UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Computational methods in Biology: cancer biomarkers, protein networks and lateral gene transfer

Henry Heberle

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Henry Heberle

Computational methods in Biology: cancer biomarkers, protein networks and lateral gene transfer

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Rosane Minghim

USP – São Carlos
January 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

H445c

Heberle, Henry
Computational methods in Biology: cancer biomarkers, protein networks and lateral gene transfer / Henry Heberle; orientadora Rosane Minghim. -- São Carlos, 2019.
164 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2019.

1. Cancer Biomarker. 2. Protein Prioritization. 3. Biological Network Visualization. 4. Lateral Gene Transfer. 5. Tree Visualization. I. Minghim, Rosane, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Henry Heberle

Métodos computacionais em Biologia: biomarcadores de câncer, redes de proteínas e transmissão lateral de genes

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Rosane Minghim

USP – São Carlos
Janeiro de 2019

ACKNOWLEDGEMENTS

To my family and friends: it would not be possible without you.

Thank you mother and father, Lucia and Jerônimo Heberle, for all you have made for me. Without your care and hope, none of this would have happened.

Paulo Falco Cobra: you pushed me to see the world through different eyes. Gabriela Vaz Meirelles: you bridged from Biology to Computer Science, from Systems Biology to the Florence's Art Academy; you motivated and inspired me. You were there to make me not give up, to make me see we are more than some-people may think. Rosane Minghim: you believed on me, supported and allowed me to be here where I am today; you oriented me to lighten my life with mindfulness. Keeping a healthy mind during a doctorate is not an easy task. Cintia Diniz: when we had our first contact, I was full of over-thinking and being hard to myself. With your help I changed my life. Kyle Paul Ouellette: you showed me art, kindness and appreciation when I needed the most. Thank you.

Thank you my friends, Geandro Cason, Daniel Filizola, Michel Ferrari, Mônica Camacho, Rogério Colaço da Silva, Gabriela Carosio, Pedro Paulo Aquilante Junior, Otávio Siqueira, Luís Carlos Leva Borduchi, Felipe Hilário, Flávio Contrera, Fabio Vieira, Carlos Eduardo G. Bassetti, Rodrigo Bela, Nil Soares, Nícolas Roque, Anderson Felix and Alexandre Pinchemel. You gave my life meaning.

A special gratitude goes to researchers from LNBio, UNICAMP and Dalhousie University who helped me and believed on me for this journey: Dr. Adriana Franco Paes Leme, Romenia Ramos Domingues, Dr. Daniela Campos Granato, Dr. Carolina Moretto Carnielli, Dr. Guilherme Pimentel Telles and Dr. Robert Beiko; and to ICMC-USP, CAPES, and the Government of Canada.

This doctorate was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The supertree project was developed with the support of the Government of Canada.

“Leonardo’s delight at combining diverse passions remains the ultimate recipe for creativity. So, too, does his ease at being a bit of a misfit: illegitimate, gay, vegetarian, left-handed, easily distracted, and at times heretical. His life should remind us of the importance of instilling, both in ourselves and our children, not just received knowledge but a willingness to question it—to be imaginative and, like talented misfits and rebels in any era, to think different.”

- Leonardo Da Vinci (2017) by Walter Isaacson

ABSTRACT

HEBERLE, H. **Computational methods in Biology: cancer biomarkers, protein networks and lateral gene transfer**. 2019. 164 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Molecular Biology is a branch within Science of great importance. Despite the fact it studies microscopic entities, the volume and complexity of information are great. The applications are varied and can be of global interest, such as the spread of antibiotic resistance genes among bacteria and new methods for diagnostic and prognostic of cancer. By understanding biomolecular mechanisms, scientists can define treatments for diseases, support the decisions made by patients, identify the influence of intestinal microbiota over physical and psychological conditions, find cause and source of microbial antibiotic resistance, among many other applications. Computer Science plays key roles in this context, such as enabling complex data analyzes by specialists, creating models that simulate biological structures and processes, and by providing algorithms for extracting information encoded in biological data.

During my doctorate, we explored those mechanisms in three main levels: quantification of proteins from cells, analysis of interactions that happen inside cells, and the comparison of genomes and their genetic history. This manuscript reports different projects, four of them already published in scientific journals. They comprise the discovery of candidate proteins for cancer biomarkers, the visual analysis of protein-protein interaction networks and the visual analysis of lateral gene transfer in bacterial phylogenetic trees. Here, we explain these projects and the main findings associated with the use of computational methods. Among the results are the evaluation of stability of ranking and signature methods applied to discovery proteomics data, a new approach to select candidate proteins from discovery to targeted proteomics, lists of candidate biomarkers for oral cancer, and new techniques for the visualization of biological networks and phylogenetic supertrees.

Keywords: Cancer Biomarker, Protein Prioritization, Biological Network Visualization, Lateral Gene Transfer, Tree Visualization.

RESUMO

HEBERLE, H. **Métodos computacionais em Biologia: biomarcadores de câncer, redes de proteínas e transmissão lateral de genes**. 2019. 164 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

A Biologia Molecular é um ramo da Ciência de grande importância. Apesar de estudar entidades microscópicas, o volume e a complexidade das informações são imensos. Suas aplicações são variadas e podem ser de interesse global, como a disseminação de genes de resistência a antibióticos entre bactérias e novos métodos para diagnóstico e prognóstico de câncer. Entendendo os mecanismos biomoleculares, cientistas podem definir tratamentos para doenças, apoiar as decisões tomadas pelos pacientes, identificar a influência da microbiota intestinal sobre as condições físicas e psicológicas, encontrar causas e fontes de resistência microbiana aos antibióticos, entre muitas outras aplicações. A Ciência da Computação desempenha papéis-chave nesse contexto, como permitir análises complexas de dados por especialistas, criar modelos que simulam estruturas e processos biológicos e fornecer algoritmos para extrair informações codificadas em dados biológicos.

Durante meu doutorado, exploramos esses mecanismos em três níveis principais: quantificação de proteínas a partir de células, análise de interações que ocorrem dentro das células e comparação de genomas e seus históricos. Este manuscrito relata diferentes projetos, quatro deles já publicados em revistas científicas. Eles compreendem a descoberta de proteínas candidatas a biomarcadores de câncer, a análise visual de redes de interação proteína-proteína e a análise visual da transferência lateral de genes em árvores filogenéticas bacterianas. Aqui, explicamos esses projetos e as principais descobertas associadas ao uso de métodos computacionais. Entre os resultados estão a avaliação da estabilidade dos métodos de ranqueamento e assinaturas aplicados aos dados proteômicos de descoberta, uma nova abordagem para selecionar proteínas candidatas desde a descoberta até proteômica direcionada, listas de candidatos a biomarcadores para câncer oral e novas técnicas para a visualização de redes biológicas e supertrees filogenéticas.

Palavras-chave: Biomarcadores de Câncer, Priorização de Proteínas, Visualização de Redes Biológicas, Transmição Lateral de Gene, Visualização de Árvores.

LIST OF ABBREVIATIONS AND ACRONYMS

AMR	antimicrobial resistance
BP	biological processes
DCV	double cross validation
EC	extracapsullar
ELAP	Emerging Leaders in the Americas Program
EPS	expressed prostatic secretions
FDR	False Discovery Rate
GO	Gene Ontology
HGT	horizontal gene transfer
HHSVM	Hybrid Huberized Support Vector Machines
HPA	Humam Protein Atlas
ITF	invasive tumor front
LC	Liquid Chromatography
LFQ	label-free quantification
LGT	lateral gene transfer
LNBio	Biosciences National Laboratory
MS	Mass Spectrometry
NCF	Network-constrained forest
NDM-1	Delhi metallo-β-lactamase
NJ	Neighbor Joining
NSC	Nearest Shrunken Centroids
OC	organ-confined
OSCC	oral squamous cell carcinoma
PPI	protein-protein interaction
RF	Random Forests
RF	Robinson-Foulds
RFA	Recursive Feature Addition
RFE	Recursive Features Elimination
SPR	subtree prune-and-regraft
SRM	Selected Reaction Monitoring
SVM	Support Vector Machine

CONTENTS

CHAPTER

1

INTRODUCTION

This work applies Bioinformatics to discovery of proteins related to cancer, analysis of proteomics data, visual analysis of biological networks, and visual analysis of phylogenetic supertrees. Throughout this dissertation you will encounter concepts from Data Analysis and Data Mining, Information Visualization, Proteomics, Molecular Biology, Systems Biology and Evolution. In this chapter we summarize each project associated to this doctorate, defining their background and our collaborators, and giving an overview of each research topic and achieved results. Then we introduce the general motivation and explain how the results of this thesis can impact Science and Life; and conclude the chapter summarizing the essential contributions.

1.1 Research projects

To complete this doctorate, we had the participation of researchers from different institutions. This resulted in four major published articles and two additional that are on preparation. The main collaborators, divided by institution, are:

- Biosciences National Laboratory (LNBio): Dr. Adriana Franco Paes Leme, Dr. Gabriela Vaz Meirelles, Romenia Ramos Domingues, and Dr. Daniela Campos Granato;
- University of Campinas (UNICAMP): Dr. Guilherme Pimentel Telles;
- Dalhousie University: Dr. Robert Beiko.

We published the first work in the BMC Bioinformatics (HEBERLE *et al.*, 2015). The article describes a web tool for the comparison of sets through Venn diagrams that we designed during my Masters program. It also comprises one use case on candidate biomarkers and one on comparison of gene lists from a phylogenetic tree, both conducted during my doctorate in collaboration with Dr. Gabriela Vaz Meirelles. The work has a high impact in Biology, becoming highly cited in Computer Science according to the Web of Science after two years of publication.

Our second work was published in the journal Oncotarget (KAWAHARA *et al.*, 2015) in collaboration with LNBio, UNICAMP and other institutions, where we applied computational models to mine quantified proteins from healthy, carcinoma, and melanoma cancer cell samples. In this project, we integrated visualization, statistical, and machine learning methods to define a small list of differentially expressed proteins, establishing a pipeline to **select candidate biomarkers** to targeted validation. Three feature selection algorithms, namely, Beta-binomial, Nearest Shrunken Centroids (NSC), and Support Vector Machine - Recursive Features Elimination (SVM-RFE) , indicated a panel of 137 candidate biomarkers for carcinoma and 271 for melanoma, which were differentially abundant between the tumor classes on follow-up experimental validation. The proposed integrative analysis allowed to pre-qualify and prioritize candidate biomarkers from discovery-based proteomics to targeted Mass Spectrometry (MS) [1]. Section 2.2.1 details this work.

Then, CellNetVis, a network visualization web-system, was developed and published in the BMC Bioinformatics (HEBERLE *et al.*, 2017), having obtained the Best Paper Award during the BIOVIS workshop at ISMB in 2017, in Prague. We worked with researchers from LNBio and UNICAMP to create a unique dynamic graph layout for the analysis of biomolecular networks linking the topology to a visual representation of a cellular structure. Before CellNetVis, this type of interactive dynamic visualization of networks on the defined condition was not available. The new approach allows specialists interact with the network represented over a consistent cell diagram. Chapter 4 presents this work.

The fourth work was published in the Nature Communications, in collaboration with LNBio and other institutes (CARNIELLI *et al.*, 2018). We assessed the predictive power of sets of proteins in a targeted proteomics phase and obtained proteins and peptides signatures that may help with oral cancer prognostic. Section 2.3 details this work.

Given our experience on biomarkers discovery, we decided to evaluate methods that rank proteins and select a subset of proteins that better differentiate cancer conditions. We analyzed the usage of classical Machine Learning algorithms and examined the limitations of discovery proteomics data sets. Our collaborators from LNBio designed an experiment and tracked the expression of four recombinant proteins that were added in biological samples. We used this data set to estimate the algorithm's performance and propose an approach based on stability of proteins, which is detailed in Chapter 3.

The last work is not complete, and it is being done in collaboration with Dalhousie University where I worked during an internship funded through the Emerging Leaders in the Americas Program (ELAP) and supervised by Dr. Robert Beiko. In this project, our goal is

[1] In the discovery phase we have a small number of samples and we do not know what proteins we may find; in targeted phase, we have a defined and small list of interesting proteins that came from the discovery phase. The number of samples usually increase in the targeted phase. For these reasons, we may say that targeted proteomics is generally more controlled and precise than discovery proteomics.

to develop a visualization system for the analysis of bacteria phylogenetic supertrees [2]. With this technology, we believe that we will be able to discover lateral gene transfers and others evolutionary events. In Chapter 5, we explain how phylogenetic we can build supertrees and what we have done to visualize thousands of additional edges that turn a supertree into a phylogenetic network.

1.2 Applied Computer Science

Molecular Biology is a branch of Biology of immense relevance with a high volume of information and vast complexity. By understanding biomolecular mechanisms, it is possible to define causes and treatments for diseases, provide decision-making power to individuals predisposed to a disease condition, identify the influence of intestinal microbiota over physical and psychological conditions, identify cause and source of microbial antibiotic resistance, among many other applications.

Computer Science plays key roles in this context, such as enabling complex data analysis by specialists in biological areas, creating models that simulate biological structures and processes, and providing algorithms for extracting information encoded in biological data.

During my doctorate, we have worked on projects involving the discovery of proteins candidates for oral cancer biomarkers, the prioritization of proteins as candidate biomarkers using discovery proteomics data, the visual analysis of protein-protein interaction networks, and the visual analysis of bacterial phylogenetic trees and lateral gene transfers. This dissertation divides our contributions in the form of four central chapters: in chapters 2 and 3, we describe three results related to applying computational methods and propose pipelines to find candidate proteins for biomarkers; in Chapter 4 we describe a new approach for visualizing protein-protein interaction networks; and in Chapter 5 we present an approach for visualization of phylogenetic trees and networks in the context of bacteria lateral gene transfers, such as those linked to antibiotic resistance.

We wrote this text meaning to provide an appropriate understanding of the motivation, concepts, results and the importance of Bioinformatics for those who are not acquainted with the field. For this reason, we tried to simplify the language used in each chapter, especially when a published article with more detail is available.

[2] A phylogenetic supertree is a representation of the combination of other smaller phylogenetic trees. For instance, each gene could form a tree and the set of gene trees could form a bigger tree representing the evolution of all genes combined.

1.3 Contributions

With the projects completed in this doctoral dissertation, we advanced primarily on Bioinformatics and cancer related fields. We contributed to those making available open-source tools that allows researchers perform analyzes in a way that was not feasible before, defining approaches for analysis of biological data, reporting protein and peptide signatures that are candidates for oral cancer, reporting the stability of computational methods in the process of prioritization of proteins from discovery to targeted proteomics, and on other ways by having all the projects and findings reported in this text and on articles published in high impact journals. Below, we summarize our contributions:

- a review on current state of art computational methods and tools for solving biological problems from the fields of biomarker discovery, biological network visualization and phylogenetic tree visualization;
- pipelines and methods based on stability to prioritize proteins for targeted proteomics or further validations;
- a controlled assessment of discovery proteomics quantification and stability of methods for protein prioritization;
- a new open-source web-system for visualization of biological networks onto a cellular structure diagram;
- and a new open-source web-system for visualization of phylogenetic supertrees, shared genes and candidate lateral gene transfers.

CHAPTER

2

CANDIDATE CANCER-BIOMARKERS IN PROTEOMICS

In this chapter you will encounter the central motivation and concepts for the study of cancer biomarkers. We introduce basic biological notions and explain how we can represent biological information in data sets. Then, we demonstrate how computational methods select proteins as potential candidate biomarkers by reporting three studies on biomarkers published during this doctorate:

> KAWAHARA, R.; MEIRELLES, G. V; HEBERLE, H.; et al. Integrative analysis to select cancer candidate biomarkers to targeted validation. Oncotarget, v. 6, n. 41, p. 43635—43652, 2015. (KAWAHARA *et al.*, 2015)

> HEBERLE, H.; MEIRELLES, G.; DA SILVA, F. R.; TELLES, G. P.; MINGHIM, R. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics, v. 16, n. 1, 2015. (HEBERLE *et al.*, 2015)

> CARNIELLI, C. M.; MACEDO, C. C. S.; DE ROSSI, T.; et al. Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. Nature Communications, v. 9, n. 1, p. 3598, 2018. (CARNIELLI *et al.*, 2018)

2.1 Cancer biomarkers

Cancer is a genetic disease of worldwide concern. In 2018, the estimated cancer incidence in the world was of 18.1 million new cases. During their lifetime, one in 5 men and one in 6 women develop cancer, and one in 8 men and one in 11 women die from the disease (The International Agency for Research on Cancer, 2018). In Brazil, the estimated number of new

cases for 2018 was 600,000, from which 420,000 are non-melanoma skin cancer (Instituto nacional do câncer, 2017). With the online visualization tool *Cancer Today* (GLOBOCAN, 2018), we can explore estimates of incidence, mortality and prevalence of 36 specific cancer types, combined in 185 territories of the world in 2018, by sex and age.

Despite the treatments of cancer having improved, the death rate is still high. One factor that contributes to the number of deaths is the time that it takes for a cancer patient to be diagnosed. For instance, oral problems are prevalent in the population and patients may think a tumor is just a sore or something that will pass in a few days. As a consequence, oral cancer patients may identify their condition only when it is already in a late stage.

When a simple noninvasive test is feasible such as the human immunodeficiency virus (HIV) rapid test (GUILLON *et al.*, 2018), a person who has a high risk of cancer could test more often. Doing so, if the test is positive, the person can proceed to a more accurate test. Forthwith, the probability of completely eliminating the cancer cells increases with early tumor detection. As an illustration, a person who has cancer history in his family or has habits linked to cancer, such as daily smoking, could test himself in yearly or monthly basis.

If a person has cancer history, the colonoscopy is recommended to be done after reaching the age of 30 and repeated every 10 years (REX *et al.*, 2017). While this test is invasive, other less accurate tests are not. A patient could test with a noninvasive method more often than with a colonoscopy. With a positive result, the patient could think about anticipating the colonoscopy.

From the cited examples as well as from so many other reasons, we can understand the great importance of finding noninvasive techniques for the diagnosis and prognosis of cancer. In this doctorate, we studied computational methods applied to indicate proteins as candidate biomarkers. These proteins can potentially differentiate cancer types, cancer cells from healthy ones, or predict other outcomes such as the probability of recurrence.

A biomarker is a substance, structure or process of the body that can be measured and reproduced with precision. It is any measurement of chemical, physical or biological interaction between the biomolecular system and the potential treat (STRIMBU; TAVEL, 2010). The same way the levels of sugar in our blood are biomarkers of diabetes, specific **proteins** can be markers of cancer.

2.1.1 Proteins

A cell is formed by various molecules and carries around 17,294 protein-coding genes. Each gene encodes a different protein, and different agents can modify a protein after its production. A cell can express each protein in varied levels, that is, we can detect in it different amounts of proteins. For instance, cells from our neck express proteins that cells from our feet do not.

Proteins have particular functions, for example, structural or enzymatic (KIM *et al.*,

2014). From the amount of coded proteins in a commonly used human cell line, around 51% have functions related to a regulatory process such as cell division, cell communication and cytoskeleton organization; and 48% have core functions, like lipid metabolic process, transcription, transport, and DNA replication (BECK *et al.*, 2014). The same protein can likewise have multiple functions and take part in different biological processes.

When we determine a set that differentiates two conditions, we can search for the functions they might be expressing. These functions, pathways and other biological information about the proteins help us understand the system that makes a cell work. Thus, biomarkers are essential not only for classifying samples but also for scientists to better explain the mechanisms of biological systems, create hypotheses, and guide the future experiments. Finding a set of disease-associated proteins can lead the researches in the *right direction*. Studies such as the recent case of a new immunotherapy used to cure breast cancer (ROSENBERG *et al.*, 2018) are achievable due to the great volume of information available about the breast cancer and healthy cells' mechanisms nowadays, as well as how our systems work, e.g., the immune system.

2.1.2 Proteomes and Proteomics

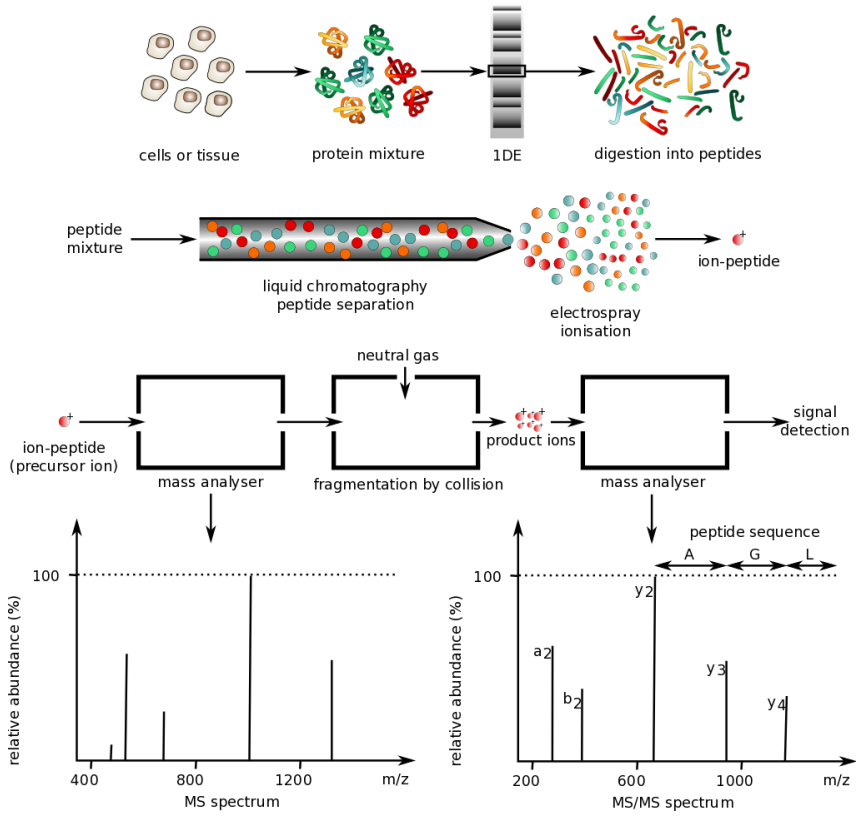
A proteome is the set of all proteins expressed by an organism. The field of Proteomics identifies, quantifies and describes those proteins, their expression and their changes under the influence of biological and chemical influences. Proteomics is divided into *expression proteomics* and *cell-map proteomics*. The first is the study of global changes in expression, while the second is the study of interactions between proteins through the isolation of protein complexes (ANDERSON; ANDERSON, 1998; BLACKSTOCK; WEIR, 1999).

In Figure 1 we illustrate a protocol for the identification and quantification of proteins from biological samples. In summary, after extracting the proteins and breaking them down into peptides, the Liquid Chromatography (LC) coupled with Mass-Spectrometer separates and identifies them. The expression of each protein is inferred with support of protein sequence databases.

In discovery-based proteomics (*discovery proteomics*), we do not have a list of proteins to be identified (reason for the name *discovery*). The enormous diversity of proteins makes challenges the detection and quantification of a specific protein. In contrast, in target-based proteomics (*targeted proteomics*), we have a (small) list of interesting proteins and more accurate quantification. For this reason, targeted proteomics can be managed to increase biological knowledge from discovery proteomics studies and achieve accurate quantification by increasing the possibility of investigation and allowing the robust quantification of peptides and proteins.

Currently, the number of samples is always smaller than the number of proteins in discovery proteomics. In Figure 2, we present an example of proteomics data where columns represent proteins and rows represent samples. Other example is the study made by Kim *et al.*

Figure 1 – A general representation of a mass spectrometry protocol. Protein mixture is extracted from cells or tissues and digested into peptides. The peptides are separated and form ion-peptides through a process called liquid chromatography and electrospray ionization. Each ion-peptide mass is quantified and used to form peptide sequences in the end of the process.



Source – Philippe Hupé (KARPIEVITCH *et al.*, 2011), licensed under CC BY-SA 3.0.

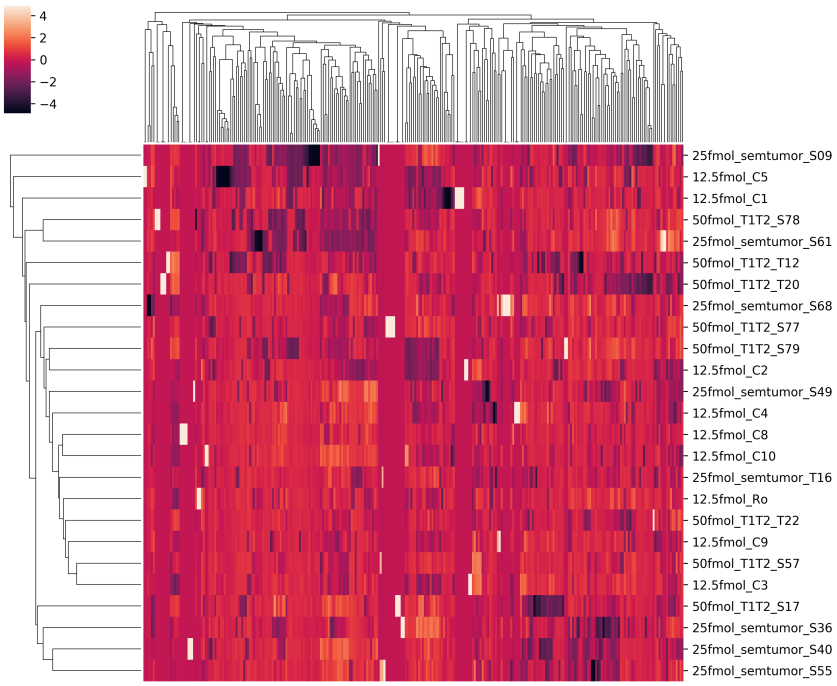
(2012), where they used a data set formed by 624 proteins and only 32 samples, divided into two classes. This is the typical scenario in discovery proteomics and has negative consequences when using statistical and computational methods. We discuss these limitations in the following sections, presenting alternatives to control the bias of selected biomarkers. In the next chapter, we present more alternatives that are being utilized primarily with transcriptomics [1] data sets.

Quinn *et al.* (2007) suggest that the number of samples per class should be at a minimum around five times the number of variables to avoid the *curse of dimensionality*. In the referred prostate data set, the ratio is 0.05 samples per protein. This is the essential challenge when using computational methods to show a panel of candidate proteins.

One characteristic of cancer cells is that they lose their original specialized functions. Additionally, cells from different cancer types differentiate less than regular cells from different organs (LODISH *et al.*, 2003). Still, through Proteomics protocols, we can find sets of differentially expressed proteins in cells from particular regions of a tumor, from different cancers, or

[1] "Transcriptomics is the study of the transcriptome - the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell - using high-throughput methods, such as microarray analysis." (Springer Nature Publishing AG, 2019)

Figure 2 – Heat map of a proteomics data set. Columns represent proteins, rows represent samples. Protein intensities are represented by colors. The number of variables is greater than the number of samples. Dendrograms represent hierarchical clustering of euclidean distances between rows and between columns.



Source: Elaborated by the author.

even from secretion of different cells (e.g. with and without tumor).

2.1.3 Finding candidate biomarkers

Defining a robust biomarker is a complex and long process and requires scientists from different fields. Here, we use the term biomarker *candidates* to point out that there are validation steps before considering them biomarkers, including the clinical trials. The following steps simplify the Discovery and Test Validation processes defined by Guillon *et al.* (2018):

1. define/collect the target samples, for instance, from cancer and healthy groups;
2. identify the proteins and their expression in each sample;
3. use computational/statistical methods to define the proteins that can discriminate the groups of samples and analyze the link with clinical data;
4. validate the markers using an independent set of samples;
5. validate the markers in clinical trials.

According to the WHO (2001), a biomarker must be relevant and valid. The **relevance** refers to the capacity of a biomarker answer important health questions, being of particular public

concern. The **validation** refers to the study of how true a marker is, revealing, for instance, the false positive rate. We can divide validation into three categories:

1. **measurement validity**: suggests how true are the expression of proteins we find;
2. **internal study validity**: refers to the controlled tests performed with the available data sets, for instance, the precision and recall of a classifier;
3. **external validity**: the extent to which the results found can be generalized to other populations, for instance, clinical trials.

We limited the validation of candidate biomarkers in this dissertation to the *internal study validity* and, in a way, we adopted the *measurement validity* idea to define the fundamental study of Chapter 3. Paczesny (2013) defines the *internal study validity* by different steps:

1. discovery: mass-spectrometer combined with sequence data identify proteins and their expression; computational methods report the best (or a good) set of proteins that discriminate the classes, computed based on data matrices that represent the quantity of each protein in each sample;
2. validation: the results found in the discovery part are validated using statistical and computational methods, literature review about the selected proteins, metabolic pathway analysis, protein-protein interaction networks analyzes, and others related to the knowledge about the biological system;
3. qualification: more experiments are performed targeting the selected proteins.

In the first phases we have the smallest sample sizes (e.g. 40 samples) and the greatest number of proteins (e.g. 2,000 proteins). In the further phases this characteristic inverts, changing to a small number of proteins and a bigger cohort of patients. For instance, we could have a clinical trial with 1,000 patients with which we investigate the power of quantifying 5 specific proteins to predict the chances of recurrence of cancer.

2.1.4 Discovery-to-targeted analysis

To discover a small list of key proteins is the main goal of using computational methods with the proteomics data. Thousands of proteins are reduced to dozens or units. The proteins we seek must follow some requirements, such as being differentially expressed in the studied groups or being capable of differentiating them if a predictor is built based on them, representing a logic formed by distinct proteins.

The most popular way to select candidate proteins is by ranking them, prioritizing the most likely to discriminate groups accurately. We can divide ranking methods into two main

groups: filter and wrappers. Some common filters that are adopted in Proteomics are *Student's t test* (HAYNES, 2013), *Wilcoxon rank sum test* (DENG; MA; PEI, 2004), *Kruskal-Wallis rank sum test* (BOULESTEIX; STRIMMER, 2007), and *beta-binomial test* (PHAM *et al.*, 2009). Proteins are ranked based on the p-values from the statistical tests, using the value of 0.01, 0.05 or 0.10 as a cutoff. Other methods are based on multivariate supervised models, the wrappers, e.g. NSC (TIBSHIRANI *et al.*, 2002) and SVM-RFE (THURSTON *et al.*, 2011) that can output a list of important features or a rank.

The use of multiple methods is crucial to increase the validity of the rankings. Due to the small number of samples and the differences between methods, it is difficult to determine what is the best method to discover candidate biomarkers. Many studies report new methods suggesting they outperform the state of art techniques, but they also use small data sets.

Some studies use biological knowledge to validate the results, such as the number of selected proteins that are already known as liked to cancer. Despite it makes sense to search for proteins that are biologically related to cancer, these may not be the best proteins to build classifiers. There are two concepts that we need to consider: the quantification and the biological meaning. If a protein is identified as a cause of cancer, it does not mean that we will be able to differentiate the healthy and the cancer conditions based on the expression values of this protein. Imagine a situation where this protein is found with the same expression in different conditions, but in cancer samples it was changed by post-translation modifications. The current Proteomics technology is not capable of identifying such modifications in high-throughput scale, since it breaks all proteins before quantifying them, thus, loosing such type of information. On the other hand, a protein that is highly differentiated between the two classes might be just a product with no important function to the surveillance of the cancer cells. Thus, the number of reported articles associating a protein to a disease may indeed be used as a validation in the sense of biological function or cause, but it is not the best validation criteria for the prediction power of a set of proteins.

In a context, additional information increases the strength of the evidences and different outputs must be critically analyzed before further high cost validation steps. We consider that with the available technologies it is not possible to select a method as being the best one. Under the circumstance of a few samples, many techniques should be used to prioritize proteins. Then, the proteins must be analyzed by specialists from the different areas (e.g., computer scientists, biologists, statisticians and others) in critical assessments. Tools from Machine Learning and Information Visualization are fundamental in this phase.

Despite all the mentioned limitations, the researches on this field are important to the development of life quality and human health and cannot wait the quantification technologies to be improved. Researchers need to design experiments in a way that they can discover important information about cancer cells and advance health sciences. They need have in mind that the results are biased and, thus, the false positive and false negative rates are great. This means that

they might find results in the clinical studies that contradict the results found by the discovery and targeted proteomics. Showing the variability of methods' results clearly is one of the contributions of this dissertation.

In the following sections we report two different analyzes in discovery proteomics and one in targeted proteomics, published during this doctorate. We start explaining the double cross validation scheme proposed by Christin *et al.* (2013). It was used in combination with three other ranking methods in two studies. In the first, we selected proteins that differentiated healthy, carcinoma, and melanoma cells (Section 2.2.1). In the second, we illustrated the use of InteractiVenn (HEBERLE *et al.*, 2015), a tool for comparison of lists of elements developed during my Masters (Section 2.2.2). Then, we present a study where we have found a set of proteins that are candidate biomarkers for oral cancer, discovered by the use of saliva samples (Section 2.3). In the latter, both discovery and targeted proteomics were adopted to reduce the number of proteins, but the prediction power of signatures were measured particularly in the targeted phase.

2.2 Double cross validation for biomarkers discovery

The small number of samples makes it difficult to evaluate methods using the common cross validation schemes such as K-Fold and Leave-One-Out cross validation. Ideally, we first divide the data set into **train** and **test**, and cross validate the train set. The test set, commonly called the **independent test**, is used to verify if the estimates of scores computed within the cross validation can be reproduced.

Due to the data set size, we can have two worse scenarios when we randomly split the samples. The first happens when the test samples are too easy to classify, implying the score to be overoptimistic. The second is the opposite case where the test samples have characteristics that could be considered out-liars in comparison with the remaining samples in the train data. If a cross validation pipeline is executed only one time, one could fall into one of the cases.

Two possible alternatives for a critical evaluation of estimated scores are (1) to repeat the cross validation pipeline and (2) a cross validation inside another cross validation. For example, Christin *et al.* (2013) used the double cross validation (DCV) to evaluate different sizes of data sets and different classifiers for feature selection in proteomics critically. Based on a set of 50 samples, they derived more data sets of different between- and within-class variability, well illustrating the use of double cross validation and limitations of biomarker discovery in proteomics. Two of the methods used by them were based on defining cutoff values and thresholds. In contrast with a simple cross validation, a double cross validation is less overoptimistic. A cutoff value is utilized, for instance, to define the selected proteins as the top-N elements in a rank. The DCV score is, formerly, the mean score of the independent test sets' scores - the ones that were not used to select proteins and define Ns or thresholds.

In the following subsections we report one study made in collaboration with the LNBio (2.2.1), and one study created to illustrate use of InteractiVenn (2.2.2). Both are example of candidate biomarkers selection using proteomics data.

2.2.1 Integrative analysis to select cancer candidate biomarkers to targeted validation

This section contains text adapted, summarized, and simplified for the purpose of inter-disciplinary contribution, from the following article which is licensed under Creative Commons Attribution 3.0 (<http://creativecommons.org/licenses/by/3.0/>). For results and findings in the Cancer domain, specially if citing our work, please read the original article.

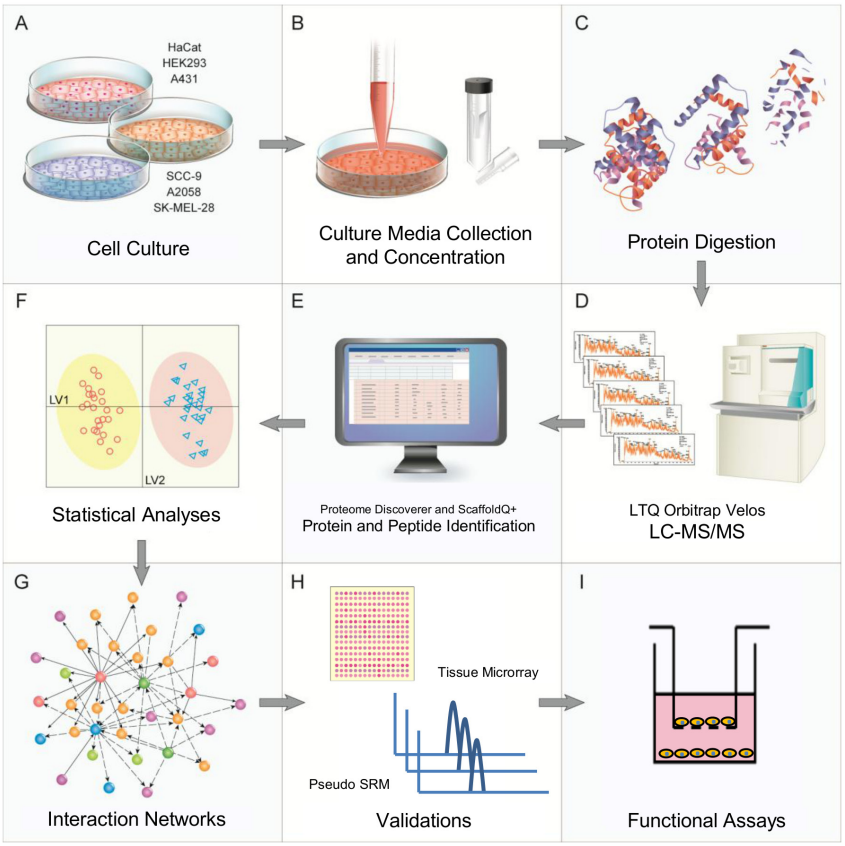
> KAWAHARA, R.; MEIRELLES, G. V; HEBERLE, H.; et al. Integrative analysis to select cancer candidate biomarkers to targeted validation. Oncotarget, v. 6, n. 41, p. 43635—43652, 2015. (KAWAHARA *et al.*, 2015)

In a work in collaboration with the LNBio, we defined an experimental pipeline (Figure 3) that is a bridge between discovery MS and targeted MS. This pipeline comprises four steps: discovery proteomics, feature selection analyzes, biological information analysis and targeted validation (Figure 3). As a proof of concept, melanoma (A2058 and SK-MEL-28), skin and tongue-derived carcinoma (A431 and SCC-9, respectively) and non-cancerous cell lines (HaCaT and HEK293) had the protein content of their secretome collected, concentrated, trypsin digested and analyzed by mass spectrometry. State-of-the-art univariate and multivariate methods were later employed to identify the most differentially abundant proteins among the three classes. By compiling these data into integrative networks, they revealed cancer-specific biological information. These networks were able to characterize both carcinoma and melanoma cell archetypes and to point out pathways that could be potentially altered in each condition. Protein expression by tissue array in carcinoma and melanoma patients' samples and by saliva samples, as well as gene silencing and functional experiments in cell lines provided validation for the proposed pipeline.

The data set is formed by 18 samples divided into 3 classes (normal, carcinoma and melanoma), resulting in 6 samples per class, 3 samples per cell line. A total of 1,697 proteins were identified after proper proteomic protocol, which was processed to measure spectral counts. In comparison to intensity-based quantification, spectral counts requires less complex signal processing and numbers are readily available after protein identification (PHAM *et al.*, 2009). In Figure 4, we show the result of neighbor joining tree visualization technique applied to the counts, before and after the feature selection described in the next paragraphs.

We performed an unsupervised hierarchical clustering implemented in the MetaboAnalyst platform with the 1,697 proteins, which segregated the samples into two main classes. One is

Figure 3 – Experimental workflow and overview of the proteomics and bioinformatics analyzes, validations and functional assays. Cell cultures are used to extract protein to digestion into peptides (A to C). These are then quantified by LC-MS/MS (D) and the sequences and abundances are used for identification of proteins and peptides (E). Statistical analysis and networks analysis (F and G) are followed by experimental validations (H and I).

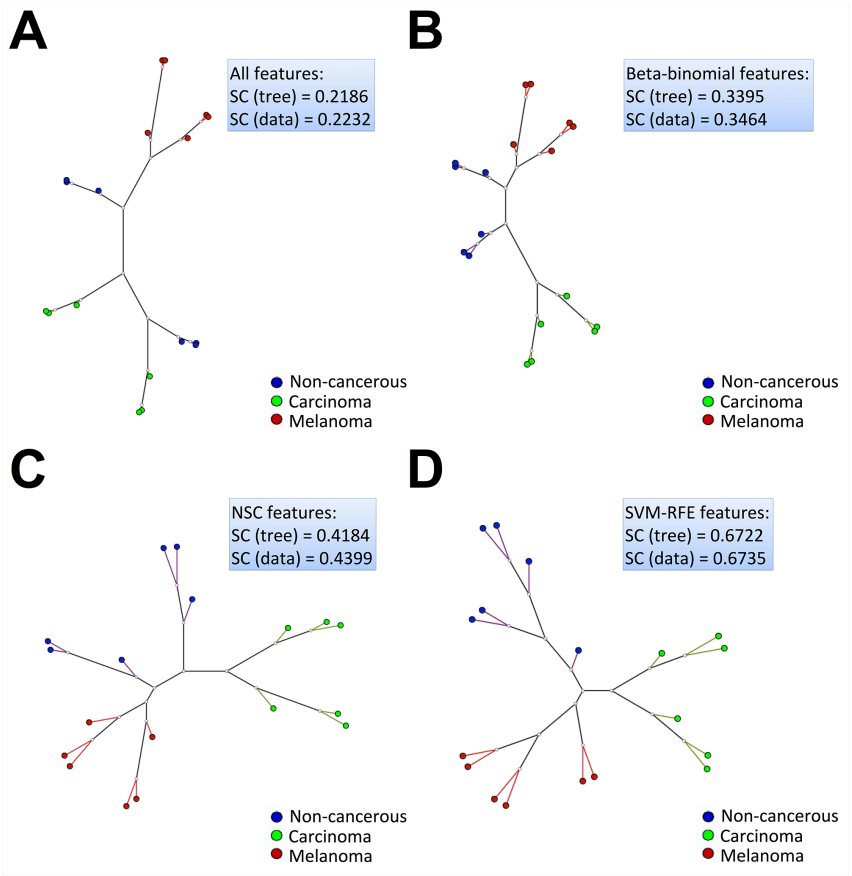


Source – (KAWAHARA *et al.*, 2015), licensed under CC Attribution 3.0

composed exclusively by melanoma cell lines and the other that is composed by carcinoma and non-cancerous cells (Figure 5A). Interestingly, the basal cluster segregated the cells according to their tissue of origin: from the epithelium-derived cell lines (SCC-9, A431 and HaCaT), from the skin-derived melanoma cells (SK-MEL-28 and A2058) and from the human kidney non-cancerous cells (HEK293), although a perfect group segregation for either non-cancerous or cancer cell lines was not observed. The result of this exploratory analysis indicated that melanoma's secretome is radically distinct from that produced by carcinoma and non-cancerous cells.

Aiming to compute the importance of features, we conducted the ranking analysis using particular techniques: univariate with Beta-Binomial (PHAM *et al.*, 2009), semi-multivariate with NSC (TIBSHIRANI *et al.*, 2002) and multivariate with SVM-RFE (THURSTON *et al.*, 2011). The Beta-Binomial test was created to test the significance of proteins to differentiate classes. It was specifically developed as a statistical tool for proteins spectral counts in label-free mass spectrometry-based proteomics (PHAM *et al.*, 2009). SVM-RFE is demonstrated experimentally

Figure 4 – Neighbor joining (NJ) clustering calculated from a Euclidean distance matrix of the secretome data set samples, considering **(A)** all features (1,697 proteins), **(B)**. Beta-binomial (601 proteins), **(C)** NSC (130 proteins) and **(D)** SVM-RFE (13 proteins) features. SC (tree) stands for silhouette coefficient calculated from the NJ tree and SC (data) stands for silhouette coefficient calculated directly from the original data of each analysis.

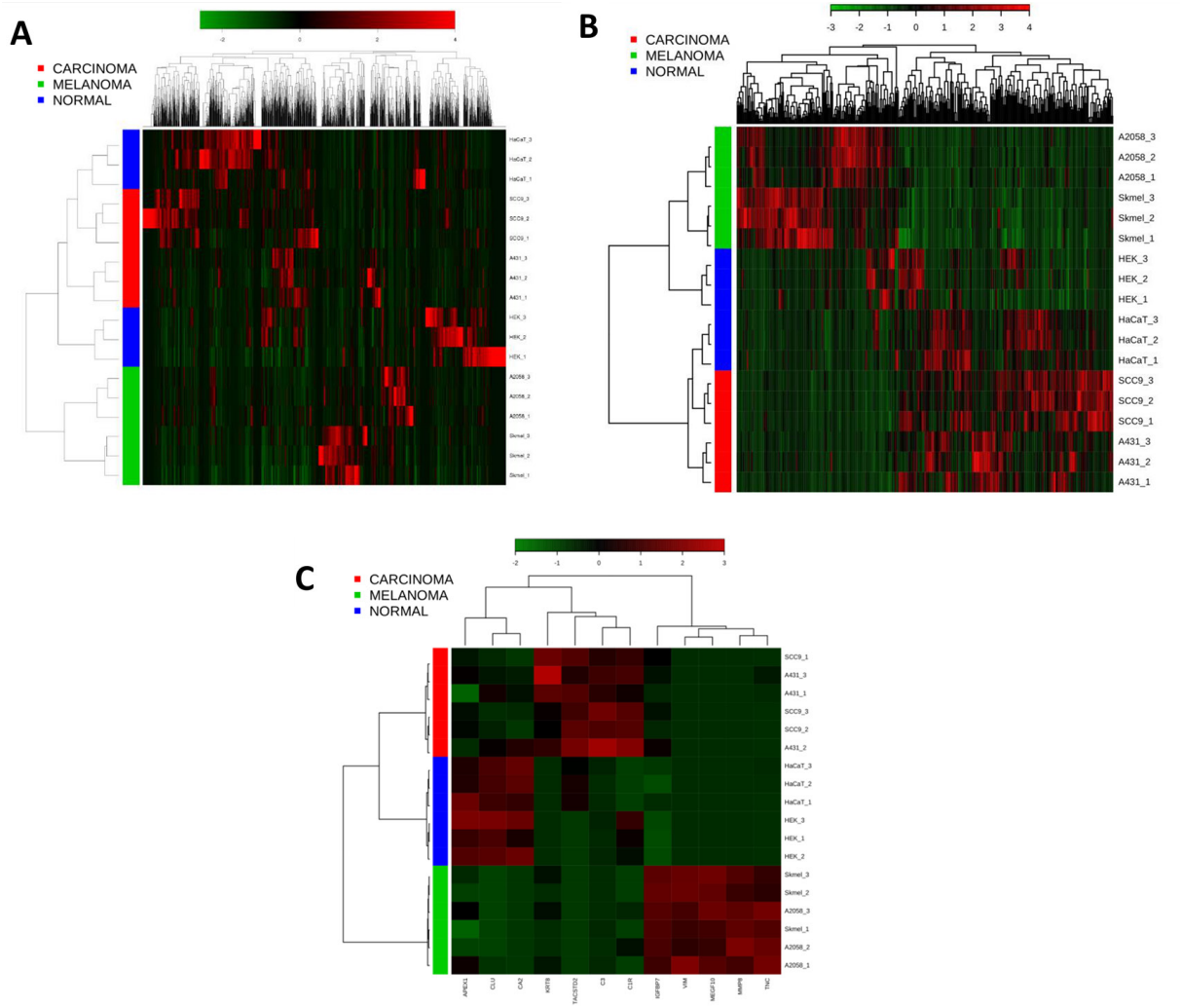


Source – (KAWAHARA *et al.*, 2015), licensed under CC Attribution 3.0

to select genes that improve the classification performance and have biologically relevance for cancer research. The idea is to create a model considering all features, remove the feature with the smaller weight in the SVM model, and repeat it, forming a final rank of features. The top features are the ones that survived to the recursive elimination for a longer time (THURSTON *et al.*, 2011). NSC was also demonstrated to be effective for selecting genes linked to cancer. It is based on the Nearest Neighbors algorithm, where for each centroid, the values of each feature are divided by its within-class variance and, then, if resulted values crosses zero, the respective features are eliminated (TIBSHIRANI *et al.*, 2002).

The SVM-RFE and NSC methods had their performance assessed in terms of double cross validation accuracy. These models presented 94.4% and 100% accuracy, respectively. We computed the final list of proteins using a cutoff of 0.05 for the Beta-Binomial p-value, a value of N from the double cross validation scheme to select the top-N features for SVM-RFE final rank, and a threshold value also from the DCV to use with NSC. Both algorithms were performed using the complete train set while their parameters were defined in the DCV. The scheme for

Figure 5 – Comparison of the three feature selection methods (Beta-binomial, SVM-RFE and NSC) used to identify differentially abundant proteins among carcinoma, melanoma and non-cancerous cells. **(A)** Clustering of the entire secretome data set before applying feature selection methods. From the 2,574 proteins identified and quantified by spectral counts, 1,697 (65.9%) compose the heat map. The 877 remaining proteins exhibited ≤ 2 spectral counts and were excluded from the analysis. **(B)** Clustering after applying feature selection methods. 603 significant differentially abundant proteins among melanoma, carcinoma and non-cancerous classes selected by Beta-binomial, NSC and SVM-RFE analyzes compose the heat map. **(C)** Clustering of the 12 significant differentially abundant proteins among melanoma, carcinoma and non-cancerous classes identified in the intersection of Beta-binomial, NSC and SVM-RFE analyzes. The secretome data set is composed by non-cancerous cells (HaCaT and HEK293), carcinoma (A431 and SCC-9) and melanoma (A2038 and SK-MEL-28) cell lines.



Source – (KAWAHARA *et al.*, 2015), licensed under CC Attribution 3.0

the combination of DCV and estimation of a good N/threshold was detailed by Christin *et al.* (2013). We re-implemented the scripts for defining N/threshold using DCV as an open-source script written in R language, for both SVM-RFE and NSC.

The Beta-Binomial, NSC, and SVM-RFE models retrieved 601, 130 and 13 proteins, respectively, that were differentially abundant among the three secretome classes. These proteins

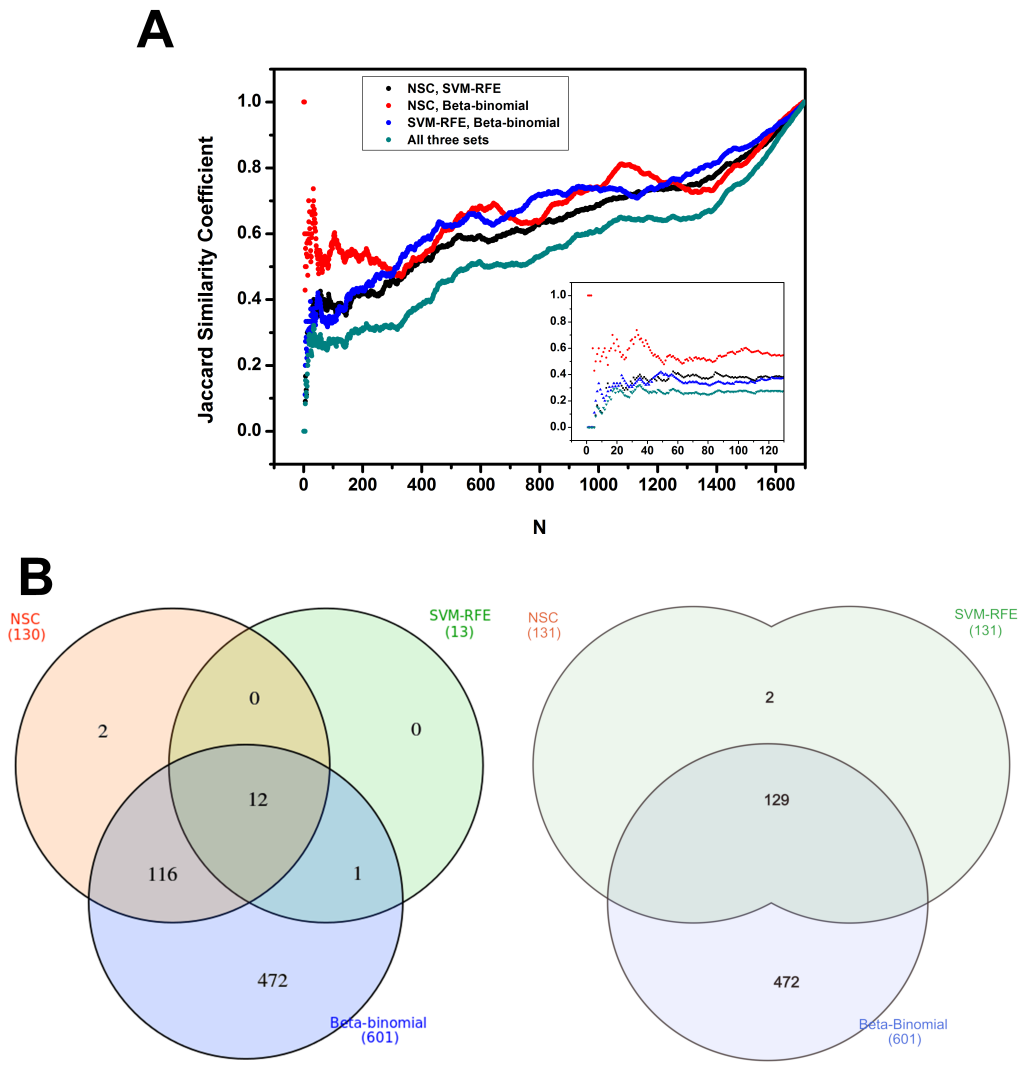
were further associated with each class after a decision boundary step (Supplementary Table S4 from (KAWAHARA *et al.*, 2015)). In this procedure, we link each protein to the class where it appears with a different expression in comparison with the other two classes. For instance, in Figure 5C, we can see that in the firsts proteins (APEX1, CLU, and CA2), the higher expressions are concentrated in the normal samples, indicating that if this protein is down-regulated in a cell, it is probably a cancer cell. The proteins in the middle (KRT8, TACSTD2, C3, and C1R) are more expressed in the carcinoma cells, while the proteins in the end (IGFBP7, VIM, MEGF10, MMP8, and TNC) are more expressed in melanoma cells.

