on interpreting a graph in terms of node and edge embeddings [21–23], we started to work in the direction of adding causality relations for the inference of regulatory interactions using deep learning [24] with impressive results in terms of reconstruction of known transcription factor to target interactions. These approaches represent the future of the field of network reconstruction in systems biology and are a natural extension of the work presented in the thesis. Especially promising are recent advances proposed in some seminal works on model formulation [25–27], and efforts to propose methods able to scale for large networks and successfully integrate multiple data from different sources [28–30].

As mentioned earlier, while far from being an accurate description of the dynamic behavior of a biological system, relevance networks can be useful in many systems biology applications, especially to facilitate interpretable patient classification. Definition of patient groups or sub-populations plays a fundamental role in adopting a precision medicine approach. PIMKL [31] [1], included in Chapter 4, exploits them to achieve a molecular process-aware classification of samples. By considering a selection of pathways, we partition a known molecular interaction network and construct *pathway-induced* kernels that are combined using a multiple kernel learning (MKL) algorithm to predict a phenotype of interest. Working at pathway level enables us to shift from a common biomarker paradigm towards a composite biomarker paradigm. This allows us to interpret classification results in the light of the molecular processes enriched by the weights obtained for each kernel from applying the MKL algorithm, providing a biologically relevant explanation for PIMKL predictions. PIMKL surpasses state-of-the-art methodologies benchmarked [32] in the task of predicting biochemical recurrence on a collection of microarray data from six independent breast cancer cohorts. In this validation we use as the prior knowledge source the network used as benchmark, namely a combination of KEGG [18, 19] and Pathway Commons [33], and the *hallmark* gene sets from MSigDB [34] to define the pathways. Besides a superior performance, PIMKL infers molecular signatures that are stable across the different cohorts, highlighting pathways that are linked to breast cancer development. For example, we find that heme metabolism is significantly enriched in all cohorts, indicating a strong relation between this pathway and relapse-free survival. Interestingly a link between this pathway and cancer progression has been observed in lung cancer cell lines and animal models [35], supporting the validity of the findings. These results show that PIMKL can achieve high performance in terms of specificity and sensitivity while providing insights about relevant pathways for the phenotype of interest. Additionally, we validated the potential of PIMKL for transfer learning tasks by using the molecular signatures learned in the six independent cohorts to predict biochemical recurrence-free survival in METABRIC [5]. Performance in terms of area under the curve remain unchanged when comparing the transferred signature to a re-trained one, proving PIMKL generalization power. Finally, we demonstrate how PIMKL can be easily extended to handle multiple data modalities by simply adding additional kernels to the mixture. By applying PIMKL on METABRIC and considering gene expression and copy number alteration data, we observe that the algorithm is able to fuse the modalities and discard noisy kernels while maintaining the high performance. This feature is important when it is necessary to deal with heterogenous data for which we do not have any prior information about the quality and information content of

[1] `https://sysbio.uk-south.containers.mybluemix.net/pimkl/`, as of November 2018

the different datasets, a common case when we are confronted with the analysis of multi-omic datasets. Despite the flexibility of PIMKL, some shortcomings are still present. In the first place, considering interaction networks present in databases and defining pathways by limiting ourselves to annotated gene sets might bias the analysis towards what is already known, and thus reduce our ability to discover new molecular processes linked to a specific phenotype. To overcome this problem, at a network level, we can use interaction networks inferred from independent datasets using different reconstruction algorithms and consensus approaches, like in COSIFER. Regarding the gene sets used to identify cellular processes, an approach to avoid depending too much on prior knowledge is represented by applying community detection algorithms [36] on the molecular interaction network considered. Finding gene sets with similar properties in a data-driven way can really increase the potential of a method like PIMKL to discover novel groups of interacting genes and allows completely new pathways to be defined in an assumption-free fashion. Another issue to consider, common to most supervised algorithms for patient stratification, is the appropriate treatment of label uncertainty. This is a major problem since it has been shown in multiple studies how clinical labels, especially in cancer grading diagnosis, can be noisy and can exhibit a low agreement between different specialists [37, 38]. Besides keeping a supervised approach by explicitly modeling label uncertainty [39–43], we can adapt the algorithm to work in an unsupervised setting. Extending PIMKL by making use of an unsupervised MKL methodology [44], will improve its usability and help overcome the noisy labels problem, while allowing patient groups to be defined without any supervision. This gives a major advantage in a precision medicine perspective since we would be able to define novel patient groups associated with a potentially different prognosis and possibly a different treatment.