The candidate biomarkers were further applied to perform the hierarchical clustering in heat maps using the MetaboAnalyst platform (Figure 5B). By this analysis considering only the selected features, the same-class cell lines were clustered together, which confirms the set of retrieved candidate biomarkers as good discriminating proteins (Figure 5B). From this set, the Beta-binomial, NSC and SVM-RFE models retrieved 135, 32 and 4 characteristic proteins for carcinoma and 269, 78 and 6 proteins for melanoma, respectively (Supplementary Table S6 from (KAWAHARA *et al.*, 2015)).

Furthermore, we compared the lists of candidate biomarkers using Venn diagrams (HEBERLE *et al.*, 2015) and plot of the Jaccard similarities between their ranks (Figure 6). The total number of proteins selected by each method is 601 for Beta-Binomial (p-value), 130 for NSC (threshold) and 13 for SVM-RFE (top-N). This comparison showed that the SVM-RFE optimal feature subset is almost completely shared by the NSC and Beta-binomial models (12 out of 13 proteins) and that the NSC optimal feature subset is practically completely shared by the Beta-binomial model (128 out of 130 proteins) (Figure 6B). Moreover, based on the Jaccard similarity coefficient, the comparison of protein rankings resulting from the three models is almost linear, not showing large variances in the similarity coefficient from the 10th to the 130th position in the ranking (green line, inset of Figure 6A). This means that the three models have almost a constant similarity coefficient (0.3) from the 10th to the 130th position in the ranking. From the 130th to the 200th position there is an increase in the slope of the curve reflecting an increase in the similarity coefficient, which would naturally occurs at a specific point when increasing N.

Notably, the SVM-RFE model was able to discriminate the three classes based on the smallest set of only 13 proteins (gene names: C3, CLU, MEGF10, MMP8, BANF1, VIM, APEX1, CA2, TACSTD2, KRT8, TNC, C1R and IGFBP7), of which only BANF1 was not retrieved by the other two methods. In contrast, as expected for a univariate method, the Beta-binomial model yielded the largest set of differentially abundant proteins, covering all the proteins that were retrieved by the two multivariate methods (except for two proteins from NSC) (Figures 6B and 6C. Notably, using only the 12 candidate biomarkers retrieved by the three feature selection methods, a perfect segregation among the carcinoma, melanoma and non-cancerous classes was also observed (Figure 5C).

Figure 6 – Comparison of the three feature selection methods (Beta-binomial, SVM-RFE and NSC) used to identify differentially abundant proteins among carcinoma, melanoma and non-cancerous cells. **(A)** Jaccard similarity coefficient vs. the optimal feature subset (N) retrieved by each method. **(B)** Venn diagrams showing the intersections among the optimal feature sets retrieved by the three methods and the union of NSC and SVM-RFE sets in comparison to Beta-Binomial.



Source – Figure adapted from (KAWAHARA *et al.*, 2015), licensed under CC Attribution 3.0

In addition to the hierarchical clustering and heat map analysis, we created similarity trees using Euclidean distances for the 18 samples. Figure 4 shows that the Neighbor Joining (NJ) trees were capable of showing the most similar elements of the set in agreement with the presented heat maps. In this work, a reasonable separation of the three classes was formed when the entire data set was considered in the NJ tree construction (Figure 4A) (silhouette coefficient, SC > 0.2). However, as shown by the previous unsupervised hierarchical clustering for the complete data set, the melanoma samples were the only ones that clustered together in the same or nearby branches connected to the same node, separated from the carcinoma and non-cancerous samples, which were distributed in different branches and did not show a perfect

segregation in their respective classes. On the other hand, the improvement in the NJ clustering and silhouette coefficients after each feature selection method demonstrates their performance in selecting candidate biomarkers (Figure 4B, 4C and 4D). The high DCV accuracy is explained by these trees, since the closer to 1 is the silhouette coefficient, indicating that the classes are homogeneous and different from each other, the easier is to a classifier perform well, even in a DCV.

By the end of feature selection, further experiments and analysis was performed to evaluate the importance of the selected features. For instance, most of the selected proteins have already been demonstrated to be associated with cancer. By searching them in the IPA and Human Protein Atlas Database, we found out respectively that 32% and 23% of the carcinoma candidates and 28% and 22% of the melanoma candidates were previously found to be associated with cancer.

The selected proteins were also investigated by querying protein-protein interaction networks and performing enrichment analysis and literature curation. A protein network anticipated a potentially important role for the set of candidate biomarkers in the carcinoma, which was especially related to the complement and coagulation cascades, whereas in melanoma, the pathways associated with the cell cycle, cell adhesion and ubiquitin-mediated proteolysis were highlighted as being among the most altered in this pathologic condition.

Our collaborators further tested the strength of the pipeline in selecting candidate biomarkers by immunoblotting, human tissue microarrays, label-free targeted MS and functional experiments. It is noteworthy that the proteins Complement Factor B (CFB) and Complement C3 (C3) were identified in significantly increased levels in oral squamous cell carcinoma (OSCC), compared to the adjacent normal tissue. Moreover, CFB knockdown decreased both the migration in the skin-derived epidermoid carcinoma (A431) cell line and chemotaxis in human macrophages.

The same feature selection analyzes were also performed for a published proteomics data set on prostate cancer (KIM *et al.*, 2012) to validate our approach. This additional study was published as a use case of our InteractiVenn web-based tool (HEBERLE *et al.*, 2015) and described in Section 2.2.2.

In conclusion, the proposed integrative analysis based on a discovery-to-targeted pipeline was able to pre-qualify potential candidates from discovery-based proteomics to targeted MS and can contribute to the next phases of biomarker development in translational initiatives to drive either patient stratification, decision making or intervention. Despite the DCV performed well for the studied data set, we demonstrate why this conclusion may not apply for other cases in Chapter 3.

2.2.2 Comparing feature selection methods for the discovery of candidate prostate cancer biomarkers

> HEBERLE, H.; MEIRELLES, G.; DA SILVA, F. R.; TELLES, G. P.; MINGHIM, R. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics, v. 16, n. 1, 2015. (HEBERLE *et al.*, 2015)

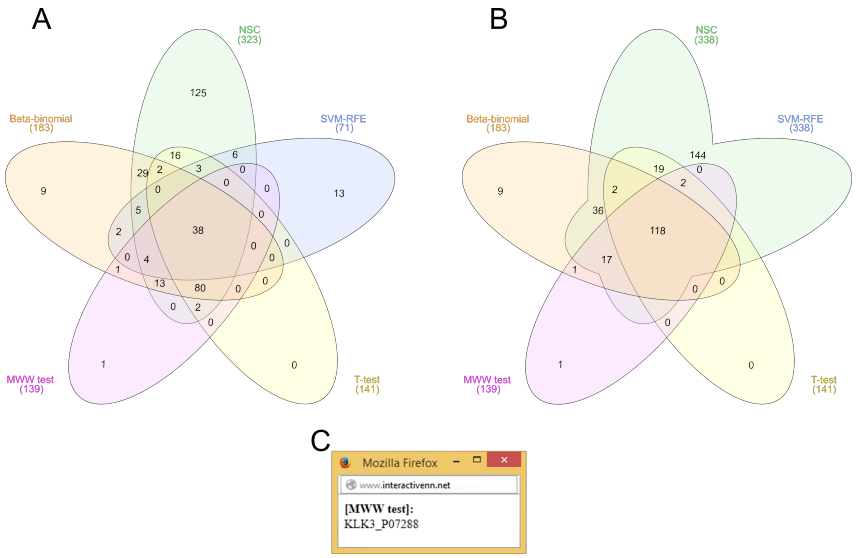
InteractiVenn (<http://www.interactivenn.net/>)[2] is a technique and a system for the analysis of sets through Venn diagrams that we developed during my Masters, in collaboration with researchers from LNBio, UNICAMP and Embrapa. In 2017 and 2018, it was announced as a highly cited paper in Computer Science by the Web of Science and exhibited by the library of the Instituto de Ciências Matemáticas e de Computação (ICMC) (Biblioteca Achille Bassi, 2018a; Biblioteca Achille Bassi, 2018b). Here, we replicate the use case described in the original article. We demonstrate how InteractiVenn can be adopted to support the analysis of lists of candidate biomarkers.

To show the usefulness of our tool, we have analyzed the published prostate cancer proteomic data set (KIM *et al.*, 2012), searching for candidate biomarkers through feature selection analyzes. Here, in order to generate lists of proteins sorted by relevance in discriminating the two classes in the data set (organ-confined and extracapsular prostate cancer cells), five methods were applied, including the three used before in the discovery-to-target pipeline (KAWAHARA *et al.*, 2015), the t test and the MWW test.

Based on the confidence level (p-value ≤ 0.05) for the univariate methods (Beta-binomial, t test and MWW test) and on the double cross validation procedure for the semi and multivariate methods (NSC and SVM-RFE), the top-N final ranked lists of candidate biomarkers resulted from each model were compared. In total, all five methods have shown 349 different proteins (union code: *ABCDE*). Figure 7A shows that all methods retrieved 38 common proteins, while the semi and multivariate methods have, in general, more exclusive proteins than the univariate ones. We can also see that the semi and multivariate methods exclusively share 6 proteins, whereas among the univariate methods, only one protein is exclusively shared by the Beta-binomial and MWW tests. **Union operations** allow us to see different patterns, for instance, by using the code BC to trigger the union of sets B (NSC) and C (SVM-RFE), we see that there are only 144 proteins in the semi and multivariate methods (Figure 7B). The approach adopted by InteractiVenn preserves the position and shape of the sets, allowing a smoother exploration. Other unions are also possible as well.

[2] Fork and contribute: <https://github.com/heberleh/interactivenn>

Figure 7 – Comparison of ranked lists of candidate biomarkers by five feature selection methods. (A) 38 proteins are shared by all methods, whereas the semi and multivariate methods show more exclusive proteins than the univariate ones; (B) 144 proteins are exclusively shared by the semi and multivariate methods; (C) KLK3 was retrieved as an exclusive protein by only the MWW test.



Source – (HEBERLE *et al.*, 2015), licensed under CC Attribution 2.0

Furthermore, seven proteins identified as candidate biomarkers in the prostate cancer cells in the work by Kim *et al.* (2012) were also verified in the same work by experimental biochemical methods and were searched in the Venn diagram sets built using the InteractiVenn tool: KLK3 (PSA), ACPP (PAP), SFN, MME, PARK7, TIMP1 e TGM4. Notably, from these proteins, KLK3 was the only one not validated as a candidate biomarker and, using InteractiVenn, we could observe that it was retrieved as an exclusive protein only by the MWW test (Figure 7C). Out of the other six validated candidates, four (ACPP, SFN, MME e TGM4) were located in the intersection among the three methods used in the discovery-to-target pipeline (KAWAHARA *et al.*, 2015), one (PARK7) was found in the intersection between Beta-binomial and NSC, and another one (TIMP1), in the intersection between NSC and SVM-RFE. Interestingly, none was located exclusively by the t test, suggesting that the three methods used in the pipeline described by Kawahara *et al.* (2015) could retrieve the best potential candidate biomarkers in their intersections. We explain this study in more details in Chapter 3.

2.2.3 A priori knowledge

The small number of samples limit the cross validation estimates. For this reason, it is also difficult to define a small set of proteins that better discriminate the conditions that are being studied with a low false positive/negative rate. If we generate all possible combinations of proteins, we will find many that have the same prediction power.

Some approaches integrate biological information to the multivariate models to reduce

random effect by applying biological bias. For instance, we could define that the lists of selected proteins (**signatures**) must have connections in the network that represent the biological interactions inside the cell. This would reduce the domain of possible signatures that would be tested.

Another approach would be filtering the initial set of proteins by means of clinical and biological information. Integrating the clinical data, for example, with the protein quantification data, one could filter the initial proteins, what would reduce the random effect in the feature selection process. Also, the final signature(s) would be, *a priori*, liked to clinical outcomes. This is the case of the project described in the following section (Section 2.3).

2.3 Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer

This section contains text adapted, summarized, and simplified for the purpose of interdisciplinary contribution, from the following article which is licensed under Creative Commons Attribution 4.0 (<http://creativecommons.org/licenses/by/4.0/>). For results and findings in the Cancer domain, specially if citing our work, please read the original article.

> CARNIELLI, C. M.; MACEDO, C. C. S.; DE ROSSI, T.; et al. Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. Nature Communications, v. 9, n. 1, p. 3598, 2018. (CARNIELLI *et al.*, 2018)

The most common type of head and neck malignant tumor is the oral squamous cell carcinoma (OSCC), ranked the eighth leading cause of cancer worldwide. OSCC exhibits a high prevalence and morbidity, with 300,000 new cases and 145,000 deaths per year worldwide (TODUA *et al.*, 2015). Standard multimodal management of OSCC is based on the tumor-node-metastasis (TNM) classification (EDGE; COMPTON, 2010), in which the tumor size and location and the presence of metastasis are used to define OSCC prognosis and treatment in the clinical setting (CHEN *et al.*, 2017). However, this system has several flaws, such as patients with the same TNM stage exhibit different clinical behaviors, different treatment responses and substantial variability in clinical outcomes (ALMANGUSH *et al.*, 2015; SILVA *et al.*, 2012). Despite efforts to improve imaging and therapeutic modalities, OSCC prognosis, including survival rates, remains poor and may widely vary, even in the early stages of the disease, e.g., 20-40% of occult metastases are detected at the initial diagnosis (SILVA *et al.*, 2012; WAAL, 2013; GANLY; PATEL; SHAH, 2012; ROBINSON *et al.*, 2016). Furthermore, OSCC recurrence rates range from 18 to 76% in patients undergoing standard treatment, and local relapse represents a clinical challenge for therapeutic management (SILVA *et al.*, 2012). Thus, the identification of complementary biological signatures that assist in the prognostic prediction of patients with OSCC is needed. Saliva testing may represent a promising noninvasive

tool to validate prognostic biomarkers, such as proteins, lipids, mRNA, miRNA and exosomes, and better classify patients into low- and high-risk groups (WINCK *et al.*, 2015; KAWAHARA *et al.*, 2016; CHEN, 2015; WANG; WANG; HUANG, 2015).

In collaboration with LNBio and others institutes, we explored the domain of possible signatures of candidate proteins for prognostic of OSCC, that is, identified good prognostic candidate signatures for OSCC patients. In this case, in contrast with applying computational methods in the discovery phase, we applied them in the target phase. For this reason, the characteristics of the data set are very different from the other two studies reported in this chapter. In this case, we have two different data sets: one represents the intensity of each peptide identified by targeted proteomics from 40 samples; the other represents the intensity of proteins that those peptides' intensities form.

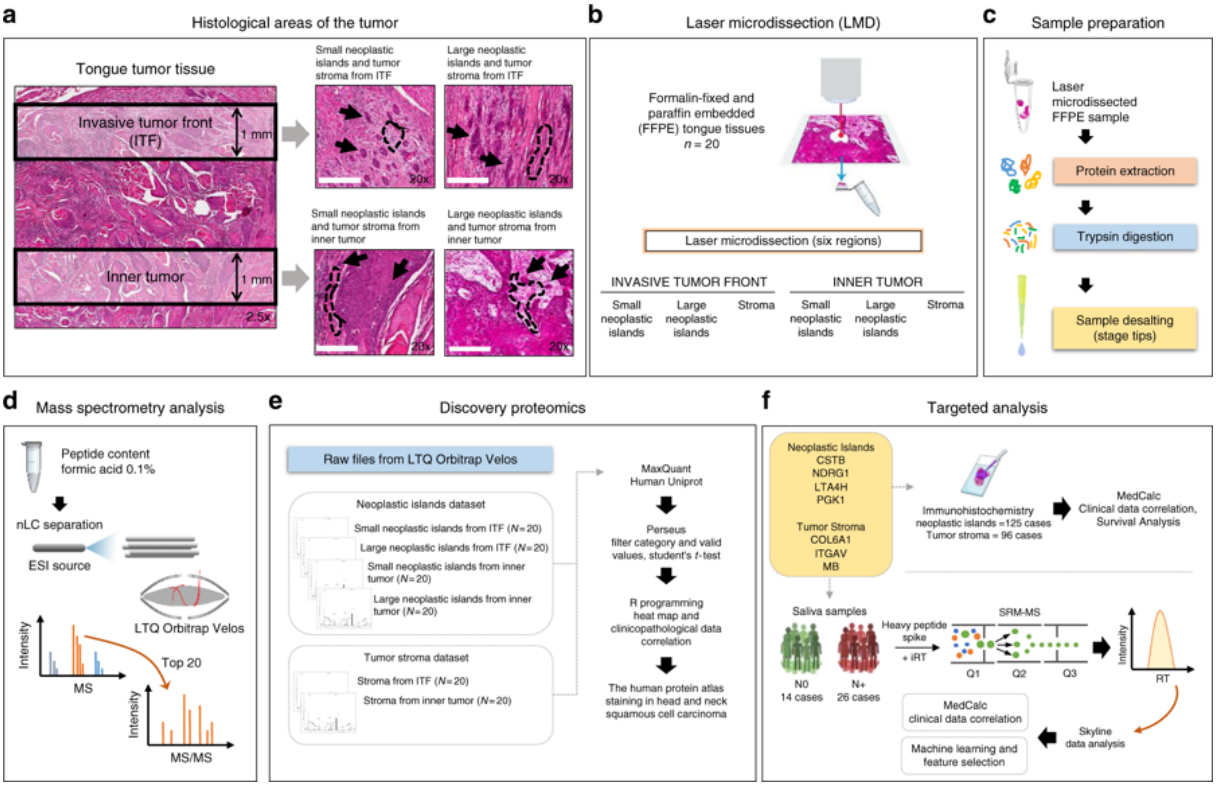
The small number of proteins is due to *a priori* clinical information and experiments that were capable of selecting a panel of most linked proteins for the prognosis of OSCC in previous discovery proteomics. We evaluated the prediction power of all possible signatures from both intensity matrices. One news articles and two news videos about this project were published by *Globo Comunicação e Participações S.A.*, explaining to the public the findings and their importance for society (Globo News, 2018; Globo Comunicação e Participações S.A., 2018) (in Portuguese). The paper was also awarded with the "Prêmio de Inovação do Grupo Fleury", which comprised the use of computational methods in Biology.

In the initial discovery phase, in which our collaborators were the main contributors, we integrated clinical and biological knowledge in the analysis of discovery proteomics as *a priori* information to pre-select the most prominent proteins for the prognostic of oral cancer patients. For instance, our collaborators used histopathology, discovery proteomics analysis of formalin-fixed paraffin-embedded OSCC tissues, and clinical features of patients to find proteins that fill some criteria. The entire pipeline is illustrated in Figure 8. The steps before targeted phase are summarized in the next section.

2.3.1 Discovery phase

Specific analysis based on the invasive tumor front (ITF) have demonstrated reliable predictive value for OSCC prognosis and is considered a key region in the dynamic progression of malignant tumors. The presence of neoplastic islands, classified as large or small according to the number of cells in the ITF, has been described as the most aggressive pattern compared to tumors with a more uniform growth pattern, as tumor invasion occurs in a more widespread manner as cellular islands or single cells (ALMANGUSH *et al.*, 2015). Furthermore, there is evidence that components of the tumor stroma critically influence carcinogenesis and the malignant phenotype in multiple stages of tumor development and progression (CURRY *et al.*, 2014; TLSTY; COUSSENS, 2006). The complex interactions between tumor cells and the various types of cells and matrix components within the microenvironment play important roles

Figure 8 – Experimental design. **(a)** The ITF was delimited as a 1mm depth from the edge of the tumor slice, and the inner tumor was defined as up to 1mm from the epithelial tumor tissue origin. In more detail, the ITF and inner tumor with small and large neoplastic islands (arrows) are surrounded by the tumor stroma (dashed lines) (scale bars, 200μm). **(b)** Laser microdissection of the six regions of interest. **(c, d)** Protein extraction from microdissected tissues and trypsin digestion. The peptide mixture was desalted in stage tips and analyzed by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS). **(e)** To select protein targets, MaxQuant and the Perseus package were adopted to identify and quantify proteins, and R software was used for statistical analysis of the clinicopathological parameters and for analysis of the proteins with positive staining for OSCC tissues in The Human Protein Atlas. **(f)** Targeted proteins were evaluated following two different strategies by immunohistochemistry in OSCC tissues and by SRM of saliva samples from OSCC patients with (N+) or without (N0) lymph node metastasis.



Source – (CARNIELLI *et al.*, 2018), licensed under CC Attribution 4.0.

in cancer onset, progression, invasion and metastasis (TURLEY; CREMASCO; ASTARITA, 2015, 2015). This information guided to the design of the present discovery methods.

The different regions and sizes of tumors from 20 OSCC patients were analyzed in the discovery proteomics phase, revealing exclusive and shared proteins. The analyzed proteomes were:

1. Small neoplastic islands from the ITF;
2. Large neoplastic islands from the ITF;
3. Small neoplastic islands from the inner tumor;

4. Large neoplastic islands from the inner tumor;
5. Small neoplastic islands from the inner tumor;
6. Stroma from the ITF;
7. Stroma from the inner tumor.

The quantification for neoplastic cells resulted in 2,049 proteins, and for the tumor stroma resulted in 1,733 proteins. After filtering for proteins that have at least ten label-free quantification (LFQ) intensity values in one group, the numbers were reduced to 799 and 704, respectively. Then, we applied the Student's t test (p-value < 0.05) for proteins testing neoplastic islands samples from ITF against inner tumor and found 32 proteins with differential abundances; for the proteins from stroma, we applied paired Student's t test (p-value < 0.05), and found 101 proteins with differential abundances. The difference in the statistical tests is justified by the high variance among the neoplastic islands in opposition to the lower variance in the stroma group.

The biological processes (BP) of those proteins were also investigated using the Gene Ontology (GO) database (<http://www.geneontology.org/>) (The Gene Ontology Consortium, 2000), identifying the enrichment level of each BP in selected proteins. For instance, *cellular metabolic* processes were more enriched in the neoplastic island proteins, and *cellular adhesion* processes were more enriched in the tumor stroma proteins.

Finally, we used linear regression to analyze the proteome LFQ data set and clinicopathological data to identify the proteins associated with patient features (Table 1 and Supplementary Figure 4 from (CARNIELLI *et al.*, 2018)). The majority of proteins (ACTR2, CSTB, LTA4H, PGK1, NDRG1, FSCN1, ITGAV, THBS2) significantly associated with clinical parameters showed lower expression in the ITF of the tumor stroma or neoplastic islands, except COL6A1, COL1A2, S100A8, S110A9, and MB. In the next steps, we prioritized these proteins, selecting a smaller group and, then, evaluated their predictive power to differentiate patients with and without lymph node metastasis.

2.3.2 Targeted proteomics

The targeted proteins evaluated in the subsequent steps of verification using immunohistochemistry (IHC) in a 125-patient cohort and SRM in an independent 40-patient cohort were elected if they filled the following criteria:

1. Only proteins with different protein abundances between the ITF and the inner tumor in the discovery phase (Student's t test, P-value < 0.05);
2. Only proteins that present a significant association with clinical characteristics of patients (Linear regression, P-value < 0.05, R < -0.7 or R > 0.7 and R2 > 0.4);

3. Only proteins with positive staining of squamous cell carcinoma in HNSCC in The Human Protein Atlas (<https://www.proteinatlas.org/>) (UHLEN *et al.*, 2010);

4. Only proteins not cited or cited only in limited studies related to oral cancer.

Cystatin-B (CSTB), leukotriene A-4 hydrolase (LTA4H), protein NDRG1 (NDRG1), and phosphoglycerate kinase 1 (PGK1) from the neoplastic island data set and collagen alpha-1(VI) chain (COL6A1), integrin alpha-V (ITGAV) and myoglobin (MB) from the tumor stromal data set were prioritized (Figure 3 and Supplementary Figure 5 from (CARNIELLI *et al.*, 2018)). All these proteins, according to the literature and to the domain predictions performed here, are nonclassically secreted (Supplementary Data 18 from (CARNIELLI *et al.*, 2018)).

2.3.3 Candidate biomarker signatures

Assessing the protein profiles of large and small neoplastic islands and their surrounding stroma by combining laser microdissection (LMD) and proteomics reveals several proteins - including CSTB, NDRG1, LTA4H, PGK1, COL6A1, ITGAV and MB - with distinct expression patterns between ITF and inner tumor, suggesting a potential prognostic value by clinicopathological association analysis. In the subsequent targeted phase, we use two follow-up approaches to verify these signatures in two independent patient cohorts. First, analysis of clinical significance and immunohistochemical staining were performed in a 125-OSCC patient cohort, indicating CSTB, at low expression levels in the ITF, as an independent marker for local recurrence. Second, Selected Reaction Monitoring (SRM) Mass Spectrometry [3] is applied to study the abundance of the above-mentioned seven proteins in saliva samples from an independent 40-OSCC patient cohort.

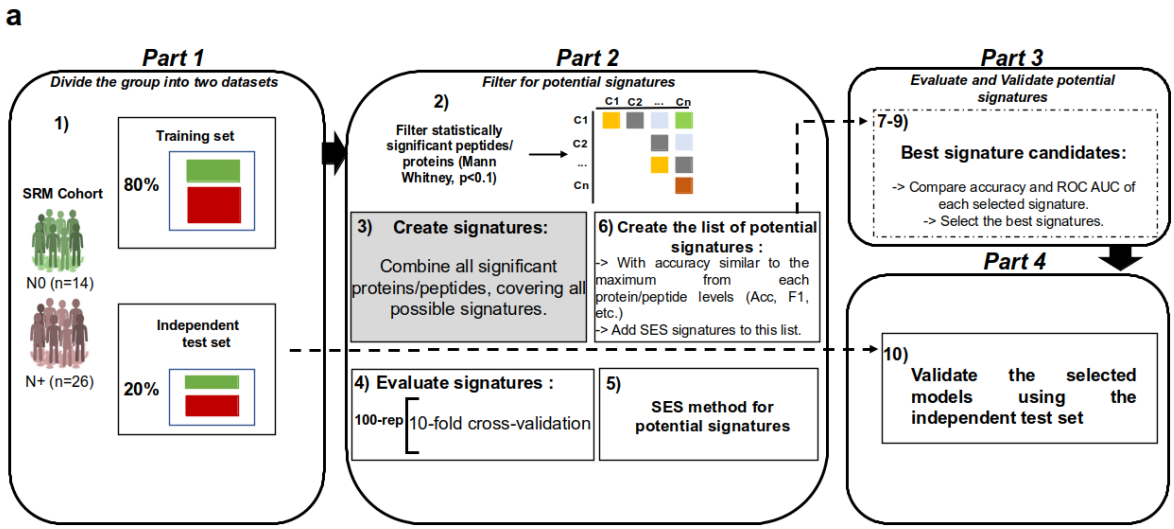
Further, we evaluated the predictive power of individual and groups of peptides and proteins to distinguish the patient with lymph node metastasis (N+) from the patient without lymph node metastasis (N0). For this, we used two distinct data sets formed by the 40 samples, one with prioritized proteins and the other with their peptides. The selected proteins that were quantified by the targeted proteomics are: CSTB, LTA4H, NDRG1, PGK1, COL6A1, ITGAV and MB.

Both data sets were used in the same pipeline presented in Figure 9. First, we split the samples into train (80%) and test (20%) sets. Then we generate all possible combinations of peptide/protein signatures and perform a 100-repeated stratified 10-fold cross validation, estimating their accuracy, ROC AUC, precision, specificity and sensitivity. We ranked the signatures and evaluated their scores, finding 4 main signatures that performed well and were formed by proteins/peptides that were very frequent in signatures with the highest scores. The signatures S1: (Pep8_LTA4H, Pep12_CSTB), S2: (Pep8_LTA4H, Pep9_COL6A1, Pep12_-CSTB), S3: (Pep8_LTA4H, Pep9_COL6A1), and S4: (LTA4H) are the most relevant signatures

[3] The mass-specrometry technique used for targeted proteomics.

(Si) considering accuracy and AUC (Figure 10). The sequence of each peptide is reported in the original article.

Figure 9 – Workflow for Machine Learning approach to measure the predictive power of peptides and proteins. (Part 1) Patients data is first divided into independent test and training sets. (Part 2)The training set used in a repeated cross-validation scheme, comprising filtering and comparison of predictive power of protein signatures. (Part 3) The most powerful signatures are compared in terms of ROC AUC and other measures to indicate a panel of best signatures. (Part 4) The final signatures are validated with the independent test set separated in Part 1.



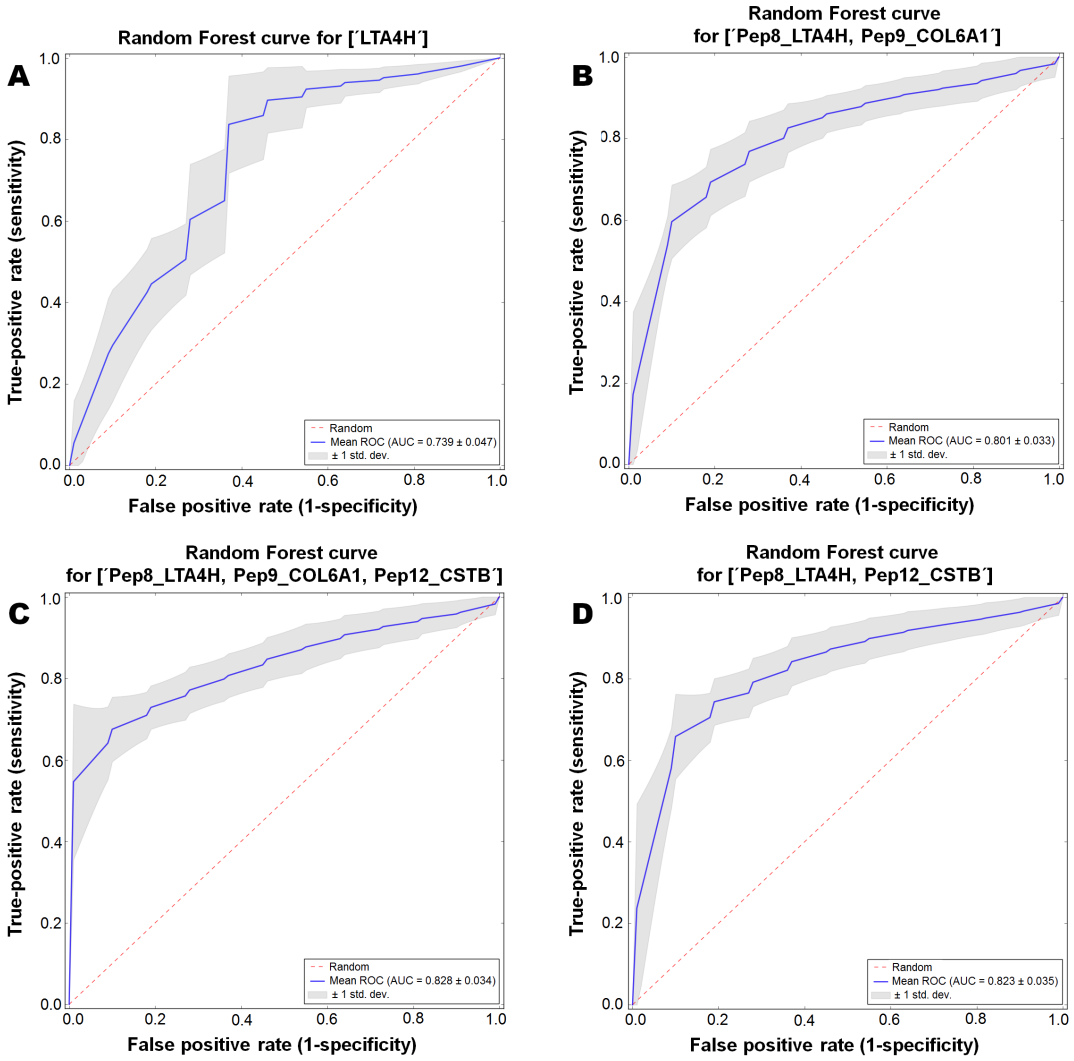
Source – Figure 7 (a) from (CARNIELLI *et al.*, 2018), licensed under CC Attribution 4.0.

Our initial purpose was to find a protein signature for the OSCC prognoses. By testing the performance of the peptide intensities, however, we found out that peptides had prediction performance greater than the protein signatures. The ROC AUC of the peptide level is considerably higher than that of the protein level, 82.8% (S2, Figure 10C) compared with 73.9% (S4, Figure 10A). Only the signature S4 (LTA4H) was selected at the protein level, with an AUC of 73.9%, as other signatures have AUCs lower than 62.5%. Further, balancing the training subsets with the SMOTE technique also increased the overall prediction performance. We represented all scores from the tested signatures as box plots in Figure 11. The position of S1, S2, S3 and S4 are indicated by arrows in the diagram.

Furthermore, the signatures S1 and S2 at the peptide level and S4 at the protein level are the best candidates for both types of cross validation, using imbalanced and balanced classes. On the other hand, the signature S3 had its performance decreased in the over-sampled cross validation. The signature S2 was selected as the best features to discriminate N0 and N+ of OSCC. Interestingly, both S1 and S2 are formed by the best 2 and best 3 peptides from the rank of individual peptide scores (Figure 12).

Taken together, our results identify a prognostic signature that may assist in the clinical decision-making process leading to appropriate treatment, improving the prognosis and survival

Figure 10 – Cross validation estimated ROC curves of the best protein and peptide signatures. **(A)** S4: LTA4H. **(B)** S3: Pep8_LTA4H and Pep9_COL6A1. **(C)** S2: Pep8_LTA4H, Pep9_COL6A1, and Pep12_CSTB. **(D)** S1: Pep8_LTA4H and Pep12_CSTB).



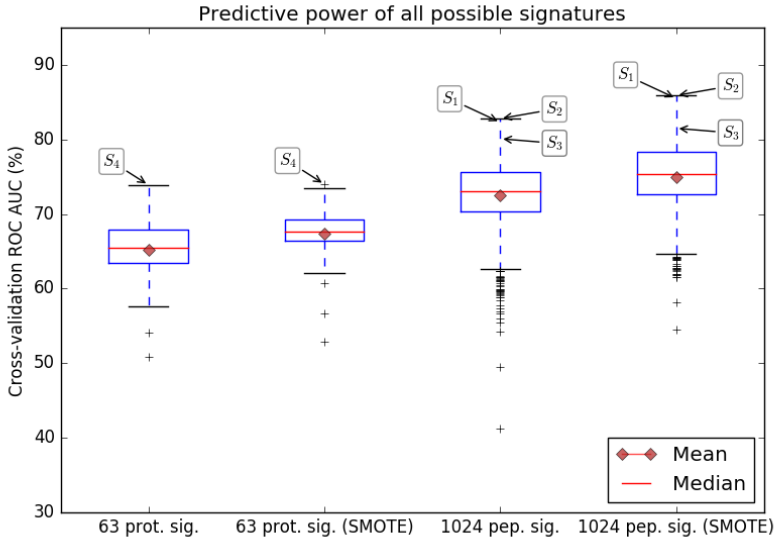
Source – Adapted Figure 7 (d) from (CARNIELLI *et al.*, 2018), licensed under CC Attribution 4.0.

of patients with OSCC.

2.4 Lists of candidate proteins are not stable

Despite we have one example of a good scenario in the oral cancer study described in this chapter, the problem of the small number of samples still remains. We have found good signatures of candidate biomarkers but the estimated prediction power of each signature can still be different in the next studies where they must be evaluated in clinical tests. This can happen because the number of samples used in the multivariate analysis is small and may not represent the population as we believe it does. The *a priori* studies that separated a very specific set of proteins, with a bigger number of samples and paired analysis, increase the chances of this

Figure 11 – Box plots representing the AUC of all possibilities of signatures for both imbalanced and balanced (SMOTE) cross validation. At the peptide level, 1024 signatures were tested. At the protein level, 63 signatures were tested. Signatures formed by peptides from different proteins S1 Pep8, Pep12 and S2 Pep8, Pep9, Pep12 have approximately 10.5% higher AUC than the peptide signature formed by LTA4H (S4). S2 peptide signature outperformed both S1 and S4 signatures, being S1 and S2 very similar. The candidate signatures are indicated by labels: S1, S2, S3, and S4.

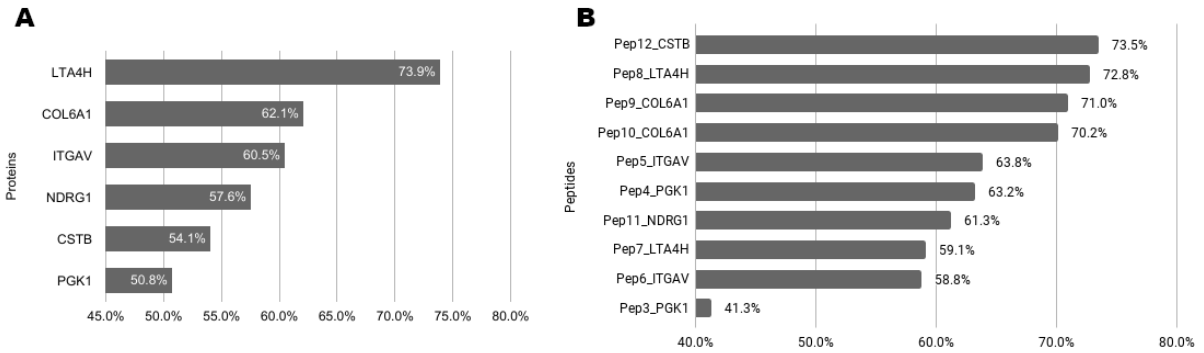


Source – Adapted Figure 7 (e) from (CARNIELLI *et al.*, 2018), licensed under CC Attribution 4.0.

estimate be indeed true. This is not the case for many studies where we have the common setting of a much greater number of variables than samples. Even with *a priori* information reducing the number of initial proteins, depending on the settings and type of cancer that are being studied, we could end up with a list of hundreds of proteins.

In the next chapter, we discuss the problems linked to the small number of samples and what we can do to better understand the results. We explain how this problem have been studied and how stable the list of proteins from different projects are. We use particular data sets with specific characteristics to evaluate ranking and signatures methods. One of these data sets contains four proteins that were added to samples before the discovery proteomics. With this approach, we tracked true positive markers through each step of feature selection and signatures evaluation executed in a double cross validation fashion.

Figure 12 – The predictive relevance of individual proteins **(A)** and peptides **(B)** to distinguish N0 from N+ patients is represented by a bar chart indicating their cross validation ROC AUC (100 repetitions of stratified 10-fold cross validation). The most relevant protein and peptide ordered by the AUC is LTA4H and Pep12_CSTB, respectively.



Source – Adapted Figure 7 (b) and (c) from (CARNIELLI *et al.*, 2018), licensed under CC Attribution 4.0.

CHAPTER

3

METHODS FOR PROTEIN PRIORITIZATION FROM DISCOVERY TO TARGETED PROTEOMICS

As mentioned in the Chapter 2, due to the small number of samples it is inappropriate to base all the decisions on statistical and Machine Learning methods. In the discovery proteomics, or even in the targeted proteomics when we have a list of proteins bigger than the number of samples, the final lists of candidates vary much with small changes in the train data set.

Previously, we studied the behavior of a few Machine Learning methods and the frequency that they selected each protein, in collaboration with Domingues (2017). In that project, our collaborators designed the experiments in a manner that we could seek for candidates that discriminate three groups and track four true positive candidates.

To understand the behavior of ranking methods, four different proteins were added in the biological samples, with three different concentrations, before their quantification. These four proteins are expected to be selected as true positives.

Given the proteomics protocol described in Chapter 2 (Figure 1), when we add the same amount of proteins to a biological sample, we may quantify different intensities of peptides. This means that we can end up with a protein intensity matrix where the **added protein A** has different intensities than the true ones (in the biological sample). Samples can have different amounts of proteins that contain some peptides (or similar peptides) of A and different samples could already have the protein A expressed. For this reason, we usually chose unique peptides of each possible protein. Still, due to other limitations in the proteomics protocol, such as inferring peptides sequences based on mass spectrum, the final distribution of our added proteins may not be uniform. The LC-MS/MS quantification protocol and limitations are not scope of this dissertation.

Three main attributes for validation of candidate proteins are: predictive power, stability and interpretability. There are many approaches to evaluate each of these attributes. Among them we can mention for predictive power: the simple and double cross validation; for stability: the evaluation of frequency of selected proteins varying the data set by using different quantification techniques or permuting the intensity matrix; for interpretability: the enrichment of each biological process or protein function, network analysis, literature curation and data bases queries. These concepts are detailed along this chapter.

The related works presented here do not use the same approach for validation. There is much to be explored in this way, being a target area for meta-analysis and replication studies. For instance, most of works use the common cross validation and compare biased overoptimistic predictive power estimates of signatures, what is actually not a valid comparison due to the limited sample size.

Here, with the controlled data set, we evaluated different methods for ranking and signature discovery, and proposed an approach to identify interesting proteins for further targeted proteomics, biochemical experiments or clinical trials. We combined different ranking methods with different classical machine learning algorithms and evaluated the position of the added proteins, verifying their rank stability. Then we identified sets of proteins with *good* and *best* prediction scores and checked if they were included in these signatures, all in a double cross validation fashion. The main research questions we answered are:

> Given the noise in **discovery proteomics' data**, would the true candidate biomarkers be identified by methods based on ranks, signatures and frequencies? Would these proteins be positioned similarly by rankers? What happens if instead of using a single training set to rank proteins, we use randomization to generate multiple training sets and combine with multiple ranking methods to find a consensus-rank?

In the end of this chapter, we repeated all the analyzes using one additional data set and reported our conclusions on computational methods and stability of biomarkers discovery in discovery proteomics. All results, research methods and the main concepts about the ranking and signatures algorithms are described in the next sections.

3.1 Intra- and inter-stability

The term stability refers to the consistency in selecting proteins. That is, if we make small changes in the data set, the proteins that we select from the ranking/signature models should be similar. In the case of ranks, we can monitor the position or the score of each protein in each computed rank. The higher is the variance of the position or the score of a protein, the lower is its stability. For the stability of a rank (not a single protein), then, we can compare the mean frequency of top-N proteins from each rank. For methods which result is a list of proteins

(not a rank), the protein stability can be measured by the frequency of each protein in the resulted signatures.

To analyze the stability of a method or protein, we can measure the intra-stability and the **inter-stability**. We refer to **intra-stability** when we use only samples from the same data set in the analysis. For instance, using one data set and applying sub-sampling, cross validation, double cross validation and varying the ranking methods. The **inter-stability** refers to a more general and complex evaluation, where we compare the findings of a disease by using different sources of data sets. For instance, samples from Brazil versus from Canada for a study comparing the same two groups (cancer vs. non-cancer), or samples quantified in Brazil but with a different setting of quantification. In this dissertation we discuss the intra-stability and refer to it by using only the terms stability and instability.

In the previous project where we first proposed the use of synthetically added proteins for the comparison of ranking methods (DOMINGUES, 2017), we performed an intra-stability evaluation. In one part of the study, we analyzed the variation of p-values from the Kruskal-Wallis test simulating small changes in the train data set. We identified that the added proteins were more stable than the other proteins, confirming that they were picked as true positives when the cutoff of 0.05 was considered. An example of inter-stability analysis is reported by Kapoor and Dass (2002). They proposed a new technique for gene biomarker selection and tested it using two different data sets on breast cancer (GLINSKY *et al.*, 2004; WANG *et al.*, 2005a). The proposed technique resulted in 12% of intersecting genes, which is better than the 1.3% of intersection from lists reported by the two original articles (GLINSKY *et al.*, 2004; WANG *et al.*, 2005a).

Many computational models use biological information to reduce the instability of the methods applying intentional bias (HWANG *et al.*, ; DUTKOWSKI; IDEKER, 2011; SANAVIA *et al.*, 2012; CUN; FRÖHLICH, 2012a; BARTER *et al.*, 2014; CUN, 2014; WEILAND, 1989) in such a way that if possible solutions do not follow the predefined biological patterns, they are not tested. In general, this category of models have improved stability and equal or lower prediction power than methods that do not use a priori biological information. They also have greater biological interpretability (CUN; FRÖHLICH, 2012a; SANAVIA *et al.*, 2012; WEILAND, 1989).

Studies tried to define the best methods for biomarker discovery mostly in transcriptomic, but the small number of samples limit the conclusions. Also, there are many ways to create biological networks and query biological information, being the complexity of defining the set of a priori data a second problem. For instance, most of cancer biology still remains without a robust structure (GOOD *et al.*, 2013). The biomarkers discovery based on integrative Omics is new and needs to be more explored and improved.

Most decisions will still be difficult while the quantification technologies and other facts limit the number of quantified biological samples. Studies have been highlighting that for a computational or statistical method be powerful enough to find cancer biomarkers by means

of statistical tools, hundreds or thousands of samples would be necessary. That is, if we want to base our analysis and believe in the measured prediction power or the selected signature as true right away from the computational analysis, a much bigger number of samples would be necessary. Meanwhile, filters and specialized knowledge need to be used to guide the experiments and computational tools in a critical assessment with good validation criteria (BARKER, 2003; PUNTMANN, 2009) to increase the number of true positives that follow up to the next biochemical experiments and clinical trials. Both considered of high cost.

3.1.1 Biological interpretability

Analyzing the biological interpretability consists in investigate how a method is capable of finding genes related to the mechanisms that can influence the cancer cells, that were associated to cancer or that are known as biomarkers. As an example, Cun and Fröhlich (2012b) evaluated the interpretability of methods analyzing the enriched pathways using KEGG and the association of genes to cancer using FunDO (OSBORNE *et al.*, 2009), which is based on statistical tests to relate genes to diseases by means of annotations from the Disease Ontology (OSBORNE *et al.*, 2009).

Wang, Li and Fang (2012) analyzed the interpretability of their method by verifying if the found top-50 genes were associated to cancer, e.g. around 20% of the genes were linked to cancer. In a further analysis they extended it to literature review finding information about them and counting how many articles support the linkage of genes to cancer.

Among the methods with high biological interpretability are the RRFE (ZIZAS *et al.*, 2012) and the *average pathway expression* (GUO *et al.*, 2005). The first is based on the GeneRank (MORRISON *et al.*, 2005) algorithm which makes hub proteins be selected. Since they are hubs in the network that represents the biological system, a change in this protein could change the entire system. These are usually well-studied proteins and some are linked to diseases. The second example uses complete pathways in its construction, thus, the interpretability is there by definition.

3.2 Methods for biomarkers discovery

The identification of potential candidates in proteomics are based in the differences of spectral intensities or spectral counts between the control and disease groups. In this section we will review methods that are used to prioritize proteins. For methods developed in the context of transcriptomics data we use the word *gene* and the word *protein* for the ones developed to or tested with proteomics. Some methods that use only the quantified proteins were already exemplified in Chapter 2, e.g., SVM-RFE and NSC.

There are two main approaches to find sets of proteins with good prediction performance: (1) test all possible combinations (NP-Complete problem Burke (2000)) and (2) use some

heuristic to explore part of the domain of possible combinations. The method (2) is the one that is used if the number of proteins are not small because the (1) is time demanding, being inconvenient due to cost.

Among the most used heuristics are the ranking methods. These algorithms rank the proteins according to their relevance to discriminate the conditions that are being studied. It is a popular approach because the differentiation of each protein among conditions needs to be verified by biochemical experiments of high cost and, thus, having a ordered list to follow helps. For this reason, some algorithms also focus on reducing the false positive rate and on finding results more biologically plausible. It is a very common task to do a literature review about the top-N proteins of a rank, seeking to reveal information that already have been discovered.

The two main types of feature selection algorithms are the filters and wrappers. The main difference between them is in the way they give the importance to the features. The **wrappers** test subsets of features with supervised models. The **filters** have much lower computational complexity using simple criteria for validation. One example of filtering is the calculation of variance of each feature to eliminate the ones that are almost constant. Filters are also used as a first step before wrappers because reducing the number of features reduces the time to execute a wrapper. This is very common in biology, for instance, when we consider only proteins with p-value < 0.05 for a given statistical test (filter) and, then, apply a ranking method such as SVM-RFE (wrapper). A filter algorithm can also be used to rank proteins without assumptions of classification power.

3.2.1 Statistical tests

Among the most used filters in transcriptomics and proteomics are the statistical tests. Each protein is individually tested to identify the difference between the mean of different conditions. For instance, the values of one protein in cancer samples are tested against the values from control samples, with the null hypothesis that they are equal. A low p-value indicates that the protein is differentially expressed. All p-values are ordered to form the rank of proteins. Sets of proteins from this rank can also be tested using classifier models to check the prediction power of the top-ranked ones (LI; ZHANG; OGIHARA, 2004). The test can be used as a filter to consider in the following steps only proteins that are differentially expressed or to rank the proteins for additional validations. One example is the Beta-Binomial test for spectral counts data sets, as reported in the studies of Chapter 2.

Guo *et al.* (2009) used Student t test (HAYNES, 2013) in a proteomics study to rank proteins that discriminate colorectal cancer for its diagnose in an initial stage. Sets formed by the first N proteins in the rank were evaluated in a 10-fold cross validation using artificial neural networks models. The cancer condition was formed by 62 samples and the control by 31. The study reported 94% sensitivity and 96% specificity with the final selected top-N proteins.

The use of t test is very common in biology and variations were proposed to find discriminant genes and proteins (YAN *et al.*, 2005; WANG; LI; FANG, 2012). The test has one requirement that is hard to verify when the number of samples is small: the conditions' distributions should be nearly normal. The test based in the sum of the difference of ranked values, the Wilcoxon rank sum test, demonstrated to be better than the t test and does not require the classes to be normal distributed (DENG; MA; PEI, 2004). For more than two classes, the Kruskal-Wallis rank sum test was evaluated. Both had good performance to select discriminating genes (LEE *et al.*, 2005; BOULESTEIX, 2007).