As said, determining patient sub-populations is fundamental in precision medicine, since it defines patient groups that can be used to answer specific biological questions and suggest personalized therapeutic options. The research activities described in Chapter 5 and Chapter 6 are enabling the development of such approaches. In Chapter 5 the implementation of a framework for hardware acceleration of Boolean model simulations using FPGA cards has been described. Boolean models and more generally logical models are a powerful tool to analyze and simulate a complex biological system [45, 46], but currently their application has been limited to the analysis moderate-size networks, up to a few hundred of genes or proteins. One of the main limitations is the lack of a scalable way to simulate the dynamics and find the steady states of the system. The hardware accelerator we propose addresses this problem by exhibiting consistent speedups between 10x and 1,000x for a selection of models of variable sizes and describing different molecular systems. The framework presented enables a

thorough analysis of the state space and system's attractors. The results obtained show how the hardware implementation of the simulator permits the simulation of molecular systems including thousands of nodes. This implies that, by using multiple FPGA cards, we are able to easily simulate a molecular system up to genome-wide scale offering a unique opportunity to model arbitrarily large signaling pathways and fully exploit high-throughput omic datasets. Being capable of simulating network dynamics at such scale will for example facilitate the simulation of personalized patient models and the analysis of their response to different perturbations, such as known effects of a drug or a treatment. Besides the clear advantages of the framework, a shortcoming is evident: the need for specific, if not unique, requirements in terms of hardware architecture in order to be able to profit from the acceleration based on FPGA. This could be addressed by making the platform available and hosting a cloud acceleration service for Boolean model simulation and attractor analysis accessible to the scientific community. Such a service would represent a unique resource to simulate logical models at an unprecedented scale and promote the consistent usage of these personalized model-based approaches.

In Chapter 6, another algorithm for personalized modeling, implemented during this research work, called Chimaera [47] [1] has been presented. Chimaera is a method for inferring the clonal composition of tumors using mutation data obtained by profiling multiple regions of a cancer at a single time point. The approach we adopt is based on a commonly used methodology: we define clones by grouping co-occurring mutations, in terms of mutation frequencies, across different regions and define the phylogenetic clonal evolution by finding a partial order between them. The two main assumptions are: first, spatial resolution can help sample a tumor's subclones that appeared at different time points; and, second, mutations are accumulated over time, namely higher frequency mutations are older and define ancestral clones. Many methods, using a similar procedure, have been proposed [48–50] in recent years, but none of them is able to properly account for genome instability at copy number level, a common phenomenon in different cancer types. Chimaera explicitly accounts for copy number alterations by correcting somatic mutation frequencies. Our algorithm exhibits superior performance on multiple synthetic data generated from different phylogenies, noise rates and levels of copy number alterations when compared to other state-of-the-art methods. We also prove Chimaera's potential by analyzing multiple cancer types. First, we used Chimaera to infer clonal evolution for a CRPC patient using WES data from ten regions, showing how we can define an evolution trajectory for the tumor development in the considered patient. We also applied Chimaera on an hepatocellular carcinoma cohort published in a recent study [51].