Moore *et al.* (2008) analyzed the effects of combining new candidates to a biomarker that was already been used to predict ovarian cancer. The biomarker CA125 has limited specificity because it also has an elevated level in individuals with gastrointestinal cancer, fallopian tube cancer, lung cancer, breast cancer, endometrial cancer, and more. Thus, they tried to combine this marker with others proteins ranked by Wilcoxon rank sum test and measured the prediction power by using Logistic Regression models and cross validation. "As a single tumor marker, HE4 had the highest sensitivity for detecting ovarian cancer, especially Stage I disease. Combined CA125 and HE4 is a more accurate predictor of malignancy than either alone" (MOORE *et al.*, 2008). The main goal was to improve the specificity. Combining more proteins did not improve the estimated score.

Statistical test and Logistic Regression models were used by Zhang *et al.* (2010) in combination with statistical tests to identify pancreas cancer by saliva samples. The transcriptomic of 42 healthy samples, 42 pancreatic cancer samples and 30 chronic pancreatitis patients were quantified. They split the data set creating a discovery set with only 24 samples, 12 healthy and 12 with pancreatic cancer, and a validation set with 90 samples, the remaining 30 samples from each class. The candidate biomarkers found using the 24 samples were tested with 100-leave-one-out cross validation and Logistic Regression. The three classes were compared two by two, all reporting ROC AUC > 97%. For a reason that we could not find or explain, they used Wilcoxon signed rank test, which is a paired version of Wilcoxon rank sum test. If the healthy and cancer samples were from the same patients, the signed test would be appropriate. Both studies from Moore *et al.* (2008) and Zhang *et al.* (2010) tested limited combinations of proteins and genes, respectively.

Yan *et al.* (2005) explain that many available statistical tests are actually based on the t statistics and have one bad characteristic associated to this fact and the way they discriminate the conditions (EFRON *et al.*, 2001; TUSHER; TIBSHIRANI; CHU, 2001; PAN; LIN; LE, 2003): they are based on the comparison of means. T test, Wilcoxon and Kruskal-Wallis tests are examples. Given this problem, Yan *et al.* (2005) proposed a method to detect differentially expressed genes based on relative entropy called SDGREE. It combines kernel density estimation models and can detect differences between two classes. This method does not require any distribution format nor perform differences between means.

3.2.2 Multivariate models

The statistical tests can be used in combination with multivariate models or as a pre-processing step. In this subsection we report studies that focused more on the multivariate studies than on statistical measurement. This is the case of NSC and SVM-RFE, explained in Chapter 2.

Evaluate all the possible combination of proteins is in general computationally impracticable. Heuristics are used to explore a sub-set of combinations. This is the case of rank-based approaches that test the top-N proteins from a given rank. Other approaches are more dynamic and expand the signatures' domain dynamically. That is, given a set of best signatures, the algorithm creates more signatures according to some criteria that may have information related to these current solutions. Examples of this approach are the genetic algorithms (LI *et al.*, 2001a), the clustering analysis and rankings methods to select discriminant genes (WANG *et al.*, 2005b) besides many other algorithms that have been proposed in the last years (WANG; LI; FANG, 2012).

Machine learning models can have a property that is not consistent to biology: they may remove redundancy and, in that way, remove good candidates if they are correlated to others. Wang, Zhu and Zou (2008) proposed a solution for this attribute of SVM-RFE models. Their Hybrid Huberized Support Vector Machines (HHSVM) makes correlated features to be positioned close in the outputted rank. It works both for the best features as well for the worst features. Despite it is a model that is considered purely based on quantification data, it is somehow biased to a biological principle: proteins that participate in the same biomolecular activity tend to be correlated (KAVITHA; KANNAN, 2016).

Li *et al.* (2001a) analyzed the combination of genetic algorithms with k-Nearest Neighbor (KNN) classifiers, which has similarities with NSC such as having the features transformed into distances between samples. The proposed method identifies groups of genes by genetic algorithms and track the genes' frequency. They studied colon cancer and used a data set formed by 40 cancer samples, 22 control samples and 2 thousand genes. It resulted in 6,348 subsets of discriminating genes. In another study, Li *et al.* (2001b) tested the robustness of their algorithm and pointed out that despite their good results, the selection of genes are highly dependent on the training samples and that, in general, "gene selection may be less robust than classification". A similar method based on genetic algorithms was proposed by Liu *et al.* (2005), this time, using SVM models.

These are only a few examples that were applied to find genes or proteins that are candidates for biomarker of cancer. Many other algorithms that are based only in the quantification data have been proposed, such as the Significance Analysis of Microarrays (SAM) (TUSHER; TIBSHIRANI; CHU, 2001), the Heuristic Breadth-first Search Algorithm (HBSA) (ZHANG *et al.*, 2010), the Sequential Forward Search (SFS) and Sequential Forward Floating Search (SFFS) (XIONG; FANG; ZHAO, 2001), the Markov blanket-embedded (ZHU; ONG; DASH,

2007) and the SCAD-SVM (smoothly clipped absolute deviation penalty) (ZHANG *et al.*, 2006). The limitations and the rapid evolution of proteomics and transcriptomics require more comparative studies and tools that allow researchers to find candidate biomarkers with a high probability of being a true positive to make progress in this area and to reduce costs.

Most of the algorithms were proposed considering transcriptomic data sets, which is different from proteomics in many levels. For instance, not all expressed RNA (transcriptomics) are going to be translated into proteins (proteomics) and proteins can be combined into complexes or have post-translational modifications. In the next section we report some techniques that explicitly combine biological information into the multivariate models.

3.2.3 Combining biological information with multivariate models

The most common methods used when seeking for important proteins or good signatures for prediction of samples are the ones that output ranks and the ones that output signatures. Due to the small number of samples, methods to prioritize proteins are unstable.

Despite Christin *et al.* (2013) compared the performance of protein selection methods in the studied mentioned in Chapter 2, the choice of the best approach is still limited by the data set size. This limitation is exemplified when a different number or set of features are selected by the same method using sets sampled from the same data.

Another approach to improve the stability of ranking and signature methods, also to increase the probability of high biological interpretability, is to combine biological information inside the multivariate models. It works limiting the domain of subgroups of genes/proteins that are going to be tested, creating criteria that reduce the domain of proteins, or applying transformations on linear models. One example is to create a Random Forest model where the proteins of a tree necessarily participate of the same biological process. In this example, it is as if instead of considering the proteins as markers we are now considering the biological process, the pathway or function, the real marker. This concept goes to the direction of finding what is not being executed or what new is being executed in the cell that makes it different from a healthy cell.

In some studies, researchers proposed to find what are the metabolic pathways that are linked to a specific condition (DONIGER *et al.*, 2003; PAVLIDIS *et al.*, 2004; TIAN *et al.*, 2005). For instance, Tian *et al.* (2005) created a statistical model that identifies pathways significant related to a defined phenotype, such as diabetes, Alzheimer, and more. The two main databases used in this study were Gene Ontology (GO) (ASHBURNER *et al.*, 2000) and KEGG (KANEHISA *et al.*, 2012).

Guo *et al.* (2005) proposed the dimensionality reduction of an expression matrix that represents samples versus genes to a matrix that represents biological function and samples. They used Gene Ontology to identify the genes' pathways annotations and based on the expres-

sion of the genes, find the most enriched functions. They reported good prediction power and interpretability of the biomolecular mechanisms associated to the studied disease.

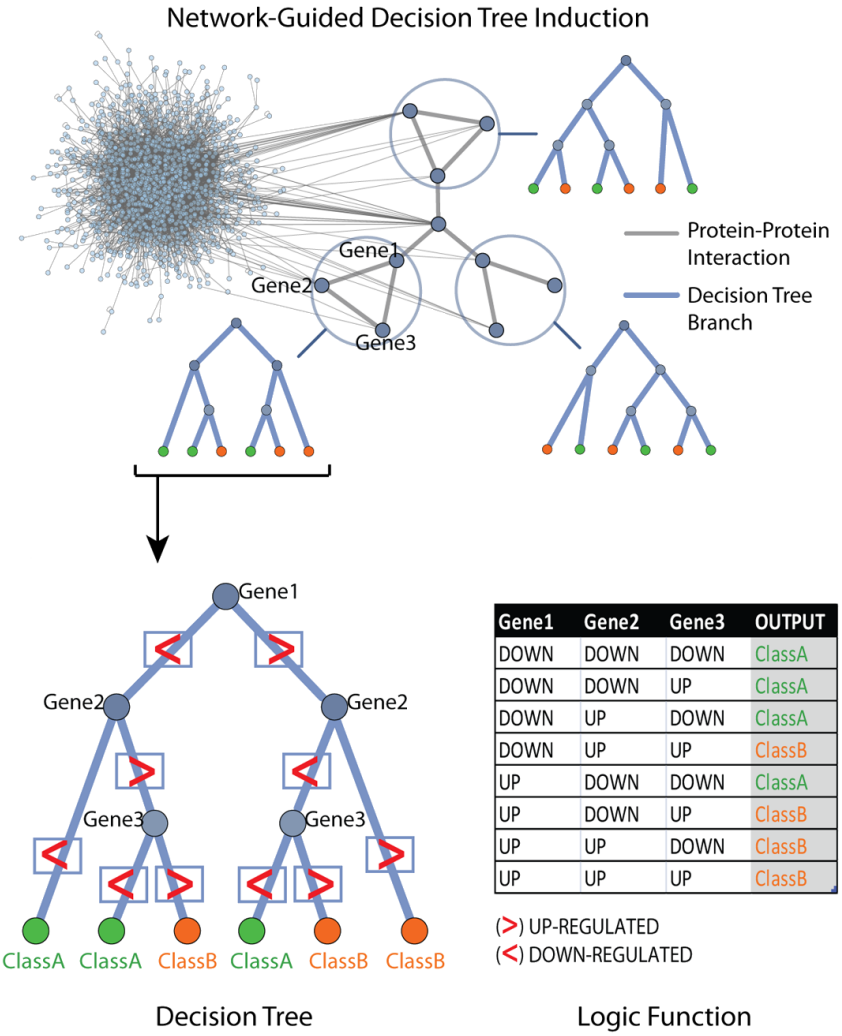
With access to great protein-protein interaction networks, researchers started proposing to identify sub-networks that could discriminate different conditions. One important example is the study made by Kavitha and Kannan (2016), where they searched for signal and regulation circuits based on interaction networks. They used simulated annealing (VASILIJEVIC; JAMBROSIC; VUKIC, 2018) to identify sub-networks and evaluated them with statistical methods to identify proteins which expression most differentiate among the classes (conditions).

Kapoor and Dass (2002) proposed a similar method to identify breast cancer markers. This time, the method identified modules (sub-network) that were considered a discriminant of the samples' conditions, based on statistical methods. Each sub-network became a feature in the samples expression matrix. The value given for each feature is the average of expression of the genes in that sample, given the genes from the respective sub-network. Instead of having protein ranks, we now have a rank of sub-network that may represent biological modules. The technique demonstrated to be more stable when executed on breast cancer data sets (GLINSKY *et al.*, 2004; WANG *et al.*, 2005a) (12% of intersecting genes), in contrast to the original studies from Glinsky *et al.* (2004) and Wang *et al.* (2005a) (1.3% of intersecting genes).

Fold-change (ratio) is a concept commonly used to express if genes or proteins are up- or down-regulated. With the idea of finding protein-protein interaction (PPI) network modules that discriminate conditions, Dutkowski and Ideker (2011) proposed a method to find logic involving the up/down information. For instance, if protein A is up-regulated and B is down-regulated then the sample is with cancer. An illustration of the idea is shown in Figure 13. The logic is represented in a Decision Tree (MURTHY, 1998), that is, each module is actually a tree that codes the logic that separates up and down expression in each bifurcation. By creating many trees, the method end up with a Forest (VERIKAS; GELZINIS; BACAUSKIENE, 2011) that classify new samples and can indicate important biological modules that differentiate the conditions.

Most of knowledge-driven multivariate models have improvements on stability but no improvement in the prediction power. It is expected that due to the small number of samples and great number of features, there are many subsets of features that can result in the same *best* prediction score. The ROC AUC reported by Dutkowski and Ideker (2011) is the same for the regular Random Forests and for the Network-Guided Forests. Weiland (1989) proposed a similar method named Network-constrained forest (NCF) and claimed that they could find better prediction scores than Random Forests (RF). By testing many classification tasks using Mathews correlation coefficient, they reported the averages as follows: NCF with 0.71, RF with 0.56, SVM with 0.69 and Naive-Bays with 0.42. Despite the score of Random Forests were indeed improved by the network information, the SVM outperformed the NCF and RF in many tasks and the averages are similar, 0.69 against 0.71. Still, given the additional information from the modules and decision rules that are found, models guided by networks represent a powerful tool

Figure 13 – Inducing Decision Trees using a PPI network and the gene expressions. Given a starting gene, the algorithm search for neighbors and creates the logic trees. One tree is detailed in the sub-figure at the left-bottom, which describes each gene in the Decision Tree. The branches of the tree have an associated symbol of > and < representing the test for up- or down-regulation used when a sample is being tested by the Decision Tree. In the right there is a table that represents all decision rules of the tree. Each rule maps a path from the root to a leaf.



Gene1	Gene2	Gene3	OUTPUT
DOWN	DOWN	DOWN	ClassA
DOWN	DOWN	UP	ClassA
DOWN	UP	DOWN	ClassA
DOWN	UP	UP	ClassB
UP	DOWN	DOWN	ClassA
UP	DOWN	UP	ClassB
UP	UP	DOWN	ClassB
UP	UP	UP	ClassB

Source – Adapted figure from (DUTKOWSKI; IDEKER, 2011), licensed under CC Attribution 4.0.

and a great example of the future of integrative omics.

Different categories of biological information were compared by Sanavia *et al.* (2012). They used the information associated to each gene to create similarity matrices, one for each type of information. These matrices are used to transform the gene expression matrix in a way that the more similar genes are, more they are going to be similar weighted. Doing so they compared the influence of PPI network topology, semantic similarities by functional annotations and genes correlation. They pointed out that the semantic and topological similarities resulted in more stable candidate lists than by using correlation or not using any a priori information to build the classifier models.

Cun and Fröhlich (2012b) compared many recent methods that uses the quantification matrix and some that use the biological knowledge. The ones that do not use a priori information are the Prediction Analysis of Microarray data (PAM) (TIBSHIRANI *et al.*, 2003), the *significance analysis of microarray* (SAM) (TUSHER; TIBSHIRANI; CHU, 2001), the SCAD-SVM (ZHANG *et al.*, 2006), the SVM-RFE (THURSTON *et al.*, 2011) and the HHSVM (WANG; ZHU; ZOU, 2008). Among the biological information guided are the Reweghted Recursive Feature Elimination (RRFE) (ZIZAS *et al.*, 2012), the *graph diffusion kernels* para SVMs (RAPAPORT *et al.*, 2007), the *p-step random walk graph kernel* to SVMs (GAO *et al.*, 2009), the Network-based SVM (ZHU; SHEN; PAN, 2009), the Pathway Activity Classification (PAC) (LEE *et al.*,), the *average pathway expression* (GUO *et al.*, 2005), the *classification by significant hub genes* (TAYLOR *et al.*, 2009) and the pathBoost (BINDER; SCHUMACHER, 2009). They analyzed the stability and prediction performance of all the mentioned methods using six different transcriptomic data sets varying the number of patients from 159 to 286.

The biological knowledge-driven algorithms did not have improved prediction power but revealed better biological interpretability of selected genes, specially the ones proposed by Zizas *et al.* (2012) and Guo *et al.* (2005). This behavior is an expected effect of the regularization of multivariate models to prioritize signatures or genes that follow biological patterns, thus restricting the domain. While some methods had improved stability, some of had a decrease in the prediction performance. Under these circumstances, the choice for using or not methods with biological knowledge will depends on the purpose of the proteins search. While prioritize biological information, others may prefer to find signatures with the best prediction power as possible. Most of the methods were analyzed with transcriptomic data, therefore more studies and methods need to be proposed for proteomics data.

A problem that could be cited about this category of methods is that we may not know all biological information about a system and, thus, we could eliminate true positives not because they did not follow a biological patter in real life, but because it does not contain that information in the database yet. Another problem is the use of these data sets and scientific articles as a base for the comparison of biological interpretability of different methods. For instance, the hub genes are the well-studied ones and, thus, will contain much information registered in the databases,

while some other gene may be the cause of a disease and still be a novel not studied gene, or a gene with just a few annotations.

3.2.4 Closing remarks

In the previous sections we explained how stability, prediction power and biological interpretability are key to find good candidate biomarkers. While in some cases the interpretability is crucial, in other cases what matter is the prediction power. In a way or another, the stability of methods are crucial to be able to believe that the selected proteins have potential to be true positives or work as diagnostic or prognostic markers.

The current state of quantification technologies limits not only the methods and their estimations but also the comparison of methods when researchers evaluate and try to find the best ones. Due to the small number of samples, comparing many methods and evaluating more than the computational results is a crucial procedure to do in the biomarker discovery phase.

In this project, we evaluate different ways of ranking proteins based on classical classifiers. We also propose a way to test a set of potential good signatures and evaluate their importance based on ranking and cross validation scores. We also propose the evaluation of protein stability based on the frequency in the first positions of each rank and on good and best signatures.

We created a data set that contains proteins that are considered true positives and allowed us to evaluate their behavior of these proteins in the discovery proteomics protocol and the behavior of rankers and classifiers. With the knowledge we obtained by these extensive experiments and literature review, in future work, we intend to continue the research and apply the state of art biomarker methods in the data sets analyzed here, including the methods formerly created for trancriptomic data.

In the following sections we introduce our approach explaining each step. We also describe the two data sets that are explored later. In the end, we present the results and discuss the performance of methods on these two cases and the applicability of our method or the general idea of it as a guide for more reliable results and critical assessment in biomarker discovery phase based on quantification data.

3.3 Material and Methods

For the evaluation and comparison of methods, we decided to use two different data sets. The main data set (D1) is the one proposed by Domingues (2017), in which we have four proteins that were added before the discovery proteomics quantification and are considered positive candidates. The second data set (D2) is the one about prostate cancer, before studied and reported by Kim *et al.* (2012), Kawahara *et al.* (2015), Heberle *et al.* (2015).

The dimensions and class distribution of the two data sets are described in Table 1. In

the following subsections we detail each data set, their important characteristics and the general pipeline used to compute the ranks, signatures, their predictive power.

Table 1 – Summary of data sets.

ID	Condition	Size	Class distr.	Proteins
D1	Oral	28	9/10/9	276
D2	Prostate	32	16/16	624

3.3.1 Data set D1 - Oral cancer samples with four additional proteins

This data set is formed by samples from three conditions (classes) related to oral cancer: people with no tumor (C), patients that had a tumor removed (TR) and patients with tumor (T). We will refer to the samples without tumor as *control* (C) in this manuscript.

Four additional proteins were added to be monitored along the analysis and quantification through discovery mass spectrometry. For each class, they were added using the same concentration: 12.5 fmol for C, 25 fmol for TR and 50 fmol for T.

The added proteins are: Enolase 1 (P00924_ENO1) and Alcohol dehydrogenase 1 (P00330_ADH1) from *Saccharomyces cerevisiae* (baker's yeast); Glycogen phosphorylase, muscle form (P00489_PYGM) from *Oryctolagus cuniculus* (rabbit); and Serum albumin (P02769_ALB) from *Bos taurus* (bovine). These are *model* proteins in mass spectrometry, due to their physico-chemical characteristics that allow a good tryptic digestion and good ionization, thus reducing biases when they are read in the equipment. We do not find these proteins in the human saliva proteome, what reduces the false positive rate (quantification) since the probability of identifying shared peptides is reduced.

This data set was created by Domingues (2017) who developed a study to qualify discovery proteomics and analyze how filtering by statistical tests influence machine learning algorithms. In a collaboration, we developed analyzes of frequency of proteins being selected by the statistical tests through sampling many subsets. The work developed in her masters can be seem as our firsts results and the start of our collaboration in this specific project. The main purpose of it is to analyze the output of the mass-spectrometer considering the discovery proteomics, which is less precise than the targeted proteomics protocol. In this way, controlling the behavior of the four added proteins can give us insights about the use of multivariate models for protein prioritization. These are some questions that guided our decisions: would the added proteins inside a class have equal/similar intensities in the quantified intensity matrix? Would they be placed close by ranking methods? Would they be selected by statistical tests? How do small changes in the training set influence the results? In this work, we extend the analyzes associated to these questions.

3.3.2 Data set D2 - Prostate cancer biomarkers

Kim *et al.* (2012) published biomarkers candidates for prostate cancer and let the data set available. In their article they stated the motivation as follows:

> Current protocols for the screening of prostate cancer cannot accurately discriminate clinically indolent tumors from more aggressive ones. One reliable indicator of outcome has been the determination of organ-confined versus nonorgan-confined disease but even this determination is often only made following prostatectomy. This underscores the need to explore alternate avenues to enhance outcome prediction of prostate cancer patients. Fluids that are proximal to the prostate, such as expressed prostatic secretions (EPS), are attractive sources of potential prostate cancer biomarkers as these fluids likely bathe the tumor (KIM *et al.*, 2012).

The data describes the organ-confined (OC) and the extracapsular (EC) prostate cancer cells. The two classes comprises 16 samples each, with 624 quantified proteins. From those proteins, fourteen were indicated as candidate biomarkers and seven were validated by experimental biochemical methods as mentioned in Chapter 2, Section 2.2.2. Among the 14 proteins are P07288 (KLK3, PSA) and P15309 (ACPP, PAP) which are well-known as linked to prostate cancer. Five out of the 14 were successfully verified to correlate with the discovery data: P31947 (SFN), P08473 (MME), Q99497 (PARK7), P01033 (TIMP1) and P49221 (TGM4). The verification was performed with a study on "clinical outcome information related to whether the patient had biochemical recurrence or no evidence of recurrence within a two-year period post-prostatectomy" Kim *et al.* (2012). The other proteins were selected by the computational methods but they did not find the correlation with the small cohort of clinical outcome: Q9UHI8 (ADAMTS1), P36955 (SERPINF1, PEDF), Q13332 (PTPRS), P02787 (TF), P04083 (ANXA1), P12277 (KCRB, CKB) and P35579 (MYH9).

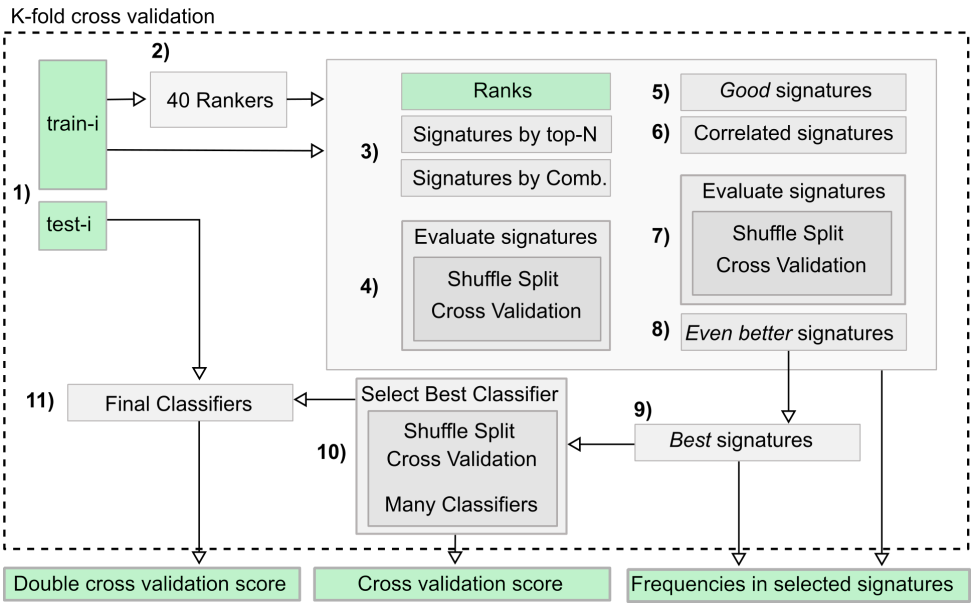
3.3.3 General pipeline

In this project we defined an analytic pipeline combining approaches for ranking features, identifying candidate signatures evaluated by simple and double cross validation. The general idea of the pipeline is illustrated in Figure 14, and detailed bellow:

1. Identify and separate correlated proteins $\alpha > thresholddefinedineachexperiment$;
2. Compute ranks;
3. Create signatures from ranks and small combinations;
4. Evaluate classifiers and select the best one;
5. Evaluate signatures using the selected classifier;

6. Rank signatures and select the **good signatures**;
7. Evaluate the good signatures, including correlated proteins;
8. Rank good signatures to form a set of **best signatures**;
9. Evaluate the *best signatures* using different classifiers;
10. For each best signature, select the best classifier and form a signature-pair;
11. Compute frequency of proteins in top-10 from ranks;
12. Compute frequency of proteins in *good*, *even better* and *best* signatures;

Figure 14 – General pipeline for candidate biomarkers assessment. **1)** The original data set is divided into train and test sets in each iteration *i* of the outer loop of the double cross validation (k-fold cross validation). **2)** The train set is used to create 40 ranks of proteins. **3)** Signatures are created by selecting the top-N proteins and by creating all or random combinations. **4)** Each signature is tested in a shuffle split cross validation. **5)** Signatures with *good* score are selected. **6)** Highly correlated proteins that were removed between steps 1) and 2) take place of their reference to form new signatures. **7)** All signatures are tested again in a shuffle split cross validation, now with a greater number of splits than before. **8)** With the new scores, signatures are selected again by a criterion that consider the score, the standard deviation and the frequency of proteins in *good* signatures. **9)** The first round of evaluation of signatures outputs the best signatures, formed by the signatures with the highest scores among the *even better* signatures. **10)** The best signatures are tested again by using different classifier models, including the Bagging and Voting classifiers. **11)** The best pairs (signature, classifier) of each signature is tested with the independent test set. In the end of the DCV, the independent test scores are averaged to form the DCV score.



Source: Elaborated by the author.

After each inner loop, the signatures are evaluated with the test sets from the outer-loop. The DCV frequency of each protein is given by the mean frequency of the ones calculated in the inner loops. The DCV score is the mean of scores computed in the outer loop (Figure 14-11).

Since the number of samples of the data sets D1 and D2 is small, we did not separate an independent test set. All samples were used in the double cross validation, which was originally proposed for such situations. Despite we use the DCV approach, we also used the results of the first DCV outer loop for a simple cross validation analysis, to compare the results of one fold (simple cross validation) against all DCV folds and to make the explanations easier to understand.

Each outer loop separates an independent test set and a train set (Figure 14-1). The train set is the input of a regular cross validation (Figure 14-2..8). In this chapter, we first report the fold-0, explaining how the results are interpreted, and the performance associated to the independent test set (regular cross validation). Then, we proceed to the double cross validation analysis.

3.3.4 Filtering

A **filtering** step was applied in each outer loop of the DCV, keeping only differentially expressed proteins. Each protein was tested with the Wilcox Rank Sum test (data set D2) or Kruskal-Wallis test (data set D1). An additional False Discovery Rate (FDR) was considered to adjust the p-values. In our experiments with the data set from (DOMINGUES, 2017), the Benjamini-Hochberg multiple test with FDR of 5% reduced the feature space to around 0 to 5 proteins. It was also inconsistent among the cross validation splits, varying too much the selected proteins. In fact, proteins with a distribution similar to the ones from the added proteins (DOMINGUES, 2017) were more stable and were selected more frequently in the outer loops from the DCV in many repetitions. Still, allowing a FDR of only 5% created situations where no protein was selected.

The too-small number of selected proteins leads us to a new problem: a high false negative rate. Due to this problem, and considering the discovery proteomics limitations, we considered a different FDR for the Benjamini-Hochberg multiple tests in a way that the number of the selected proteins were not too small as 7 and not bigger than selecting proteins by just using the raw Wilcoxon/Kruskal p-value < 0.05. We defined the FDR cutoff to be of 25%. In fact, defining a strict false positive rate is valid only when you need to reduce extremely the chances of a false positive in the cost of increasing the number of false negatives. John H. McDonald wrote in his *Handbook of Biological Statistics*:

> An unfortunate byproduct of correcting for multiple comparisons is that you may increase the number of false negatives, where there really is an effect but you don't detect it as statistically significant. If false negatives are very costly, you may not want to correct for multiple comparisons at all (John H. McDonald, 2014).

High **correlated** proteins were identified and removed from each train set in the outer loop. Pairs of proteins with the absolute value of Pearson correlation higher than 0.95 were grouped. For each group, only one arbitrary protein was selected to represent the others. If a signature pass for the final steps, that is, is indicated as a good signature, the equivalent signatures formed by exchanging the original proteins for correlated ones are also tested (Figure 14-6). Two high correlated proteins (>0.95) do not appear in the same signature.

3.3.5 Ranking proteins

There are many methods for ranking features. We implemented six categories of ranking approaches combined with univariate and multivariate algorithms defined in the Sklearn library (Scikit-learn developers, 2017). The first category, named ***Type 1*** - *Univariate*, is composed by univariate methods, such as t test and information gain. The other categories are based on multivariate models:

- ***Type 2*** - *Single Score* creates classifiers considering only one protein (for each available protein) and rank them based on the accuracy from a k-fold cross validation (k equal to 7 and 6 for D1 and D2, respectively);
- ***Type 3*** - *Attributes' Weights* only trains classifiers and rank the features based on default feature weighting method, e.g. linear model's coefficients;
- ***Type 4*** - *Recursive Feature Elimination* ranks the features based on the weights, but repeating the process N times, adding the worst feature to the end of the rank, as explained for the case of SVM-RFE;
- ***Type 5*** - *Stability Selection* ranks the features based on the frequency that they are selected as best in a bootstrap sampling scheme. The method creates many training sets using bootstrap, for each train set the algorithm computes the weight of each feature and select them if they have a weight greater than the mean weight. Frequently selected features are considered stable and are positioned in the beginning of the rank (MEINSHAUSEN; BÜHLMANN, 2010; SHAH; SAMWORTH, 2013; Thomas Huijskens, 2018);
- ***Type 6*** - *Decrease of Accuracy* computes the k-fold cross validation accuracy using all features, than, remove one feature at a time and recompute the accuracy. The bigger is the difference from the base accuracy and the accuracy after removing a feature, the more relevant is the feature;
- Finally, we proposed the ***Type 7*** - *Recursive Feature Addition* which is similar to Type 4. In this case, instead of creating a rank from the worst to the best, it creates a rank by removing the best feature from train set and adding it to the beginning of the rank, repeating the process until all features were ranked.

We created the Recursive Feature Addition (RFA)[1] based on the fact that the proteins are usually correlated. Even though the ones with absolute correlation greater than 0.95 are not considered together in our experiments, there may be correlations near 0.90 or 0.80 that could indicate great chance of dependency [2]. If proteins A and B are highly correlated, they should be positioned similarly in the rank. On the other hand, some models could compute a high weight for A and a low weight for protein B if they carry redundant information, e.g. if they are correlated. This is the case of methods like the Recursive Feature Elimination combined with linear models that can eliminate the protein B at first and keep protein A until the end of the elimination process. Thus, we believe that when creating the rank by "removing" the best feature in each step, A and B have the chance of being positioned near each other in the rank. Oshiro, Perez and Baranauskas (2012) originally proposed a different approach with the same name. Their RFA algorithm first rank the features based on individual accuracy (accuracy of each gene using the train set), then, recursively add genes to a final set based on one of five possible strategies. The algorithm also differs from our proposed RFA in the sense that their outputs is a list of selected feature while ours outputs a rank containing all features. Further in this chapter, we discuss and compare our approach in contrast with Recursive Elimination, analyzing the position of selected proteins and the position of the correlated proteins in RFA and RFE ranks.

Each of the multivariate-based ranking methods mentioned above was used in combination with different classifiers. We implemented the scripts[3] in python and considered classical machine learning tools from scikit-learn library (MARTÍNKOVÁ *et al.*, 2004). Only classifiers that implement the features weights (importance) were considered: Linear SVC, Decision Tree, Random Forest, Lasso, Ridge and Linear Discriminant Analysis. Before executing the ranking methods, we executed a grid search to pick the best value of some parameters. For Linear SVC, Lasso and Ridge the values for the C parameters were 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100 and 1000; for Decision Tree, the parameter max_depth were tested with integer values from 1 to 20; for Random Forest, the max_features was tested with values 0.5, 0.75 and 1. A total of 40 ranks were generated by different ranking approaches for each outer-loop's training set.

3.3.6 Scaling and normalization

All the features from the intensity matrix were normalized by z-score before fitting every classifier model. In any cross validation, the train set was normalized and, then, each test sample was transformed according to the mean and standard deviation computed using the train samples.

Each ranking method gave us a different range of values that represent the proteins importance. We scaled the values in the range $[0, 1]$ to allows us to compare the ranks and the

[1] Fork and contribute: <https://github.com/heberleh/recursive-feature-addition>
[2] Independent variables is an assumption for most of ML methods.
[3] Fork and contribute: <https://github.com/heberleh/proteins-stability>

ranking methods.

3.3.7 Creating signatures

There are different ways to create signatures to have their predictive power evaluated. One common way was explored in the assessment of double cross validation made by (CHRISTIN *et al.*, 2013). It consists in testing the top-N features from a given rank. Doing so, a total of N signatures is tested, with N being the maximum number of features. There are other ways to explore the domain of signatures, such as using random walks in a protein-protein interaction network, approaches based on the greedy algorithm or simulated annealing, and others approaches used when the feature space is big enough to not permit the evaluation of all combinations of proteins.

Here, we created signatures in two different ways: (1) using the top-N proteins from each rank, where N is less or equal to max_N - the minimum among 20, the number of features and the number of samples; (2) computing all combinations of size 1, 2, 3 and 4, and 1000 and 1200 random combinations of size 5 and 6, all considering the max_N. For data sets D1 and D2 the max_N is equal to 20 and, thus, the signature creation procedure was applied to each of the 40 ranks, for each of the training sets. This means that, for instance, in the case of D1 which DCV has outer-k equal to 9, a total of 360 ranks were used to form signatures using their top-20 proteins differently combined. We believe that the frequency of proteins in such signatures can indicate better biomarkers candidates for targeted proteomics. The parameters mentioned here could be set to higher values if a cluster of computers is used to run the scripts.

3.3.8 Test size for cross validation

The cross validation used in the general pipeline requires a testing-size for the Stratified Shuffle Splits and a K value for the K-Fold cross validation. We defined the K to be equal to the number of samples in the smaller class for internal validations (inner), such as for benchmark parameters and for computation of scores of proteins in ranking methods. The value of K for the outer loops (outer) from DCV is defined such each fold contains 15% of the samples, stratified. The test size to create the stratified shuffle splits used in the ranking process and in the signatures evaluation is defined as follows: if the number of samples in the smallest class is less than 13, the test size is equal to the number of classes, otherwise the test size is equal to 20% of the samples. The **outer** test size for D1 is 4 and for D2 is 6. The **inner** test-size for D1 is 3 and for D2 is 5.

3.3.9 Selecting the best classifier

Signatures created with the methods (1) and (2) (Section 3.3.7) are pre-evaluated using a single classifier and, then, the signatures with a good score are re-evaluated, using many different classifiers and a greater number of train and test sets. To find the *good signatures*, giving the high

demanding computational time of the tests, it is common to pick one unique classifier to compare the signatures' scores. One approach to choose the classifier would be using all features and a k-fold cross validation. Here, we decided to create Bagging Classifiers in a 32-Stratified Shuffle Splits cross validation, one for each of the following *regular* classifiers: Linear SVC, Radial SVC, Decision Tree, Lasso, Ridge, Linear Discriminant Analysis and Nearest Centroid. Each Bagging classifier is formed by 128 models. For example, the Decision Tree was used to create one Bagging Classifier formed by 128 Decision Trees, for each of the 32 different train/test sets drawn from the stratified shuffle splits.

The Bagging method trains many classifier models varying the samples, the features, or both, usually by bootstrapping. We considered that using Bagging Classifiers and bootstrapping only the features we could compare the *regular* classifiers in a way more close to the signature discovery process. Therefore, it would give us more information about the predictive power when using multiple feature spaces (128 estimators) than by creating *regular* classifiers using a space with all the features (1 estimator) since each Bagging classifier creates random signatures internally. In the end, the script selects the Bagging classifier that better predict the test sets from the 32 stratified splits, comparing their mean weighted F1 scores. When the subsets of samples are drawn as random subsets of the features, the Bagging algorithm is known as Random Sub-spaces (HO, 1998).

3.3.10 Finding good signatures

Once the best Bagging classifier was chosen, we used its simple classifier to compare the prediction power from each possible signature, according to the methods (1) and (2) from Section 3.3.7. The signatures are tested using a 32-Stratified Shuffle Split cross validation. Then, we rank the signatures according to their mean weighted F1 score. Signatures are considered good if their score difference to the top-1 signature is at most 0.10.

The *good signatures* sets are expanded to include all possible correlated signature. For each signature, we form all possible permutations using the correlated proteins and add the resulting new signatures to the *good signatures* set. All these signatures are evaluated as described in the last paragraph, with one difference: the scores are adjusted in a way that they are more penalized the lower is the mean frequency of signature's proteins and the higher is the score standard deviation (Equation 3.1). The mean frequency can change the cross validation (CV) score in up to 0.1. The standard deviation of the CV score is used to get the less optimistic score from the interval $[score - 2 * std, score + 2 * std]$.

$$adjusted\_score = ((signature\_mean\_freq/10) + 0.9) * (score - 2 * std(score)) \quad (3.1)$$

After computing the adjusted score, the signatures are ranked again and only the ones with a difference of at most 0.05 from the maximum adjusted score and from the maximum

original score remain in the set of *even better signatures*.

3.3.11 Selecting the best signatures

To select the best signatures, more filtering steps are executed. We defined the filtering based on the idea that we want to prioritize stable scores (low standard deviation), high differentiation on protein expression (low p-values), and, still, a good cross validation score. If a signature satisfies the criteria (1) and (2) from the list below, which is applied in the listed order, it is added to the *best signatures* set. The terms in the following steps refer to the average (mean_score), the maximum (max_score) and the standard deviation (std_score) of the non-adjusted score, and the average (p_value) and the minimum average (min_p_value) of FDR adjusted p-value of proteins in each signature. If more than ten signatures are selected after (1) and (2), they are filtered again by criteria (3). If there are more than ten signatures, they are filtered by criteria (4). If there are still more than ten signatures, the filtering steps are repeated reducing the accepted range. The remaining signatures form the *best signatures* set.

1. $signature.std\_score < min\_std + 0.025$;
2. $signature.mean\_score < max\_score - 0.025$;
3. $signature.mean\_p\_value < min\_p\_value + 0.02$;
4. $signature.mean\_score > max\_score - 0.05$;

3.3.12 Evaluating the best signatures

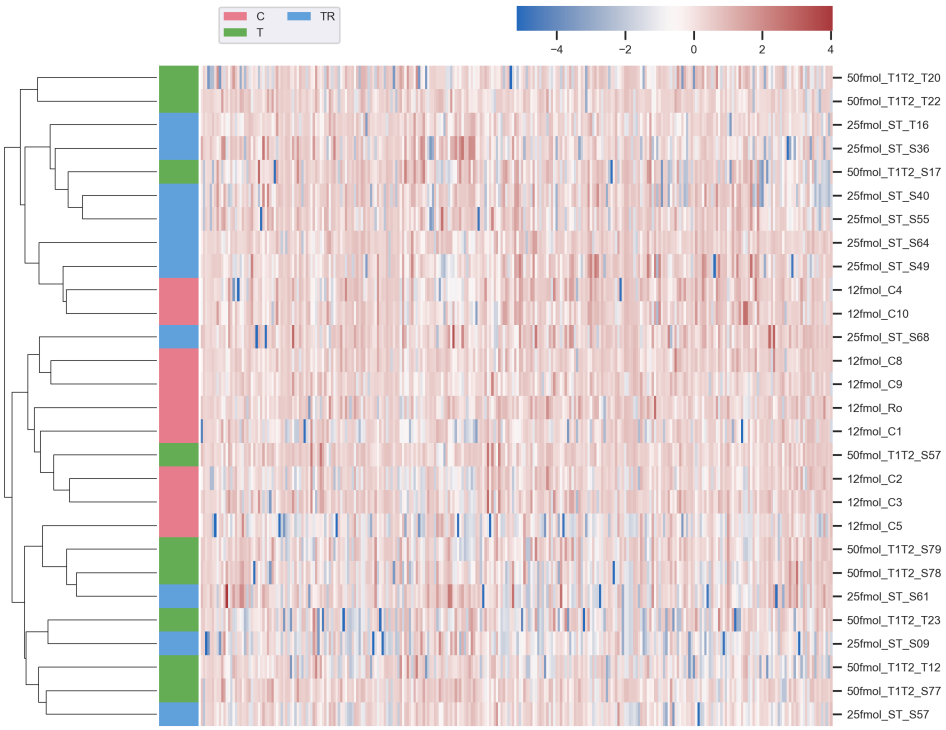
Since a signature can have different prediction power on different classifiers, we extensively evaluated them in a 64-Stratified Shuffle Splits cross validation. We also considered five additional classifiers other than the *regulars* cited in Section 3.3.9. One is the Random Forest (64 trees). Three are Bagging Classifiers created using 64 instances of the selected regular classifier (Section 3.3.9) in three different approaches enumerated in the list below. The last is a Voting Classifier with soft vote, it is formed by the three approaches of Bagging Classifiers with equal weights in the voting decision.

1. Bagging (BREIMAN, 1996): bootstrap samples;
2. Random Sub-spaces (HO, 1998): bootstrap features;
3. Random Patches (LOUPPE; GEURTS, 2012): bootstrap features and samples.

3.4 Case 1 - Monitoring the added proteins

In this section we analyze the results of the candidate biomarkers study on the data set D1. In Figures 15 and 16 we show the distribution of proteins in each sample and the $\log_2 fold-change$ (ratio) between different groups, respectively. As we can see, the samples were not correctly grouped by the clustering algorithm and only a few proteins have $|\log_2 ratio| \geq 1$. $\log_2 ratio$ is commonly used in the biological areas because the ratios are symmetric in zero and easier to understand than the ratio itself. For instance, a $\log_2 ratio = 1$ (or -1) indicates that one group is twice greater than the other and $\log_2 ratio = 2$ (or -2) indicates one is four times greater.

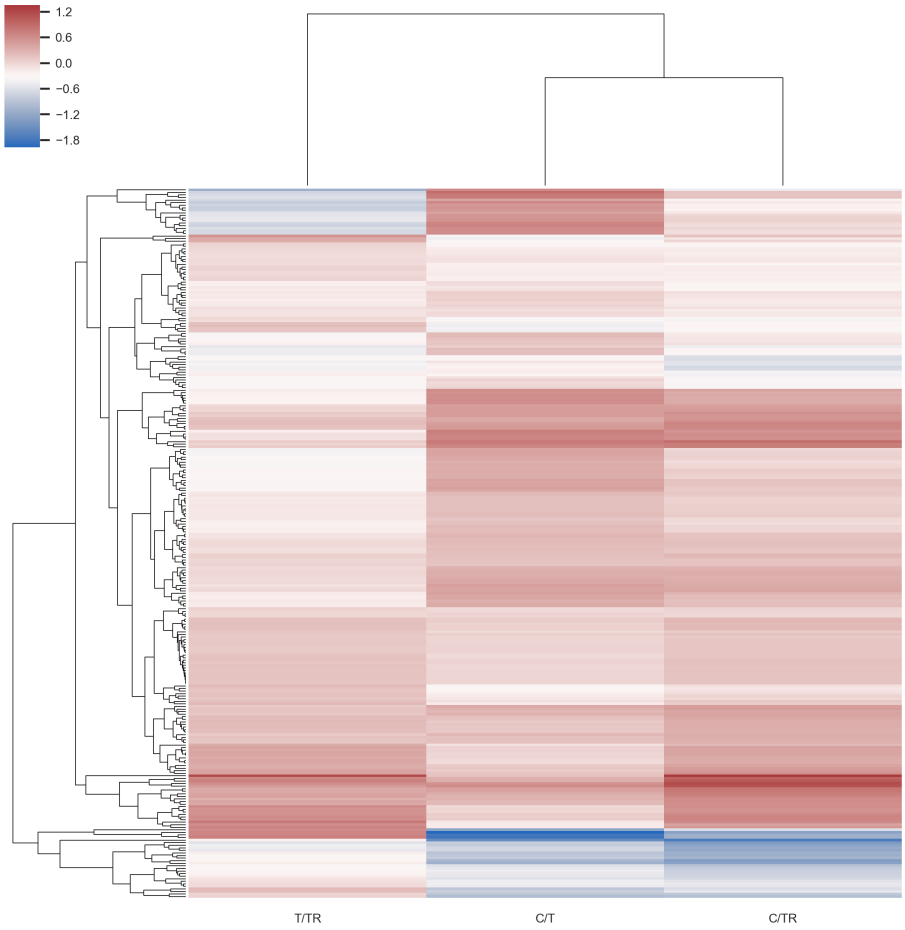
Figure 15 – Heat map of all samples clustered by hierarchical algorithm based on correlation. The control, tumor-removed and tumor samples were not separated by the algorithm. The Euclidean distance was tested and resulted in a worse hierarchy. Proteins were normalized by the standard score.



Source: Elaborated by the author.

In both Figures we see that groups TR and T have more similarities between them than with C. The mean intensity in each class also reveals some patterns that could indicate good candidate biomarkers if the actual difference between distributions of intensities agree with the mean difference. For instance, some proteins seem to be up-regulated in T and TR and down-regulated in C; some seem to be up-regulated in C and down-regulated in TR and T; some seem to be up-regulated in C and down-regulated in T and TR; among other patterns that could be used to create a logic of up- or down-regulation to predict an outcome.

Figure 16 – Heat map of $\log_2 ratio$ values between classes. Each protein was scaled with the MinMaxScaler (Scikit-learn developers, 2017) to be in the interval $[0.00000001, 1]$ before computing the ratios, avoiding $\log_2 0.0$.



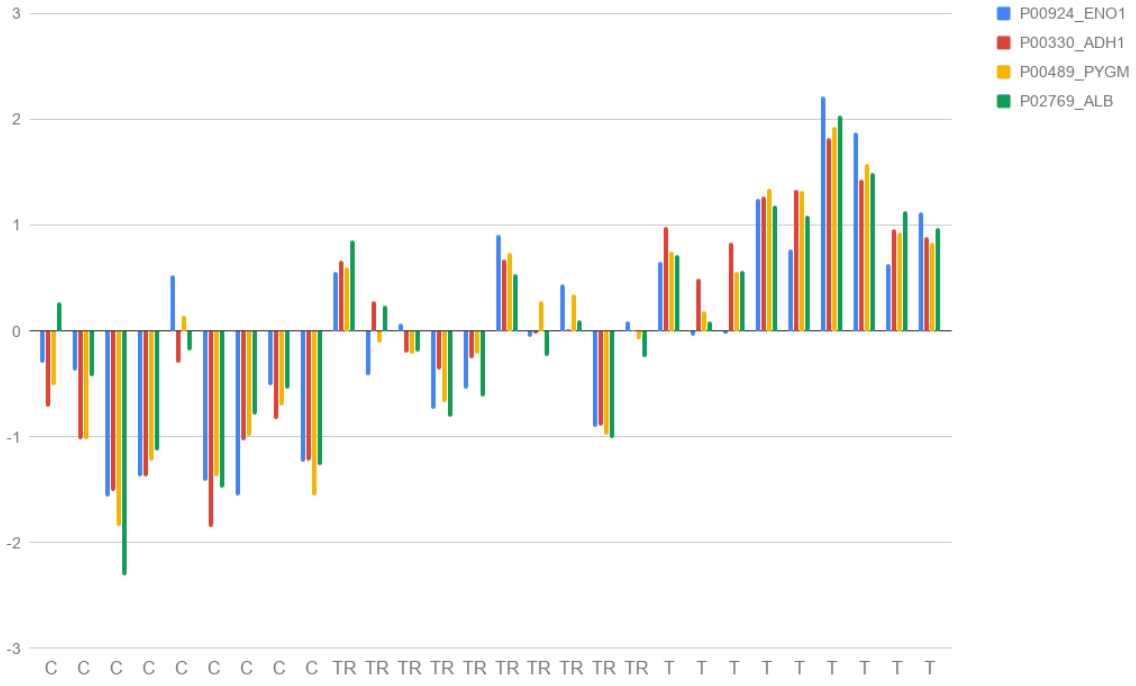
Source: Elaborated by the author.

While analyzing the averages can be simplistic, we expect that the multivariate approach proposed here can reveal proteins which distributions are indeed different among classes and that are good candidates to create a future prediction model. Despite our added *true positives* represent different uniform distribution in the biological samples of different classes, the discovery proteomics' limitations make the data distribution not uniform (Figure 17). Despite of the noise in the quantification data, we can still see the differences between classes. This represents well the reason why we want to check if discovery proteomics can reveal such *perfect* markers, considering the protocol's limitations.

3.4.1 Analysis of one fold from DCV

Here we report the results of one split of the data set into train and test sets. The splits of the double cross validation are done randomly and, thus, we selected the fold number 0 to describe in this section. After analyzing the single fold, we discuss the stability among all folds from the DCV, giving a general idea about how results change among different data set splits

Figure 17 – Distribution of standardized intensities of added *true positive* markers. Discovery proteomics quantification introduced noise in the three uniform distributions (O, TR and T).



Source: Elaborated by the author.

and how stable proteins are.

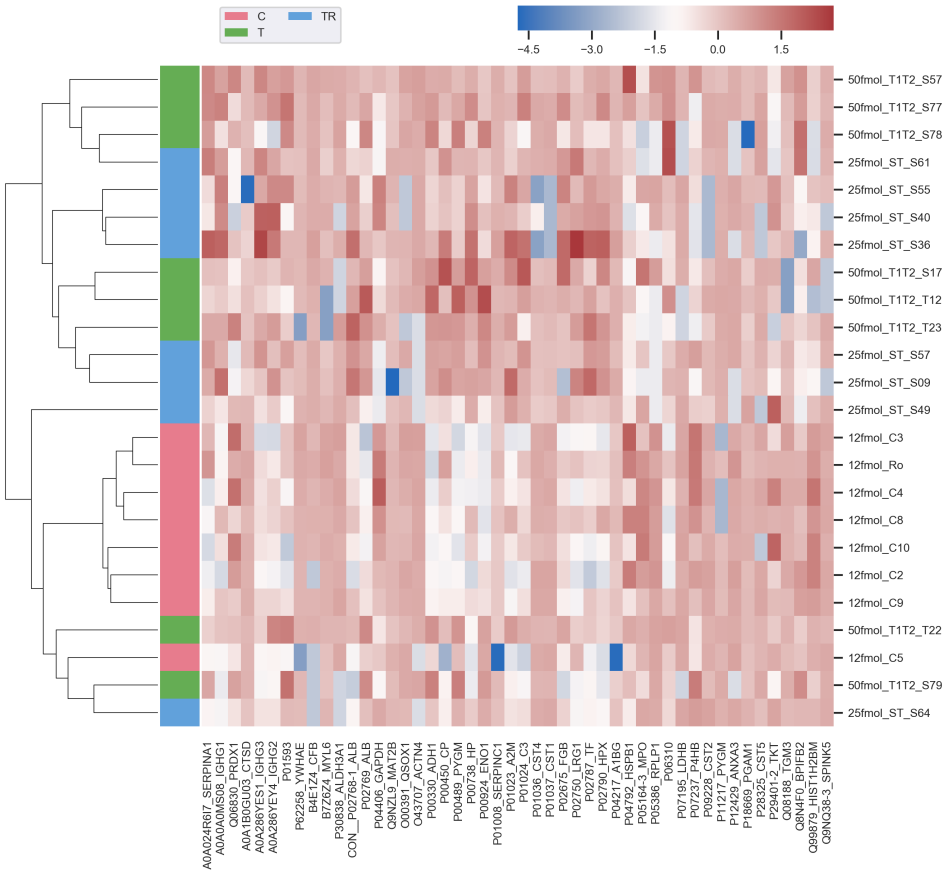
3.4.1.1 Filtering

After filtering the proteins by Kruskal Wallis test (p-value < 0.05), the dimensions were reduced from 276 to 48. The Figure 18 shows the features' values of each sample normalized by the standard score. The Kruskal filter was not enough to group the classes in the hierarchical clustering.