[1] `https://sysbio.uk-south.containers.mybluemix.net/chimaera/`, as of November 2018

For each patient we have been able to build phylogenetic trees in agreement with the results of the study and recapitulating known biological behavior. Moreover, Chimaera not only enabled the study of cancer evolution at the single patient level, but also when comparing mutations in clones estimated as ancestral, highlighted a significant enrichment of the WNT-signaling pathway. Interestingly, WNT-signaling is one of the usual suspects for tumor initiation and proliferation and has been tested as a drug target in pre-clinical studies [52]. Results on HCC show that Chimaera can be effectively used to describe patient-specific tumor evolution and suggest potential therapeutic interventions. Despite its flexibility, Chimaera presents some limitations. First, the assumptions made on somatic mutation and copy number behavior might be too strict and not consistent with the actual biology underlying cancer evolution. Second, the resolution given by DNA sequencing from multiple regions might not be enough to sample all cell populations, leading to an incomplete description of the clonal architecture and forcing the algorithm to perform the inference on extremely noisy data. Regarding the first issue, while on one hand we have seen cases in the cohorts analyzed where for some genomic locations the assumptions were violated, on the other hand mutations exhibiting this behavior were limited in number and represented only a small portion of the SNVs considered. Regarding the low resolution problem, the most promising option seems to be single-cell DNA sequencing. With this technology we can call mutations for each cell and potentially detect every single cell population present in a sample. Unfortunately some technical limitations, like quantity of material needed and noise, are preventing us from relying on these measurements for the time being. Nevertheless, extending Chimaera to work with these data types, possibly obtained by profiling multiple regions at the single-cell level, will surely expand our ability to infer tumor evolution. However, even by working with standard DNA sequencing data, Chimaera infers phylogenetic trees representing tumor development that can be used to design therapeutic interventions tailored for a specific patient, representing a perfect example of an algorithm design in the era of precision medicine.

This research work has produced a set of valuable tools that can be used to study biological systems and analyze multi-omic data at multiple levels. The graph-based approaches implemented, COSIFER and INtERAcT, are the pillars used to support the analysis of high-throughput molecular data by summarizing information about molecular interactions and mechanisms governing cell behavior. On top of these pillars we have developed PIMKL that, with its interpretable phenotype prediction, allows clinically relevant patient groups to be defined while using a prior interaction network to highlight active molecular processes. Once patients are stratified, we can use the proposed hardware-accelerated Boolean models to potentially simulate network dynamics and find attractors at genome-wide scale for specific patient

groups. Going towards a personalized medicine perspective, Chimaera has been implemented to integrate multiple genomic data to infer patient-specific tumor evolution.

In conclusion, we believe that the algorithms implemented in the course of this thesis will help the systems biology community accelerate a consistent adoption of precision medicine approaches.

References

[1] D. Marbach, J. C. Costello, R. Küffner *et al.* "Wisdom of crowds for robust gene network inference". *Nature Methods* **9**:8 (2012), pp. 796–804. arXiv: arXiv:1511.08814v1.

[2] S. M. Hill, L. M. Heiser, T. Cokelaer *et al.* "Inferring causal molecular networks: Empirical assessment through a community-based effort". *Nature Methods* **13**:4 (2016), pp. 310–322. arXiv: 15334406.

[3] T. Schaffter, D. Marbach, and D. Floreano. "GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods". *Bioinformatics* **27**:16 (2011), pp. 2263–2270.

[4] C. G. A. Network *et al.* "Comprehensive molecular portraits of human breast tumours". *Nature* **490**:7418 (2012), p. 61.

[5] C. Curtis, S. P. Shah, S.-F. Chin *et al.* "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups". *Nature* **486**:7403 (2012), p. 346.

[6] A. Liberzon, C. Birger, H. Thorvaldsdóttir *et al.* "The molecular signatures database hallmark gene set collection". *Cell systems* **1**:6 (2015), pp. 417–425.

[7] J. Felipe Lima, S. Nofech-Mozes, J. Bayani, and J. Bartlett. "EMT in breast carcinoma—A review". *Journal of clinical medicine* **5**:7 (2016), p. 65.