Aiming to see how a statistical test for FDR would perform, we used the Benjamini-Hochberg multiple tests procedure with cutoff of 0.05 and only 5 proteins were selected. Among them were the four added *true positives* and HIST1H2BM (Q99879), with adjusted p-value < 0.029. Given that the uniform distribution of the added proteins make them *perfect* candidates, the FDR applied in the noisy discovery proteomics data resulted in a reliable list.

Despite the result obtained by the FDR filter indicated five proteins, we believe that given the small number of samples a strict cutoff may increase the false negative rate drastically. Also, it limits the proposed approach defined here, once we aim to compare multiple ranks and signatures. For this reason, we decided to allow a FDR of 0.25 and the number of selected proteins was increased to **35**.

Figure 18 – Heat map showing the intensity values of proteins with Kruskal Wallys p-value < 0.05. The dendrogram shows the results of hierarchical clustering using average linkage and correlation between samples.



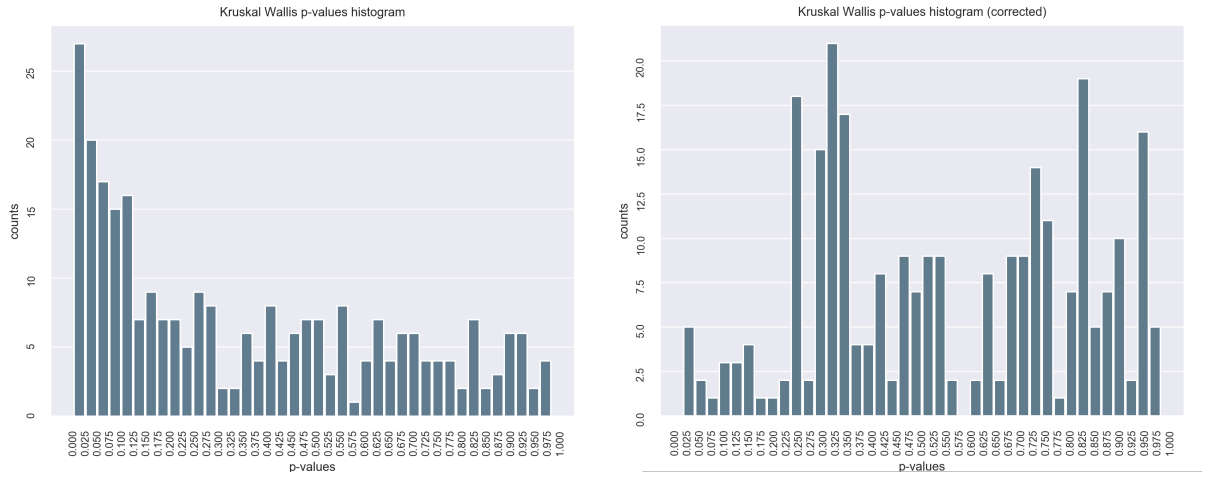
Source: Elaborated by the author.

The Figure 19 shows the p-values histogram for the Kruskal and the adjusted p-values by the FDR test. When filtering proteins by Kruskal Wallis p-value < 0.05, 48 proteins remain; in contrast, when defining an acceptable false discovery rate of 25%, the number dropped to 35. In the Figure 20 we note that the proteins selected by FDR < 25% did not improve the hierarchical clustering.

The $\log_2 ratio$ between classes' mean is represented in Figure 21. Again, we can see that the greatest differences (darkest colors) are in the ratios that compare the class C against T and TR. The intensities' mean of the 4 added proteins agree with the biological concentrations, that is, $\bar{\mu_C} < \bar{\mu_{TR}} < \bar{\mu_T}$.

From the remaining 35 proteins, 7 proteins were highly correlated to another and formed 3 groups. One protein of each group remained in the training data set while the other 4 proteins were removed: **P00330_ADH1**, **P00489_PYGM**, P04217_A1BG, P09228_CST2. Henceforth, the added proteins P00489_PYGM and P00330_ADH1 were represented by P02769_ALB in the training set. The other correlated proteins were represented as follows: P01037_CST1 represented P09228_CST2; and P01008_SERPINC1 represented P04217_A1BG. All pairs match the criteria

Figure 19 – Histograms of Kruskall Wallis p-values and adjusted p-values (Benjamini-Hochberg) of all proteins.



Source: Elaborated by the author.

of having correlation greater than 0.95. In Figure 22 we see the absolute value of correlation between proteins.

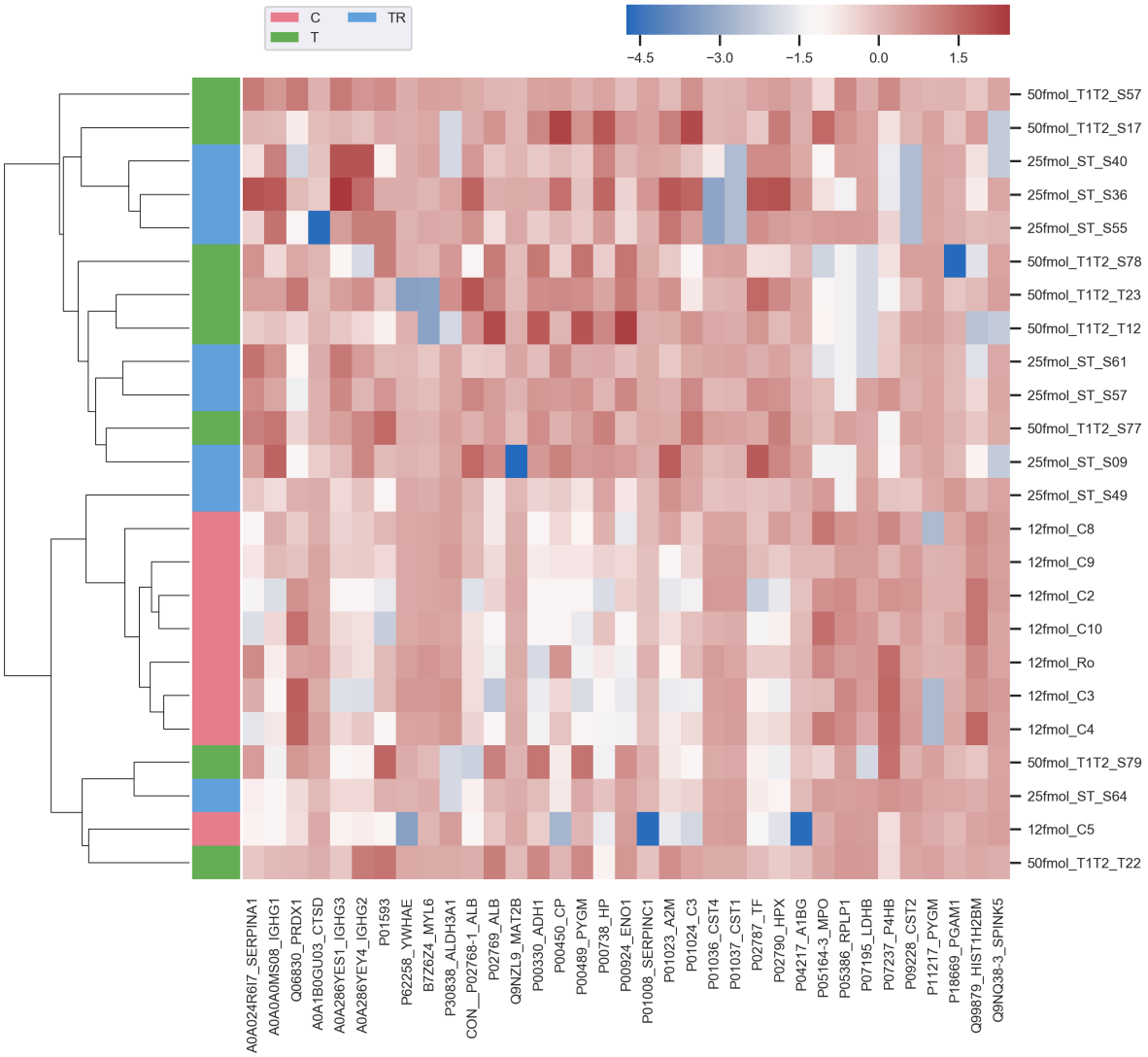
3.4.1.2 Ranking

The 7 types of rankers comprising 6 different classifiers and statistical methods were applied to the train set. Figure 23 presents the heat map of all proteins' scores (scaled in [0,1]). The rankers disagree and give different scores for the same protein but some proteins were more frequently found with good scores than others. Figure 24 focuses the visualization on the top-10 proteins of each rank, showing that the variance is high for all proteins but lower for the added proteins and some others. The high variance of scores indicates that the proteins should not be prioritized considering only one method and highlights the importance of studying different results about the same data set.

We analyzed the top-10 proteins from the ranks selecting the ones that appeared in more than 33% of ranks and creating the box plots of Figures 25 and 26. In Figure 25 we can see that the scores varied from 0 to 1 even for the best proteins. Despite some proteins had evidences on being a top-protein, e.g. P11217_PYGM, we cannot guarantee that the other proteins that had a high score on less than 33% of ranks are not good candidates. Hence, in our pipeline, we considered proteins indicated among the top-10 proteins of any rank to create the signatures. Both ENO1 and ALB appeared with good mean scores and variance in comparison to other proteins.

In view of the variation among ranking methods, we have shown that selecting proteins by top-N is not always a good approach. After all, combining the score and position information as shown in the Figure 26 can lead us to a better comprehension of the importance of each protein

Figure 20 – Heat map showing the intensity values of proteins with FDR < 0.25. The dendrogram shows the results of hierarchical clustering using average linkage and correlation between samples.

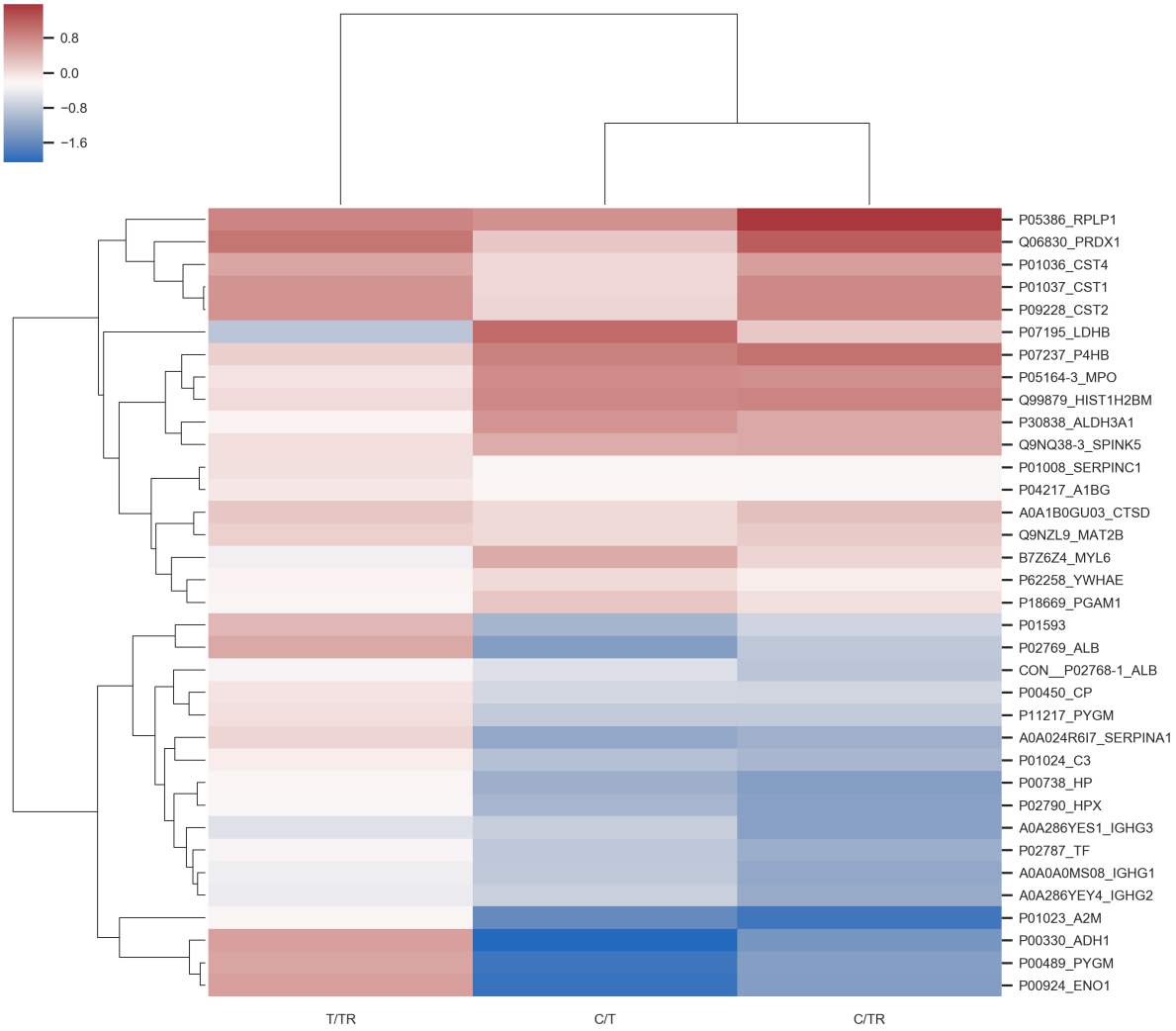


Source: Elaborated by the author.

to differentiate the samples of the train set. Selecting the 50% higher scores and proteins in the top-10 list of at least 33% of the ranks is consistent with the hypothesis that there are good and bad rankers. In the same fashion, selecting all top-10 from each ranker gives the proteins similar opportunity to be selected to further analysis and, at the same time, introduce variance in the signatures that were tested in further steps of the proposed pipeline.

The added proteins and P11217_PYGM had a score greater than 0.70 for most of the rankers. However, 29 of 40 ranks attributed a score < 0.70 for at least one of these 5 proteins. The types of rankers and the associated multivariate models that scored them as bad candidates varied for from protein, even for the equally concentrated added proteins. Thus, it suggests that it would not be correct to establish what were the overall worst rankers by means of discovery proteomics data.

Figure 21 – Heat map of $\log_2 ratio$ values between classes. Each of the 35 proteins filtered by FDR < 0.25 was scaled with the MinMaxScaler (Scikit-learn developers, 2017) to be in the interval $[0.00000001, 1]$ before computing the ratios, avoiding $\log_2 0.0$.

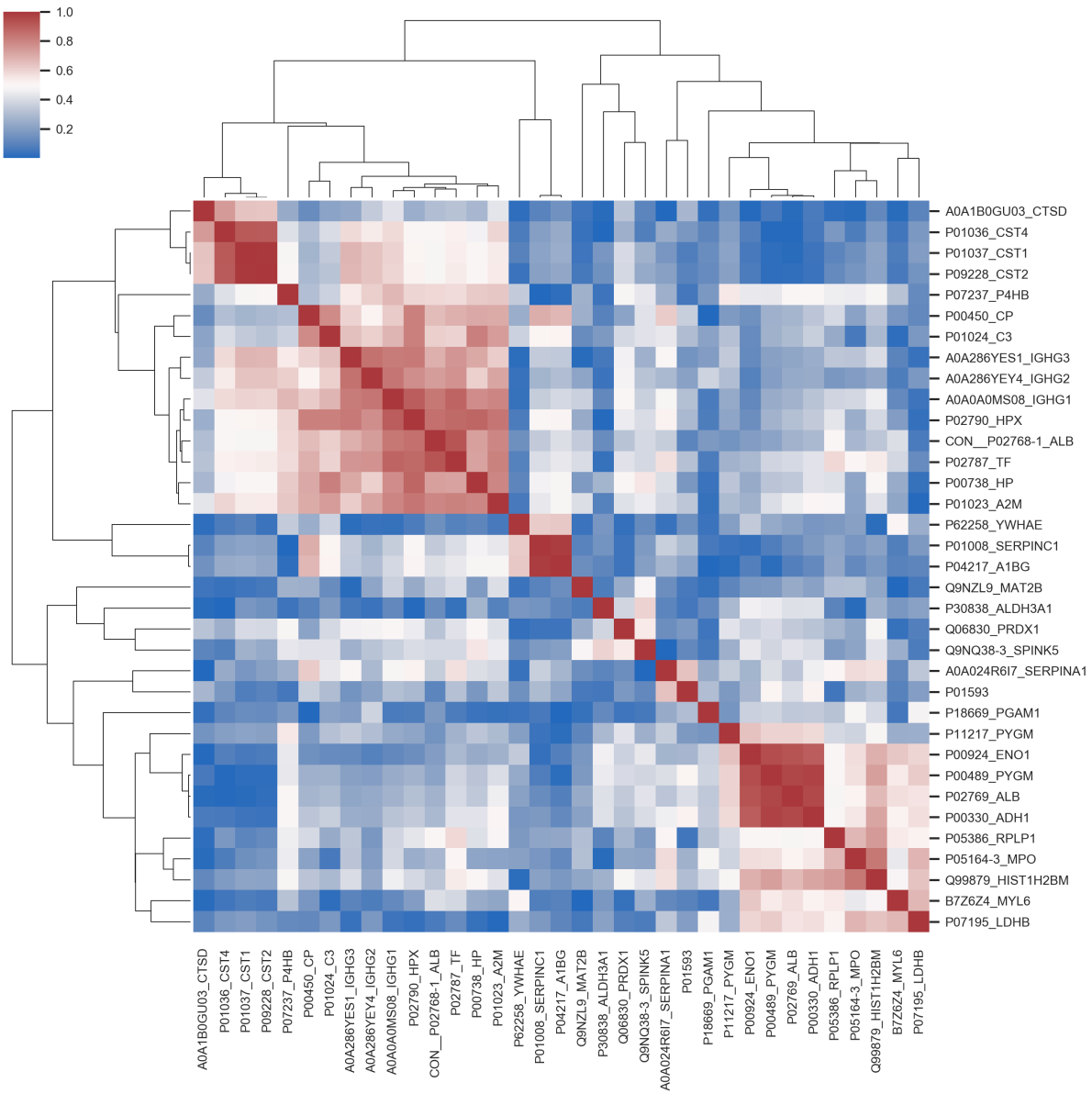


Source: Elaborated by the author.

Considering the two added proteins, 26 out of 40 rankers scored them as < 0.70, being 9 exclusives of ENO1, 1 exclusive of ALB and 16 in the intersection. ENO1 had more rankers scoring it with values below 0.7, which could mean that the rankers are not a good measurement of importance or that ENO1 was quantified with much noise from the mass-spectrometer. By all means, ENO1 presented a different distribution and was not considered highly correlated (0.95) to the other three proteins. In essence, we believe that ENO1 have more noise and that the rankers can be used to identify good candidates from discovery proteomics for further experiments.

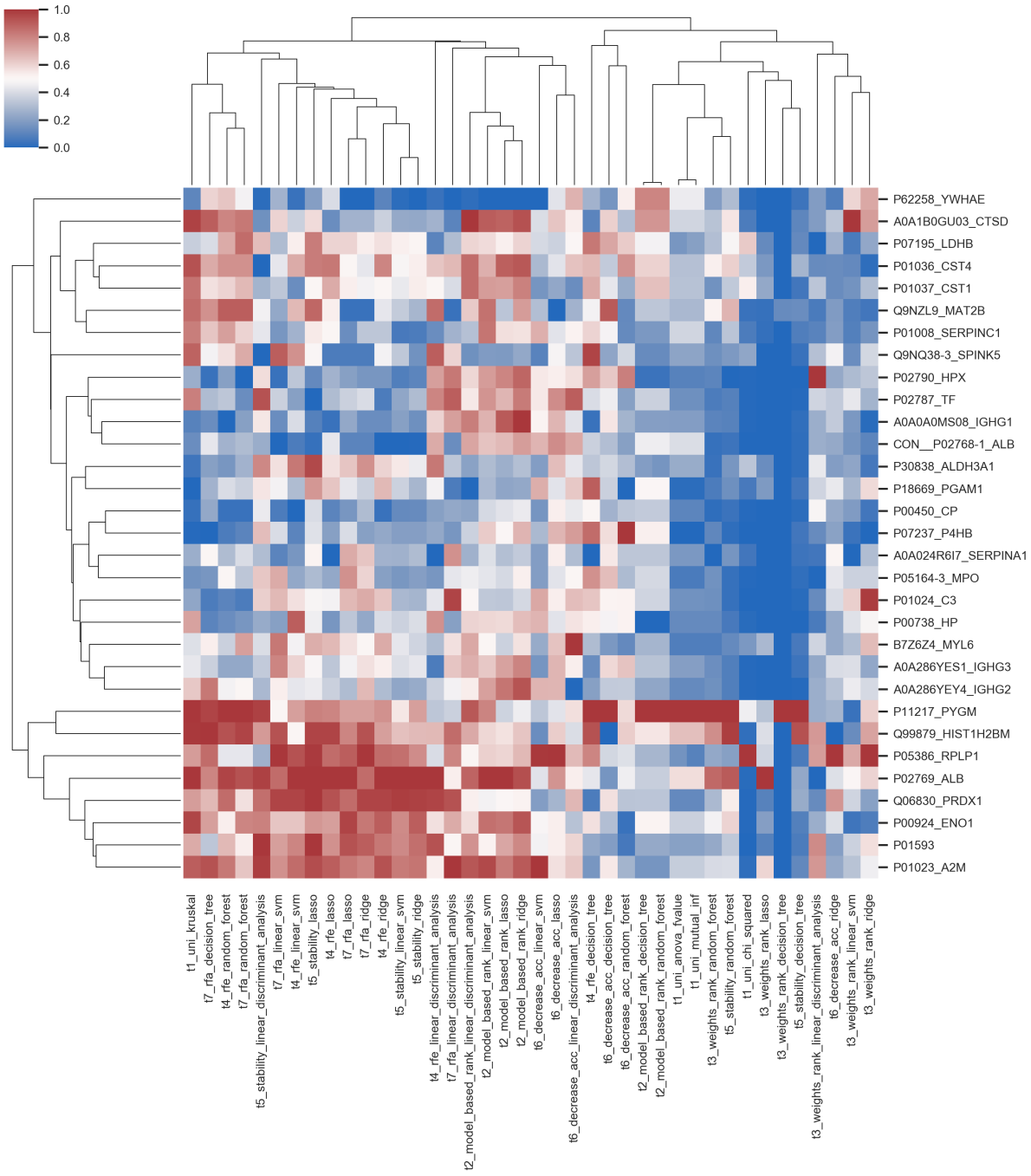
The top-10 frequency of proteins demonstrated to be a good measurement for the comparison of importance of proteins. Both ALB and ENO1 had a good top-10 frequency in the 40 ranks (Figure 27): 0.72 and 0.57, respectively. Between them there are two proteins with good frequency: P11217_PYGM (0.70) and P01023_A2M (0.67). In particular, ALB and

Figure 22 – Heat map showing the absolute value of correlation between proteins with FDR < 0.25. The dendrogram shows the results of hierarchical clustering using average linkage and correlation between the correlation distributions.



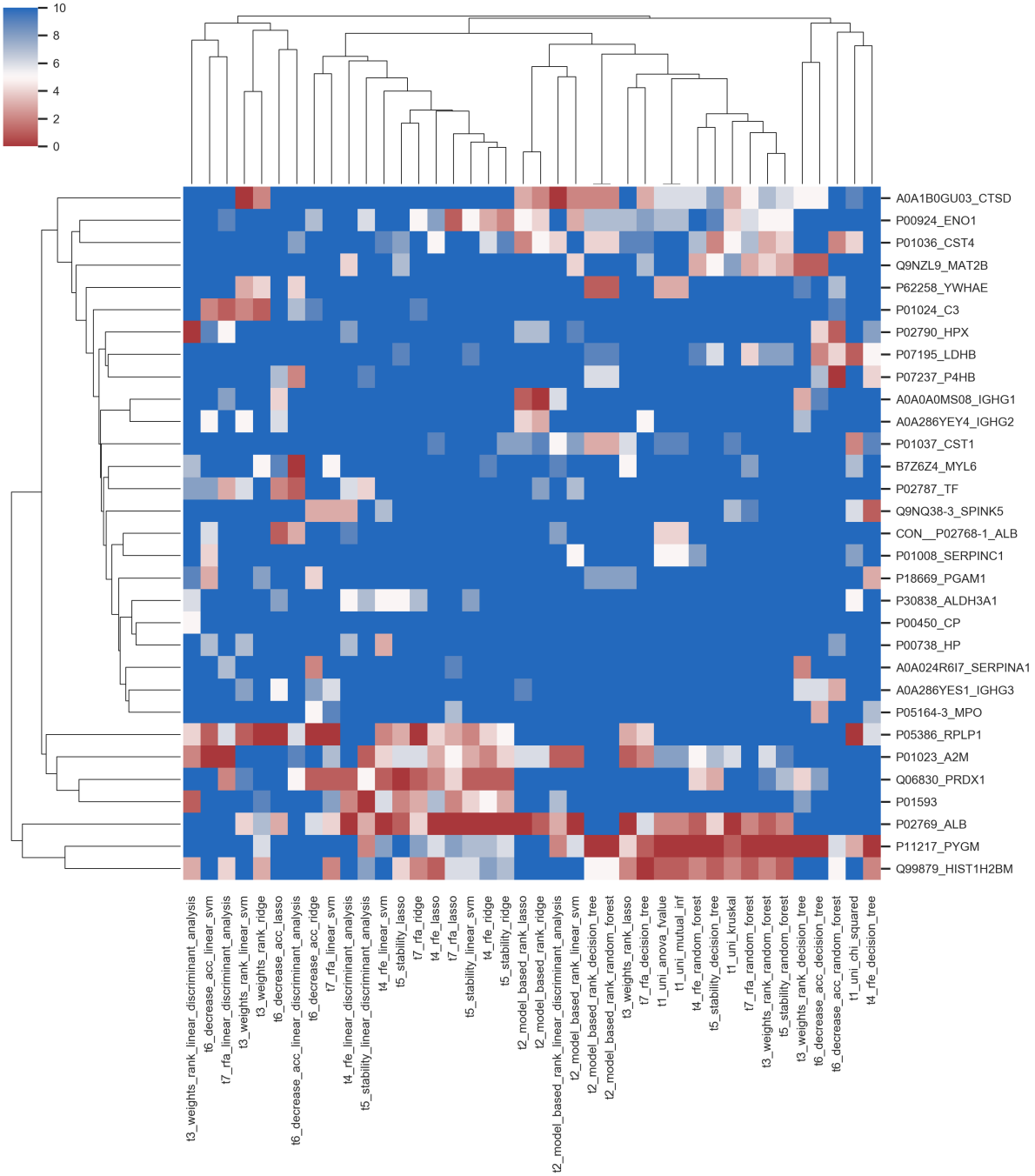
Source: Elaborated by the author.

Figure 23 – Heat map showing the scores of proteins in 40 different ranks. The dendrogram shows the results of hierarchical clustering using average linkage and Euclidean distance between the score's distributions.



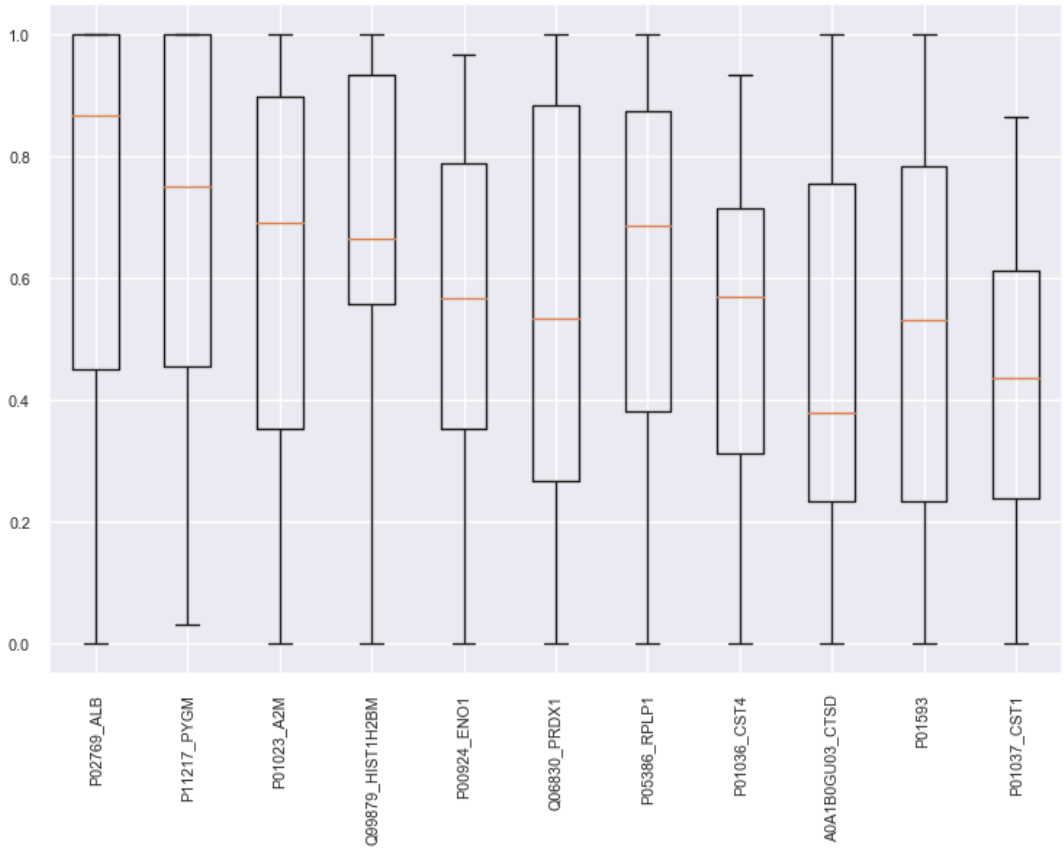
Source: Elaborated by the author.

Figure 24 – Heat map showing the rank position of proteins that appeared in the top-10 proteins of 40 different ranks. If protein had position greater than 10, it was set to 10 to highlight the frequency of each protein among the top-10 lists. The dendrogram shows the results of hierarchical clustering using average linkage and Euclidean distance between the position's distributions.



Source: Elaborated by the author.

Figure 25 – Boxplot of scores from proteins that appear as top-10 in more than 33% of ranks.



Source: Elaborated by the author.

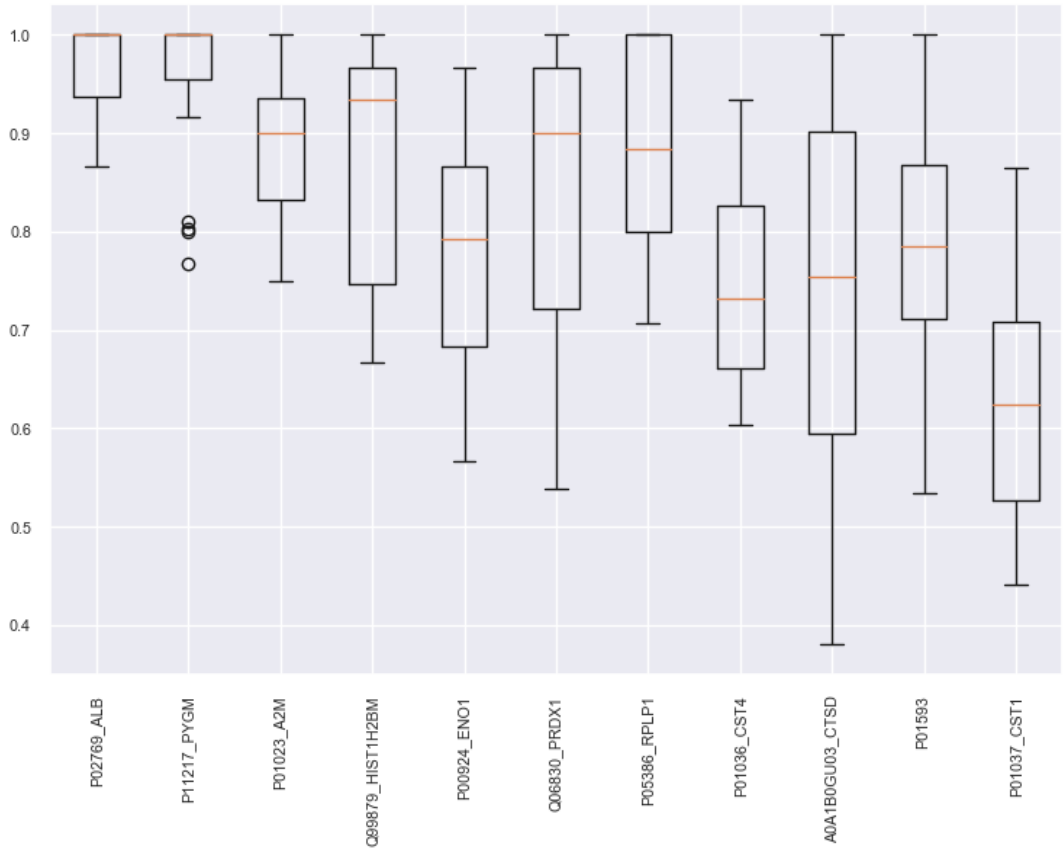
P11217_PYGM had equal scores' mean but, while ALB had higher variance, P11217_PYGM was considered having outliers indicating big variance in a few ranks (Figure 26).

3.4.1.3 Cross validation

Lasso was the selected classifier for the cross validation part of the pipeline. Including the correlated signatures, a total of 1,803 were selected as good candidates. From these, only 26 signatures remained in the *even better* list.

The Table 2 shows the 9 final best signatures from fold 0, selected based on the cross validation and criteria stated in Section 3.2. Notably there was a great variance of prediction scores in each iteration of the cross validation applied for each signature (see standard deviation). This was expected to happen since the distribution of intensities were not uniform in each class. For instance, in the case of the Oncotarget work illustrated in Figure 5, we see that the selected proteins can clearly discriminate the classes. Such pattern does not happen with the data set studied here. The samples studied in that project do not represent a common configuration because they were controlled model cells while the common case is the analysis of different samples from patients. The data set described in this section represents the configuration of

Figure 26 – Boxplot of the 50% highest scores from proteins that appear as top-10 in more than 33% of ranks.



Source: Elaborated by the author.

most discovery proteomics data sets; to put it another way, a data set with noisy intensities as exemplified in Figure 17.

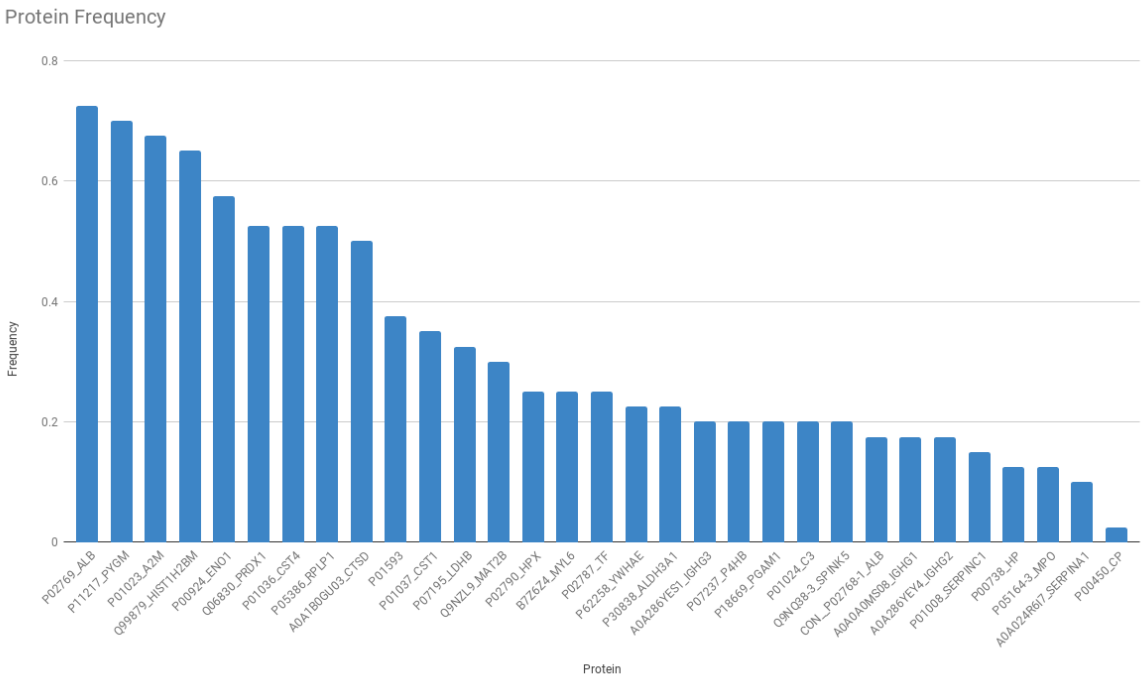
Another negative side of selecting the best signatures was that our added true positive markers were not so frequent. This may have happened due to the fact that, for this fold, there were better markers and, thus, these turned to be less frequent. As you will see in the next sections that explain the results of all folds, this is a particular case of fold-0. That is, the most frequent proteins in the best signatures vary much with small changes in the training sets. So a protein may be frequent in this fold but if it is not in the other folds, this protein should not be prioritized, it is not stable.

3.4.1.4 Closing remarks

In this section we analyzed one split from the double cross validation. We have shown the most important features found by the pipeline, their scores and rank stability, and that only part of the ranking methods identified the added proteins as biomarkers candidates.

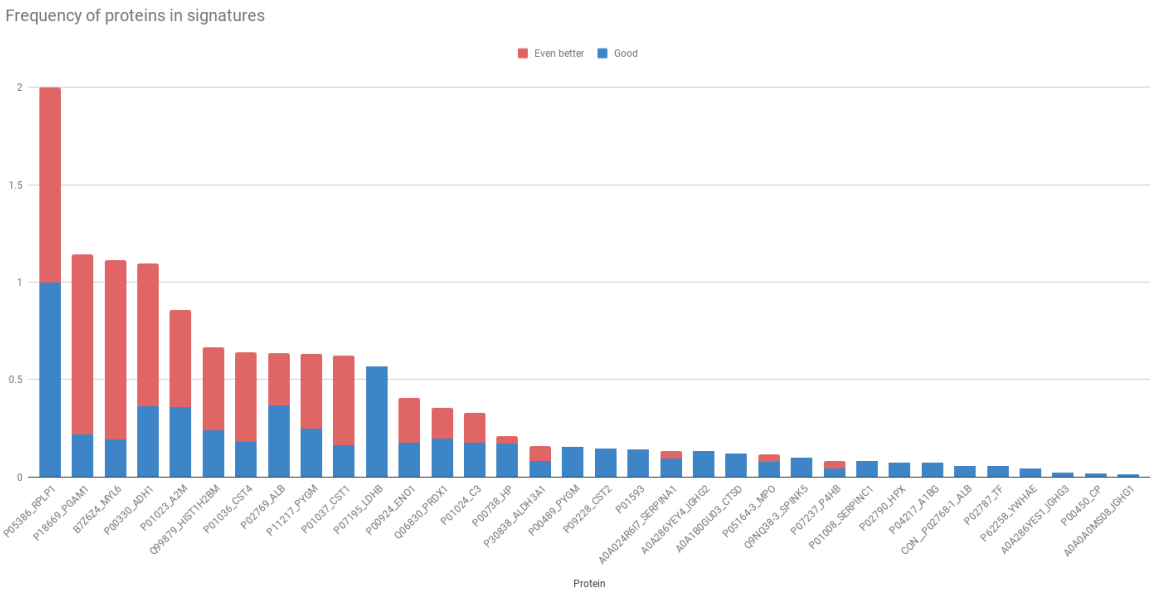
The use of many methods and a percentage of the highest scores demonstrated to be a

Figure 27 – Plot of the frequencies of proteins in top-10 lists of the 40 ranks. There are 12 proteins with a frequency greater than 0.3 and 8 proteins with frequency greater than 0.5.



Source: Elaborated by the author.

Figure 28 – Plot of the frequencies of proteins in *good* and *even better* lists of signatures.



Source: Elaborated by the author.

Table 2 – Best signatures and scores from fold 0 of the DCV.

	S0	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
RPLP1	✓	✓	✓	✓	✓	✓	✓	✓	✓
A2M	✓	✓	✓	✓	✓	✓	✓	✓	
HIST1H2BM	✓	✓	✓	✓	✓	✓	✓	✓	
MYL6	✓	✓	✓	✓	✓	✓		✓	✓
P11217_PYGM	✓	✓	✓	✓	✓	✓	✓		✓
PGAM1	✓	✓	✓	✓	✓	✓			✓
CST4	✓	✓	✓	✓			✓		✓
ADH1	✓		✓		✓		✓	✓	✓
CST1	✓	✓	✓	✓	✓	✓			
ENO1	✓	✓	✓	✓	✓	✓			
ALB		✓		✓		✓			
PRDX1							✓	✓	
ALDH3A1	✓	✓							
Cross valid.	91.7%	89.8%	91.7%	91.7%	91.7%	91.7%	91.7%	91.7%	91.7%
Stand. Dev.	17.3%	20.9%	17.3%	17.3%	17.3%	17.3%	17.3%	17.3%	17.3%

reliable approach in contrast with the usage of just a few rankers, in view of the fact that the 40 rankers positioned the proteins differently and that we cannot not choose specific ones being the truth. Analyzing the top-N proteins of each rank and their frequency is a way to understand and select stable proteins for further investigation.

Given that some classifiers may consider as important only one protein and return bad scores for its correlated proteins, we believe that the lowest scores should not be considered as valid. For instance, SVM-RFE usually position correlated proteins in very different positions of the rank. Thus, while selecting the best scores seems to be a good approach, it may indirectly increase the false negative rate.

Considering this data set and the added proteins, our results suggest that the greater is the number of rankers that give a high score to a protein, the higher is the probability of the hypotheses being true. Selecting proteins that are in 33% of the top-10 proteins from each rank also demonstrated to help in the identification of interesting features.

In the next section we analyze the variance of results among different splits from the double cross validation. We discuss again how the small number of samples makes the results unstable and how we can handle this problem to select a set of proteins for targeted proteomics, similarly to this section.

3.4.2 Analysis of all data splits from DCV

This section reports the results of the other folds in a comparison analysis. We describe similar subsections and concepts of the Section 3.4.1, now basing our results and conclusions on average information from the complete double cross validation.

3.4.2.1 Ranking

Figure 29 shows the minimum score ($mean - 2 * std$) for each ranker, considering the DCV loops. We can see two lines with many scores above 0.5 at the top and another group at the bottom. Most of the proteins had many unstable scores and, as a result, a great *std* value. These bad scores are represented mostly by blues (< 0.3). Again, despite all proteins have at least one *blue score*, there are groups, such as the top and bottom ones, that shows a considerable amount of *red scores*, while most of the other lines are completely white and blue (< 0.55).

The Figure 30 shows the average of the 50% highest scores among the DCV folds. We can see that proteins P11217_PYGM and ALB were considered with great scores (near 1.0) in all the 9 folds. Besides, protein ADH1, ENO1 and CST1 also have the good scores and got score 0.0 in fold-0 and fold 4, respectively, because they were removed from the training set due to high correlation. In the case of proteins GC, CFB, SERPINC1, which are in the bottom *red group*, they did not pass the FDR filter, which is the case of most of the proteins in the *blue group*. The *blue group* represent the most unstable proteins, since their p-values varies much with small changes in the data set (simulated by the DCV).

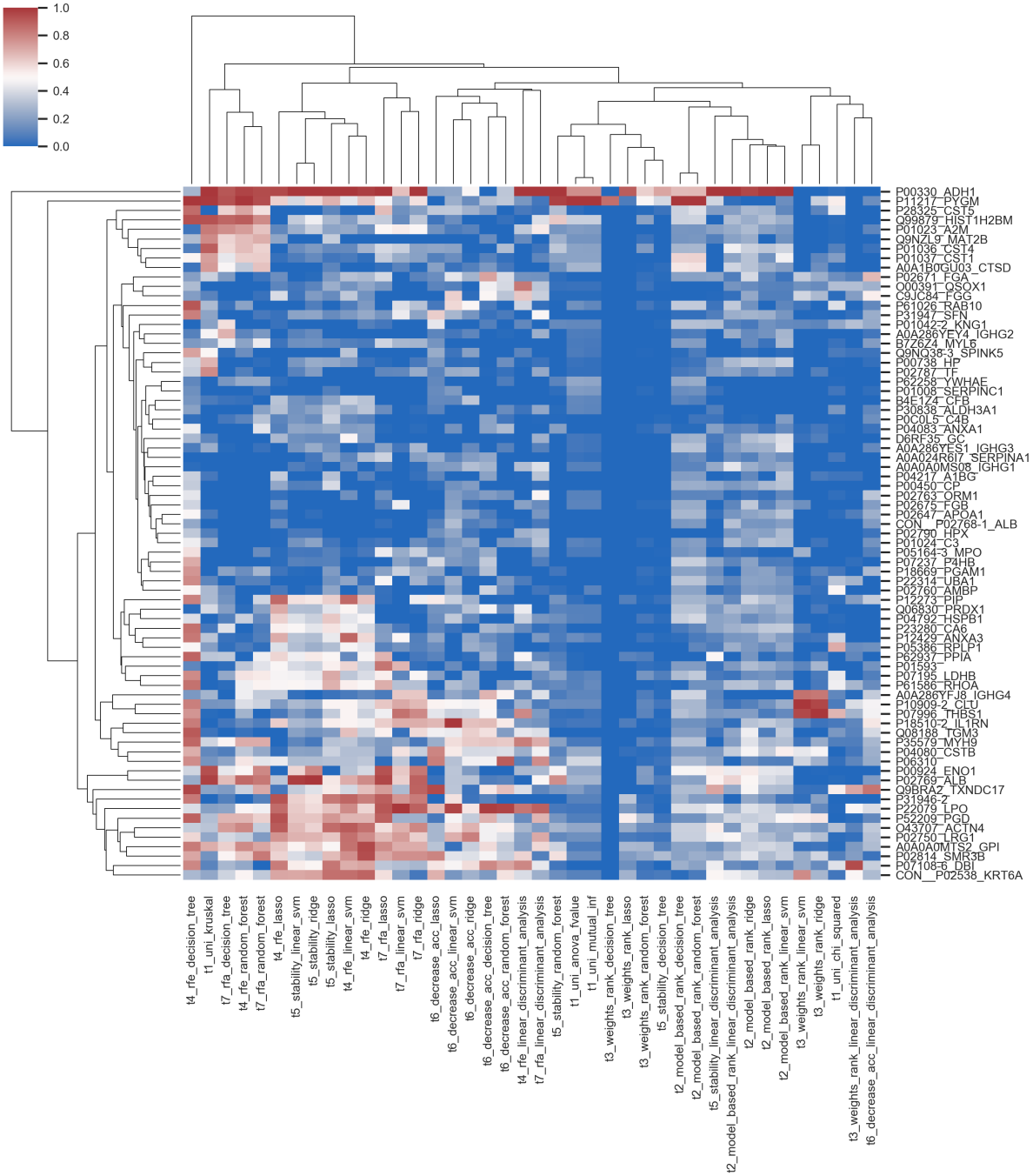
The Figure 31 also shows the 50% highest scores of the top-10 proteins of each fold, considering only proteins that appeared in at least 33% of the folds and in at least 33% of the ranks in each fold. If a protein is not found in a rank, it is scored with a small negative value. We can see that ENO1 and P00489_PYGM were not selected for this plot, once it is considered as highly correlated and, thus, did not appear in the training set of many folds DCV. All these proteins had high scores and are the best candidates according to the ranking methods since they were also very frequent among the top-10 proteins in the $9 * 40$ ranks.

The same idea of checking the frequency among the top-10 proteins of all folds was applied to create the Figure 32. Here, we can see that ENO1 was selected in only 4 folds and P00489_PYGM was not selected at all (due to high correlation). We can say that the 4 markers, thus, are among the best candidates according to these criteria, including ENO1 and PYGM that when not selected for analysis were highly correlated (>0.95) to ADH1 or ALB.

The proteins were again analyzed comparing their frequency in *good* and *even better* signatures from all folds in the DCV (Figure 33). Here, it is important to note that the proteins highly correlated and removed from the training set were not used to form the *good signatures* that were further expanded by forming the correlated signatures. Henceforth, as a limitation of the pipeline, proteins ENO1 and specially P00489_PYGM may have had a decrease in the frequency in the *good* and *even better* signatures. Be that as it may, the two proteins were scored similarly to protein ALB, which always represented the highly correlated group when it happened; that is, it was never removed from training sets.

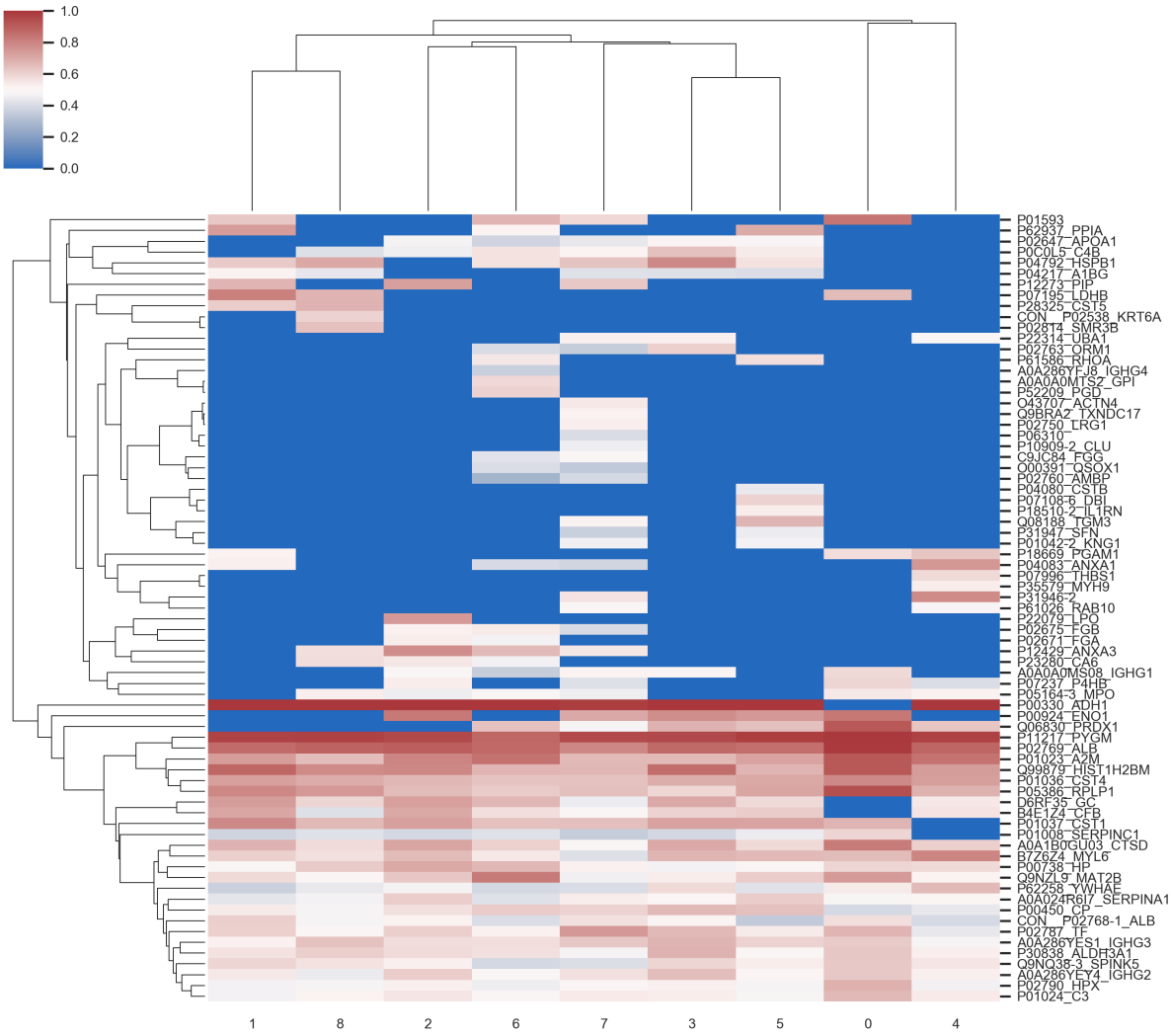
Not every protein that had the highest frequencies in *good* signatures had the highest frequencies in *even better*, and vice versa. Hence, giving the high variance in the cross validation

Figure 29 – Lowest scores ($mean - 2 * std$) calculated with the mean and the standard deviation of scores of each protein by each ranker among folds. Negative values were set to 0.0.