[8] F. Liu, L.-N. Gu, B.-E. Shan, C.-Z. Geng, and M.-X. Sang. "Biomarkers for EMT and MET in breast cancer: An update". *Oncology letters* **12**:6 (2016), pp. 4869–4876.

[9] L. C. Freeman. "A set of measures of centrality based on betweenness". *Sociometry* (1977), pp. 35–41.

[10] M. L. Gallo, A. Sebastian, R. Mathis *et al.* "Mixed-Precision In-Memory Computing". *Nature Electronics* **1**:4 (2017), pp. 246–253. arXiv: 1701.04279.

[11] M. Manica, R. Mathis, and M. R. Martínez. "INtERAcT: Interaction Network Inference from Vector Representations of Words" (2018). arXiv: 1801.03011.

[12] M. Torii, C. N. Arighi, G. Li *et al.* "RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information". *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **12**:1 (2015), pp. 17–29. (Visited on 07/04/2017).

[13] D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. "A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature". *PLoS Computational Biology* **6**:7 (2010). (Visited on 07/06/2017).

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". *arXiv:1301.3781 [cs]* (2013). arXiv: 1301.3781. (Visited on 05/01/2017).

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and Their Compositionality". *Proceedings of the 26th International Conference on Neural Information Processing Systems*. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119. (Visited on 04/02/2017).

[16] J. Lin. "Divergence measures based on the Shannon entropy". *IEEE Transactions on Information Theory* **37**:1 (1991), pp. 145–151.

[17] D. Szklarczyk, A. Franceschini, S. Wyder *et al.* "STRING v10: protein–protein interaction networks, integrated over the tree of life". *Nucleic Acids Research* **43**:Database issue (2015), pp. D447–D452. (Visited on 07/06/2017).

[18] M. Kanehisa and S. Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Res* **28**:1 (2000), pp. 27–30. (Visited on 12/13/2016).

[19] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. "KEGG as a reference resource for gene and protein annotation". en. *Nucleic Acids Research* **44**:D1 (2016), pp. D457–D462. (Visited on 12/13/2016).

[20] M. E. Ahsen, R. Vogel, and G. Stolovitzky. "Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions" (2018). arXiv: 1802.04684.

[21] B. Perozzi, R. Al-Rfou, and S. Skiena. "DeepWalk". *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. New York, New York, USA: ACM Press, 2014, pp. 701–710.

[22] A. Grover and J. Leskovec. "node2vec". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, 2016, pp. 855–864.

[23] Z. Gao, G. Fu, C. Ouyang *et al.* "edge2vec: Learning Node Representation Using Edge Semantics" (2018). arXiv: 1809.02269.

[24] R. Mathis, M. Manica, and M. Rodriguez Martinez. "DeepGRN: Deciphering gene deregulation in cancer development using deep learning" (2017).

[25] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel. "Neural Relational Inference for Interacting Systems" (2018). arXiv: 1802.04687.

[26] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende, and K. Kavukcuoglu. "Interaction Networks for Learning about Objects, Relations and Physics". *Nips* (2016), pp. 4502–4510. arXiv: 1612.00222.

[27] P. W. Battaglia, J. B. Hamrick, V. Bapst *et al.* "Relational inductive biases, deep learning, and graph networks" (2018). arXiv: 1806.01261.

[28] W. L. Hamilton, R. Ying, and J. Leskovec. "Inductive Representation Learning on Large Graphs" (2017). arXiv: 1706.02216.

[29] J. Chen, T. Ma, and C. Xiao. *FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling*. 2018.

[30] R. Ying, R. He, K. Chen *et al.* "Graph Convolutional Neural Networks for Web-Scale Recommender Systems" (2018). arXiv: 1806.01973.

[31] M. Manica, J. Cadow, R. Mathis, and M. R. Martínez. "PIMKL: Pathway Induced Multiple Kernel Learning" (2018). arXiv: 1803.11274.

[32] Y. Cun and H. Fröhlich. "Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge". *BMC bioinformatics* (2012).

[33] E. G. Cerami, B. E. Gross, E. Demir *et al.* "Pathway Commons, a web resource for biological pathway data". *Nucleic Acids Research* **39**:SUPPL. 1 (2011).

[34] A. Liberzon, C. Birger, H. Thorvaldsdóttir *et al.* "The Molecular Signatures Database Hallmark Gene Set Collection". *Cell Systems* **1**:6 (2015), pp. 417–425.

[35] J. Hooda, M. Alam, and L. Zhang. "Evaluating the association of heme and heme metabolites with lung cancer bioenergetics and progression". *Metabolomics* **5**:3 (2015), p. 1000150.

[36] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig. "Community detection in networks: A multidisciplinary review". *Journal of Network and Computer Applications* **108** (2018), pp. 87–111.

[37] L. Ferejohn. "That other story". *Thomas Wolfe Review* **36**:1-2 (2012), pp. 144–151.

[38] D. S. Gomes, S. S. Porto, D. Balabram, and H. Gobbi. "Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast". *Diagnostic Pathology* **9**:1 (2014), p. 121.

[39] G. Stempfel and L. Ralaivola. "Learning Kernel Perceptrons on Noisy Data Using Random Projections". *Algorithmic Learning Theory*. Springer, Berlin, Heidelberg, 2007, pp. 328–342.

[40] G. Stempfel, L. Ralaivola, and F. Denis. "Learning from Noisy Data using Hyperplane Sampling and Sample Averages" (2007).

[41] G. Stempfel and L. Ralaivola. "Learning SVMs from sloppily labeled data". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 5768 LNCS. PART 1. Springer, Berlin, Heidelberg, 2009, pp. 884–893.

[42] K. Crammer and D. D. Lee. *Learning via Gaussian Herding*. 2010.

[43] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. "Learning with Noisy Labels". *Advances in neural information processing systems*. 2014, pp. 1196–1204.

[44] J. Mariette and N. Villa-Vialaneix. "Unsupervised multiple kernel learning for heterogeneous data integration". *Bioinformatics* **34**:2009 (2017).

[45] S. Pandey, R. S. Wang, L. Wilson *et al.* "Boolean modeling of transcriptome data reveals novel modes of heterotrimeric G-protein action". *Molecular Systems Biology* **6** (2010), p. 372.

[46] S. Li, S. M. Assmann, and R. Albert. "Predicting essential components of signal transduction networks: A dynamic model of guard cell abscisic acid signaling". *PLoS Biology* **4**:10 (2006). Ed. by J. Chory, pp. 1732–1748.

[47] M. Manica, P. Chouvarine, R. Mathis *et al.* "Inferring clonal composition from multiple tumor biopsies". *arXiv* (2017). arXiv: 1701.07940.

[48] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael. "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data". *Bioinformatics* **31**:12 (2015), pp. i62–i70.

[49] S. Malikic, A. W. McPherson, N. Donmez, and C. S. Sahinalp. "Clonality inference in multiple tumor samples using phylogeny". *Bioinformatics* **31**:9 (2015), pp. 1349–1356.

[50] N. Niknafs, V. Beleva-Guthrie, D. Q. Naiman, and R. Karchin. "SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing". *PLoS Computational Biology* **11**:10 (2015). Ed. by Q. Morris, e1004416.

[51] D. C. Lin, A. Mayakonda, H. Q. Dinh *et al.* "Genomic and epigenomic heterogeneity of hepatocellular carcinoma". *Cancer Research* **77**:9 (2017), pp. 2255–2265.

[52] J. N. Anastas and R. T. Moon. "WNT signalling pathways as therapeutic targets in cancer". *Nature Reviews Cancer* **13**:1 (2013), pp. 11–26. arXiv: 9809069v1 [arXiv:gr-qc].

8 Copyright and Contributions

The following describes the contributions by chapter.