Source: Elaborated by the author.

Figure 30 – Heat map showing the mean top-50% scores from each loop. If a protein was not selected in a loop by the FDR filter, it is scored with 0.0.

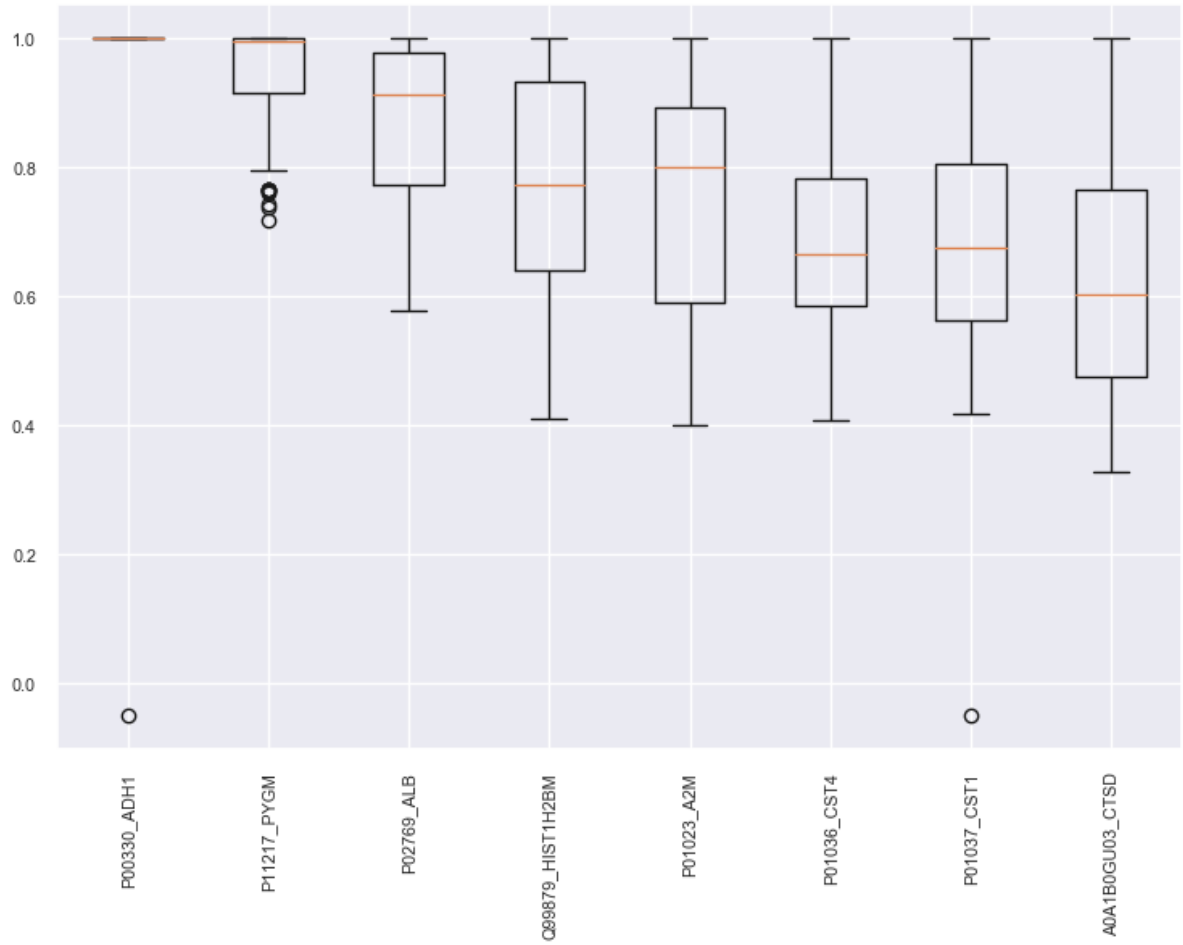


Source: Elaborated by the author.

prediction scores, it is interesting to analyze both scores. For this reason we combined the frequencies of being in *good* and *even better* signatures as shown in Figure 33. The plot demonstrates the variance of scores and shows how the rank may change depending on what approach we chose to make decisions.

Regarding the RFE and RFA methods, we compared the scores of ADH1, ENO1 and ALB given by each ranker. We show the box plots of values attributed by the two methods in Figure 34. For this data set, the RFE method is more unstable than the RFA. Considering that these proteins are true markers, RFA also resulted in greater scores than RFE. Once the median and mean are close, in each distribution, we tested using a paired two-sided T test (17 degrees of freedom), resulting in a p-value of 0.00347. Thus, we reject the hypothesis that the average differences are equal to zero and conclude that we are 95% confident that the RFA scores were increased by 6.42% to 27.54% in comparison to RFE scores on average. This method must

Figure 31 – Boxplot of 50% highest scores of each protein in each loop, considering the proteins that appeared at least in 33% of the folds being a top-10 in at least 33% of the internal ranks.



Source: Elaborated by the author.

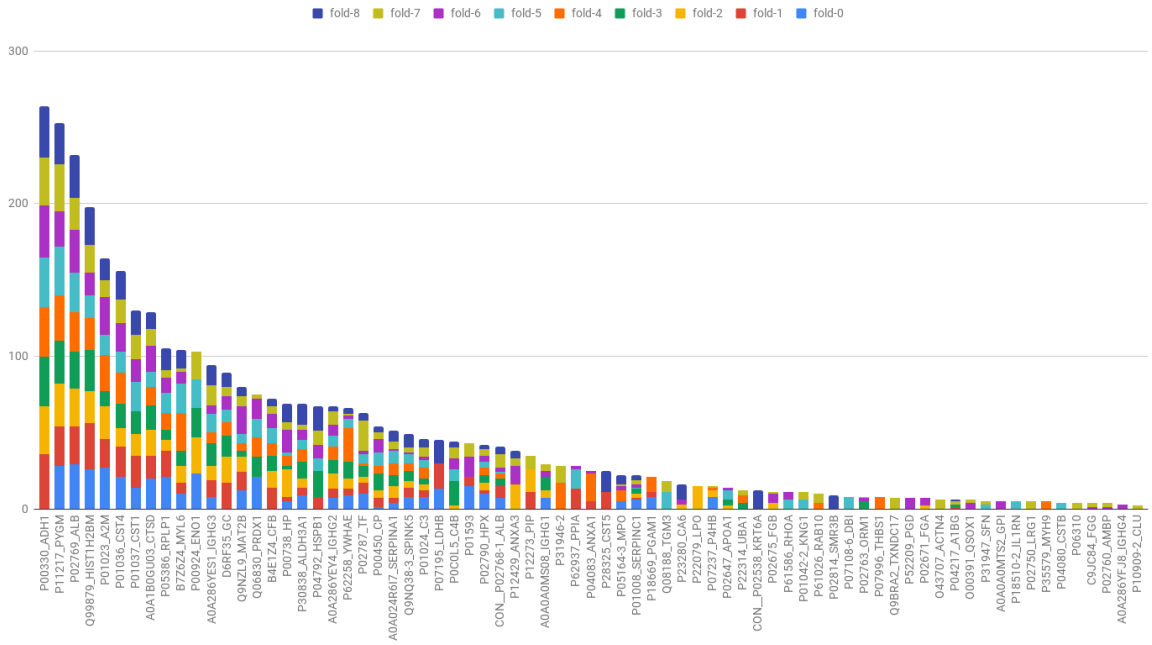
be extensively assessed in future studies. Its implementation in Python is publicly available at <https://github.com/heberleh/proteins-stability>.

3.4.2.2 Potential signatures

If the data set were homogeneous, and the rankers were stable for the frequently selected proteins, we would expect that the final signatures from each train set are similar. Actually, their similarities remained on the proteins that form the signatures and not in the signatures themselves. That is, despite there are some proteins that were frequently found in signatures from different rankers and different train sets, the actual signatures did not repeat among the DCV outer loops.

In total, the pipeline identified 52 best signatures. The average simple (mean) cross validation score of the best signatures in all outer-training sets of DCV is 89.57% with 6.20% standard deviation. Assuming a t distribution for the mean CV score, we would expect it to be in the interval $89.57 \pm 1.72\%$ with 95% confidence (52 samples and 51 degrees of freedom). In

Figure 32 – Plot of the number of times that each protein appeared among the top-10 of each rank in each fold.



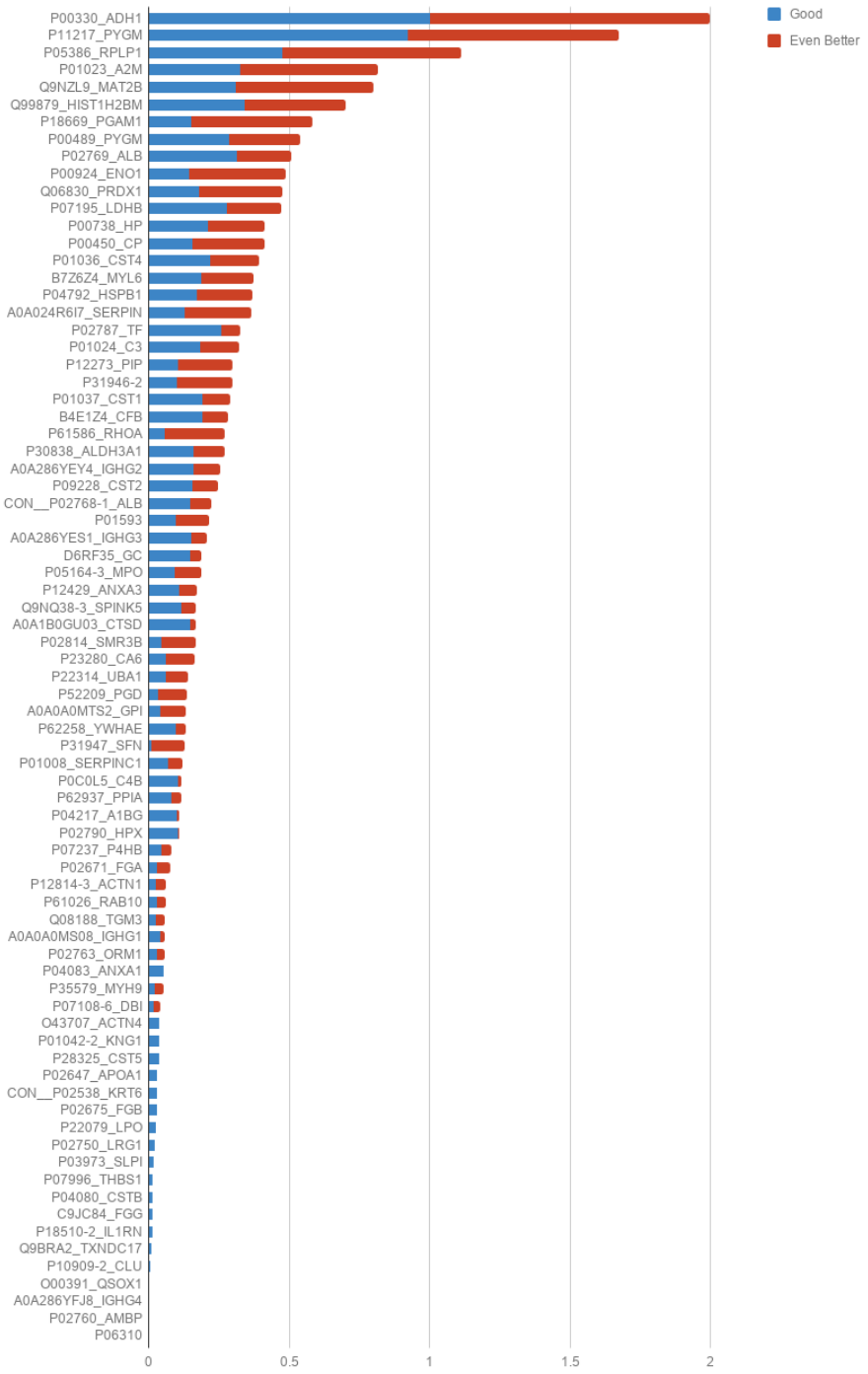
Source: Elaborated by the author.

contrast with the good estimate of CV, the DCV resulted in poor score of 44.80% and a standard deviation of 23.65%. Assuming a t distribution, we would expect the average DCV score to be in the interval $44.80 \pm 18.18\%$ with 95% confidence (9 samples and 8 degrees of freedom). This could mean that we should increase the number of samples if possible before starting more expensive validation processes. Another point of view would be that the criteria chosen to select the best signature and define frequencies are not appropriate for discovery proteomics. On the other hand, the position of the 4 added markers say the opposite. Taken the points together, the hypothesis of using too few samples for this study is strong, making the classifier models over-fitted but not necessarily making the ranks based on frequencies false.

As has been noted, the variance and standard deviation of the cross validation scores indicates that we should not pick one of these signatures to be an exclusive choice. On the other hand, if the data set represents well the population, the DCV score is high, instead of selecting one signature from the *best signatures* list we could select the union of the proteins for two reasons: (1) some signatures are too small for a targeted proteomics study and (2) in a targeted protocol, the quantification of around 10 to 40 proteins is currently not a problem.

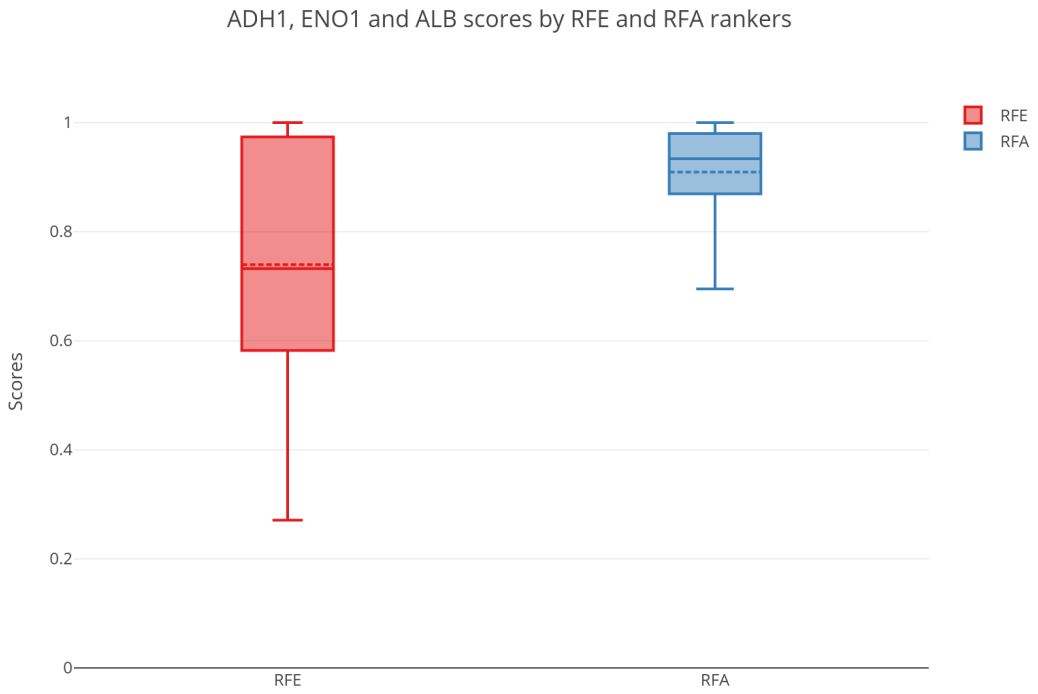
The results from fold-0 illustrated in Figure 28 are radically different from the DCV results illustrated in Figure 33. Proteins positioned in the first positions of the frequency rank from fold-0 are basically a particular case of the training set formed in that fold. While it could be true, it goes against the added true positive markers, indicating it as a *worst case scenario*.

Figure 33 – Satck plot of the frequency of proteins in *good signatures* and *even better signatures*.



Source: Elaborated by the author.

Figure 34 – Box plot of scores of ADH1, ENO1 and ALB given by the RFE and our proposed RFA approach. For this data set, RFA is more stable and attributed higher scores for these true positive markers.



Source: Elaborated by the author.

Also, the general results considering the simulation of small changes in the training data sets applied in the DCV scheme positioned the added proteins in good positions and also identified other candidates, which would be expected since the original samples are indeed from different conditions (C, TR and T). To sum up, we believe that such negative variation among folds represents the high levels of noise that discovery proteomics may lead - in combination with the chaotic nature of cancer cells.

The ranks of proteins can be used to decide what are the proteins that are going to be analyzed in the targeted phase. Different criteria may define the final set of proteins, such as selecting the top-N proteins of a stacked-rank, the intersection or the union of the top-N from different ranks, where N is chosen based on the cost of analyzes in the targeted phase. Different analysis could be integrated here. As an example, one could compare lists of candidate proteins from a clinical-based analysis and ML-based analysis.

Other information may guide to pick the cutoff value N. For instance, the signature with the highest adjusted score (0.94) is formed by proteins P00330_ADH1, P12273_PIP, P31946-2, A0A024R6I7_SERPINA1 and P05386_RPLP1. According to the rank defined in Figure 35, these proteins have the following positions: 1, 29, 23, 4 and 30, respectively. Consequently, if

resources are limited for the next phase, it would be interesting to define N as close as possible to 30. If plenty resources are available, then, we could check what is the greater position in the second best signature, and so on.

Given those points, we stacked the DCV frequencies of proteins in *good*, *even better* and *best* signatures and ordered the values to visualize in which positions are the proteins selected in best signatures. After all, only a few proteins that are not frequent in *good* and *even better* signatures appeared in the *best signatures*. On the other hand, their frequency in the *best* is very low too.

Altogether, combining and visualizing the different ranks created by frequencies of our pipeline demonstrated to be a good approach to select candidate biomarkers for targeted proteomics, in view of that the added true positive markers were well positioned in the frequency ranks and not always well positioned by simple ranks or when selecting particular signatures. Moreover, the top-10 proteins from the overall rank allowed clustering the conditions (C, TR and T) well, as shown in Figure 36. This was not achieved when we clustered the samples using all filtered proteins.

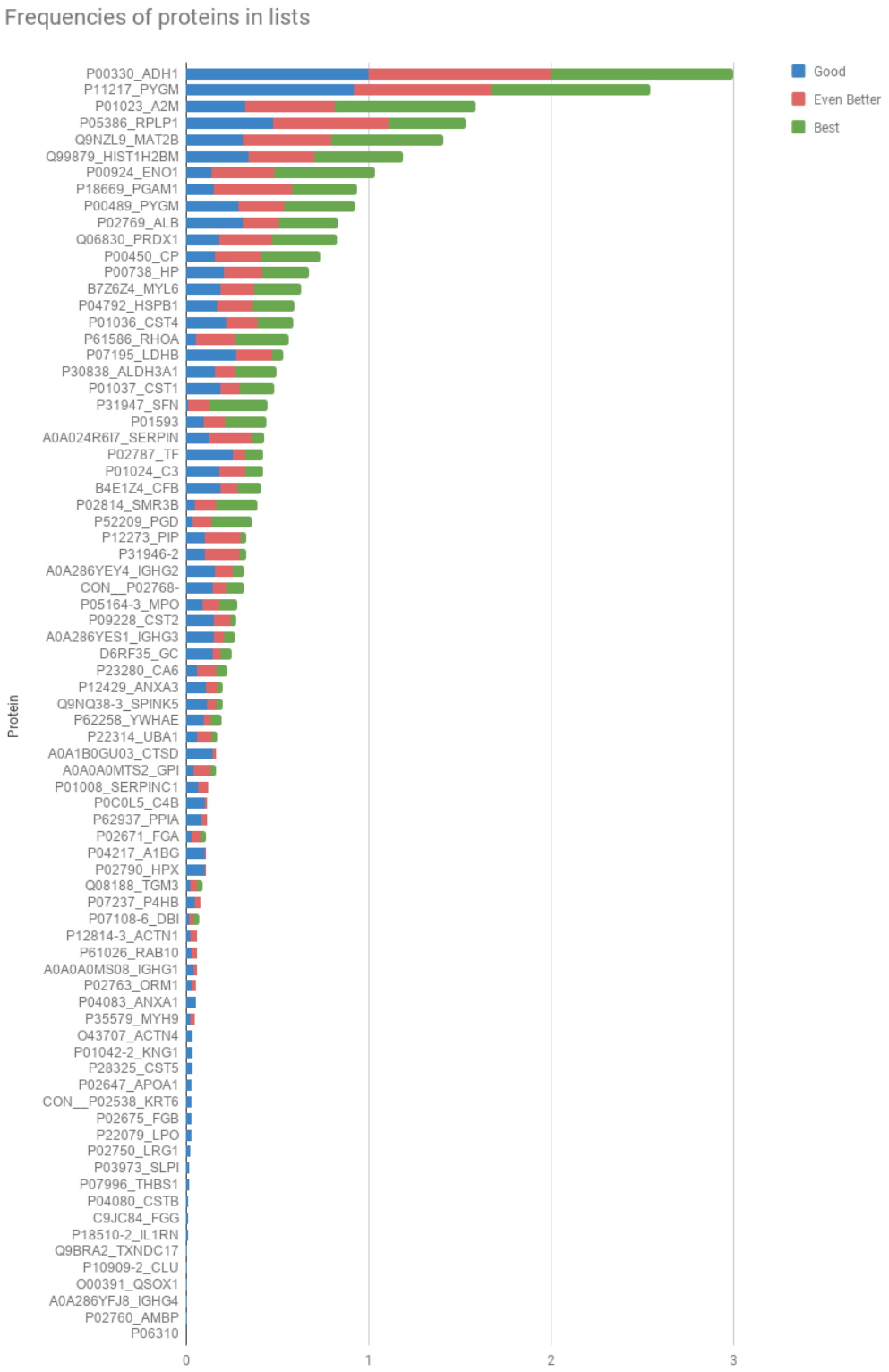
3.5 Case 2 - Prostate cancer biomarkers

This section compares the results of the pipeline applied to the data set D2 from the study of Kim *et al.* (2012). Fourteen proteins from their study are used as a reference to track their position in rankings and frequency in candidate signatures. The proteins are divided into three categories which refers to proteins that are well studied and suggested as biomarkers of prostate cancer, proteins which proteomic discovery intensities were correlated with clinical outcome and proteins that were not correlated. We refer to these proteins as reference markers in the analysis of this section.

- Well-known candidates: P07288 (KLK3, PSA) and P15309 (ACPP, PAP);
- Verified candidates: P31947 (SFN), P08473 (MME), Q99497 (PARK7), P01033 (TIMP1) and P49221 (TGM4);
- Inconsistent candidates: Q9UHI8 (ADAMTS1), P36955 (SERPINF1, PEDF), Q13332 (PTPRS), P02787 (TF), P04083 (ANXA1), P12277 (KCRB, CKB) and P35579 (MYH9).

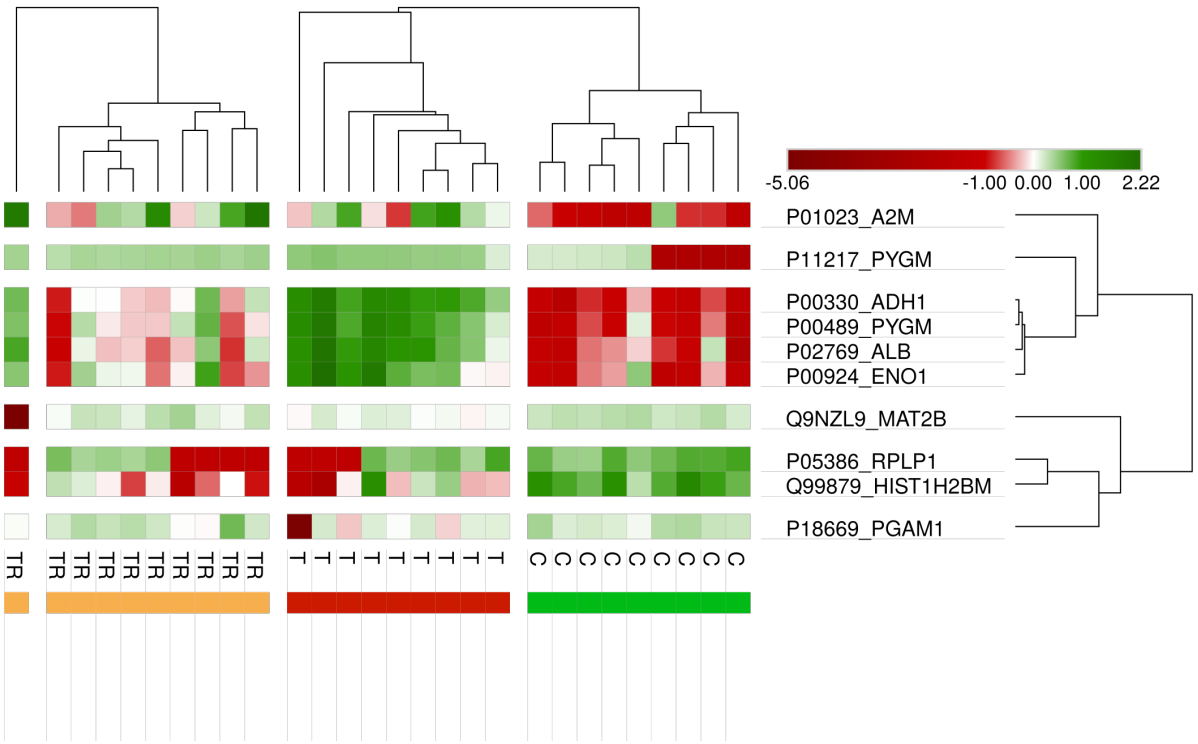
This data set contains duplicates that biased the cross validation results reported in this manuscript since one sample can be found in the train set while its duplicate could be in the test set. Each class (OC and EC) is formed by 8 patients, which turn to 16 samples per class after duplication. We considered that using only 8 samples would be too few for a double cross validation and continued the study with 32 samples. Thus, one should consider the prediction performance described in this section as overoptimistic, even for DCV. Still, the estimations can

Figure 35 – Satck plot of the frequency of proteins in *good signatures*, *even better* and *best* signatures.



Source: Elaborated by the author.

Figure 36 – Heat map of the top-10 proteins considering overall frequency scores. Created with Morpheus, <https://software.broadinstitute.org/morpheus>.



Source: Elaborated by the author.

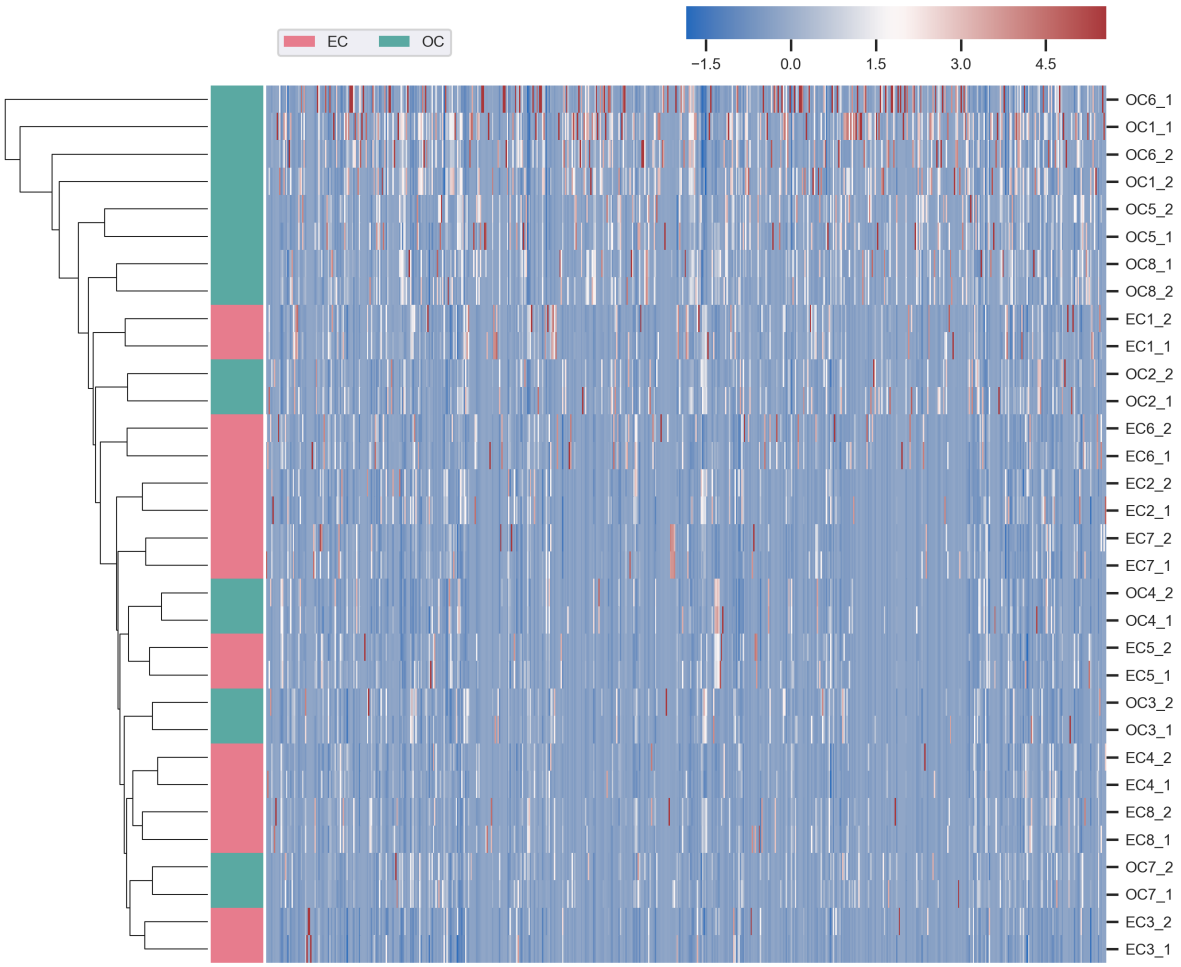
be used to compare signatures and proteins internally. As shown in Figure 37, duplicates have some extreme dissimilarities.

The original study reduced the number of proteins by defining criteria as stated in the original article:

> Proteins were ranked initially based on six features by assigning a ≤ 1 value to a given protein with the following annotations (0 if the term could not be assigned): 1) present by mass spectrometry in six out of eight samples corresponding to the risk group that the candidate was found to be higher in by spectral counting; 2) overexpression of candidates genes in the prostate as demonstrated by at least a threefold above median mRNA expression in the normal prostate tissue compared with 22 different normal human tissues; 3) presence of predicted TMDs; 4) presence of predicted SP sequences; 5) cellular localization assignment to “cell surface” (GO: 0009986) and “extracellular” (GO:0005576) by Gene Ontology annotation; and 6) cancer-associated protein as available from the Human Protein Atlas database.

The above criteria reduced the number of proteins to 60, thirty two of which are cancer-associated. The ranked proteins were again reduced, now to 14, “because of limitations in antibody availability and resources” (KIM *et al.*, 2012). Here, we focus on comparing the

Figure 37 – Prostate cancer data set heat map.



Source: Elaborated by the author.

proteins that are considered good and were verified by Kim *et al.* (2012) and the best candidates found by our proposed Machine Learning methods.

3.5.1 Analysis of fold-0 from DCV

The analysis of this section refers to one fold of the DCV scheme. The data set is formed by 26 samples in the train set and 6 samples in the test set, with 53 proteins. The test set is formed by samples EC8_2, OC1_1, EC7_2, EC7_1, OC7_2 and OC7_1. This means that one duplicate of EC8 (EC8_1) and one duplicate of OC1 (OC1_2) is in the train set.

3.5.1.1 Filtering

After filtering the proteins by Wilcoxon rank sum test (p-value < 0.05), the dimensions were reduced from 624 to 78; which is very different from when applying to the entire data set where the reduction is to 139. By using the Benjamini-Hochberg multiple tests and accepting FDR of 25%, the number of proteins dropped to 43. From the 14 reference proteins, only 7

passed the FDR test (PAP, PARK7, ANXA1, PTPRS, MYH9, CKB and TF). Two proteins were found as highly correlated to each other (P62937 and Q00610) what makes the number of proteins in the train set drops to 42; none of the two is a reference marker.

The intersection of the filtered proteins with the reference markers is small, specially when considering the verified candidates. Despite we "lost" 5 verified proteins (only two verified passed), our candidate proteins could be verified in future experiments and they could be considered true positives, e.g., with better prediction power. The opposite idea also can be true and candidates selected by machine leaning could be mostly true negatives. Only clinical trials may confirm.

3.5.1.2 Ranking

The top-10 most frequent proteins that appear in more than 33% of ranks are: RPS27A (P62988 [4], 77.5%), VAT1 (Q99536, 65%), IGHA1 (P01876, 65%), TIMP2 (P16035, 57.5%), HSP90AA1 (P07900, 57.5%), **ACPP** (P15309, 47.5%), CFB (P00751, 42.5%) and YWHAG (P61981, 35%). Figure 38 shows that these proteins and the distribution of their 50% higher scores. Despite sharing 7 proteins with the reference markers, only the Prostatic acid phosphatase (ACPP, P15309) was identified as stable among ranking methods. ACPP is the 3rd most stable protein. Also, these proteins were not among the *inconsistent candidates*.

In fact, the proteins that did not pass the verification step by Kim *et al.* (2012) had a bad top-10 frequency score. The percentages of each is indicated as follows: ANXA1 (P04083, 22.5%), CKB (P12277, 15%), TF (P02787, 15%), PTPRS (Q13332, 10%) and MYH9 (P35579, 12.5%). As we can see, these values are much lower than ACPP (47.5%) and the others indicated in the first paragraph, which reached 77.5%.

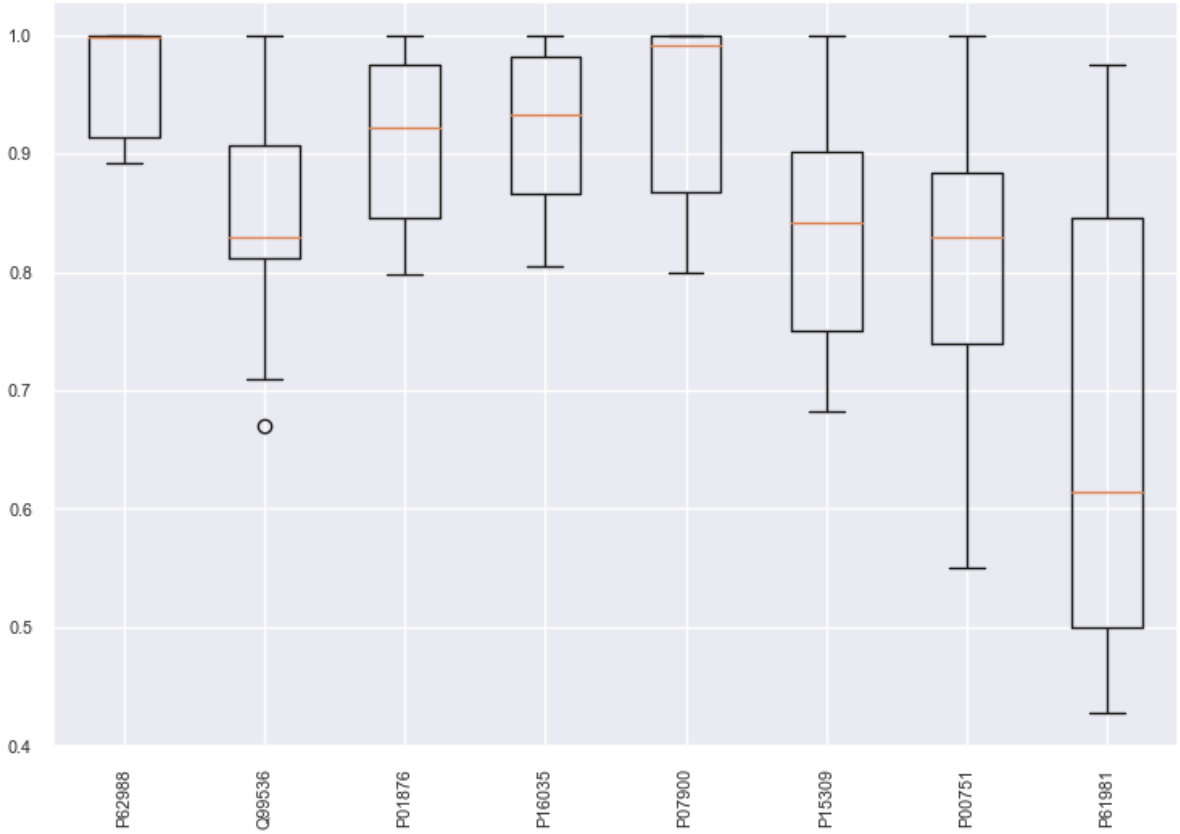
3.5.1.3 Frequency in good signatures

The top-10 proteins from the rank of frequent proteins in the 40,896 good signatures had a score above 16%. The proteins with at least 16% of frequency are listed below, where top-10 most frequent proteins that appear in more than 33% of ranks are in bold. **TIMP2** (P16035, 53.64%), **VAT1** (Q99536, 36.66%), **RPS27A** (P62988, 30.87%), CNDP2 (Q96KP4, 26.63%), **HSP90AA1** (P07900, 20.59%), **IGHA1** (P01876, 18.25%), HSPA8 (P11142, 17.80%), BTD (P43251, 17.45%), EIF5A (P63241, 16.90%), CHIT1 (Q13231, 16.21%) and CFL1 (P23528, 16.08%).

The other proteins that appeared in good signatures had scores below 16%, decreasing by 1% each 2 or 3 proteins until reach the lowest value (2.45%). Protein ACPP was in only 13.07% of the signatures.

[4] On August 10, 2010 this entry became obsolete, now found as secondary accession in P0CG47, P0CG48, P62979 and P62987; RPS27A is found as P62979.

Figure 38 – Box plots of the 50% higher scores of each protein that has top-10 frequency in ranks higher than 33%. The proteins RPS27A (P62988), IGHA1 (P01876), TIMP2 (P16035) and HSP90AA1 (P07900) have the highest scores, specially RPS27A (P62988). The Prostatic acid phosphatase (ACPP, P15309) does not have the best score distribution but its median and maximum value are good. Note: the diagram is presenting the interval [0.4, 1.0].



Source: Elaborated by the author.

3.5.1.4 *Frequency in even better signatures*

After filtering the signatures by "adjusted score > max-0.05" and "mean score > max-0.5", in this order, the rank of proteins changed. These are the proteins with frequency > 30% in a set of 36 signatures: TIMP2 (P16035, 88.89%), CFL1 (P23528, 55.56%), CHIT1 (Q13231, 41.67%), LDHA (P00338, 41.67%), IGHA1 (P01876, 41.67%), BTD (P43251, 38.89%) and CNDP2 (Q96KP4, 30.56%).

3.5.1.5 *Best signatures*

The final best signatures are:

- S1: {TIMP2 (P16035), CHIT1 (Q13231), HSPA8 (P11142)};
- S2: {TIMP2 (P16035), CHIT1 (Q13231), HSPA8 (P11142), BTD (P43251)}.

The mean prediction score for these signatures are 97.59% in cross validation and 82.85% on the independent test set. The signatures represent three proteins that were highly frequent in the *even better proteins* set (TIMP2, CHIT1 and BTD). The protein HSPA8 had a frequency equal to 13.88% - in 5 of 36 signatures.

3.5.2 Double cross validation

Thirteen out of 66 proteins were found in more than 10% of the *even better signatures* set as follows: **IGHA1** (P01876, 61.96%), **TIMP2** (P16035, 57.93%), **RPS27A** (P62988, 44.97%), **ANXA1** (P04083, 33.29%), **SERPINF1** (P36955, 32.60%), **CD109** (Q6YHK3, 21.74%), IGHA2 (P01877, 20.43%), **BTD** (P43251, 20.37%), IGHV1-69 (P01742, 18.08%), PI15 (O43692, 15.01%), **PKM** (P14618, 14.45%), **HEXB** (P07686, 11.57%) and **ACPP** (P15309, 10.13%). In bold are the proteins that appeared among the top-14 proteins in the final *best signatures* found in the complete DCV.

The proteins that appeared in the best signatures (31) are listed as follows: **IGHA1** (P01876, 83.87%), **RPS27A** (P62988, 77.42%), **TIMP2** (P16035, 41.94%), **BTD** (P43251, 38.71%), **ANXA1** (P04083, 35.48%), **SERPINF1** (P36955, 29.03%), **HEXB** (P07686, 25.81%), **PKM** (P14618, 22.58%), **ACPP** (P15309, 16.13%), HSPA8 (P11142, 16.13%), CKB (P12277, 12.90%), GGT1 (P19440, 9.68%) and **CD109** (Q6YHK3, 9.68%).

In a comparison among the 13 first proteins from both lists, we have three proteins from the *even better* list that are not among the top-13 from the *best signatures*. Still, they appeared in some of the best signatures, having their frequency as follows: IGHV1-69 (P01742, 6.45%), IGHA2 (P01877, 3.23%) and PI15 (O43692, 3.23%). Three proteins from the best signatures were not in the top-13 from the *even better* list but had an acceptable frequency in comparison with the other 40 proteins with frequency below 5% - 25 from these are bellow 1%: CKB (P12277, 9.87%), GGT1 (P19440, 5.79%) and HSPA8 (P11142, 5.47%).

3.5.2.1 Potential signatures

As expected, there were a high variance in the final signatures from each DCV fold, even with the biological duplicates induced bias. In total, there are 31 best signatures. From these, IGHA1 and RPS27A have the greatest frequencies. It is difficult to choose a signature based on the scores of each fold since choosing one signature would mean to choose a single training set. Still, these results help to decide what are the proteins that are going to be verified or validated in further steps.

The best signature with the highest mean protein frequency in good signatures is {IGHA1, RPS27A, ANXA1, CKB} (P01876, P62988, P04083, P12277), with 99.6% of prediction score. The signature with the highest average adjusted prediction score is {RPS27A, IGHA1, IGHV1-69, CD109} (P62988, P01876, P01742, Q6YHK3), with 100% of prediction score (non-adjusted).

The average simple (mean) cross validation score of the best signatures in all outer-training sets of DCV is 98.44% with 1.58% standard deviation. Assuming a t distribution for the mean CV, we would expect it to be in the interval $98.44 \pm 0.58\%$ with 95% confidence (31 samples and 30 degrees of freedom).

The independent test set from each of the 7 folds resulted in 100%, 100%, 100%, 100%, 82.86%, 73.33% and 73.33% of weighted F1 score. This give us the DCV score of 89.93% and 11.99% of standard deviation. Assuming a t distribution, we would expect the average DCV score to be in the interval $89.93 \pm 11.08\%$ with 95% confidence (7 samples and 6 degrees of freedom). In contrast with the DCV of the data set D1, this good score could be considered overoptimistic due to bias caused by the use of biological duplicates. Supposing that the bias caused by duplicates did not exist or that, by biological or clinical reasons, it could be considered irrelevant, the high value of DCV would indicate that the proteins found in the best signatures have a high probability of being true positives.

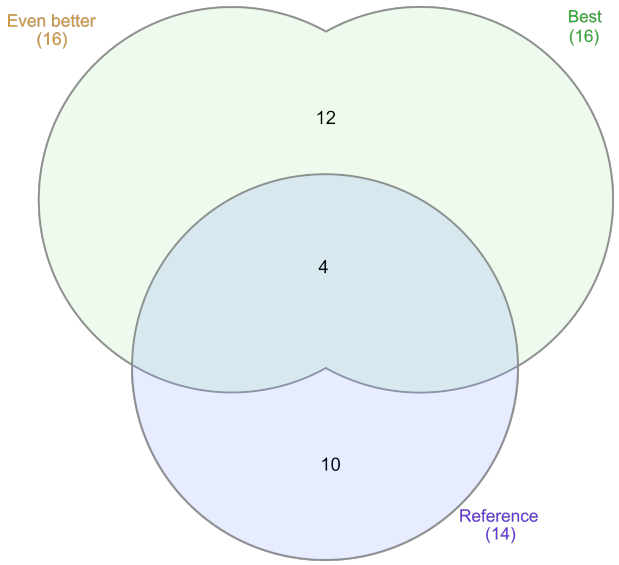
As expected, selecting signatures with maximum mean scores is a more unstable approach than selecting proteins by frequency. None of the training sets, again, resulted in any equal best signature. On the other hand, the top-frequent proteins in signatures from the *even better* sets and from the *best signatures* sets are similar. There are 10 proteins in the intersection of the top-13 *even better* and *best* proteins lists.

The proteins that appeared in more than 30% of the best signatures are: **IGHA1** (P01876, 83.87%), **RPS27A** (P62988, 77.42%), **TIMP2** (P16035, 41.94%), **BTD** (P43251, 38.71%) and **ANXA1** (P04083, 35.48%). From these, proteins **IGHA1**, **RPS27A** and **TIMP2** are in more than 50% of *even better* or *best* signatures.

Signatures S1 and S2 from fold 0 contains proteins TIMP2, BTD, CHIT1 and HSPA8. Only TIMP2 maintained its high frequency in *even better* signatures in comparison to all DCV folders. They had a frequency of 88.89%, 38.89%, 41.67% and 13.88% that were changed to 57.93%, 20.37%, 5.95% and 5.47% considering all folds, respectively. Despite 5% seems to be very low, it is greater than the frequency of 40 other proteins. From these, TIMP2, BTD and HSPA8 are in the top-10 most frequent proteins in the best signatures (out of 27 proteins). On the other hand, CHIT1 still appears in some of the best signatures, being the top-17 most frequent protein.

Figure 39 shows the intersection between the union of the *even better* and *best* top-13 proteins and the 14 proteins verified by Kim *et al.* (2012). Only four proteins that they found by their methods were selected by the multivariate methods applied here. As shown in Section 3.5.1, many of their proteins were removed by the 25%-FDR filter applied in each of the training sets.

Figure 39 – InteractiVenn diagram illustrating the union of the top-13 proteins from the *even better* and *best* signatures (DCV) in comparison with the 14 proteins verified by Kim *et al.* (2012) (*reference*). In the union (12): IGHA1, TIMP2, CD109, IGHA2, BTD, IGHV1-69, PI15, PKM, HEXB, HSPA8 and GGT1. In the intersection (4): ANXA1, SERPINF1, ACPP and CKB. Exclusives of the *reference*: KLK3, SFN, MME, PARK7, TIMP1, TGM4, ADAMTS1, PTPRS, TF and MYH9.



Source: Elaborated by the author.

3.5.2.2 Biological Interpretability

We searched in the Humam Protein Atlas (HPA) - Phatology - for the proteins that appeared in most of the *even better* or *best* signatures (>50%) (TIMP2, IGHA1 and RPS27A) - three exclusive proteins of our multivariate approach. The protein TIMP2 (P16035) is the Metalloproteinase inhibitor 2. It can participate on biological process linked to the characteristics of a tumor, such as having cells dying later and replicating more than usual: *negative regulation of cell proliferation*, *negative regulation of mitotic cell cycle* and *aging*; also to other process that may be interesting to treatment response: *response to drug*, *response to hormone* and others. The high expression of this protein is found in most of the patients of many cancers. At least 80% of *testis*, *renal* and ***prostate*** patients have high or medium expression of this protein. It is expressed in 231 organs, with the highest expression level in decidua (forms the maternal part of the placenta). This could indicate the high or low expression of TIMP2 in prostate as a candidate biomarker. In fact, according to train set tested here, the TIMP2 has log2 fold-change of -27 (EC/OC), meaning that its mean expression in the extracapsular cancer is extremely lower than in organ-confined cancer. This is the second highest difference in expression found, together with protein P43251 (-28) BTD. In fact, 100% of EC patients had values 0.0 for TIMP2 while OC patients have a mean of 1.6 and 1.4 standard deviation. Six OC samples out of 16 have values 0.0, but only one patient have 0.0 for the duplicates. For instance, patient OC7 have values 1.97 in one replicate and 0.00 in the other.

The protein **IGHA1** (P01876), the Immunoglobulin heavy constant alpha 1, was not found in the HPA. Despite that, this gene encodes a constant (C) segment of Immunoglobulin A heavy chain (IgA1). IgA1 is an antibody that plays a critical role in immune function in the mucous membranes. This means that its function is to defend the body against infections and foreign antigens to the immunologic system. Welinder *et al.* (2016) published that "intra-tumor IgA1 is common in cancer and is correlated with a poor prognosis in bladder cancer" (muscular sac in the pelvis). Pratt (2002) wrote that "homogenous group of patients characterized by a high incidence of chromosome 13 abnormalities, a higher incidence of IgA subtype, and a poor prognosis" (AVET-LOISEAU *et al.*, 2002). The plasma cell myeloma is a cancer in a type of white blood cell normally responsible for producing antibodies. This protein is nearly twice more expressed in EC samples (log2 fold-change of 0.9) from the train set. Interestingly the Immunoglobulin heavy constant alpha 2, IGHA2 (P01877), also had a high frequency in the *even better* signatures and 1.44 log2 fold-change (EC/OC). This protein was not found in HPA.

The Ubiquitin-40S ribosomal protein S27a (RPS27A, P62988) was found with medium expression in many cancer types, including prostate cancer. It is a prognostic marker in renal cancer and liver cancer, with the lower expression linked to a higher survival probability. It has 1.95 log2 fold-change (EC/OC) meaning that it is almost 4 times more expressed in EC than in OC.

The Prostatic acid phosphatase (ACPP, P15309) appeared in many good candidates, being in 10.13% of *even better* signatures and 16.13% in the best. The protein ACPP, Prostatic acid phosphatase, is linked to prostate cancer, having 30% of patients high expression and 70% medium expression of this protein according to HPA. This protein participates in biological processes such as *nucleotide metabolic process*, *regulation of sensory perception of pain*, ***desphosphorylation*** and others. It "acts as a tumor suppressor of prostate cancer through dephosphorylation of ERBB2 and deactivation of MAPK-mediated signaling" and is "used as a diagnostic tool for staging metastatic prostatic cancer" according to Uniprot. This protein is almost 4 times less expressed in EC samples (log 2 fold-change of -1.7).

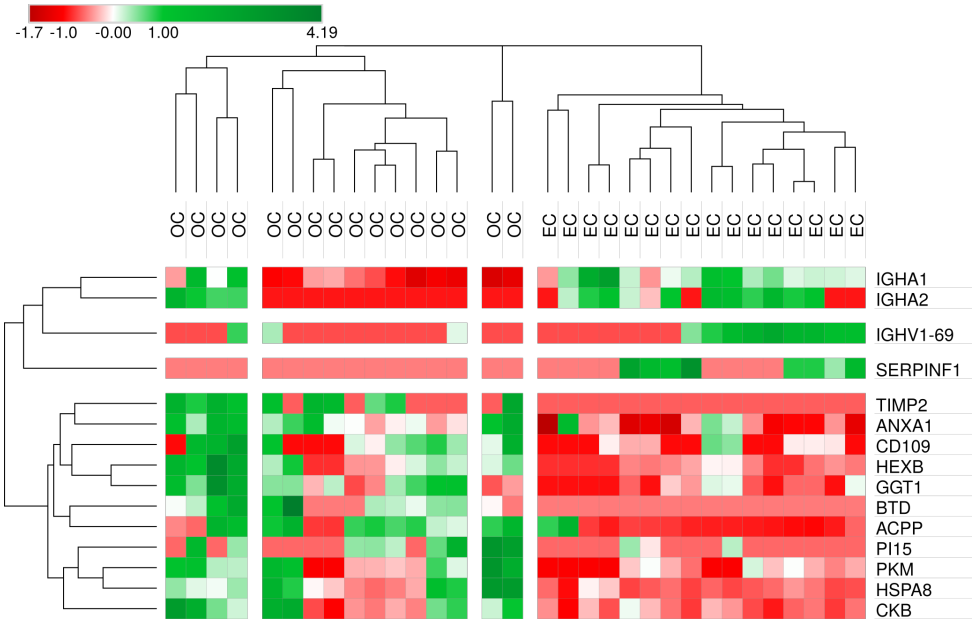
Finally, we performed an evaluation considering the cBioPortal (BOCK *et al.*, 2012; CERAMI *et al.*, 2012) and queried the proteins IGHA1, IGHA2, TIMP2, BTD, RPS27A, ANXA1 and ACPP. As expected, ACPP is the most altered in prostate cancer. Considering the genes' alterations rates in patients with prostate cancer, we have the following rank: ACPP with 22.86% of amplification; BTD with 20.0% of amplification; TIMP2 with 15.71% of amplification; ANXA1 with 11.43% of amplification; RPS27A with 7.14% of amplification, IGHA1 with 0.2% of mutation, deep deletion or amplification; and IGHA2 with 0.18% mutation, deep deletion or amplification. Thus, in an opposite direction of this proteomics multivariate analysis, the genomic studies indicate that the IGHA1 and IGHA2 may be a false positive biomarker candidate and, thus, additional research should be performed before further experiments with this protein.

The IGHA1 has its values varying much among the samples, making the differentiation

not as homogeneous as in the case of IGHA2, which was not detected or down-regulated (value 0.0) in 6 out of the 8 OC patients and in only 1 out of the 8 EC patients. Despite this fact, IGHA1 was much more selected in the best signatures than IGHA2. This highlights that multivariate models do not seek for a common characteristic interesting for biomarkers discovery: a more homogeneous differentiation pattern. Thus the importance of filtering with statistical methods and biological or clinical criteria before performing the multivariate analysis.

The Figure 40 shows the distribution of intensities of final interesting proteins prioritized by DCV. Most of proteins are more concentrated in EC cells. The dendrogram shows the Euclidean distances between samples and proteins. We can see the improvement in the clustering of samples, in comparison with Figure 37.

Figure 40 – Heat map of interesting proteins' intensity.



3.5.2.3 *Closing remarks*

The approaches (ours versus Kim *et al.* (2012)) to select prostate cancer candidates compared in this section are based on different information. Kim *et al.* (2012) defined a final list of candidate biomarkers using statistical methods and other criteria that could bring up interesting proteins to be validated in further experiments. They also limited the 60-proteins set to 14 justifying the choice based on limitations of cost of the further phases.

While in our approach only 4 of the verified proteins were in the *even better* and *best* sets of DCV, we found proteins that are differentially expressed and had the best multivariate prediction scores in signatures to distinguish OC from EC samples among the tested signatures. Again, we consider that the approaches using machine learning can be used to indicate interesting proteins and give us a hit about the prediction power of their combinations, but are not the best

method alone to indicate a panel of proteins for validation or clinical trials, requiring the combination of specialized knowledge and filtering criteria.

3.6 Conclusion

In this chapter we reported an approach to select candidate biomarkers from discovery to targeted proteomics and exemplified how the analysis could be done by detailing two studies. Instead of using a few ranks of proteins or selecting a few signatures, our approach is based on sampling the original data set and computing frequencies of proteins in candidate lists.

A candidate list can be the top-N proteins of many different ranks, the N most frequent protein in signatures with good prediction score, among other possibilities. Our approach consists in applying this idea in a double cross validation scheme, in a way that the specialist will have different outputs to make decisions such as what proteins are going to be used in a new quantification based on targeted proteomics. The double-cross validation helps specialists to decide if they can trust the results or if they should consider increasing the number of samples when possible.

We demonstrated the effect of small changes in two proteomics data sets and how defining too specific criteria may led to great disagreements. We also showed how the protein positions vary in ranks based on the frequency in top-10 lists of proteins (top-10 proteins from 40 different rankers) and on frequency in *good*, *even better* and *best* signatures. In essence, the lists of candidate proteins of these ranks are very similar and, thus, represent a more stable approach than approaches based on too specific criteria such as running an RFE method and defining the cutoff in the minimum position of maximum prediction score.

As has been noted, the RFE method tend to ignore redundancies and, thus, place similar important proteins distant in the rank. Given this nature of the method, we considered that creating a rank based on adding the best protein in the first position of the rank could eliminate the mentioned negative characteristic of RFE - in opposition to adding the worst protein to the last position, as RFE does. We demonstrated that our control true positive markers were indeed better positioned by our RFA approach.

In **future works**, we expect to increase the parameters such as the maximum size of signatures and others, by using a cluster of computers to run the pipeline. One characteristic not considered in the current methods is that the bigger is the signature, the bigger is their influence over the proteins frequencies. This could be addressed by averaging the proteins frequencies in signatures of the same size. We could also weight the frequencies by signature size.

Despite we separated proteins highly correlated to address the assumption of non-dependent variables in most of ML models, the definition of “highly correlated” is not precise and correlation does not imply dependency. There are statistical methods for verification of

collinearity and **multiple collinearity** that could be tested with proteomics data to address this problem. Also, the results from the methods tested here should be compared with the ones based on biological information, in a double cross validation fashion. Bigger data sets should be also considered, seeking to find better DCV scores.

In the analyzes of the data set D1, we identified two types of PYGM proteins as biomarker candidates. This should be investigated to understand if it is a consequence of a mistake in the quantification design or if the human PYGM is actually a potential candidate, together with the rabbit PYGM, given that their sequences are very similar.

CHAPTER

4

BIOLOGICAL NETWORKS ONTO CELLULAR STRUCTURE

HEBERLE, H.; CARAZZOLLE, M. F.; TELLES, G. P.; MEIRELLES, G. V.; MINGHIM, R. CellNetVis : a web tool for visualization of biological networks using force-directed layout constrained by cellular components. BMC Bioinformatics, v. 18, n. 10, p. 395, 2017. (HEBERLE *et al.*, 2017)

In the following sections we describe the system we designed for the visualization of biomolecular networks contextualized by cellular structure. This work gained the Best Paper Award in the BioVis symposium that happened during the Intelligent Systems for Molecular Biology (ISMB) conference at Prague, in 2017.

4.1 Introduction

With the advent of "omics" science, analyzes performed from screening a wide range of physical, genetic and chemical-genetic interactions have brought new perspectives to contemporary biology, as they provide new clues in protein/gene function, help to understand how metabolic, regulatory and signaling pathways are organized and facilitate the validation of therapeutic targets and potential drugs. Biomolecular interaction networks are simple abstract representations where the components of a cell (e.g. genes, proteins, metabolites, miRNAs etc.) are represented by nodes and their interactions are represented by edges. An appropriate display of the data is crucial for understanding such networks, particularly regarding high-throughput analysis.

Since different regions of the cell are related to specific activities, visually organizing network nodes into cellular components can help understand the biological system and its relationship to the distribution of network elements over the cell structure. The position of nodes can unveil, for instance, patterns of relations among different cellular components. Additionally, it is common to query just a subnetwork of an entire interactome, so when users query specific pathways by a list of their units (e.g. gene symbols) they can easily see, by using a proper layout, where these pathways may occur in the cell.

Many tools are available to visualize and explore network models but most of them are not designed to partition networks into a cell structure. Among those are Graphviz (ELLSON *et al.*, 2002), Gephi (BASTIAN; HEYMANN; JACOMY, 2009), Pajek (BATAGELJ, 1999), PEx-Graph (MARTINS *et al.*, 2012), Cystocape (SHANNON *et al.*, 2003) and Tulip (AUBER, 2004). They were created for a generic purpose, being applied in problems ranging from social network analysis to biology. Cytoscape is the most popular tool in biology and counts with many plugins for Systems Biology in particular, including two that work with cellular partitions: Cerebral (BARSKY *et al.*, 2007) and Mosaic (ZHANG *et al.*, 2012). Other software systems, like Extended LineSets (PADUANO; FORBES, 2015), Entourage (LEX *et al.*, 2013), and ReactionFlow (DANG *et al.*, 2015), focus on the analysis of pathways and their mechanisms.

Garcia et al. describe an extension to the force-directed layout to place nodes according to their connection and class structure (GARCIA *et al.*, 2007). In their method, the cellular component annotations can define the class structure and approximate nodes of the same class. The approach, however, does not represent cellular components. Other approaches that group nodes in two-dimensional space have been proposed, such as constrained force-directed layout (DWYER, 2009), constrained projections (DWYER; ROBERTSON, 2010), hierarchical graph placement (PANUCCIO *et al.*, 2016; MUELLER; DEHMER; EMMERT-STREIB, 2013; SCHUHMACHER, 2015) and others (BAUR; BRANDES, 2008; DOGRUSOZ *et al.*, 2009; ARCHAMBAULT; MUNZNER; AUBER, 2011; ALTARAWNEH; SCHULTZ; HUMAYOUN, 2014). Despite their good performance even for large networks, the cell structure is not taken into consideration in either of those cases. Also, they are not adapted to display networks in an explicitly full cell diagram.

Only a few tools provide the capability of displaying networks organized by cellular components. Biographer (KRAUSE *et al.*, 2013) is a web-based tool to edit and render reaction networks. It implements features for visualization based on Systems Biology Graphical Notations (SBGN). The user can manually create shapes of type “compartment” and position nodes inside them. Mosaic (ZHANG *et al.*, 2012) is a Cytoscape plugin and can represent a network divided into cellular partitions automatically, duplicating nodes when there is more than one cellular component annotation. It uses force-directed layout, but it does not update the layout when nodes are moved. Also, the display was designed to show small subnetworks. Cerebral (BARSKY *et al.*, 2007; FRIAS *et al.*, 2015), originally designed as a Cytoscape (SHANNON *et al.*, 2003) plugin

and extended to work with Cytoscape.js, can automatically divide the network into subcellular regions represented by parallel rectangles, one over the other, which is not consistent with the standard graphical representation of a cell. Kojima et al. developed a grid layout that may be applied over a full cell diagram, representing the cellular components properly (KOJIMA; NAGASAKI; MIYANO, 2008). The new version, Cell Illustrator Online (NAGASAKI; LI, 2010), is a tool that enables drawing, visualization and modeling of biological pathways. It produces layouts that more closely resemble a consistent cell diagram and displays a network across cellular components. However, that tool is more focused on the mechanisms rather than on the network overview and exploration, the structure is manually defined by the user, and it is neither free nor open-source.

Despite the capability of drawing networks organized by cellular components, Mosaic (ZHANG *et al.*, 2012), Cerebral (BARSKY *et al.*, 2007), CerebralWeb (FRIAS *et al.*, 2015) and Cell Illustrator (KOJIMA; NAGASAKI; MIYANO, 2008) do not provide real-time automatic layout modifications for dynamic exploration. Even for small networks, with hundreds of nodes, these tools cannot reposition the network while the user is interacting, exploring the layout and manually repositioning the nodes. Many biological networks are dense causing the "hairball" problem, what makes the analysis of links, flows and topology difficult. Interactively moving nodes or organelles can increase readability and understanding, clarifying the flow of edges between them and letting the user explore the view to better understand the network dynamics.

We have developed a web tool called CellNetVis that tackles most of the mentioned drawbacks. It is meant for easy and dynamic display and exploration of biological networks over a full cell diagram. We adapted an iterative force-directed algorithm to produce a dynamic layout for the entire network where nodes are positioned into movable cellular components. The input for the tool is a properly annotated network in the XGMML format. The tool displays the network over a standard cell graphical representation showing the main partitions and organelles according to the Gene Ontology (GO) cellular component database (ASHBURNER *et al.*, 2000). It also provides interactive features such as search, selection, drag and drop of organelles and nodes, as well as the capability of displaying nodes annotation information.

CellNetVis allows certain features, essential to current biological network analysis needs, not provided by other tools, such as, at the same time, being web-based, supporting large networks and providing automatic display of nodes inside their cellular components. Additionally, the particular implementation of the force-directed algorithm provides a balance between processing time and visual understanding of network structure with layout flexible to adapt to user's manipulation. We discuss these issues in contrast with available tools in the section *Comparison with available tools*.

4.2 Implementation

CellNetVis (<http://bioinfo03.ibi.unicamp.br/lnbio/IIS2/cellnetvis/>)[1] was written in Javascript and HTML and is a free and open-source software. It loads networks constructed using the XGMML format (PUNIN; KRISHNAMOORTHY, 2001). The only requirement is that the network nodes must have an attribute named either "Selected CC" or "Localization", which corresponds to a unique selected cellular component (CC), such as the one generated by the IIS (CARAZZOLLE *et al.*, 2014) and the InnateDB (BREUER *et al.*,). As the majority of proteins are described as acting in more than one subcellular compartment in GO, IIS and InnateDB apply a priority filter to assign the most specific cellular component to each protein, which is then used by CellNetVis to position the nodes in the cell diagram. Other strategies for assigning a single cellular component to each node may be adopted as well.

As shown in Table 3, InnateDB specifies in the XGMML file five possible compartments, while the IIS specifies twenty-one. CellNetVis works with all these 21 compartments. Additionally, the tool supports the retrieval of cellular components for human, mouse and bovine genes from the InnateDB web service. In this case, nodes must have an attribute that identifies the gene or protein ID in the Ensembl (YATES *et al.*, 2016), Entrez (MCENTYRE, 1998), InnateDB (BREUER *et al.*,) or UniProt (GONZÁLEZ, 2011) format.

Table 3 – Cellular components specified in the XGMML file by IIS and InnateDB.

IIS	InnateDB
Extracellular, cell wall, plasma membrane, mitochondrion, endoplasmic reticulum, Golgi apparatus, endosome, centrosome, microtubule organizing center, lysosome, vacuole, glyoxysome, glycosome, peroxisome, amyloplast, apicoplast, chloroplast, plastid, cytoplasm, cytosol and nucleus.	extracellular, cell surface, plasma membrane, cytoplasm and nucleus.

Source: Carazzolle *et al.* (2014), Breuer *et al.* ().

A few decisions guided the construction of the cellular design in CellNetVis. Figure 41 shows an example of a small network displayed over the cell diagram. The cell is drawn aiming to highlight the separation between the main subcellular compartments: extracellular region, cell wall, plasma membrane, cytoplasm and nucleus. Cell contour lines are drawn using lighter colors as they serve only as a reference. In contrast, network nodes contour lines are displayed with darker colors by default, and if nodes are selected, then the remaining ones are shown with transparency to improve contrast. Regarding the organelles, their contour lines are drawn with less contrast to reduce visual density, since typically these are regions with many edge crossings. The cell diagram is colored using a ColorBrewer (HARROWER; BREWER, 2011) "BrBG" diverging scheme, characterized by colors that can be easily differentiated.

[1] Fork and contribute: <https://github.com/heberleh/cellnetvis>

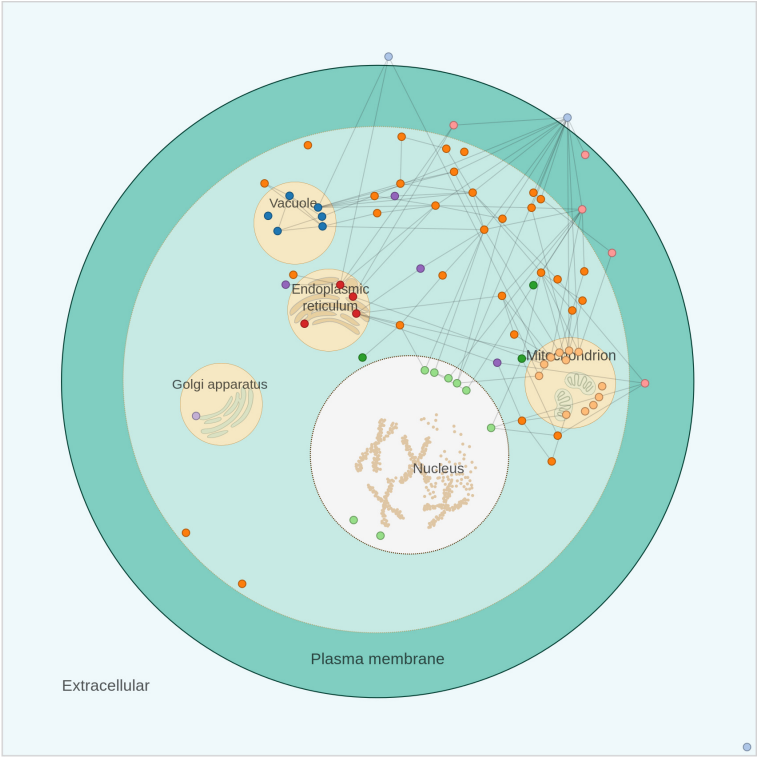


Figure 41 – Display of a small network over a cell diagram.

If the compartment attribute is annotated with any other value not specified on CellNetVis or is empty, it will be positioned in the cytosol. If a node is also annotated with a "Cellular Component" multivalued attribute, all compartments in the list will be highlighted when the node is selected. For instance, if the protein A is drawn on *nucleus* (Selected CC) and has "nucleus, cytosol, mitochondrion" annotated in the "Cellular Component" attribute, all these components will be highlighted when the user selects this vertex. The user can also change the value of "Selected CC" during the visualization process.

The network is drawn over the cell representation by the force-directed layout adapted from the algorithm implemented in D3 (BOSTOCK; OGIEVETSKY; HEER, 2011) version 3.0 (D3, 2015). Our layout has the important advantage over existing tools based on grid-layout of enabling dynamic plotting and interaction with complex networks. We have modified the force-directed layout to constrain the movement of each node to the area of its respective cellular component.

Since the constraints computation in the force-directed layout is computationally expensive, the cell diagram is drawn using only circles, instead of other shapes that are commonly used to create a cell diagram. Complex shapes increase the time to check if each node is in the correct region and, given its current position, recalculate the new position according to the respective component shape. Circles simplify these verifications and position calculations. Another thing that reduces calculations is allowing movement of organelles and their content to the extracellular region. The control over the cell structure consistency is left to the user's discretion.

During each iteration of the force-directed algorithm, the position x, y of each node n is updated. How x, y is recalculated depends on the Selected CC ($n.cc$) of n, as described in the pseudo-code below.

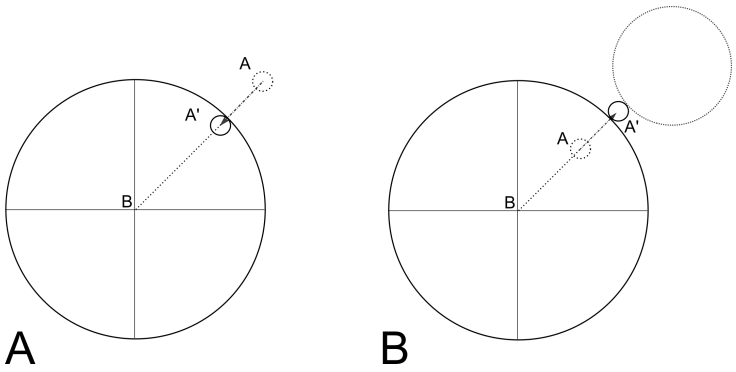
Algorithm 1 – Constraint algorithm that keeps nodes inside the correct cellular compartment.

```
procedure CONSTRAIN(nodes)
    for each node n in nodes do
        if n.cc == cytoplasm then
            if n is out of cytoplasm then
                pull n to cytoplasm inner-border
            else if n is inside an organelle O then
                push n to outer-border of O
            end if
        else if n.cc == extracellular then
            if n is inside cell then
                push n to cell outer-border
            end if
        else if n.cc == p_membrane then
            if n is in extracellular then
                pull n to p_membrane inner-border
            else if n is inside cytoplasm then
                pull n to p_membrane inner-border
            end if
        else  % is in an organelle
            if n is out of n.cc then
                pull n to inner-border n.cc
            end if
        end if
    end for
end procedure
```

When the node is in an organelle, the algorithm checks the distance R between the center of a node (point A: x, y) and the center of its corresponding cellular compartment's circle (point B: cx, cy) (Figure 42A). The node is then placed in the new position, point A' (x', y') calculated by $x' = \frac{r}{R}(x - cx)$ and $y' = \frac{r}{R}(y - cy)$, where $R = \sqrt{(x - cx)^2 + (y - cy)^2}$ and r is the organelle radius. When the node is in the cytosol, the computation is similar, but in the opposite direction (Figure 42B). When the node is in the cell wall or in the plasma membrane, two constraints are checked since there is an outer limit (cell wall or extracellular regions) and an inner limit (cytoplasm or plasma membrane).

When nodes are constrained to specific cellular regions, edges cross at higher rates in the layout. If the network is large, there will probably be too many nodes in the organelles and forces pulling nodes in the same region of the compartment, resulting in overlap of nodes. This limitation is not a feature of CellNetVis, but a deep problem in graph drawing. To reduce this effect, a new constraint was implemented in CellNetVis. The algorithm identifies whether a node is colliding with another one. If so, the nodes are repositioned. This verification is done after

Figure 42 – Diagram of the force-directed layout constraint algorithm. The diagram represents the basic concept about how nodes' positions are redefined by our constraining algorithm during the force-direct layout iterations. It shows how a node is moved from cytosol to the inner-border of an organelle defined in its Selected CC attribute (A) and how a node that should be in the "cytosol" is moved from an organelle to its outer-border (B).



Source: Elaborated by the author.

each iteration of the D3 force placement and queries a quad-tree data structure (SAMET,).

A user-controlled parameter, named repulsive is used to support overlap reduction procedures. If repulsive is large, layout stability is lower but the visual separation of nodes is faster. If repulsive is small, visual stability is higher, but the nodes need more time to separate. Smaller repulsive values do not guarantee that nodes will not overlap. When a large network is loaded, this procedure is disabled by default. Other parameters of the force-directed algorithm can be configured. For instance, setting the *charge* of each node to a more negative value will make nodes more separated. All parameters available in CellNetVis are further explained in the Help page.

The user has four additional options to improve the network layout: moving organelles, constraining nodes to a specific position, hiding unfocused nodes (filter function) and turning on edge bundling that is used to decrease cluttering from crossing edges. Organelles that are not annotated in any node of the network are removed from the view. We integrated the Corneliu Sugar implementation (SUGAR, 2015) of Force-Directed Edge Bundling (HOLTEN; WIJK, 2009) in CellNetVis.

Highlighting neighborhoods of selected nodes, displaying labels, calculating network topology measures and the possibility to color nodes according to different attributes were implemented. Counting of nodes per cellular component was also implemented as a donut chart. The cell diagrams can be exported as a bitmap (.PNG) or vector (.SVG) image.

To allow integration with other systems and publication of a network view in the form of a URL, CellNetVis provides a special parameter named "file", which receives the URL of a XGMML file. When this parameter is used, an asynchronous call is executed by CellNetVis and, after the successful download, the file is parsed and processed the same way as a regular input.

The external XGMML server provider must have the CORS header 'Access-Control-Allow-Origin' set (W3C, 2014).

The response time of CellNetVis depends on the time taken for the construction of the network structure by the Javascript code, the SVG rendering time taken by the web browser and, if the URL approach is used to load the XGMML file, the time to download the network. All the computation is done on the client-side, so the time needed to display a network and interact with the system depends only on the user's computer.

4.3 Results and discussion

CellNetVis is capable of displaying information related to complex networks, nodes and edges as well as their relations with cell partitions. Figure 43 shows the CellNetVis interface. To analyze a network in the cell diagram, the user starts by uploading a network as a XGMML file (Figure 43A). The network will be loaded in the cell diagram area (Figure 43G) and the nodes will be distributed inside each subcellular localization according to its annotation. Alternatively, the user may create an URL that specifies the "file" parameter, that is, the CellNetVis URL plus the XGMML file URL. The force-directed algorithm starts automatically when a network is loaded and will resolve the positioning of nodes within each cellular component. It may be interrupted and restarted at any time (Figure 43D). Nodes and organelles may be manually positioned along the display. When that is done, the neighboring nodes or nodes inside the moving cellular components will be moved accordingly (Figure 43G).

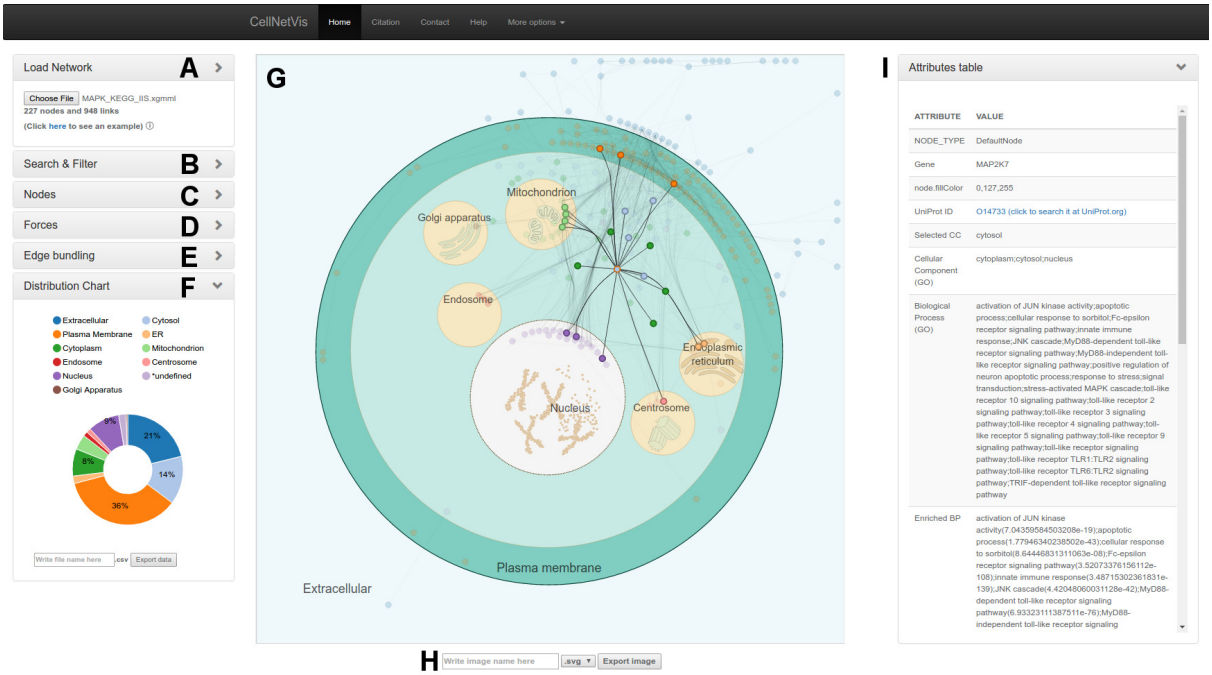
Edge bundling may be applied to the network (Figures 43E e 43G). The effect is to group and smooth edges that flow along the same region of the display. Bundling edges typically reduces the visual density of the network layout, providing a clearer view of the relations among groups of nodes.

CellNetVis allows searching for nodes by label (Figure 43B). The tool then highlights the nodes matching the search. It is possible to hide unselected nodes using the filter functionality ("Filter" button). This allows users to focus the analysis in a fraction of the network.

Node attributes may be displayed by our tool in a tabular fashion that includes a link to the UniProt website whenever the proper accession number is available (Figure 43I). Moreover, three network topology measures can be computed and added to each node: degree, betweenness, and clustering coefficient. The colors of the nodes can also be changed by using the drop-down list of nodes attributes (Figure 43C).

After loading a network, a donut chart showing the counts and percentages of nodes per cellular component is displayed on the bottom left side of the cell diagram (Figure 43F). Such chart may be exported in CSV, SVG and PNG formats. The cellular diagram may be exported as an SVG or PNG file.

Figure 43 – The CellNetVis interface. The main interface sections are indicated by lettering as follows: (A) Load network section, (B) Search section, (C) Nodes section, (D) Force-directed algorithm section, (E) Edge bundling section, (F) Donut chart section, (G) Cell diagram section, (H) Cell diagram export section, and (I) Node attributes table section.



Source: Elaborated by the author.

The following sections describe two use cases and an additional comparison of CellNetVis against other available tools. In all cases, we used the same desktop computer with the following configuration: Chrome web browser version 56 (64-bit), Ubuntu 16.04 (64-bit), Intel Core i7-2600K 3.4 GHz (launch date: 2011), GeForce GTX 750 Ti, and 8GB DDR3 RAM.

4.3.1 Use Case 1: Comparison of GO and HPA subcellular compartments annotations on a Homo sapiens high-throughput network

We used 2097 proteins from the Human Protein Atlas (UHLEN *et al.*, 2010) supportive data (Additional File 2 from (HEBERLE *et al.*, 2017)) to construct a first neighbors network on IIS. A final large network containing 1942 nodes and 17498 links was then exported from IIS to CellNetVis to test the program capacity of handling large networks for a proper visualization and analysis (Figure 44A). Organelles were manually moved to improve the layout (Figure 44B) and edge bundling was turned on (Figure 44C). With these steps, the existence of edges and their frequency between cellular compartments became clearer. As expected, by comparing the donut chart information to the HPA data, the GO annotations ranking by the percentage of nodes distributed in each cellular component was similar to the HPA annotations ranking, particularly concerning the top (nucleus followed by cytoplasm, including cytoskeleton and

cytosolic proteins) and bottom (microtubule organizing center) terms of the ranking (Additional File 3 from (HEBERLE *et al.*, 2017)). This network is available through the CellNetVis Help page, and can be downloaded and uploaded or directly visualized at CellNetVis.

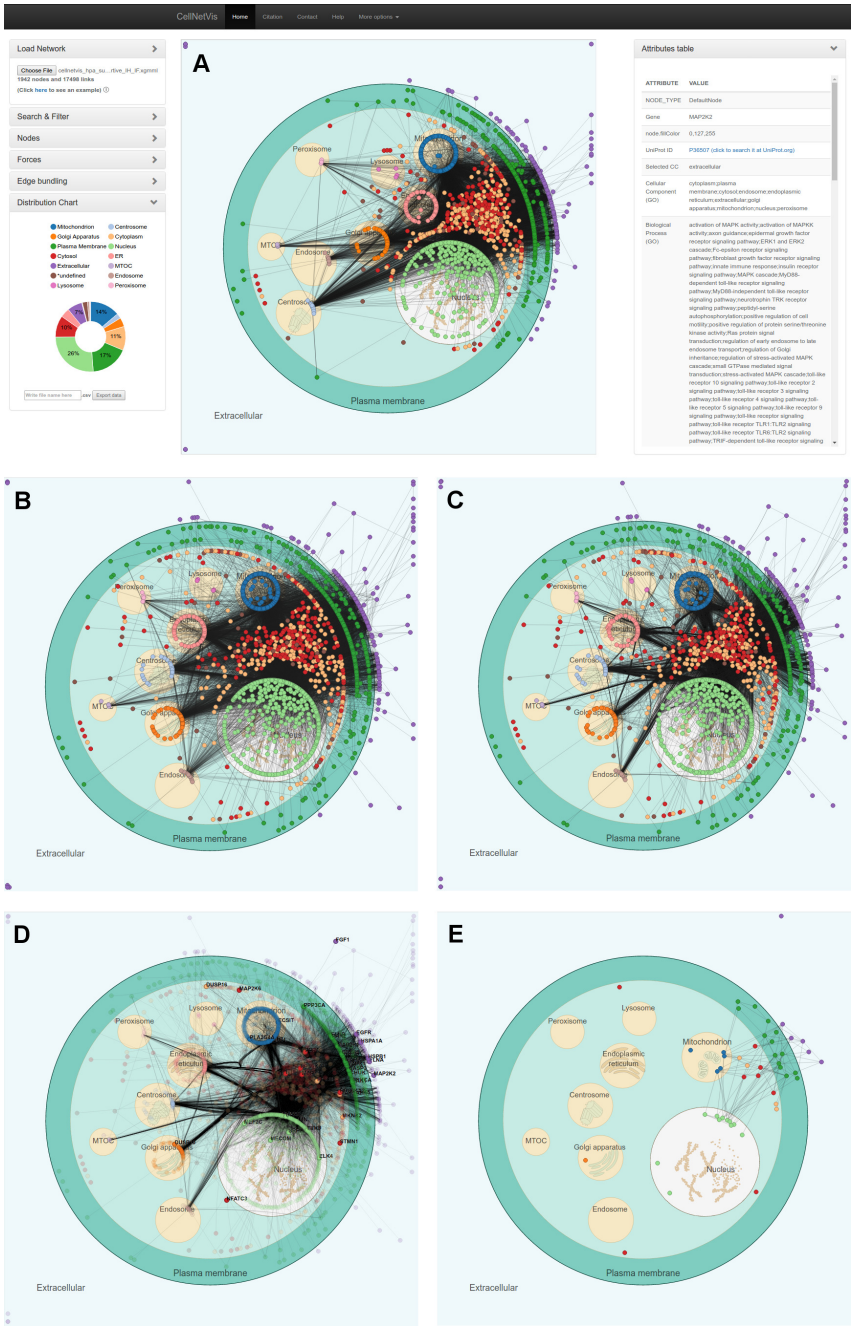
Besides being useful to connect the network to information regarding subcellular compartments, CellNetVis is also useful to analyze their interactions and pathways by setting node colors according to, e.g., the GO biological processes or KEGG (GERMERAAD; HOPPING; MULLER, 1968) pathways, or by highlighting only the nodes annotated for a particular process/pathway, such as the MAPK signaling pathway (Additional File 4 from (HEBERLE *et al.*, 2017)) depicted in Figure 44D.

From the 257 proteins annotated as involved in the MAPK signaling pathway in the KEGG database (Additional File 4 from (HEBERLE *et al.*, 2017)), only a fraction of them was found in the HPA network. Filtering enables this fraction of nodes to be visualized as a separate network, so that the user can more accurately analyze only the interactions pertinent to this specific pathway (Figure 44E). The force-directed algorithm may be restarted, and the layout computed considering only visible nodes.

CellNetVis handled 1942 nodes and 17498 edges, although still showing the hairball effect that most node-link approaches have. Despite the clutter, the user can see the distribution of nodes and edges in cellular components and has an overview of the network. Edge bundling also helps in the overview phase. The filtering function is important in exploration as it allows the user to focus on areas and edges of interest while hiding everything else. The force-directed layout affects only visible nodes and the filtering function can be turned off at any time. Further techniques to change the visualization approach and reduce the hairball problem, e.g. Nodetrix (HENRY; FEKETE; MCGUFFIN, 2007) and Power Graphs (WANG; THILMONY; GU, 2014), are scope for future work.

One limitation of CellNetVis is clear in this use case: although the system response was fast, the edge bundling took six minutes to complete and the non-overlap functionality (repulsive force guided by *repulsive* value) had to be disabled. One alternative to the non-overlap functionality is to set a higher negative *charge* to nodes, which also has the effect of separating them. In our tests, Firefox browser loaded and showed the network three times faster than Chrome. Despite the good loading time, the system response on Chrome was much better than on Firefox. We tested the system response changing the network sizes (number of nodes and edges). According to our analysis, CellNetVis has a smaller and more stable response time on Chrome compared to Firefox (Additional File 5 from (HEBERLE *et al.*, 2017)).

Figure 44 – CellNetVis interface showing a human high-throughput network distributed on a cell diagram. (A) Visualization of the first neighbors network queried from IIS platform using 2097 proteins from the HPA supportive IH and IF data as input. The nodes' colors were set to be displayed according to the "Selected CC" attribute. The MAP2K2 node was selected to show its attributes, as an example, on the table on the right side of the diagram. (B) Some organelles were moved and the force-directed algorithm was stopped. (C) Edge bundling was computed and displayed. (D) Proteins annotated to the human MAPK signaling pathway from KEGG database were searched in the network using the Find button and are shown as highlighted nodes. (E) Only the previously highlighted nodes corresponding to proteins annotated to the human MAPK signaling pathway are visible.



Source: Elaborated by the author.

4.3.2 Use Case 2: Visualization of the Homo sapiens MAPK signaling pathway organized in cellular compartments

We used 257 proteins from the human MAPK signaling pathway in the KEGG database (Additional File 4 from (HEBERLE *et al.*, 2017)) to construct a first neighbors network on IIS. A final small network containing 227 nodes and 948 links was then exported from IIS to CellNetVis (Figure 45A). This file is also available on CellNetVis Help page to be downloaded and then uploaded or directly visualized at CellNetVis. Every time the user loads a different network, only the organelles corresponding to the GO cellular components annotations of that network are loaded in the cell diagram. Therefore, differently from the previously applied filter step on a larger network (Figure 44E), only the organelles annotated for the MAPK signaling pathway proteins are shown in this case. Due to the size of the network, the system response was good both on Chrome and Firefox, with Chrome still showing a larger speed.

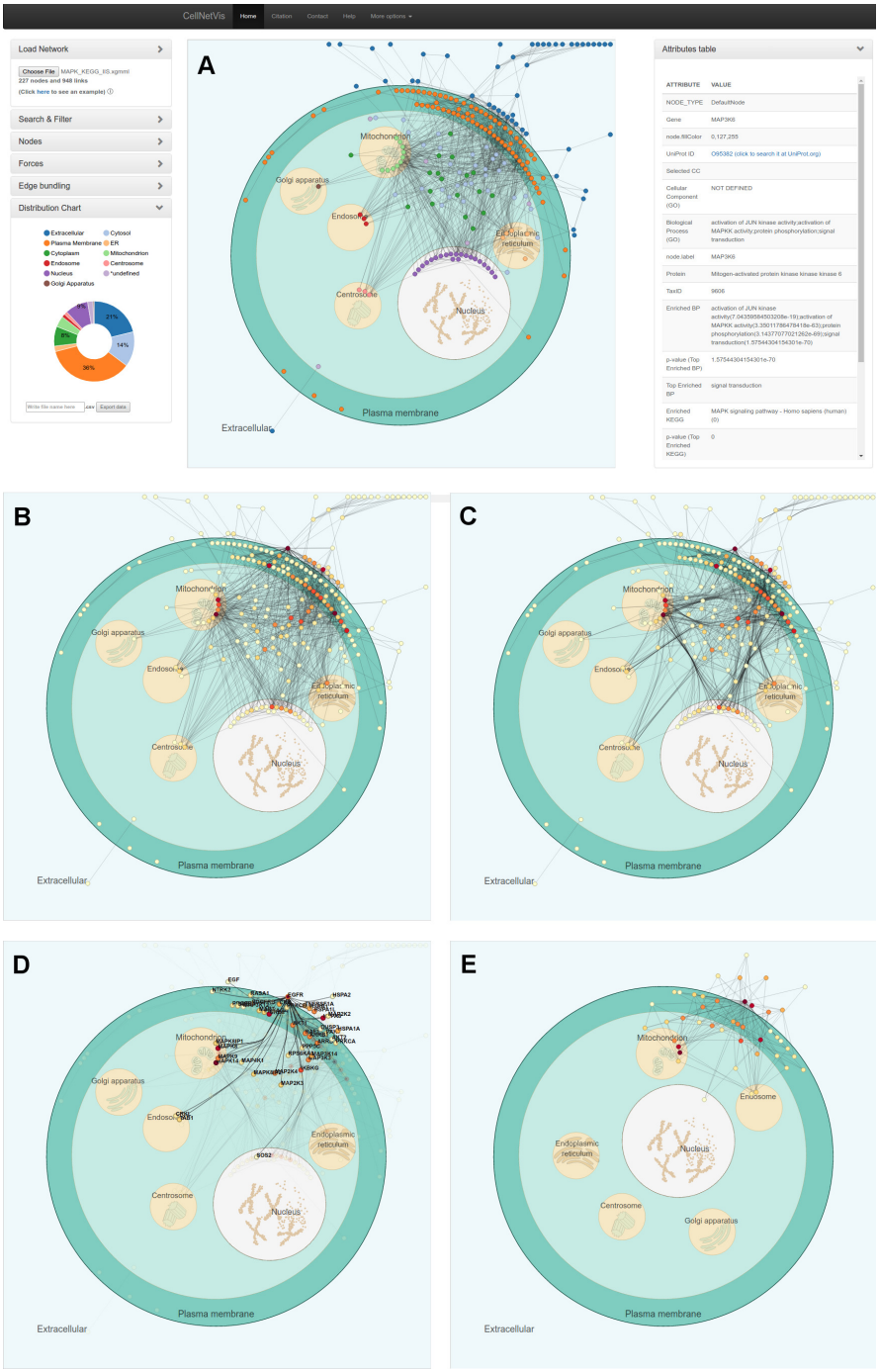
The nodes were colored by their degree, in order to show the hubs (nodes with the highest connectivity), representing the proteins responsible for the major signal integration and transduction in the pathway (Figure 45B). Edge bundling was applied for a better visualization of the main paths of signal flow in the network (Figure 45C). From this analysis, we observe that the main paths occur between the extracellular region and plasma membrane, between the plasma membrane and mitochondrion, endoplasmic reticulum, endosome, centrosome or nucleus, and between the cytosol and the previously mentioned organelles. We can also observe that the hubs (dark red) are mainly located in the extracellular region, plasma membrane and mitochondrion.

By clicking on the node with the darkest color (the highest degree), its label appears (EGFR), the table is updated to show EGFR node attributes on the right side of the diagram, and only the first neighbors of EGFR are highlighted in the network (Figure 45D). This analysis showed that EGFR interacts with proteins on the extracellular region, plasma membrane, cytosol, mitochondrion, endosome and nucleus. By looking at the "Cellular Component (GO)" line on the nodes attributes table, we observe that EGFR is not annotated to localize at mitochondria. This suggests that EGFR may interact with those mitochondrial proteins at other subcellular compartments where they also exist, such as the case of MAPK14, which interaction may occur in the cytoplasm or nucleus. In Figure 45E, organelles were moved and the force-directed layout restarted to create a layout that focuses on the subnetwork topology instead of on concentration and flow of interactions through the cell compartments.

4.3.3 Comparison with available tools

A comparison was performed between the force-directed layout of CellNetVis, the multiple force-directed layout of Mosaic (ZHANG *et al.*, 2012) plugin, and the grid-layout of Cerebral (BARSKY *et al.*, 2007) plugin and CerebralWeb (FRIAS *et al.*, 2015). Although Cell Illustrator Online (CIO) (NAGASAKI; LI, 2010) is capable of showing networks inside a

Figure 45 – CellNetVis interface showing the human MAPK signaling pathway distributed in a cell diagram. (A) Visualization of the first neighbors network queried from IIS platform using 257 proteins annotated to the human MAPK signaling pathway from KEGG database as input. The nodes' colors were set to be displayed according to the "Selected CC" attribute. The attributes of MAP3K6 are shown on the table on the right side of the diagram. (B) The nodes' colors were set to be displayed according to the "[degree]" attribute. The force-directed algorithm was stopped. (C) Edge bundling was computed and displayed. (D) The node with the darkest color, EGFR, was selected. The highlighted nodes correspond to the EGFR's first neighbors, after EGFR's node selection. (E) Only highlighted nodes are visible and force-directed layout was restarted. Organelles were moved to improve the layout.



Source: Elaborated by the author.

cell diagram, the modeling and cell diagram must be manually set up, the tool focuses on the molecular mechanisms and is not freely available, thus, it was not considered in the comparison.

Our focus is freely available systems that can automatically partition the network into a cellular diagram and display a simple and interactive overview in a fast and easy way. Although Cerebral and CerebralWeb do not display a cell diagram, they can automatically separate the network into partitions. Also, CerebralWeb is freely available and can be integrated into web systems. Mosaic is not web-based, but it can automatically place nodes over a cell diagram, therefore it was also considered in the comparison. The main characteristics in contrast with CellNetVis are detailed in Table 4

Table 4 – Characteristics of CellNetVis, Cerebral, CerebralWeb and Mosaic.

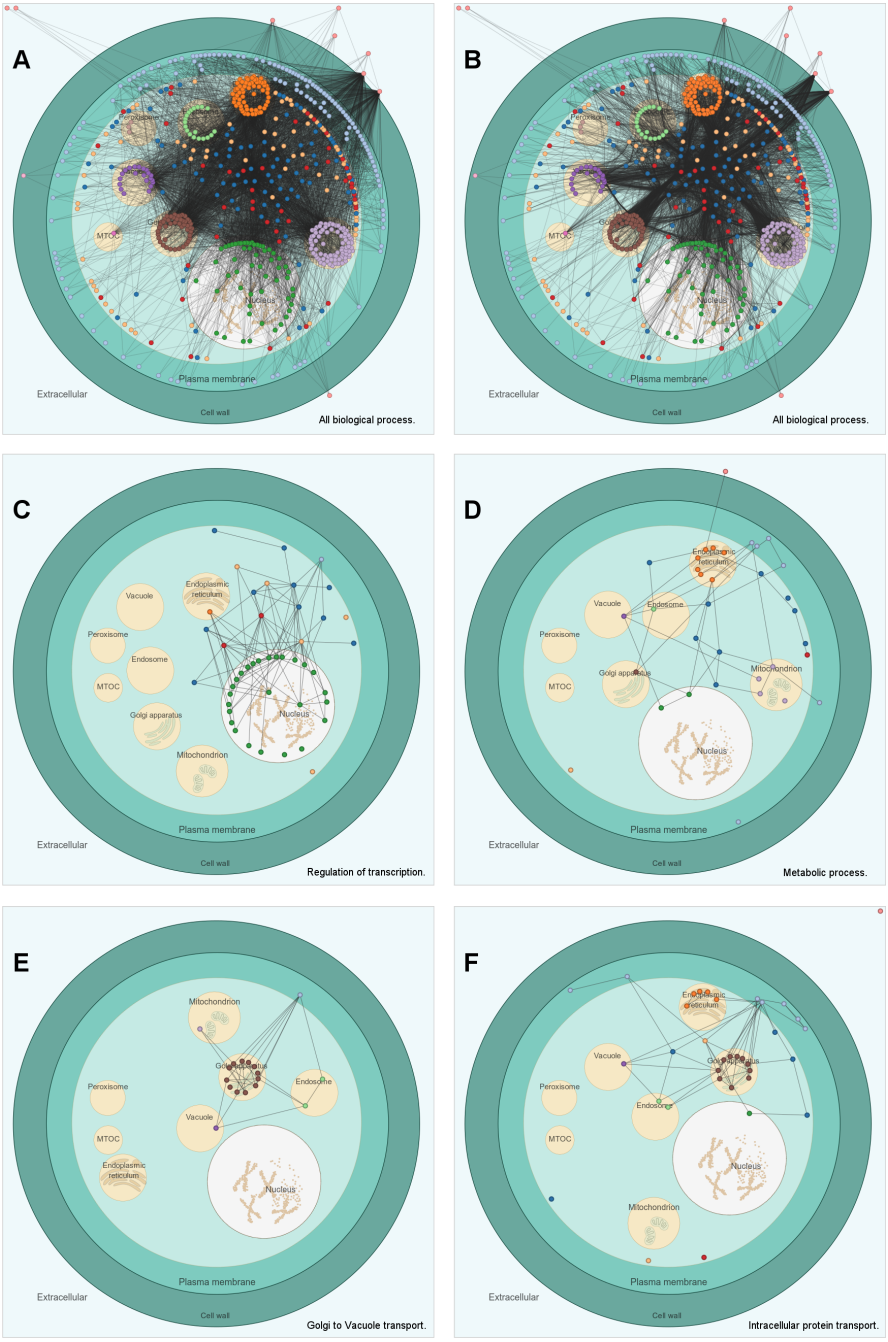
	Cerebral	CerebralWeb	Mosaic	CellNetVis
Network layout updates while interacting				×
Plots the network over a full cell diagram			×	×
Web-based system		×		×
Clutter relief by edge bundling	×			×
Movable cellular partitions				×
Highlight possible CCs of selected node				×
Supports large networks	×	×		×
Accepts pre-annotated cellular localization	×	×		×
Shareable visualization through URL				×
Downloads nodes' unique CC from InnateDB		×		×
Shows quantity of nodes that are in each CC				×
Cytoscape independent		×		×
Online-database independent	×	×		×
Ready for use (non-programming dependent)	×		×	×
Open-source and freely available	×	×	×	×

Mosaic is a Cytoscape (desktop) plugin which partitions a network into subnetworks based on GO Biological Process annotation. Each subnetwork is shown in a different cellular diagram. If a node has more than one value to this attribute, the node is duplicated. Since the tool uses the force-directed algorithm to place nodes, the layout is similar to CellNetVis. However, the system was designed to load the small subnetworks created based on nodes annotations. Overlap of nodes is very common even for small networks.

We could not replicate the results described in (ZHANG *et al.*, 2012) using Mosaic since it is out of date and could not download its required databases. Therefore we decided to create an analysis based on the Yeast example network available at Mosaic web page (MOSAIC, 2012). We created a new annotated network (642 nodes and 7785 edges) with all interactions, found by the IIS, between all the listed genes from the Yeast example. Then, we visualized on CellNetVis the network (Figures 46A and 46B) and subnetworks created by filtering the biological process annotations: 'regulation of transcription', 'metabolic process', 'golgi to vacuole transport', and 'intracellular protein transport' (Figures 46C, 46D, 46E and 46F, respectively). Using as basis

the figures (ZHANG, 2012a; ZHANG, 2012b; ZHANG, 2012c) displayed on Mosaic web page, section *Navigating the results*, CellNetVis performed better, since nodes did not overlap on any of the displayed subnetworks and their topology was clear.

Figure 46 – Visualization of Yeast subnetworks filtered by specific biological processes. (A) and (B) represent the complete Yeast network formed by 642 nodes and 7785 edges. The network was filtered according to the following biological processes: 'regulation of transcription' (C), 'metabolic process' (D), 'golgi to vacuole transport' (E), and 'intracellular protein transport' (F).

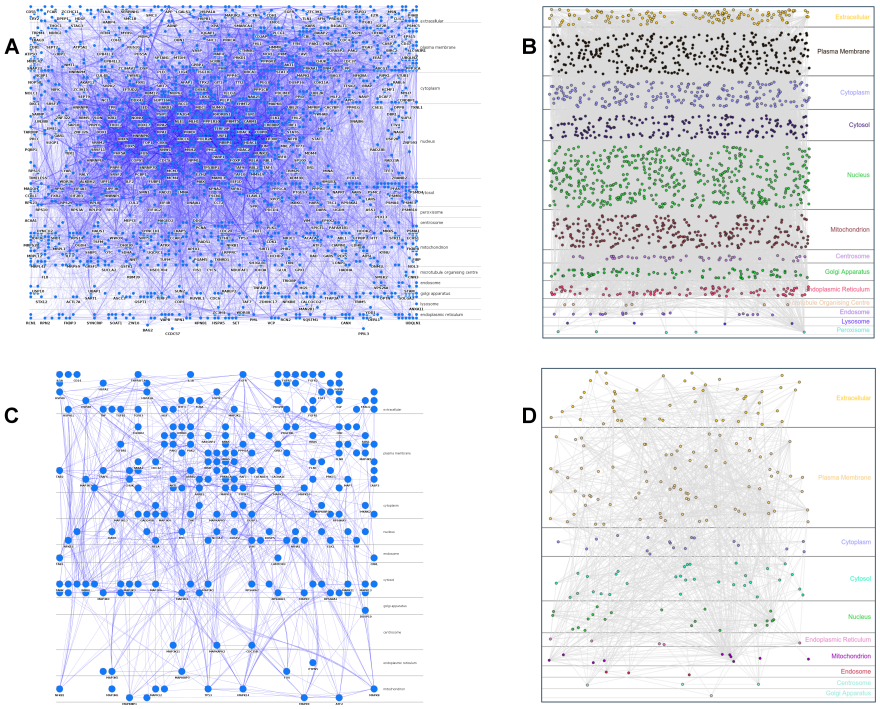


Source: Elaborated by the author.