Chapter 1: All content is contributed exclusively by me.

Chapter 2: This chapter contains an adaptation of a manuscript in preparation at the time of writing. All text content is contributed by me, Charlotte Bunne, Roland Mathis and María Martínez Rodríguez. COSIFER core has been implemented by me, Roland Mathis and subsequently extended by Charlotte Bunne. Data analysis has been conducted by me and Charlotte Bunne. Plots have been generated by me and Joris Cadow. COSIFER web service has been entirely developed by me.

Chapter 3: This chapter contains an adaptation of the pre-print of the manuscript in review at the time of writing. All text content is contributed by me, Roland Mathis and María Martínez Rodríguez. INtERAcT core has been implemented by me and Roland Mathis. Data analysis on various disease specific corpus has been conducted by me and Roland Mathis, related plots have been generated by me. Parametrical analysis and related figures has produced by me and Joris Cadow. INtERAcT web service has been entirely developed by me.

Chapter 4: This chapter contains an adaptation of the pre-print of the manuscript in review at the time of writing. All text content is contributed by me, Joris Cadow, Roland Mathis and María Martínez Rodríguez. PIMKL core, i.e., C++ code and python wrapper have been developed by me and Joris Cadow. Analysis on the different breast cancer datasets and figures have been designed and implemented by me and Joris Cadow. PIMKL web service has been developed by me and Joris Cadow.

Chapter 5: This chapter contains an adaptation of a manuscript in preparation at the time of writing. All text content is contributed by me, Mitra Purandare, Raphael Polig and María Martínez Rodríguez. FPGA hardware acceleration has been implemented by Raphael Polig. Timing experiments were conducted by Raphael Polig and Mitra Purandare. Data analysis has been conducted entirely by me.

Chapter 6: This chapter contains an adaptation of a manuscript in preparation at the time of

writing. We are analyzing an additional set of CRPC patients profiled over time to further validate the algorithm. Latest results will be included in the submitted paper. All text content is contributed by me, Pavel Sumazin and María Martínez Rodríguez. The algorithm has been designed with equal contributions by Pavel Sumazin, Philippe Chouvarine, Roland Mathis and me. Chimaera core has been implemented entirely by me. Synthetic data analysis and figures have been generated by me. Chimaera runs on real tumors have been performed by me. Real tumors results analysis has been conducted by me, Hyunjae Ryan Kim and Pavel Sumazin. Experimental work and raw data processing have been conducted by University Hospital of Zürich and Baylor College of Medicine under the supervision of Peter J. Wild and Pavel Sumazin. Chimaera service has been entirely entirely developed by me.

Chapter 7: All content is contributed exclusively by me.

Part IV.

Appendix

9 Appendix

List of Figures

List of Tables

Curriculum vitae

Matteo MANICA
Machine Learning Engineer

linkedin.com/in/matteo-manica-drugilsberg
github.com/drugilsberg
+41 78 971 44 66 +39 347 91 85 999
drugilsberg@gmail.com tte@zurich.ibm.com
Dennlerstrasse 28, 8048, Zürich
Born on 11.09.1988 in Novara, Italy

I'm a researcher in the Cognitive Health Care and Life Sciences group at IBM Zürich Research Laboratory.

I successfully defended my PhD at the end of a joint research program between IBM Research Zürich and the Institute of Molecular Systems Biology at ETH Zürich, studying the exploration of multi-modal learning approaches for precision medicine. My research focus is set on the development of integrative learning frameworks for multiple molecular and clinical data in the context of cancer medicine to improve patients stratification and allow clinicians to find personalized therapeutic interventions. I'm working on the application of machine and deep learning methods to analyze progression and development of prostate cancer in the context an H2020 EU project, PrECISE.