Regarding the Cerebral plugin and CerebralWeb, the network layout algorithm is modeled

after hand-drawn pathway diagrams, where nodes are restricted to a regular lattice grid that provides room for labels and eliminates overlapping nodes (BARSKY *et al.*, 2007). The main difference to CellNetVis is the use of a grid layout to position nodes on horizontal layers, one over the other, so as to resemble subcellular compartments. However, the use of horizontal layers for this purpose restricts cell layers to the five major subcellular compartments, which are positioned by Cerebral from top to bottom in the following order: extracellular, cell surface, plasma membrane, cytoplasm and nucleus. For instance, the majority of organelles, which are naturally localized in the cytoplasm, cannot be drawn inside the cytoplasm layer in Cerebral, only as horizontal layers on the top, bottom or between the other ones (e.g. below nucleus, as default), which is not consistent with an appropriate cellular view (Figure 47). The same happens in the web-based version of the system.

Figure 47 – Visualization of a large and of a small network on the Cerebral Cytoscape plugin (A and C) and on CerebralWeb (B and D). (A and B) Large network generated from the HPA supportive data. The drawing took approximately 3.5 min. in Cerebral (A) and 6 s. in CerebralWeb (B). (C and D) Small network generated from the human MAPK signaling pathway from KEGG database. The drawing took approximately 5 s. on Cerebral (C) and 1 s. on (D). HPA: Human Protein Atlas; MAPK: Mitogen-activated protein kinases.



Source: Elaborated by the author.

Comparing the loading and drawing times for a large network composed of 1942 nodes, Cerebral took about 4 minutes, while CellNetVis took half the time to load the network file, to check for duplicate nodes and edges, to create the data structure, to start the force-directed layout, and nearly stabilize the force system and to display a consistent layout of the network topology. For a small network composed of 227 nodes, Cerebral took 10 seconds, while CellNetVis took

approximately 1 second.

To compare the layout created by CerebralWeb and CellNetVis we created the displays for the networks from Use Case 1 (Figure 47A vs. Figure 44A) and Use Case 2 (Figure 47B vs. Figure 45A). In both cases, CerebralWeb was not capable of clearly representing the density of interaction between compartments as CellNetVis does. For instance, in Figure 44A we can see that there are more interactions between mitochondrion and nucleus than between endoplasmic reticulum and nucleus; in CerebralWeb it is not possible to see this pattern (Figure 47B). Moving the organelles on CellNetVis also allows the user to check this type of information. Considering the overview of the network on CerebralWeb (Figure 47B), the only information we can visually identify in the diagram is the distribution of nodes over the compartments. This information can be more easily identified in CellNetVis through the distribution chart (Figure 43F). Thus the overview created by CellNetVis is more informative than the one created by CerebralWeb. In contrast to Cerebral, CerebralWeb can draw large networks fast, but the layout is not as good as the layout computed by the Cerebral plugin (Figures 47A and 47C). We integrated CerebralWeb to CellNetVis system, which can be accessed through the "More options" top-menu item after loading a network. Both CerebralWeb and CellNetVis layout were displayed almost instantly after loading the network file from Use Case 2.

Another advantage of CellNetVis concerns the highlight and filtering of nodes or pathways in a complex network. As shown in Figure 44A, when a network is large there are many nodes overlapping. CellNetVis allows the user to filter nodes based on a search query. These filtered nodes can be automatically repositioned. This functionality and interactivity improves the network display and exploration and is not possible in Cerebral where the layout is pre-calculated. Cerebral only allows the highlight of neighbors for a selected node and is able to recalculate the layout as a second drawing step, but only considering all the nodes. The web version needs to be programmed to be used with these features, despite being implemented as a module of the CerebralWeb Javascript library.

One fact that could be considered a limitation of CellNetVis appears in Figure 44A, where nodes overlap at a high rate due to network size. However, the overlap of nodes is what allows the density of edges between organelles clear supporting the overview task and being more informative than the non-overlap layout created by CerebralWeb algorithm. CellNetVis can show at the overview step the connectivity among compartments (edges densities), the distribution of nodes (chart distribution), and give details according to the user interactions by search, filtering, and selection of nodes. After filtering a large network, for instance, the charges of nodes or *repulsive* value can be increased to drastically reduce overlapping effect. Considering the critical execution time that happens on general web-applications, we could say for both web-based layouts compared in this section, that CellNetVis and CerebralWeb focus on being fast enough to be used with considerably large networks. CellNetVis lets nodes overlap at a high rate when networks are large, but keeps the dynamic aspect of the layout and accentuate the

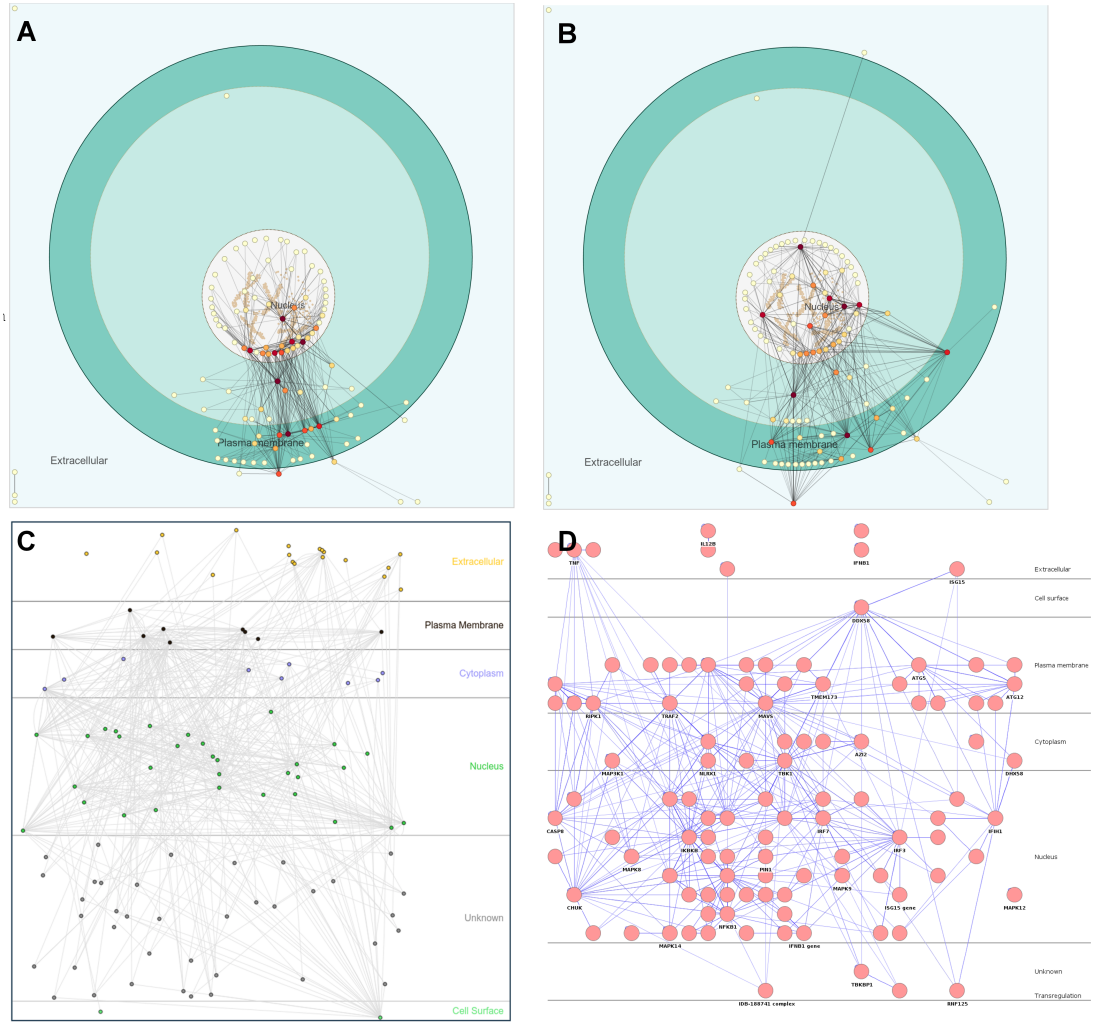
concentration of edges (Figure 44), whereas CerebralWeb layout algorithm avoids the overlap of nodes but is not dynamic and hides the overview of the network topology (Figure 47B).

The positioning algorithm of CellNetVis works well with both small and large networks and supports more directly the visualization pipeline described by Schneiderman (SHNEIDERMAN, 2003): overview, followed by zoom and filter, then details on demand. If a user modifies the position of a node or organelle in the network representation, CellNetVis is able to recalculate the position of the other nodes instantly, while Cerebral and CerebralWeb are not. Moving a node or organelle can highlight certain aspects of the data (Figure 48A). For instance, if nodes are too close inside the cell, the user can separate nodes to let the topology clear (Figure 48B). This cannot be accomplished by using CerebralWeb (Figure 48C) or Cerebral (Figure 48D). CellNetVis was shown to be a more flexible tool throughout user interaction tasks. Due to characteristics of the layout algorithm, the movement of a node in CellNetVis is not so smooth and precise, but still usable and useful.

4.4 Conclusions

CellNetVis is a free and open-source web-based software for displaying biomolecular networks in a cell diagram. It is capable of displaying complex information related to networks, nodes and edges, as well as their relations with cell partitions. While being better suited for small and medium-sized networks, CellNetVis is also capable of handling large networks. In comparison with other algorithms and tools, CellNetVis has shown to be competitive, particularly for a dynamic exploration of complex networks over a consistent representation of a full cell on the Web. CellNetVis is being used by the IIS as its main visualization system. CellNetVis may also be coupled with different annotation softwares using the XGMML format to exchange data, providing an interesting analysis layer.

Figure 48 – Visualization of the network formed by the interactions within the "RIG-I-like receptor signaling pathway (KEGG)" in *Homo sapiens*, downloaded from InnateDB (<http://innatedb.ca/interactionSearch.do?from=pw&exPathwayXref=&pathwayFilter=5713&pathwayXrefDB=&pathwayXref=&listType=interaction&coreInteractors=true>) as a XGMML file and loaded on CellNetVis, CerebralWeb and Cerebral. (A and B) Visualization of the network on CellNetVis before (A) and after (B) manually separating nodes with high degree (dark red). The same network was drawn on CerebralWeb (C) and Cerebral (D) for comparison. CellNetVis was shown to be a more flexible tool through user interaction.



Source: Elaborated by the author.

CHAPTER

5

PHYLOGENETIC TREES AND LATERAL GENE TRANSFERS

In this chapter we report a research started during my participation in the Emerging Leaders in the Americas Program. I was supervised by professor Dr. Robert Beiko from Dalhousie University, Halifax, Canada. I worked in his laboratory for six months, starting in March 2018.

In the following sections, we define real-life problems that are linked to bacteria antibiotic resistance and how phylogenetic supertrees are adopted to understand the dynamics of lateral gene transfers. During my stay in Canada, I started building a tool for the visualization and exploration of such information. We list the main results obtained so far with the firsts versions of the visualization techniques applied to the analysis of bacteria genomes. Then, we translate the supertrees problem to the biomarkers discovery question, exploring the prostate cancer data set under a different perspective.

5.1 Introduction

Investigating the evolutionary pattern of large gene sets sampled from multiple genomes, combined with other omics information, is fundamental to solve many problems in Biology (JOYCE; PALSSON, 2006). A phylogenetic tree is a common approach to present an evolutionary history, showing the transmission of DNA from parent to offspring (PAGEL, 1999). However, a phylogenetic tree can be influenced by processes such as gene loss and lateral gene transfer (LGT) - horizontal gene transfer (HGT) (MADDISON; KNOWLES, 2006; GALTIER; DAUBIN, 2008).

The horizontal transmission and proliferation of antimicrobial resistance (AMR) related genes is a major and increasing global problem (PAGEL, 1999). One example is the New Delhi metallo-β-lactamase (NDM-1) which has raised a major public health concern for the reason that this enzyme makes bacteria resistant to a broad range of β-lactam antibiotics. Thus, NDM-1

strains are capable of widespread horizontal transmission among bacterial populations, since it is often carried by plasmids, increasing the probability of rising drug-resistant bacteria (KHAN; MARYAM; ZARRILLI, 2017). Plasmids can be transferred from a bacterium to another through a process called conjugation (REDFIELD, 2001).

One tool to investigate these patterns of AMR proliferation and LGT is that of phylogenetic supertrees (BEIKO; HARLOW; RAGAN, 2005). Phylogenetic supertrees are a way of combining individual gene phylogenies covering a selection of taxa[1]. Since each gene is described as a phylogenetic tree, the supertree represents an overall evolutionary history of the taxa and individual gene trees can reveal patterns about horizontal transmission (DAVIES *et al.*, 2004; PISANI; COTTON; MCINERNEY, 2007; BEIKO; HARLOW; RAGAN, 2005).

The results of combining many trees are not trivial and specific criteria are needed to accommodate the confounding effects of LGT and other sources of phylogenetic disagreement. To understand the different phylogenetic patterns that exist in a set of trees, supertree representations must be able to map the patterns present in individual constituent gene trees.

In this work, we aim to design a visual approach for the analysis of supertrees, to allow specialists to find key information through visual exploration, regarding trees' disagreements that could not be identified using regular tree diagrams. Interactive data visualization plays an important role in tree analysis. We expect that our approach will help the analysts to find hidden information and answer important questions when analyzing supertrees, such as the major pathways and correlated LGT associated with different classes of antimicrobial resistance such as metallo-β-lactamase.

5.2 Gene trees and supertrees

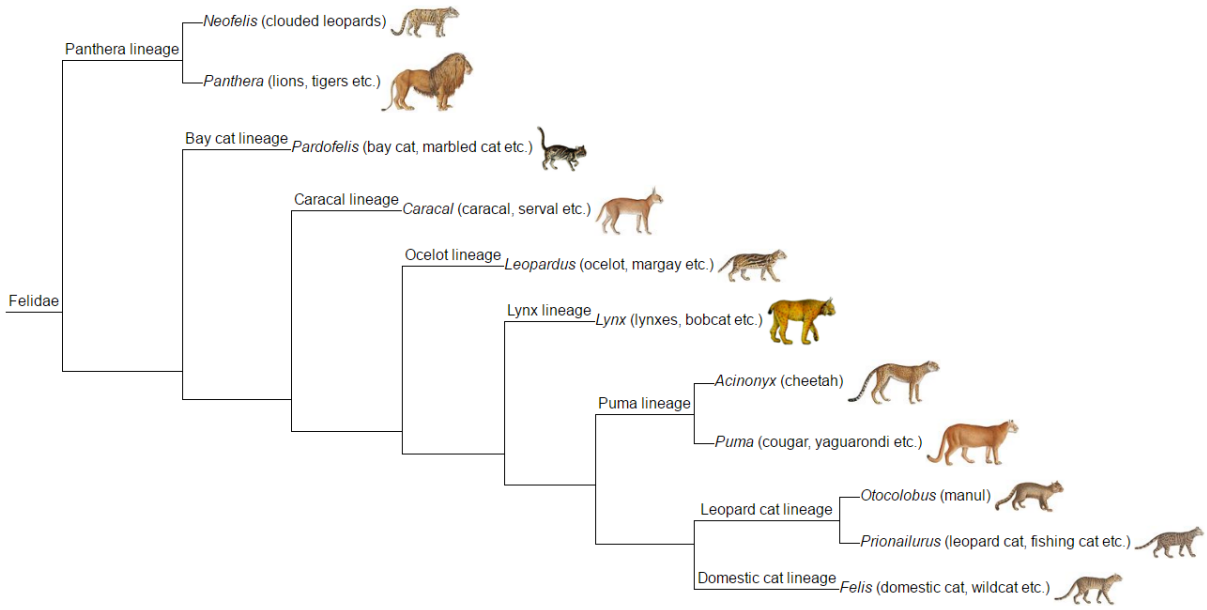
A phylogenetic tree describes the evolutionary relationships among species or other entities based on similarities and differences in, for instance, their skeleton, genes and more. Figure 49 presents a phylogenetic tree of the biological family Felidae, which are carnivorans commonly referred as cats.

In a rooted phylogenetic tree, each internal node represents the common ancestor of its descendants. The lengths of the edges can represent the estimates of time. The nodes are the taxonomic units. The internal taxonomic units are hypothetical once we cannot observe these ancestors.

One way to generate a phylogenetic tree is by computing similarities and differences between genomes. That is, based on the genetic characteristic of a taxon we can infer who the ancestors are. By linking one by one and comparing similarities and differences between taxa

[1] taxon (plural taxa) is a group of one or more populations of an organism or organisms seen by taxonomists to form a unit" (REVIEWS, 2016)

Figure 49 – Phylogeny of felidae.



Source: figure in the public domain, from (Wikimedia Commons, 2017).

and groups of taxa we can build a phylogenetic tree. Neighbor Joining is a popular technique to build such trees (SAITOU; NEI, 1987).

While some methods are based on concatenated alignments of many gene sequences to form a supermatrix, others consider the evolution of each single gene to create a supertree. On a moderate amount of misleading LGT, simulations indicated that supertrees can be more reliable than the supermatrix approach (LAPIERRE; LASEK-NESSELQUIST; GOGARTEN, 2014; WHIDDEN; ZEH; BEIKO, 2014).

The general idea of a supertree is: each gene derives a phylogenetic tree that tells us a history; a method that creates a supertree tries to maximize the agreements among these trees and minimize the disagreements. For instance, we could have 40 thousand gene trees composed by different numbers of taxa (trees of different sizes) and compute one supertree (species tree) that represents their *history* minimizing the disagreements between gene trees and supertree. One approach proposed by Bansal *et al.* (2010) is to minimize the total Robinson-Foulds (RF) distance (BOGDANOWICZ; GIARO, 2017) in this comparison. Another approach is to use the subtree prune-and-regraft (SPR) distance (HEIN *et al.*, 1995) instead. Whidden, Zeh and Beiko (2014) created "the first method to construct supertrees by controlling the SPR distance as an optimality criterion". Whidden, Zeh and Beiko (2014) explained:

> The SPR operation involves splitting a pendant subtree from the rest of the tree, and reattaching it at a different location, with the rooting of the subtree preserved. The SPR distance is the minimum number of such operations required to reconcile two trees and an SPR supertree minimizes the sum of SPR distances. A single SPR operation can accommodate a long-distance transfer, whereas the RF distance could be drastically

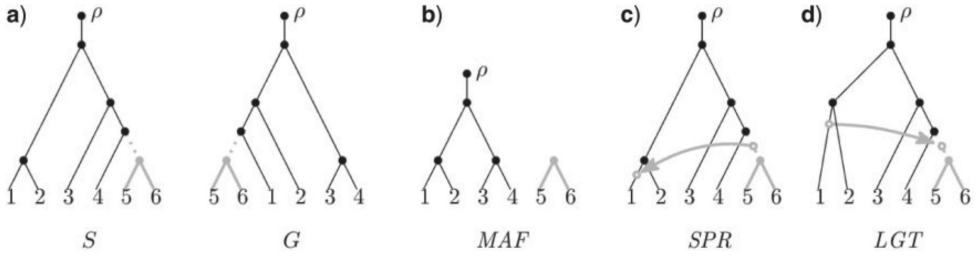
> increased by such a transfer. We therefore expect that optimizing the SPR distance will be more likely to yield the true tree, as opposed to RF, which may be unduly influenced by large topological dissimilarities in the input trees (WHIDDEN; ZEH; BEIKO, 2014).

Whidden, Zeh and Beiko (2014) compared both methods and wrote:

> Using a series of benchmark datasets simulated under plausible rates of LGT, we show that SPR supertrees are more similar to correct species histories than supertrees based on parsimony or Robinson-Foulds distance criteria. We successfully constructed an SPR supertree from a phylogenomic dataset of 40,631 gene trees that covered 244 genomes representing several major bacterial phyla. Our SPR-based approach also allowed direct inference of highways of gene transfer between bacterial classes and genera (WHIDDEN; ZEH; BEIKO, 2014).

When comparing each gene tree with the supertree, we can see where they disagree. These disagreements can be an indication of gene loss and lateral gene transfer. Whidden, Zeh and Beiko (2014) explained how an LGT can be modeled from an SPR move in Figure 50.

Figure 50 – The equivalence between the SPR distance and MAF size. (a) The species tree S and gene tree G differ particularly in the placement of the gray subtree. The roots of these trees are denoted by ρ. (b) The MAF of S and G is produced by cutting the dotted edge in both trees. (c) Each component of an MAF other than the component containing ρ represents an SPR move. A single SPR move transforms S into G by moving the gray subtree in S to its position in G. (d) Each SPR move models an LGT event in the reverse direction. From the MAF of S and G we infer that a transfer of gene G has occurred from an ancestor of taxon 1 to an ancestor of taxon 4.



Source: figure and legend from (WHIDDEN; ZEH; BEIKO, 2014), licensed under CC BY 3.0.

5.2.1 Visualizing trees

The data sets in a tree structure can be visualized in many ways by applying different techniques and layouts (SCHULZ, 2011). The Figures 49 and 50 illustrates the most common approach, that is by representing the hierarchical relationships with bifurcating or multifurcating edges.

Many techniques for tree visualization applied in different types of applications are listed in the *Visual Bibliography of Tree Visualization* web-site (SCHULZ, 2011). Some vi-

sual techniques for the visual exploration of trees are the TreeJuxtaposer (MUNZNER *et al.*, 2003), the Java TreeView (SALDANHA, 2004), the T-REX (MAKARENKOV, 2001), Treemaps (SYMEONIDIS; TOLLIS, 2005), the Interactive Tree Of Life (LETUNIC; BORK, 2016), the Dendroscope (MONNÉ; MONNÉ, 2008), the PhyloWidget (JORDAN; PIEL, 2008), the Treevolution (SANTAMARÍA; THERÓN, 2009), the CompPhy (FIORINI *et al.*, 2014), the Phylo.io (ROBINSON; DYLUS; DESSIMOZ, 2016), the Ggtree (YU *et al.*, 2017), the PhyD3 (KREFT *et al.*, 2017) and the IcyTree (VAUGHAN, 2017).

Despite there are many systems to examine phylogenetic trees, none is focused on the disagreements between thousands of gene trees and a supertree. Even though the SPR supertrees can produce reliable phylogenetic trees, analyzing the disagreements of the gene trees is a complex process due to the number of genes in the genomes. Here we define an approach to visualize both evolutionary history of taxa and possible lateral gene transfers, highlighting agreements and disagreements onto a supertree diagram. We present our main decisions and technologies used to implement our solution. Finally, we exemplify the use of our approach on two use cases, one in the original context of phylogenetic trees and the other adapting our approach to analyze the proteins and prostate cancer patients using the data set from Chapter 3.

5.3 Material and methods

In this section we present the main decisions, characteristics of the problem, visualization questions and technologies adopted in our solution. Despite the prototype of the visualization techniques defined here is functional, it is yet a work in progress.

One decision we made is that our visualization prototype would, for now, work entirely in the client-side through the web-browser. This brought us to make an additional software available that will read a supertree, gene trees and an annotations file, and output a file that is the only input for the visualization system. This software is adapted from RSPR tool developed by Whidden, Zeh and Beiko (2014). It can be utilized to construct the SPR supertree and, therefore, the input can be simply the gene trees and annotations file.

Installation is a great problem for the Omics areas. Mangul *et al.* (2018) tested the accessibility of computational biology software tools and found that “Among the tools selected for our comprehensive and systematic usability test, 49% were deemed ‘difficult to install’, and 28% of the tools failed to be installed due to problems in the implementation” (MANGUL *et al.*, 2018). In a future, both the visualization system and the RSPR may be optimized and coupled in a user-friendly web-based system with no download, installation nor compilation required.

5.3.1 Visualization questions

We defined questions that we believe are important for the understanding of the relationships between the supertree and gene trees. The questions defined how we designed our visual approach.

We state here that when we cite *edges* we are referring to the ones that connect the supertree nodes but are not part of the supertree. They can be SPR edges that represent disagreements or edges that indicate shared genes and, in this case, it may or may not represent disagreements. It is important to note that one edge can represent more than one gene tree and, therefore, more than one LGT event too. The visual analysis aimed by our approach has two purposes: for understanding the SPR supertree algorithm's behavior and for understanding the evolutionary events.

We defined bellow some questions that guided the design of the tool. The combination of these questions may represent different sequences of multiple tasks. Many other questions may be defined and the list may be improved upon to the next stages of this project.

- What are the edges that represent SPR moves? What are the edges that represent only shared genes?
- What are the more distant edges?
- What are the gene functions or biological processes associated to the genes that derived a filtered edge or set of edges?
- What are the edges that are associated to a given gene function of a biological process?
- What are the edges associated to a given gene or set of genes?
- What are the edges that appear inside a group of taxa? What are the ones that appear between different groups of taxa?
- Given filtering set up of edges (by function, numerical attributes and others), what are the genes that derived the edges?
- How is the flow of all or filtered edges among the groups of taxa?
- What are the genes that represent higher disagreements? What are the taxonomic units that they connect?
- How genes are distributed among the taxa?
- What are the taxa that have the biggest or smallest genomes? How do they connect with other taxonomic units?

5.3.2 Visualization techniques and frameworks

We decided to represent the supertree in a radial layout, separating the groups of taxa by color and indicating the density of shared genes by links's thickness. Labels of each genome/taxon/taxonomic unit are displayed around the radial layout. Genomes are likewise represented by colored bars (*genome-bars*), that indicate the number of genes. Each gene in the bars is colored according to the group where it has the maximum count.

For the shared genes and SPR edges, we decided to use regular straight and thin edges due to the rendering complexity in a web-browser. We utilized the WebGL to allow the drawing of thousands (100,000+) of edges and yet have a smooth graphic interface for visualization and interaction. For this, we used the WebGL framework PIXI.js (<http://www.pixijs.com/>).

5.3.3 Use cases

To illustrate the use of the visualization system, we reported two use cases. In the first we apply the original data set which is the main motivation of this work and studied by Dr. Robert Beiko (WHIDDEN; ZEH; BEIKO, 2014), supervisor of this project. The data represents well the characteristics of genomes with a high rate of lateral gene transfers and the complexity of real life problems, such as the research on AMR. The second use case is not related to lateral gene transfers and was created to illustrate the use of annotations together with the tree analysis once the first data set is not annotated. It also exemplifies the use of this technology in a different context.

5.4 Results

We implemented our approach as a web-tool named *sTVis - superTree Visualization*[2]. It allows the visualization and exploration of supertrees and its relations to the original gene trees. In this section we describe the system and demonstrate how our approach can be used to examine the trees disagreements through two use cases.

5.4.1 sTVis: a web-tool for visual exploration of supertrees

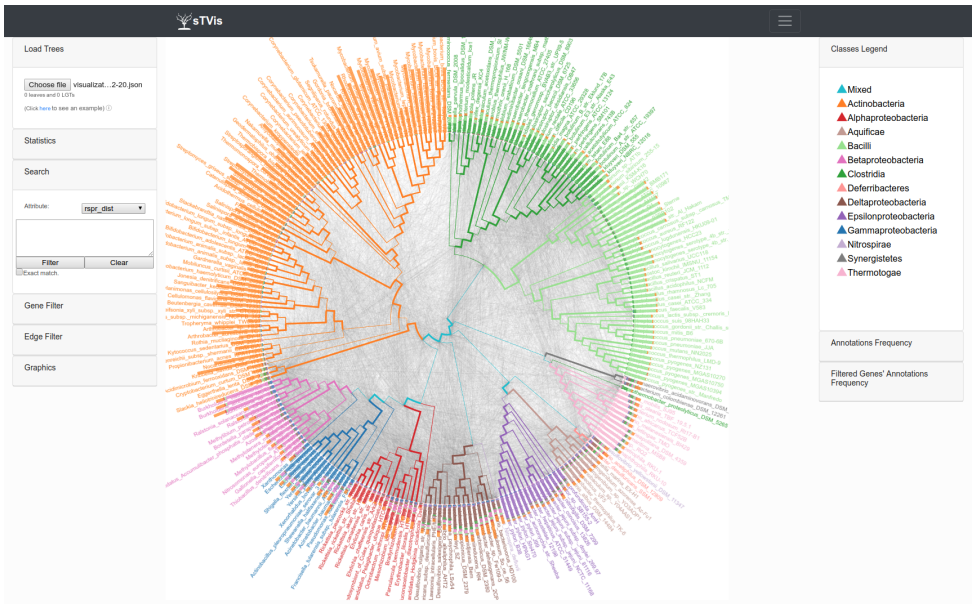
The sTVis is free, open-source and implemented in HTML, Javascript and WebGL (PIXI.js). Figure 51 shows the main interface of the system. The supertree is represented in a radial layout and the LGT edges transparent. The user can see the taxa names around the tree and the genome-bars. These three elements' transparency can be modified to display, for instance, only genome-bars, only labels or only highlighted edges.

The system is still in construction and few interaction features are implemented. For instance, the user can: click on a genome-bar to highlight edges that connect this unit to others or

[2] Fork and contribute: <https://github.com/heberleh/supertree>

the ones that connect the units in the path to the root to other units; filter edges by searching for genes that are annotated with some values (ID included) and by limiting the accepted range of numerical attributes; see the most frequent annotations from genes that are in the visible edges and click on the value to highlight the edges that contains that value. Some numerical attributes created by the supertree_rspr tool are the number of genes that support an edge, the number of genomes where a gene can be found and the SPR distances of gene trees to the supertree.

Figure 51 – The main interface of sTVis web-based tool for the visual exploration of supertrees.



Source: Elaborated by the author.

The two main requirements to run the system are: a modern web-browser that supports WebGL and a computer with a good graphic card. Depending on the number of edges that are going to be loaded in the system, the graphic card and the CPU need to have a good performance. In our tests, the system could show smoothly the 118,341 edges from use case 1 in a laptop with quad-core processor Intel Core i7-7700HQ CPU (2.80GHz), dedicated graphic card GeForce GTX 1050 Ti, operational system Ubuntu 18.04 and web-browser Google Chrome version 70.0 (64bit). The interaction such as search and filtering by numerical attributes were fast.

5.4.2 Use case 1

In this section we exemplify the use of sTVis to analyze a supertree and its gene trees studied by Whidden, Zeh and Beiko (2014). The supertree represents 244 bacteria (genomes) divided into 13 classes. Detailed information about the construction of the supertree and gene trees are not scope of this dissertation and can be found in the original article. The supertree is named “MRP Rooted Gene Trees” and the gene trees set is the “SPR-MRP-Rooting”. We chose the combination of these two sets because they represent the smallest SPR distance among

the studied gene trees and supertrees, according to Table 2 in their article (WHIDDEN; ZEH; BEIKO, 2014).

One limitation in this data set is that it does not contain the genes' names nor any gene ID and, thus, we could not carry the biological annotations nor verify if some interesting edge contains genes that are already known as LGT genes in the literature. In a **future work**, we intend to create a proper data set with genes' annotations from a study on LGTs involving Salmonella bacteria. For this reason, to show how the system can be used when there are annotations and, also, to link this chapter to the biomarker discovery studies, we created the use case 2 (Section 5.4.3), where each *protein tree* is provided with biological annotations.

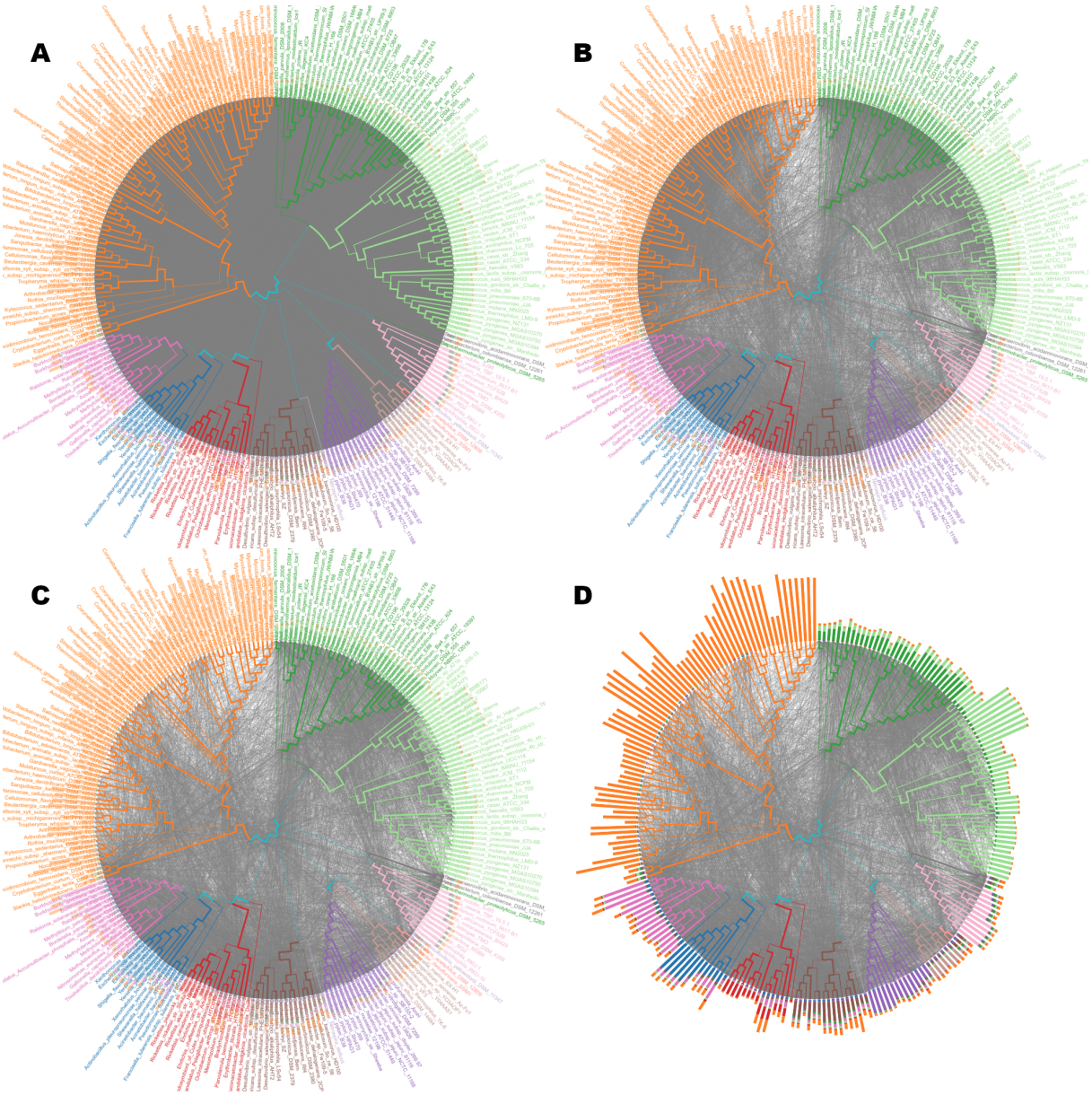
The overview of the supertree and gene trees edges is shown in Figure 52. In this figure we see the great number of edges (A) that represent the connections between units with shared genes and SPR moves. To create the shared-genes edges, we considered that each hypothetical unit of the tree is represented by the intersection of genes from their descendents. The other sub-figures display only SPR moves between all units (B) and between units of different classes (C and D). In (D) we show how the genes are distributed in the genomes considering the rule: a gene is colored according to the class where it occurs the most (greater number of genomes). With this criterion, it is expected that bigger classes *dominate* and their colors appear more frequently in all genomes. Despite that, we can see that even the smaller classes are being represented in genomes of all the other classes.

Whidden, Zeh and Beiko (2014) analyzed the SPR moves between taxa and subgroups of taxa using heat map and a summarized (clustered) network. Both are indicated in Figure 53. Despite they are great representations of the overview of LGTs, they do not support interaction nor reveals specific patterns or genes that are being part of the transfers.

In contrast with the static representations of possible LGTs events, the sTVis tool allows the user to interact and enables such analysis. The Figure 54 presents an example of interaction: when we click in a genome, we can highlight the SPR that would move this genome to connect to other regions in the supertree (A and C), so that the gene trees would agree with the supertree in the topology; we can also highlight all SPR moves that could also move its ancestors (B and D). This figure illustrates the variance of SPR moves among the taxa. In Figure 54 we compare the possible LGTs from or to *Streptococcus pyogenes* and *Bacillus clausii*.

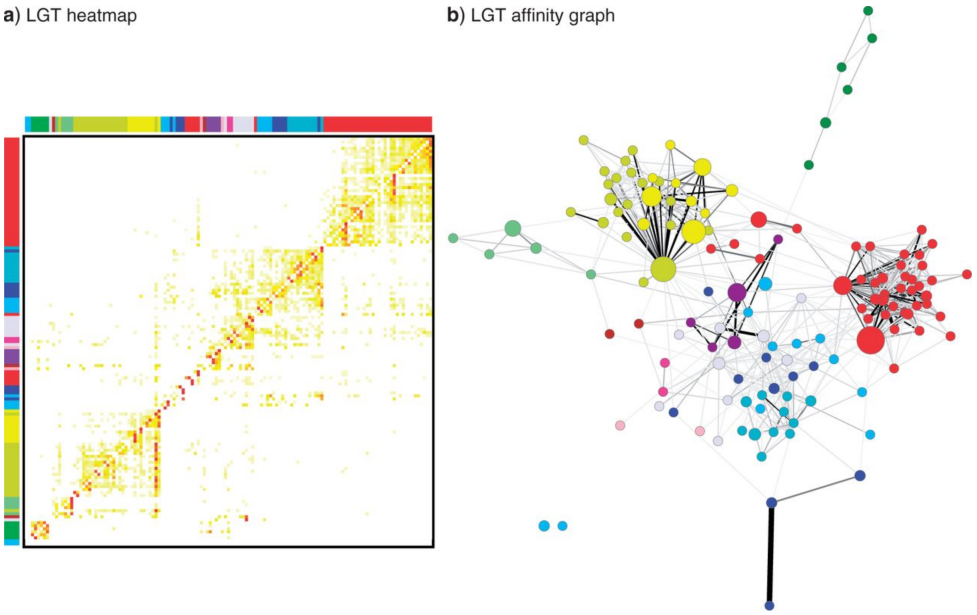
We analyzed the shared-genes edges to understand the similarities among taxa. In Figure 55 we can see that the supertree is robust in sub-figures (A) and (C) that represent shared edges with 503 to 4,327 (maximum) common genes and jaccard distance less or equal to 0.5. That is, it is expected that taxa in the same class have higher rates of shared genes. In sub-figure (B) we see edges with 211 or more shared genes and in (D) edges that connects units with jaccard distance greater or equal to 0.99. By (B) we can check what are the first taxa to connect from different classes when reducing the minimum value for shared genes filter. By (D) we understand that there are a great number of taxa that shares very few genes. In fact, the network (D) did

Figure 52 – Overview of shared-genes edges and SPR moves. (A) All possible edges being shown (more than 100 thousand). (B) All SPR edges. (C) SPR edges between different classes. (D) Genome-boxes were highlithed to show gene the class variation.



Source: Elaborated by the author.

Figure 53 – Two representations of SPR moves proposed by Whidden, Zeh and Beiko (2014). (a) An LGT heat map. (b) Each node represents a cluster of taxa scaled according to the number of genomes being represented. Edges represent the SPR moves and their thickness scales relative to the number of moves, the thicker the greater is the number (2 to 370). The graph is shown with a spring-loaded layout.



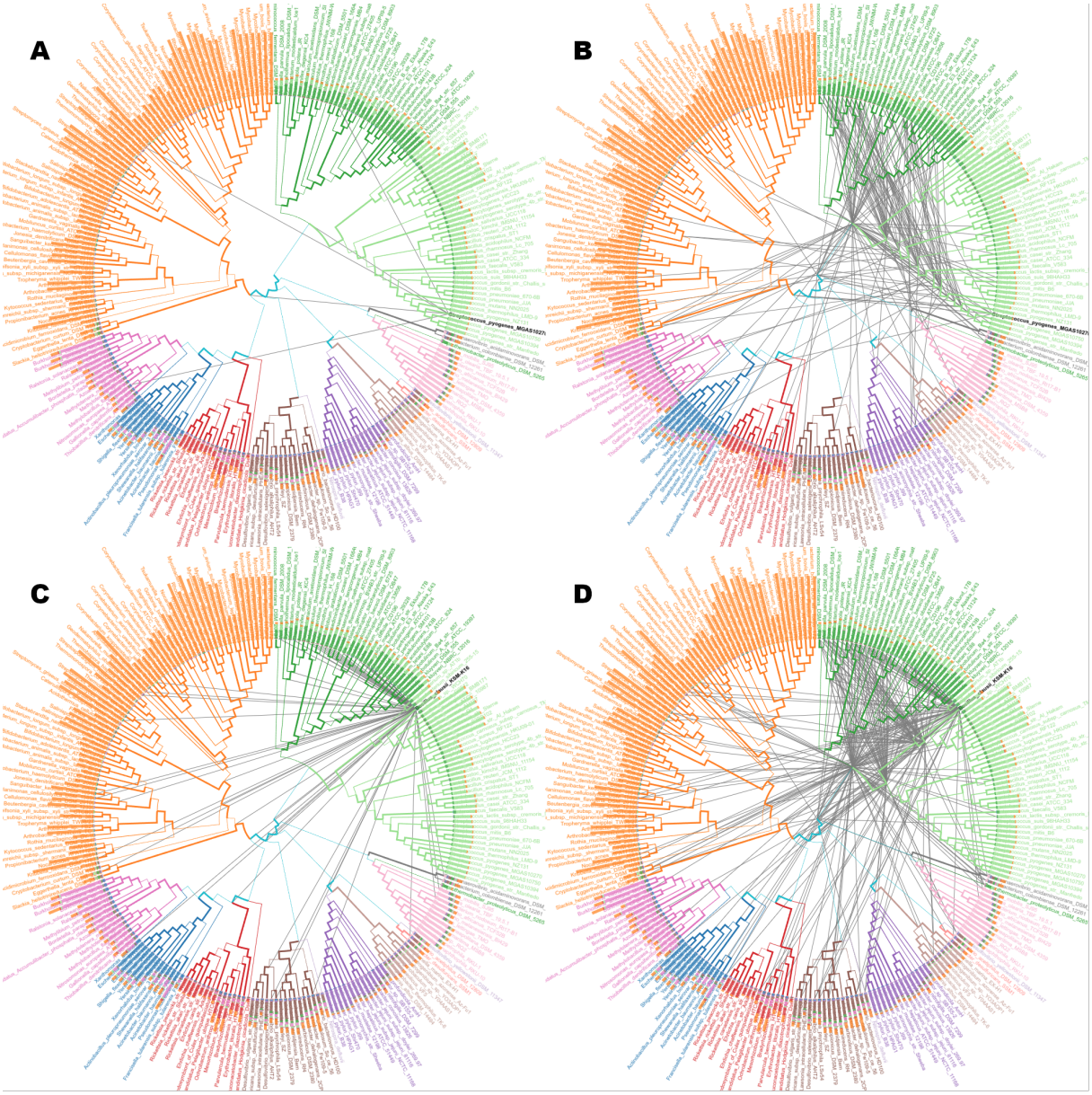
Source: adapted figure from (WHIDDEN; ZEH; BEIKO, 2014), licensed under CC BY 3.0.

not change when reducing the shared-genes filter, until it got around 94 genes. That is, the 0.99 jaccard distance is related to the intersection sets of size around 0 to 94.

If taxa positioned far in the supertree share a few genes, these could be interesting genes. Still, it is difficult to analyze this case, since the genes could come from the oldest ancestor, thus, not caused by LGT events. In fact, when filtering the edges of Figure 55 (D) to show only SPR moves, only a few last. This is shown in Figure 56, in which we highlighted one taxon as an example of possible LGT among species very distant in the supertree. Considering the filter described in the figure, there are only 2 edges that connects *Mycobacterium smegmatis*, one that happens in 3 gene trees and one that happens in 1 gene tree.

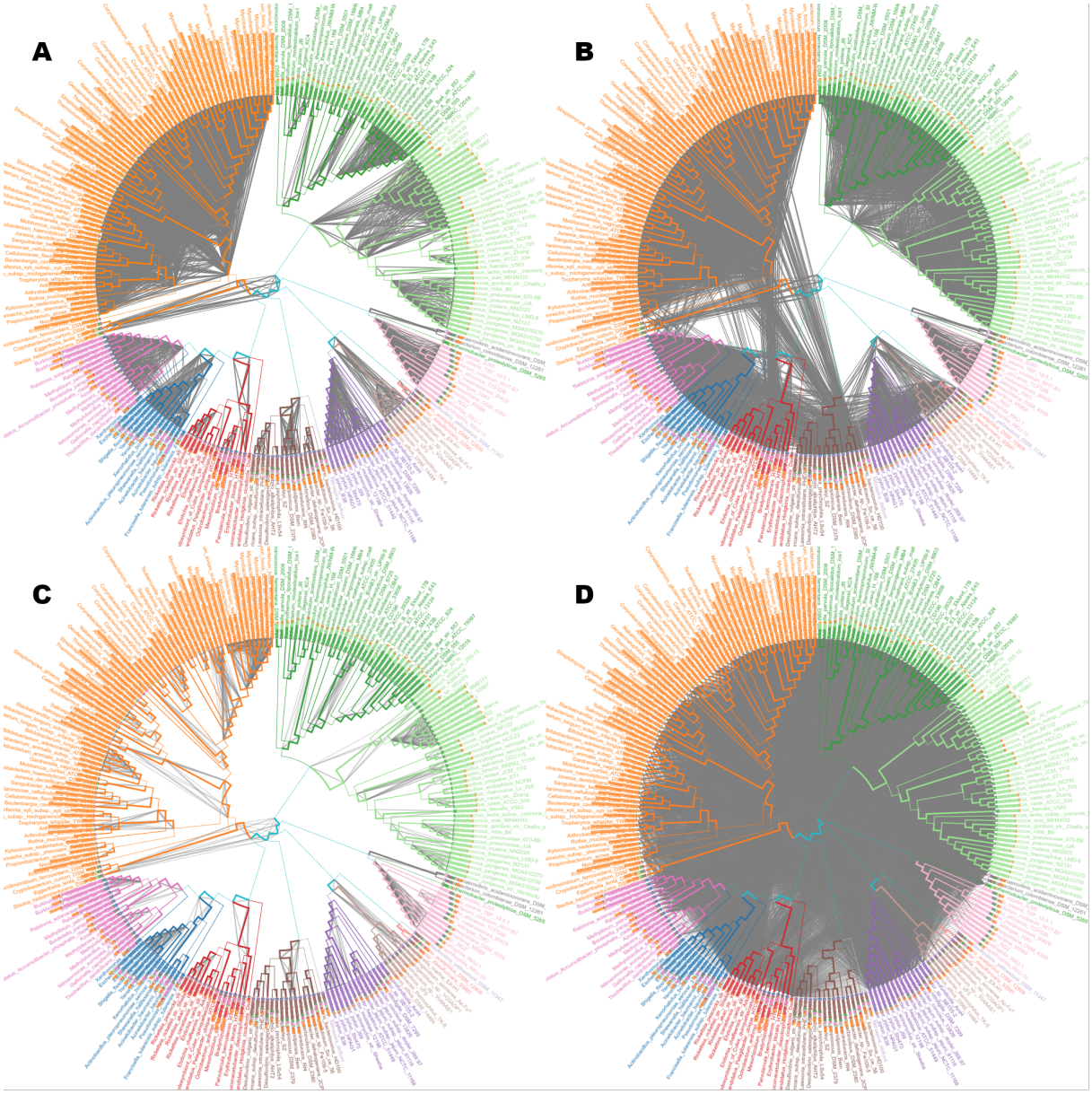
With this use case we exemplified the original purpose and use of sTVis, that is, the analysis of bacteria phylogenetic trees. Bacteria have a high rate of LGT events and, thus, the SPR Supertree is a good approach to create the evolutionary tree from taxa. The approach developed here may also be used in different contexts, for instance, where the lateral events do not happen that frequently. This is what we did to create the use case 2, where we compare Neighbor Joining trees and exemplify the use of gene annotations on sTVis.

Figure 54 – Comparison of SPR moves of different genomes. (A, B) Possible LGTs from or to the bacterium *Streptococcus pyogenes*. (C,D) Possible LGTs from or to the bacterium *Bacillus clausii*. (A, C) Only edges that connect the bacterium directly to other region of the supertree are shown. (B, D) All edges that connect the bacterium or its ancestor to other region of the supertree are shown.



Source: Elaborated by the author.

Figure 55 – Shared genes between units. (A) Edges with 503 to 4,327 shared genes (maximum). (B) Edges with 211 to 4,327 shared genes. (C) Edges with jaccard distance less or equal to 0.5. (D) Edges with jaccard distance greater or equal to 0.99. The ancestor's genes is defined by the intersection of descendents' genes.



Source: Elaborated by the author.

Figure 56 – SPR edges between units with extremely low jaccard similarity. Edges have jaccard distance greater or equal to 0.99. *Mycobacterium smegmatis* was clicked to highlight its edges. The SPR edge that connects with *Clostridium cellulovorans* contains 3 genes. The one that connects with *Geobacillus sp. WCH70* contains 1 gene.

Source: Elaborated by the author.

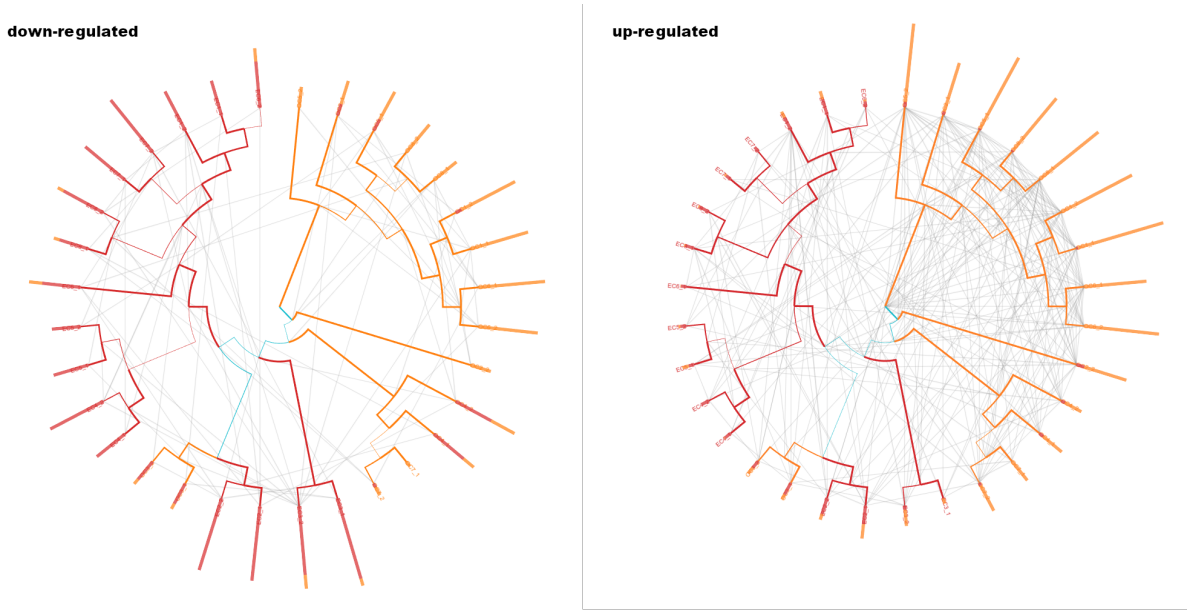
5.4.3 Use case 2

In Chapter 3 we analyzed organ-confined (OC) and the extracapsullar (EC) prostate cancer cells. The two classes comprises 16 samples each, with 624 quantified proteins. Here we revisit the data set and analyze the protein expression and sample similarities under a different perspective, specifically to illustrate the use of sTVis on a different data set and the use of annotations to guide the analysis.