Before, I worked as consultant in data science and software development with specific applications in biological fluids dynamic, digital and biological signal processing and data analysis. My main focus was on the analysis of CT angiography and MR angiography scans of abdominal aortic aneurysms (AAA). Trough image analysis, segmentation and 3D volume rendering of the abdominal aorta I contributed to create patient specific models to simulate blood flows in the vessels and to assess rupture risk of the aneurysm.

I obtained my BSc and MSc at Politecnico di Milano in Applied Mathematics and Computer Science, a course with a strong focus on numerical simulations and data analysis. In my master thesis work I developed an original model, based partial different equations for flow in porous media, to describe Medulloblastoma growth. By analysing MRIs at different time points of a given patient it was possible to fit the model trough segmentation and 3D volume rendering of the brain and the tumor mass, enabling an accurate estimate of the disease's course over time.

Education

2016-2018	Doctor of Sciences at ETH Zürich in collaboration with IBM Research.
2010-2013	MSc in Mathematics and Computer Science at Politecnico di Milano.
2007-2010	BSc in Mathematics and Computer Science at Politecnico di Milano.

Professional Experience

present — **Pre-doctoral Researcher in Cognitive Health Care and Life Sciences, IBM Research Zürich, Switzerland**

January 2016

› **PrECISE**. I work on the development of predictive computational technologies and learning frameworks for multi-modal data integration in the context of cancer medicine. The main goals are to improve patients stratification and to inform clinicians for personalized therapeutic interventions. Focus on application of machine and deep learning methods to analyze progression and development of cancer. During my work in the project besides producing publications and submitting four patent applications I implemented and deployed four open-access IBM Cloud services: PIMKL, INtERAcT, COSIFER and Chimaera.

Tensorflow | Keras | Elasticsearch | scikit-learn | pandas | Docker | Kubernetes | IBM Cloud | Travis CI | Artifactory | CMake | pip

December 2015
January 2014

Junior Analyst & Developer, Moxoff s.p.a., Italy

- Mathematical modeling and programming in mathematical biology. Performing numerical simulation of biological flows and vessels fluid dynamics.
- Analysis of geophysical and seismic images.
- Non linear optimization using heuristic and genetic algorithms.
- PIC micro-controller programming.
- Signal processing and functional data analysis.
- GUI and web application design.
- Image and video processing and analysis.
- Computer vision and machine learning.
- Implementation of tools for automatic post-processing of numerical simulations.

NetBeans | Eclipse | MPLAB 8 | OsiriX | CMake | Angular | Paraview | VTK | ITK | LifeV | Apache Spark | Jenkins | Docker | sbt | Maven | Evolving Objects

May 2014
November 2013

Intern, Moxoff s.p.a., Italy

- Mathematical modeling and programming in mathematical biology. Performing numerical simulation of biological flows and vessels fluid dynamics.
- Analysis of geophysical and seismic images.
- Non linear optimization using heuristic and genetic algorithms.

NetBeans | Eclipse | Paraview | VTK | ITK | LifeV | Evolving Objects

Software Skills

Programming	C++, C, Python, Scala, Java, R, Objective-C, Fortran, MATLAB, GNU Octave
Frameworks	Django, Swagger
Machine Learning & Data Analysis	Apache Spark, Tensorflow, scikit-learn, pandas, Keras
Cloud Platforms	IBM Cloud, AWS
DevOps	Docker, Kubernetes, Ansible, Artifactory
Continuous Integration	Travis, Jenkins
Version Control	Git, Apache Subversion

Relevant repositories

functional-cpp	Templatized functional programming library (github.com/drugilsberg/functional-cpp).
tensorflow-estimator-serving	Serving models via Docker using TensorFlow high-level API (github.com/drugilsberg/tensorflow-estimator-serving).

Cloud Services

Open-access Services on IBM Cloud

PIMKL	Pathway Induced Multiple Kernel Learning as a service for explainable phenotype prediction: https://sysbio.uk-south.containers.mybluemix.net/pimkl/.
COSIFER	Service for unsupervised network inference using a consensus of state of the art methods: https://sysbio.uk-south.containers.mybluemix.net/cosifer/.
INtERAcT	Infer interaction network between entities given trained word vectors using INtERAcT: https://sysbio.uk-south.containers.mybluemix.net/interact/.
Chimaera	Infer clonal composition of a tumor using WGS/WES measurements from multiple biopsies: https://sysbio.uk-south.containers.mybluemix.net/chimaera/.

Publications and Patents

Conferences

- Ali Oskooei, Jannis Born, Matteo Manica et al. "PaccMann: Predicting anticancer compound sensitivity with multimodal attention-based neural networks", paper accepted at the Workshop Machine Learning for Molecule and Material ad the 32nd Conference on Neural Information Processing Systems (NeurIPS), Montréal, Canada, 2018.
- Matteo Manica et al. "PIMKL: Pathway Induced Multiple Kernel Learning", poster presentation at the 17th European Conference on Computational Biology (ECCB), Athens, Greece, 2018.
- Matteo Manica et al. "Inferring clonal composition from multiple tumor biopsies", talk and poster presentation at Intelligent Systems for Molecular Biology (ISMB) and the 16th European Conference on Computational Biology (ECCB), Prague, Czech Republic, 2017.
- Matteo Manica et al. "CoDON: a learning framework for linking genomics and transcriptomics data to protein expression", poster presentation at the 15th European Conference on Computational Biology (ECCB), The Hague, Netherlands, 2016.

Journals

- Matteo Manica et al. "PIMKL: Pathway Induced Multiple Kernel Learning", *arXiv preprint arXiv:1803.11274*, under revision, 2018.
- Matteo Manica et al. "Accelerated analysis of Boolean gene regulatory networks via reconfigurable hardware", submitted, 2018.
- Matteo Manica et al. "COSIFER: Consensus Interaction Network Inference Service", under preparation to be submitted, 2018.
- Matteo Manica et al. "INtERAcT: Interaction Network Inference from Vector Representations of Words", *arXiv preprint arXiv:1801.03011*, under revision, 2018.
- Manuel Le Gallo, Abu Sebastian, Roland Mathis, Matteo Manica et al. "Mixed-precision in-memory computing", *Nature Electronics*, 2018.
- Ali Oskooei, Matteo Manica et al. "Network-based Biased Tree Ensembles (NetBiTE) for Drug Sensitivity Prediction and Drug Sensitivity Biomarker Identification in Cancer", *arXiv preprint arXiv:1808.06603*, under revision, 2018.
- Matteo Manica et al. "Inferring clonal composition from multiple tumor biopsies", *arXiv preprint arXiv:1701.07940*, under preparation to be submitted, 2018.

Patents

- "INTERACTION NETWORK INFERENCE FROM VECTOR REPRESENTATION OF WORDS", P201702217, filed.

Thesis

- Matteo Manica "Brain Tumors: mathematical modeling and numerical simulations for cancer growth and invasion", *politesi.polimi.it*, 2013.

Languages

Italian	●●●●●
English	●●●●○
German	●●○○○

References

Sandro Salsa
Full Professor, Politecnico di Milano
@ sandro.salsa@polimi.it
+39 02 2399 4553

María Rodríguez Martínez
Research Staff Member, IBM Research Zürich
@ mrm@zurich.ibm.com
+41 44 724 82 80

Interests

Music

I studied classical flute for seven years, and alto sax for one year. Then I moved my interests to jazz music, studying elettric bass and double bass for eight years with Marcello Testa. I attended various ensemble music courses with musician like Ramberto Ciammarughi, Claudio “Wally” Allifranchini and Lorenzo Cominoli. I had the pleasure to share the stage for didactic purposes with artists like Rudy Migliardi, Claudio “Wally” Allifranchini, Claudio Guida, Nicola Stranieri and Walter Calafiore. I had the honour to share the stage with the great free jazz musician Sabir Mateen. I played for eight years in Pigreco, a funk/fusion trio.