We translated the data set to the supertree problem by defining a way that the analysis would focus on up- and down-regulated patterns. To create the reference tree, we normalized the proteins with mean 0 and standard deviation 1 and transformed all values between [-0.5, 0.5] to 0.0. In this way, only information on up- and down-expression are considered to create the tree. We create two classes of analysis: one focusing on down-regulated values and the other on up-regulated values. For the first group (down), we set all values > -0.5 to NaN (not a number); for the second, value < 0.5 are set to NaN. Doing so, if a sample has NaN value for a protein, it means that this sample will not be part of this tree. The remaining samples for this protein are used to create a NJ tree, the protein tree. This is done for each protein in both up and down cases, resulting in protein trees of different sizes and shapes. For each protein, its biological annotations were carried using the IIS (CARAZZOLLE *et al.*, 2014) web-system.

In this use case it is difficult to establish what an SPR move exactly mean. Still, we can see a great difference of SPR patterns between the up- and the down-regulated versions in Figure 57.

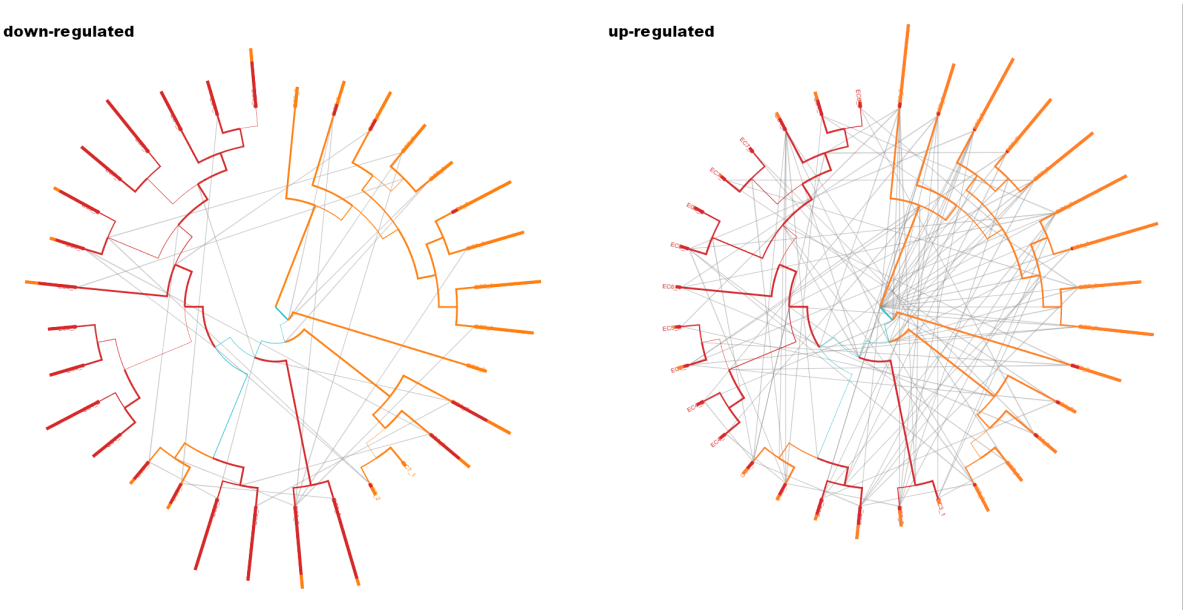
Figure 57 – SPR edges from down- and up-regulated cases.



Source: Elaborated by the author.

The genome-boxes (Figure 58, on the other hand, are very informative and reveals why two of OC (orange) samples are among the EC (red) samples in the reference tree. For both down- and up-regulated cases, their genome-boxes are much more red than the boxes of the other OC samples. Also, while there are many down-regulated proteins in both classes, we see that the up-regulated proteins are concentrated in OC samples (orange). In this figure, only SPR moves between different classes are shown.

Figure 58 – SPR edges between different classes from down- and up-regulated cases. Genome-boxes are highlighted and samples labels omitted.

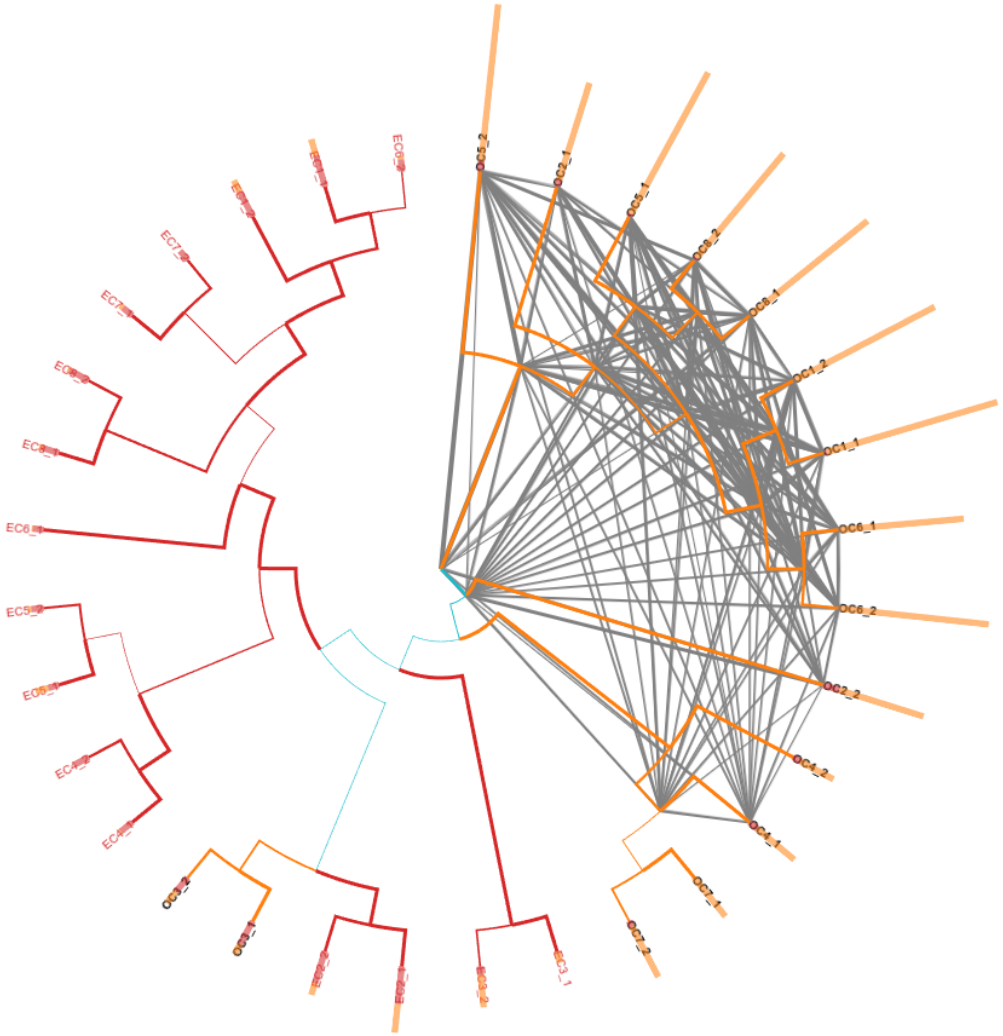


Source: Elaborated by the author.

Based on the filtering configuration of Figure 58, the system reveal differences in the protein annotations of the edges. In the down-regulated case, the top-5 annotations and the number of occurrences in proteins are: retina homeostasis (4), complement activation (3), innate immune response (3), immune response (3) and signal transduction (3). For the up-regulated they are: small molecule metabolic process (9), blood coagulation (8), platelet activation (8), platelet degranulation (8) and innate immune response (6).

By exploring the Biological Process attribute, the user could click in a value or search for it to highlight edges which genes contain that information. This is the case shown in Figure 59, where we highlighted edges that represent shared proteins between units that could participate in the *viral process*. Interestingly, all OC samples have up-regulated proteins that could express this condition, except the OC samples wrongly positioned in the reference tree (replicates from the same patient OC3). The results so far found in this use case could raise a doubt about an error in classifying the two samples as OC when they could be actually EC - or just an outlier.

Figure 59 – Shared proteins edges from the up-regulated case. Only edges with proteins that are known as agents in the *viral process* are shown. All OC samples have such proteins, except the OC samples that were positioned in the wrong group in the reference tree.



Source: Elaborated by the author.

5.5 Closing remarks

It is clear that analyzing bacteria phylogenetic networks is a complex task that requires visual techniques. In this chapter, we described a prototype that is being developed in collaboration with Dr. Robert Beiko and that is going to be applied to analyze real problems involving bacteria genomes.

The use of WebGL technology is crucial for the visualization techniques due to the great number of edges and the need for the user being able to explore the data. Despite the great number of edges can fill all the supertree background in some cases, this scenario could be improved after filtering. When we have a medium concentration of edges, edge bundling techniques could be used to improve the overview of the phylogenetic network.

Other visualization techniques could be integrated and coordinated to the supertree view, such as diagrams to analyze frequency of annotations, gene trees and other information associated to the supertree. The heat map and network presented in Figure 53 could represent alternative views to analyze LGT transfers. For instance, a multilevel network summarizing the classes and internal clusters could be used coordinated with the supertree view.

Here, we demonstrated how our approach can improve the analysis of LGT in phylogenetic supertrees and how we could translate this problem to others from different fields. We believe sTVis will allow us to find interesting patterns in the study on Salmonella and AMR and will allow understanding LGT mechanisms in many other researches.

CHAPTER

6

CONCLUSION

During this doctorate, we have handle a diversity of different problems in computational biology and shared knowledge with different experts. The challenges started right in the communication, effect of the different vocabularies employed in different disciplines. Understanding the problems faced by experts and what could be done in the field of Computer Science and its technologies to help is a complex and abstract but satisfying task.

We started back during my Masters interacting with researchers from LNBio, Embrapa and UNICAMP. InteractiVenn is a great example of a tool that may look too simple for a computer scientist point of view, but it is a powerful tool for biologists. In the course of this doctorate the tool was published and the corresponding paper is currently highly cited due to having captured a need in everyday tasks of biologists.

The problems of **biomarker discovery** are of great complexity. The simpler and smaller the intensities matrices are, the more difficult it is to have confidence in the results. We started from the problem of understanding how proteins interact, what are their function, how they are quantified and what the limitations of the quantification tools are, to simply understanding that it is a problem without an ideal solution but of extreme importance to humankind. In our first work with biomarkers, in collaboration with other researchers, we have designed and shown the use of double cross validation for prioritization of proteins from discovery to targeted proteomics.

We have extended our analysis to the evaluation of multivariate methods and suggested an approach for **prioritization of proteins** based on stability of proteins and sampling methods (randomization). We have shown how targeted proteomics differs from discovery proteomics, and how Machine Learning tools may help to find candidates for oral cancer biomarkers. The double cross validation scheme has shown to be efficient to show how much we can trust of the results of proteomics analysis. In addition, the many ways of calculating stability of proteins by cross validation demonstrated it to be a great tool to improve the understanding of the average power of proteins to discriminate the available biological samples for each condition (class).

As important as it is to identify interesting proteins, it is to understand how they work inside a cell. By employing simple and easily accessible visualization tools, biologists can perform the analyzes that permeate the understanding of how proteins interact. With that in mind, we designed visualization techniques for **networks visualization** of protein-protein interaction, which take into consideration each protein's main cellular compartment. Through an interactive interface, scientists can explore the relations between biological functions and pathways using the CellNetVis web-tool.

Finally, we have explored the area of bacterial phylogenetic trees and lateral gene transfers, in the context of high rates of spread of antibiotic resistant genes. We reported the tools used to create and track possible horizontal evolutionary events, i.e. the supertrees. By implementing a fast solution for interactively exploring the edges of a **phylogenetic network**, the sTVis, we have demonstrated how visualization may help with approaching this critical problem.

Future researches is necessary in all fields we have dealt with. Phylogenetic and protein-protein networks may encode a huge amount of information and requires fast and flexible techniques for **visual exploration**. While in this case we have successfully implemented sTVis using WebGL, the same technology could improve CellNetVis, making it faster and allowing, for instance, more precise calculations to constrain the network inside the cellular components. In the case of sTVis, many interactivity functionalities need to be implemented. It is clear that due to the complexity of the data sets, the tool requires coordinated-views and better visual techniques for edge visualization, overview and exploration.

On the **biomarkers** research, while the quantification technology and other limitations do not allow us to have data sets with more samples, we need to continue research that allows specialists to make a better decision related to choosing what proteins will be studied in further experiments. Many specialists on Biological, Medical- and Health-related areas are not equipped with enough Statistical, Mathematical and Computational skills to handle the next generation of tools by themselves. The best solution is a multidisciplinary team of experts working together at every stage of the process. For instance, we have presented one approach that can show specialists how results can vary with smalls changes in the data set, by displaying scores' distributions and multiple ranks of proteins. Methods that use a priori biological information should be compared with the ones we have presented here. After setting scripts for the double cross validation with all methods, an extensive study comprising a great number of data sets and the distinct variables prioritization methods could also be designed, being of great relevance to differentiate groups of methods.

BIBLIOGRAPHY

ALMANGUSH, A.; COLETTA, R. D.; BELLO, I. O.; BITU, C.; KESKI-SÄNTTI, H.; MÄKINEN, L. K.; KAUPPILA, J. H.; PUKKILA, M.; HAGSTRÖM, J.; LARANNE, J.; TOMMOLA, S.; SOINI, Y.; KOSMA, V. M.; KOIVUNEN, P.; KOWALSKI, L. P.; NIEMINEN, P.; GRÉNMAN, R.; LEIVO, I.; SALO, T. A simple novel prognostic model for early stage oral tongue cancer. **International Journal of Oral and Maxillofacial Surgery**, v. 44, n. 2, p. 143–150, 2015. ISSN 13990020. Citations on pages 40 and 41.

ALTARAWNEH, R.; SCHULTZ, J.; HUMAYOUN, S. R. CluE: An algorithm for expanding clustered graphs. In: **IEEE Pacific Visualization Symposium**. IEEE, 2014. p. 233–237. ISBN 9781479928736. ISSN 21658773. Available: <http://ieeexplore.ieee.org/document/6787172/>. Citation on page 106.

ANDERSON, N. G. L.; ANDERSON, N. G. L. Proteome and proteomics: New technologies, new concepts, and new words. **Electrophoresis**, v. 19, n. 11, p. 1853–1861, 1998. ISSN 01730835. Available: <http://www.ncbi.nlm.nih.gov/pubmed/9740045>. Citation on page 25.

ARCHAMBAULT, D.; MUNZNER, T.; AUBER, D. Tugging graphs faster: Efficiently modifying path-preserving hierarchies for browsing paths. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 3, p. 276–289, 2011. ISSN 1077-2626. Citation on page 106.

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene ontology: Tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25–29, 5 2000. ISSN 10614036. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract>. Citations on pages 56 and 107.

AUBER, D. **Tulip - A Huge Graph Visualization Framework**. Springer Verlag, 2004. 105–126 p. Available: <http://link.springer.com/10.1007/978-3-642-18638-7_5>. Citation on page 106.

AVET-LOISEAU, H.; FACON, T.; GROSBOIS, B.; MAGRANGEAS, F.; RAPP, M. J.; HAROUSSEAU, J. L.; MINVIELLE, S.; BATAILLE, R. Oncogenesis of multiple myeloma: 14q32 and 13q chromosomal abnormalities are not randomly distributed, but correlate with natural history, immunological features, and clinical presentation. **Blood**, v. 99, n. 6, p. 2185–2191, 3 2002. ISSN 00064971. Available: <http://www.bloodjournal.org/cgi/doi/10.1182/blood.V99.6.2185>. Citation on page 100.

BANSAL, M. S.; BURLEIGH, J. G.; EULENSTEIN, O.; FERNÁNDEZ-BACA, D. Robinson-Foulds supertrees. **Algorithms for Molecular Biology**, v. 5, n. 1, p. 18, 2010. ISSN 17487188. Available: <http://almob.biomedcentral.com/articles/10.1186/1748-7188-5-18>. Citation on page 127.

BARKER, P. E. Cancer biomarker validation: Standards and process - Roles for the National Institute of Standards and Technology (NIST). **Annals of the New York Academy of Sciences**, v. 983, n. 1, p. 142–150, 3 2003. ISSN 00778923. Available: <http://doi.wiley.com/10.1111/j.1749-6632.2003.tb05969.x>. Citation on page 52.

BARSKY, A.; GARDY, J. L.; HANCOCK, R. E. W.; MUNZNER, T. Cerebral: A Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. **Bioinformatics**, v. 23, n. 8, p. 1040–1042, 4 2007. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm057>. Citations on pages 106, 107, 116, and 120.

BARTER, R. L.; SCHRAMM, S.-J. J.; MANN, G. J.; YANG, Y. H. Network-based biomarkers enhance classical approaches to prognostic gene expression signatures. **BMC Systems Biology**, BioMed Central Ltd, v. 8, n. 4, p. S5, 2014. ISSN 17520509. Available: <http://www.biomedcentral.com/1752-0509/8/S4/S5>. Citation on page 51.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. <Gephi_Bastian-Jacomy_AAAI_2009.pdf>. In: **ICWSM**. [s.n.], 2009. ISBN 978-1-57735-421-5. ISSN 14753898. Available: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>. Citation on page 106.

BATAGELJ, a. M. V. Pajek - Program for Large Network Analysis. **Connections**, p. 1–11, 1999. ISSN 0226-1776. Available: <papers2://publication/uuid/09C6E0F6-77DF-4BE8-B808-E42191C5C44A>. Citation on page 106.

BAUR, M.; BRANDES, U. Multi-circular layout of micro/macro graphs. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Springer Berlin Heidelberg, Berlin, Heidelberg, v. 4875 LNCS, p. 255–267, 2008. ISSN 03029743. Available: <http://link.springer.com/10.1007/978-3-540-77537-9_26>. Citation on page 106.

BECK, M.; SCHMIDT, A.; MALMSTROEM, J.; CLAASSEN, M.; ORI, A.; SZYMBORSKA, A.; HERZOG, F.; RINNER, O.; ELLENBERG, J.; AEBERSOLD, R. The quantitative proteome of a human cell line. **Molecular Systems Biology**, Nature Publishing Group, v. 7, n. 1, p. 549, 4 2014. ISSN 1744-4292. Available: <http://dx.doi.org/10.1038/msb.2011.82http://msb.embopress.org/cgi/doi/10.1038/msb.2011.82>. Citation on page 25.

BEIKO, R. G.; HARLOW, T. J.; RAGAN, M. A. Highways of gene sharing in prokaryotes. **Proceedings of the National Academy of Sciences**, v. 102, n. 40, p. 14332–14337, 2005. ISSN 0027-8424. Available: <http://www.pnas.org/cgi/doi/10.1073/pnas.0504068102>. Citation on page 126.

Biblioteca Achille Bassi. Exposição destaca artigos científicos do ICMC mais citados. 2018. Available: <https://www.icmc.usp.br/noticias/3522-exposicao-destaca-artigos-cientificos-do-icmc-mais-citados>. Citation on page 38.

____. **Exposição destaca artigos científicos do ICMC mais citados**. 2018. Available: <https://www.icmc.usp.br/noticias/3522-exposicao-destaca-artigos-cientificos-do-icmc-mais-citados>. Citation on page 38.

BINDER, H.; SCHUMACHER, M. Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. **BMC Bioinformatics**, v. 10, n. 1, p. 18, 2009. ISSN 14712105. Available: <http://www.biomedcentral.com/1471-2105/10/18>. Citation on page 59.

BLACKSTOCK, W. P.; WEIR, M. P. Proteomics: Quantitative and physical mapping of cellular proteins. **Trends in Biotechnology**, v. 17, n. 3, p. 121–127, 1999. ISSN 01677799. Available: <http://www.ncbi.nlm.nih.gov/pubmed/10189717>. Citation on page 25.

BOCK, N.; ALTER, H.; KOC, E.; ROESSNER, V.; ROTHENBERGER, A.; MANZKE, T. Chronic Fluoxetine Administration during Different Postnatal Development Stages Leads to Stage Dependent Changes of Glial Fibrillary Acidic Protein Expression in Rat Brain. v. 2, n. 3, p. 292–312, 2012. Citation on page 100.

BOGDANOWICZ, D.; GIARO, K. Comparing Phylogenetic Trees by Matching Nodes Using the Transfer Distance between Partitions. **Journal of Computational Biology**, v. 24, n. 5, p. 422–435, 5 2017. ISSN 1066-5277. Available: <http://www.liebertpub.com/doi/10.1089/cmb.2016.0204>. Citation on page 127.

BOSTOCK, M.; OGIEVETSKY, V.; HEER, J. D3data-driven documents. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 12, p. 2301–2309, 12 2011. ISSN 10772626. Available: <http://www.ncbi.nlm.nih.gov/pubmed/22034350>. Citation on page 109.

BOULESTEIX, A. L.; STRIMMER, K. Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. **Briefings in Bioinformatics**, v. 8, n. 1, p. 32–44, 1 2007. ISSN 14675463. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16772269>. Citation on page 29.

BOULESTEIX, A. L. A.-L. WilcoxCV: an R package for fast variable selection in cross-validation. **Bioinformatics**, v. 23, n. 13, p. 1702–1704, 7 2007. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm162>. Citation on page 54.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1996. ISSN 0885-6125. Available: <http://link.springer.com/10.1007/BF00058655>. Citation on page 69.

BREUER, K.; FOROUSHANI, A. K.; LAIRD, M. R.; CHEN, C.; SRIBNAIA, A.; LO, R.; WINSOR, G. L.; HANCOCK, R. E. W.; BRINKMAN, F. S. L.; LYNN, D. J. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. v. 41, n. D1, p. D1228. Citation on page 108.

BURKE, H. B. Discovering patterns in microarray data. **Molecular Diagnosis**, v. 5, n. 4, p. 349–357, 12 2000. ISSN 10848592. Available: <http://linkinghub.elsevier.com/retrieve/doi/10.1054/modi.2000.19562>. Citation on page 52.

CARAZZOLLE, M. F.; CARVALHO, L. M. D.; SLEPICKA, H. H.; VIDAL, R. O.; PEREIRA, G. A. G.; KOBARG, J.; MEIRELLES, G. V. IIS - Integrated Interactome System: A web-based platform for the annotation, analysis and visualization of protein-metabolite-gene-drug interactions by integrating a variety of data sources and tools. **PLoS ONE**, v. 9, n. 6, p. e100385, 1 2014. ISSN 19326203. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4065059&tool=pmcentrez&rendertype=abstract>. Citations on pages 108 and 139.

CARNIELLI, C. M.; MACEDO, C. C. S.; ROSSI, T. D.; GRANATO, D. C.; RIVERA, C.; DOMINGUES, R. R.; PAULETTI, B. A.; YOKOO, S.; HEBERLE, H.; BUSSO-LOPES, A. F.; CERVIGNE, N. K.; SAWAZAKI-CALONE, I.; MEIRELLES, G. V.; MARCHI, F. A.; TELLES, G. P.; MINGHIM, R.; RIBEIRO, A. C. P.; BRANDÃO, T. B.; CASTRO, G. de; GONZÁLEZ-ARRIAGADA, W. A.; GOMES, A.; PENTEADO, F.; SANTOS-SILVA, A. R.; LOPES, M. A.; RODRIGUES, P. C.; SUNDQUIST, E.; SALO, T.; SILVA, S. D. da; ALAOUI-JAMALI, M. A.;

GRANER, E.; FOX, J. W.; COLETTA, R. D.; LEME, A. F. P. Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. **Nature Communications**, Springer US, v. 9, n. 1, p. 3598, 12 2018. ISSN 20411723. Available: <http://dx.doi.org/10.1038/s41467-018-05696-2http://www.nature.com/articles/s41467-018-05696-2>. Citations on pages 20, 23, 40, 42, 43, 44, 45, 46, 47, and 48.

CERAMI, E.; GAO, J.; DOGRUSOZ, U.; GROSS, B. E.; SUMER, S. O.; AKSOY, B. A.; JACOBSEN, A.; BYRNE, C. J.; HEUER, M. L.; LARSSON, E.; ANTIPIN, Y.; REVA, B.; GOLDBERG, A. P.; SANDER, C.; SCHULTZ, N. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. **Cancer Discovery**, v. 2, n. 5, p. 401–404, 5 2012. ISSN 21598274. Available: <http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-12-0095>. Citation on page 100.

CHEN, P. Increased Expression of Tissue/Salivary Transgelin mRNA Predicts Poor Prognosis in Patients with Oral Squamous Cell Carcinoma (OSCC). **Medical Science Monitor**, v. 21, p. 2275–2281, 2015. ISSN 1643-3750. Available: <http://www.medscimonit.com/abstract/index/idArt/893925>. Citation on page 41.

CHEN, Y.; LIAO, L.; QIU, J.; ZHENG, Y.; TANG, S.; LI, R.; XIE, D.; ZOU, X.; ZHENG, S.; WENG, X. T.; ZHANG, W. Research on the inhibitory effect of metformin on human oral squamous cell carcinoma SCC-4 and CAL-27 cells and the relevant molecular mechanism. **Biomedical Research (India)**, Elsevier Ltd, v. 28, n. 14, p. 6350–6354, 4 2017. ISSN 0970938X. Available: <http://dx.doi.org/10.1016/j.oraloncology.2009.01.004http://linkinghub.elsevier.com/retrieve/pii/S1368837509000050>. Citation on page 40.

CHRISTIN, C.; HOEFSLOOT, H. C. J.; SMILDE, A. K.; HOEKMAN, B.; SUITS, F.; BISCHOFF, R.; HORVATOVICH, P. A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics. **Molecular & Cellular Proteomics**, v. 12, n. 1, p. 263–276, 1 2013. ISSN 1535-9476. Available: <http://www.mcponline.org/lookup/doi/10.1074/mcp.M112.022566http://www.mcponline.org/cgi/doi/10.1074/mcp.M112.022566>. Citations on pages 30, 34, 56, and 67.

CUN, Y. Network-Based Biomarker Discovery. p. 131, 2014. Available: <http://hss.ulb.uni-bonn.de/2014/3563/3563.pdf>. Citation on page 51.

CUN, Y.; FRÖHLICH, H. Biomarker Gene Signature Discovery Integrating Network Knowledge. **Biology**, v. 1, n. 3, p. 5–17, 2 2012. ISSN 2079-7737. Available: <http://www.mdpi.com/2079-7737/1/1/5/>. Citation on page 51.

CUN, Y.; FRÖHLICH, H. F. Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. **BMC Bioinformatics**, v. 13, n. 1, p. 69, 2012. ISSN 14712105. Available: <http://www.biomedcentral.com/1471-2105/13/69>. Citations on pages 52 and 59.

CURRY, J. M.; SPRANDIO, J.; COGNETTI, D.; LUGINBUHL, A.; BAR-AD, V.; PRIBITKIN, E.; TULUC, M. Tumor microenvironment in head and neck squamous cell carcinoma. **Seminars in Oncology**, Elsevier, v. 41, n. 2, p. 217–234, 2014. ISSN 15328708. Available: <http://dx.doi.org/10.1053/j.seminoncol.2014.03.003>. Citation on page 41.

D3. **Force Layout**. 2015. Available: <https://github.com/d3/d3/wiki/Force-Layout>. Citation on page 109.

DANG, T.; MURRAY, P.; AURISANO, J.; FORBES, A. ReactionFlow: an interactive visualization tool for causality analysis in biological pathways. **BMC Proceedings**, v. 9, n. Suppl 6, p. S6, 2015. ISSN 1753-6561. Available: <http://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-9-S6-S6>. Citation on page 106.

DAVIES, T. J.; BARRACLOUGH, T. G.; CHASE, M. W.; SOLTIS, P. S.; SOLTIS, D. E.; SAVOLAINEN, V. Darwin's abominable mystery: Insights from a supertree of the angiosperms. **Proceedings of the National Academy of Sciences**, v. 101, n. 7, p. 1904–1909, 2 2004. ISSN 0027-8424. Available: <http://www.pnas.org/lookup/doi/10.1073/pnas.0308127100>. Citation on page 126.

DENG, L.; MA, J.; PEI, J. Rank sum method for related gene selection and its application to tumor diagnosis. **Chinese Science Bulletin**, v. 49, n. 15, p. 1652–1657, 8 2004. ISSN 1001-6538. Available: <http://link.springer.com/10.1007/BF03184138>. Citations on pages 29 and 54.

DOGRUSOZ, U.; GIRAL, E.; CETINTAS, A.; CIVRIL, A.; DEMIR, E. A layout algorithm for undirected compound graphs. **Information Sciences**, Elsevier Inc., v. 179, n. 7, p. 980–994, 2009. ISSN 00200255. Available: <http://dx.doi.org/10.1016/j.ins.2008.11.017>. Citation on page 106.

DOMINGUES, R. R. **Proteômica baseada em Descoberta para Busca de Alvos Terapêuticos e Biomarcadores Potenciais utilizando-se Análises Univariadas e Multivariadas**. Phd Thesis (PhD Thesis) — UNICAMP, 2017. Citations on pages 49, 51, 60, 61, and 64.

DONIGER, S.; SALOMONIS, N.; DAHLQUIST, K.; VRANIZAN, K.; LAWLOR, S.; CONKLIN, B. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. **Genome biology**, v. 4, n. 1, p. 1–12, 2003. ISSN 1465-6914. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=151291&tool=pmcentrez&rendertype=abstract>. Citation on page 56.

DUTKOWSKI, J.; IDEKER, T. Protein networks as logic functions in development and cancer. **PLoS Computational Biology**, v. 7, n. 9, p. e1002180, 9 2011. ISSN 1553734X. Available: <http://dx.plos.org/10.1371/journal.pcbi.1002180>. Citations on pages 51, 57, and 58.

DWYER, T. Scalable, Versatile and Simple Constrained Graph Layout. **Computer Graphics Forum**, Eurographics Association, Berlin, Germany, v. 28, n. 3, p. 991–998, 6 2009. ISSN 14678659. Available: <http://doi.wiley.com/10.1111/j.1467-8659.2009.01449.xhttp://diglib.eg.org/EG/CGF/volume28/issue3/v28i3pp0991-0998.pdf>. Citation on page 106.

DWYER, T.; ROBERTSON, G. Layout with circular and other non-linear constraints using Procrustes projection. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [s.n.], 2010. v. 5849 LNCS, p. 393–404. ISBN 3642118046. Available: <http://link.springer.com/10.1007/978-3-642-11805-0_37>. Citation on page 106.

EDGE, S. B.; COMPTON, C. C. The american joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. **Annals of Surgical Oncology**, v. 17, n. 6, p. 1471–1474, 6 2010. ISSN 10689265. Available: <http://www.springerlink.com/index/10.1245/s10434-010-0985-4>. Citation on page 40.

EFRON, B.; TIBSHIRANI, R.; STOREY, J. D.; TUSHER, V. Empirical bayes analysis of a microarray experiment. **Journal of the American Statistical Association**, v. 96, n. 456, p.

1151–1160, 12 2001. ISSN 1537274X. Available: <http://www.tandfonline.com/doi/abs/10.1198/016214501753382129>. Citation on page 54.

ELLSON, J.; GANSNER, E.; KOUTSOFIOS, L.; NORTH, S. C.; WOODHULL, G. Graphviz - Open Source Graph Drawing Tools. In: MUTZEL, P.; JÜNGER, M.; LEIPERT, S. (Ed.). **Graph Drawing: 9th International Symposium, GD 2001 Vienna, Austria, September 23–26, 2001 Revised Papers**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 483–484. ISBN 3540433090. Available: <http://link.springer.com/10.1007/3-540-45848-4_57>. Citation on page 106.

FIORINI, N.; LEFORT, V.; CHEVENET, F.; BERRY, V.; CHIFOLLEAU, A. M. A. CompPhy: A web-based collaborative platform for comparing phylogenies. **BMC Evolutionary Biology**, v. 14, n. 1, p. 253, 12 2014. ISSN 14712148. Available: <http://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-014-0253-5>. Citation on page 129.

FRIAS, S.; BRYAN, K.; BRINKMAN, F. S.; LYNN, D. J. CerebralWeb: A cytoscape.js plug-in to visualize networks stratified by subcellular localization. **Database**, v. 2015, p. 1–4, 1 2015. ISSN 17580463. Available: <https://academic.oup.com/database/article/doi/10.1093/database/bav041/2433173>. Citations on pages 106, 107, and 116.

GALTIER, N.; DAUBIN, V. Dealing with incongruence in phylogenomic analyses. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 363, n. 1512, p. 4023–4029, 12 2008. ISSN 09628436. Available: <http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2008.0144>. Citation on page 125.

GANLY, I.; PATEL, S.; SHAH, J. Early stage squamous cell cancer of the oral tongue-clinicopathologic features affecting outcome. **Cancer**, v. 118, n. 1, p. 101–111, 1 2012. ISSN 0008543X. Available: <http://doi.wiley.com/10.1002/cncr.26229>. Citation on page 40.

GAO, C.; DANG, X.; CHEN, Y.; WILKINS, D. Graph ranking for exploratory gene data analysis. **BMC Bioinformatics**, v. 10 Suppl 1, n. SUPPL. 11, p. S19, 2009. ISSN 1471-2105. Available: <http://www.biomedcentral.com/1471-2105/10/S11/S19>. Citation on page 59.

GARCIA, O.; SAVEANU, C.; CLINE, M.; FROMONT-RACINE, M.; JACQUIER, A.; SCHWIKOWSKI, B.; AITTOKALLIO, T. GOlorize: A Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring. **Bioinformatics**, v. 23, n. 3, p. 394–396, 2 2007. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btl605>. Citation on page 106.

GERMERAAD, J. H.; HOPPING, C. A.; MULLER, J. Palynology of tertiary sediments from tropical areas. **Review of Palaeobotany and Palynology**, v. 6, n. 3-4, p. 42–6, 1 1968. ISSN 00346667. Available: <http://www.ncbi.nlm.nih.gov/pubmed/11752249>. Citation on page 114.

GLINSKY, G. V.; GLINSKII, A. B.; STEPHENSON, A. J.; HOFFMAN, R. M.; GERALD, W. L. Gene expression profiling predicts clinical outcome of prostate cancer. **Journal of Clinical Investigation**, v. 113, n. 6, p. 913–923, 1 2004. ISSN 00219738. Available: <http://www.nature.com/doifinder/10.1038/415530ahttp://www.nature.com/articles/415530a>. Citations on pages 51 and 57.

Globo Comunicação e Participações S.A. **Estudo do CNPEM cria método para prever evolução do câncer de boca com uso da saliva**. 2018. Available: <https://g1.globo.com/sp/campinas-regiao/noticia/2018/09/24/

estudo-do-cnpem-cria-metodo-para-prever-evolucao-do-cancer-de-boca-com-uso-da-saliva.ghtml>. Citation on page 41.

Globo News. **Pesquisadores conseguem avaliar estágio do câncer de boca pela saliva**. 2018. Available: <http://g1.globo.com/globo-news/videos/v/pesquisadores-conseguem-avaliar-estagio-do-cancer-de-boca-pela-saliva/7039474/>. Citation on page 41.

GLOBOCAN. **Cancer Today**. 2018. 1–2 p. Available: <http://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf>. Citation on page 24.

GONZÁLEZ, J. M. M. Drogodependencias y trastornos de la personalidad: Variables relevantes para su tratamiento. **Papeles del Psicologo**, v. 32, n. 2, p. 166–174, 1 2011. ISSN 02147823. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3070428&tool=pmcentrez&rendertype=abstract>. Citation on page 108.

GOOD, B. M.; LOGUERCIO, S.; GRIFFITH, O. L.; NANIS, M.; WU, C.; SU, A. I.; AVE, F. P.; LOUIS, S. The Cure : Making a game of gene selection for breast cancer survival prediction. **arXiv preprint arXiv**, p. 1–9, 2013. ISSN 14388871. Citation on page 51.

GUILLON, G.; YEARWOOD, G.; SNIPES, C.; BOSCHI, D.; REED, M. R. Human anti-HIV IgM detection by the OraQuick ADVANCE ®Rapid HIV 1/2 Antibody Test. **PeerJ**, v. 6, p. e4430, 2 2018. ISSN 2167-8359. Available: <https://peerj.com/articles/4430>. Citations on pages 24 and 27.

GUO, J.; WANG, W.; LIAO, P.; LOU, W.; JI, Y.; ZHANG, C.; WU, J.; ZHANG, S. Identification of serum biomarkers for pancreatic adenocarcinoma by proteomic analysis. **Cancer Science**, v. 100, n. 12, p. 2292–2301, 6 2009. ISSN 13479032. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16755300>. Citation on page 53.

GUO, Z.; ZHANG, T.; LI, X.; WANG, Q.; XU, J.; YU, H.; ZHU, J.; WANG, H.; WANG, C.; TOPOL, E. J.; WANG, Q.; RAO, S. Towards precise classification of cancers based on robust gene functional expression profiles. **BMC Bioinformatics**, v. 6, n. 1, p. 58, 2005. ISSN 14712105. Available: <http://www.biomedcentral.com/1471-2105/6/58>. Citations on pages 52, 56, and 59.

HARROWER, M.; BREWER, C. A. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. **The Map Reader: Theories of Mapping Practice and Cartographic Representation**, v. 40, n. 1, p. 261–268, 6 2011. ISSN 0008-7041. Available: <http://www.maneyonline.com/doi/abs/10.1179/000870403235002042>. Citation on page 108.

HAYNES, W. Student's t-Test. In: DUBITZKY, W.; WOLKENHAUER, O.; CHO, K.-H.; YOKOTA, H. (Ed.). **Encyclopedia of Systems Biology**. New York, NY: Springer New York, 2013. chap. Student's, p. 2023–2025. ISBN 978-1-4419-9863-7. Citations on pages 29 and 53.

HEBERLE, H.; CARAZZOLLE, M. F.; TELLES, G. P.; MEIRELLES, G. V.; MINGHIM, R. CellNetVis: a web tool for visualization of biological networks using force-directed layout constrained by cellular components. **BMC Bioinformatics**, v. 18, n. S10, p. 395, 9 2017. ISSN 1471-2105. Available: <https://doi.org/10.1186/s12859-017-1787-5http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1787-5>. Citations on pages 20, 105, 113, 114, and 116.

HEBERLE, H.; MEIRELLES, V. G.; SILVA, F. R. F. da; TELLES, G. G. P.; MINGHIM, R.; MEIRELLES, G.; SILVA, F. R. F. da; TELLES, G. G. P.; MINGHIM, R. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. **BMC Bioinformatics**, v. 16, n. 1, p. 169, 12 2015. ISSN 14712105. Available: <http://www.biomedcentral.com/1471-2105/16/169http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0611-3>. Citations on pages 19, 23, 30, 35, 37, 38, 39, and 60.

HEIN, J.; JIANG, T.; WANG, L.; ZHANG, K. On the complexity of comparing evolutionary trees (Extended Abstract). In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [s.n.], 1995. v. 937, n. 96, p. 177–190. ISBN 3540600442. Available: <http://link.springer.com/10.1007/3-540-60044-2_42>. Citation on page 127.

HENRY, N.; FEKETE, J. D.; MCGUFFIN, M. J. NodeTrix: A hybrid visualization of social networks. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Computer Society, Los Alamitos, CA, USA, v. 13, n. 6, p. 1302–1309, 11 2007. ISSN 10772626. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4376154http://ieeexplore.ieee.org/document/4376154/>. Citation on page 114.

HO, T. K. The random subspace method for constructing decision forests. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 20, n. 8, p. 832–844, 1998. ISSN 01628828. Available: <http://ieeexplore.ieee.org/document/709601/>. Citations on pages 68 and 69.

HOLTEN, D.; WIJK, J. J. van. Force-Directed edge bundling for graph visualization. **Computer Graphics Forum**, v. 28, n. 3, p. 983–990, 6 2009. ISSN 01677055. Available: <http://doi.wiley.com/10.1111/j.1467-8659.2009.01450.x>. Citation on page 111.

HWANG, T.; TIAN, Z.; KUANGY, R.; KOCHER, J.-P. Learning on Weighted Hypergraphs to Integrate Protein Interactions and Gene Expressions for Cancer Outcome Prediction. Ieee, p. 293–302. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4781124>. Citation on page 51.

Instituto nacional do câncer. **Estimativa 2018: Incidência de câncer no Brasil.** [S.l.], 2017. Citation on page 24.

John H. McDonald. **Multiple comparisons - Handbook of Biological Statistics (3rd ed.)**. 2014. Available: <http://www.biostathandbook.com/multiplecomparisons.html>. Citation on page 64.

JORDAN, G. E.; PIEL, W. H. PhyloWidget: Web-based visualizations for the tree of life. **Bioinformatics**, v. 24, n. 14, p. 1641–1642, 7 2008. ISSN 13674803. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn235>. Citation on page 129.

JOYCE, A. R.; PALSSON, B. O. The model organism as a system: Integrating 'omics' data sets. **Nature Reviews Molecular Cell Biology**, v. 7, n. 3, p. 198–210, 3 2006. ISSN 14710072. Available: <http://www.nature.com/articles/nrm1857>. Citation on page 125.

KANEHISA, M.; GOTO, S.; SATO, Y.; FURUMICHI, M.; TANABE, M. KEGG for integration and interpretation of large-scale molecular data sets. **Nucleic Acids Research**, v. 40, n. D1, p. D109–D114, 1 2012. ISSN 03051048. Available: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr988>. Citation on page 56.

KAPOOR, K.; DASS, N. Melting temperature of helium to high pressures. **Indian Journal of Pure and Applied Physics**, v. 40, n. 12, p. 917–918, 10 2002. ISSN 00195596. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2063581&tool=pmcentrez&rendertype=abstracthttp://msb.embopress.org/cgi/doi/10.1038/msb4100180>. Citations on pages 51 and 57.

KARPIEVITCH, Y. V.; POLPITIYA, A. D.; ANDERSON, G. A.; SMITH, R. D.; DABNEY, A. R. Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects. **Ann Appl Stat**, v. 4, n. 4, p. 1797–1823, 2011. ISSN 1932-6157. Citation on page 26.

KAVITHA, V.; KANNAN, S. P. Exploring the mental lexicon and the lexical networks of the Indian learners of ESL at the tertiary level. **IUP Journal of English Studies**, v. 11, n. 3, p. 70–87, 1 2016. ISSN 09733728. Available: <http://www.ncbi.nlm.nih.gov/pubmed/12169552>. Citations on pages 55 and 57.

KAWAHARA, R.; BOLLINGER, J. G.; RIVERA, C.; RIBEIRO, A. C. P.; BRANDÃO, T. B.; LEME, A. F.; MACCOSS, M. J. A targeted proteomic strategy for the measurement of oral cancer candidate biomarkers in human saliva. **Proteomics**, v. 16, n. 1, p. 159–173, 2016. ISSN 16159861. Citation on page 41.

KAWAHARA, R.; MEIRELLES, G. V.; HEBERLE, H.; DOMINGUES, R. R.; GRANATO, D. C.; YOKOO, S.; CANEVAROLO, R. R.; WINCK, F. V.; RIBEIRO, A. C. P.; BRANDÃO, T. B.; FILGUEIRAS, P. R.; CRUZ, K. S. P.; BARBUTO, J. A.; POPPI, R. J.; MINGHIM, R.; TELLES, G. P.; FONSECA, F. P.; FOX, J. W.; SANTOS-SILVA, A. R.; COLETTA, R. D.; SHERMAN, N. E.; LEME, A. F. P. Integrative analysis to select cancer candidate biomarkers to targeted validation. **Oncotarget**, v. 6, n. 41, p. 43635—43652, 12 2015. ISSN 1949-2553. Available: <http://www.oncotarget.com/fulltext/6018>. Citations on pages 20, 23, 31, 32, 33, 34, 35, 36, 38, 39, and 60.

KHAN, A. U.; MARYAM, L.; ZARRILLI, R. Structure, Genetics and Worldwide Spread of New Delhi Metallo-{$$\beta$$}-lactamase (NDM): a threat to public health. **BMC Microbiology**, BMC Microbiology, v. 17, n. 1, p. 101, 12 2017. ISSN 14712180. Available: <http://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-017-1012-8>. Citation on page 126.

KIM, M. S.; PINTO, S. M.; GETNET, D.; NIRUJOGI, R. S.; MANDA, S. S.; CHAERKADY, R.; MADUGUNDU, A. K.; KELKAR, D. S.; ISSERLIN, R.; JAIN, S.; THOMAS, J. K.; MUTHUSAMY, B.; LEAL-ROJAS, P.; KUMAR, P.; SAHASRABUDDHE, N. A.; BALAKRISHNAN, L.; ADVANI, J.; GEORGE, B.; RENUSE, S.; SELVAN, L. D. N.; PATIL, A. H.; NANJAPPA, V.; RADHAKRISHNAN, A.; PRASAD, S.; SUBBANNAYYA, T.; RAJU, R.; KUMAR, M.; SREENIVASAMURTHY, S. K.; MARIMUTHU, A.; SATHE, G. J.; CHAVAN, S.; DATTA, K. K.; SUBBANNAYYA, Y.; SAHU, A.; YELAMANCHI, S. D.; JAYARAM, S.; RAJAGOPALAN, P.; SHARMA, J.; MURTHY, K. R.; SYED, N.; GOEL, R.; KHAN, A. A.; AHMAD, S.; DEY, G.; MUDGAL, K.; CHATTERJEE, A.; HUANG, T. C.; ZHONG, J.; WU, X.; SHAW, P. G.; FREED, D.; ZAHARI, M. S.; MUKHERJEE, K. K.; SHANKAR, S. S. K.; MAHADEVAN, A.; LAM, H.; MITCHELL, C. J.; SHANKAR, S. S. K.; SATISHCHANDRA, P.; SCHROEDER, J. T.; SIRDESHMUKH, R.; MAITRA, A.; LEACH, S. D.; DRAKE, C. G.; HALUSHKA, M. K.; PRASAD, T. S.; HRUBAN, R. H.; KERR, C. L.; BADER, G. D.; IACOBUZIO-DONAHUE, C. A.; GOWDA, H.; PANDEY, A. A draft map of the human proteome. **Nature**, Nature Publishing Group, v. 509, n. 7502, p. 575–581, 5 2014. ISSN 14764687. Available: <http://dx.doi.org/10.1038/nature13302http://www.nature.com/articles/nature13302>. Citation on page 25.

KIM, Y.; IGNATCHENKO, V.; YAO, C. Q.; KALATSKAYA, I.; NYALWIDHE, J. O.; LANCE, R. S.; GRAMOLINI, A. O.; TROYER, D. A.; STEIN, L. D.; BOUTROS, P. C.; MEDIN, J. A.; SEMMES, O. J.; DRAKE, R. R.; KISLINGER, T. Identification of Differentially Expressed Proteins in Direct Expressed Prostatic Secretions of Men with Organ-confined Versus Extracapsular Prostate Cancer. **Molecular & Cellular Proteomics**, v. 11, n. 12, p. 1870–1884, 12 2012. ISSN 1535-9476. Available: <http://www.mcponline.org/lookup/doi/10.1074/mcp.M112.017889>. Citations on pages 26, 37, 38, 39, 60, 62, 91, 93, 94, 95, 98, 99, and 101.

KOJIMA, K.; NAGASAKI, M.; MIYANO, S. Fast grid layout algorithm for biological networks with sweep calculation. **Bioinformatics**, v. 24, n. 12, p. 1433–1441, 6 2008. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btn196>. Citation on page 107.

KRAUSE, F.; SCHULZ, M.; RIPKENS, B.; FLÖTTMANN, M.; KRANTZ, M.; KLIPP, E.; HANDORF, T. Biographer: Web-based editing and rendering of SBGN compliant biochemical networks. **Bioinformatics**, v. 29, n. 11, p. 1467–1468, 6 2013. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt159>. Citation on page 106.

KREFT, L.; BOTZKI, A.; COPPENS, F.; VANDEPOELE, K.; BEL, M. V. PhyD3: A phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. **Bioinformatics**, v. 33, n. 18, p. 2946–2947, 2017. ISSN 14602059. Citation on page 129.

LAPIERRE, P.; LASEK-NESSELQUIST, E.; GOGARTEN, J. P. The impact of HGT on phylogenomic reconstruction methods. **Briefings in Bioinformatics**, v. 15, n. 1, p. 79–90, 2014. ISSN 14675463. Citation on page 127.

LEE, E.; CHUANG, H.-Y.; KIM, J.-W.; IDEKER, T.; LEE, D. Inferring Pathway Activity toward Precise Disease Classification. v. 4, n. 11, p. e1000217. ISSN 1553-7358. Citation on page 59.

LEE, J. W.; LEE, J. B.; PARK, M.; SONG, S. H. An extensive comparison of recent classification tools applied to microarray data. **Computational Statistics and Data Analysis**, v. 48, n. 4, p. 869–885, 2005. ISSN 01679473. Available: <http://www.sciencedirect.com/science/article/pii/S016794730400101X>. Citation on page 54.

LETUNIC, I.; BORK, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. **Nucleic acids research**, v. 44, n. W1, p. W242–W245, 1 2016. ISSN 13624962. Available: <http://bioinformatics.oxfordjournals.org/content/23/1/127.abstracthttp://www.ncbi.nlm.nih.gov/pubmed/17050570>. Citation on page 129.

LEX, A.; PARTL, C.; KALKOFEN, D.; STREIT, M.; GRATZL, S.; WASSERMANN, A. M.; SCHMALSTIEG, D.; PFISTER, H. Entourage: Visualizing relationships between biological pathways using contextual subsets. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 2536–2545, 12 2013. ISSN 10772626. Available: <http://ieeexplore.ieee.org/document/6634190/>. Citation on page 106.

LI, L.; DARDEN, T. A.; WEINGBERG, C.; LEVINE, A. J.; PEDERSEN, L. G.; WEINBERG, C. R.; LEVINE, A. J.; PEDERSEN, L. G. Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm / k-nearest Neighbor Method. **Combinatorial Chemistry & High Throughput Screening**, v. 4, n. 8, p. 727–739, 2001. ISSN 13862073. Available: <http://www.ncbi.nlm.nih.gov/pubmed/11894805http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1386-2073&volume=4&issue=8&spage=727>. Citation on page 55.

LI, L.; WEINBERG, C. R.; DARDEN, T. A.; PEDERSEN, L. G. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. **Bioinformatics**, v. 17, n. 12, p. 1131–1142, 2001. ISSN 1367-4803. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/17.12.1131>. Citation on page 55.

LI, T.; ZHANG, C.; OGIHARA, M. A comparative study of feature selection and multiclass classfication methods for tissue classification based on gene expression. **Bioinformatics**, v. 20, n. 15, p. 2429–2437, 10 2004. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bth267>. Citation on page 53.

LIU, J. J.; CUTLER, G.; LI, W.; PAN, Z.; PENG, S.; HOEY, T.; CHEN, L.; LING, X. B. Multiclass cancer classification and biomarker discovery using GA-based algorithms. **Bioinformatics**, v. 21, n. 11, p. 2691–2697, 6 2005. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti419>. Citation on page 55.

LODISH, H.; BERK, A.; MATSUDAIRA, P.; KAISER, C. A.; KRIEGER, M.; SCOTT, M. P.; ZIPURSKY, S. L.; DARNELL, J. Tumour Cells and the Onset of Cancer. In: **Molecular Cell Biology**. 4. ed. New York, New York, USA: W. H. Freeman, 2003. chap. 24, p. 940. ISBN 978-1-4641-0981-2. Available: <http://www.ncbi.nlm.nih.gov/books/NBK21590/>. Citation on page 26.

LOUPPE, G.; GEURTS, P. Ensembles on random patches. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 7523 LNAI, n. PART 1, p. 346–361, 2012. ISSN 03029743. Citation on page 69.

MADDISON, W.; KNOWLES, L. Inferring phylogeny despite incomplete lineage sorting. **Systematic Biology**, v. 55, n. 1, p. 21–30, 2 2006. ISSN 10635157. Available: <https://academic.oup.com/sysbio/article/55/1/21/2842934>. Citation on page 125.

MAKARENKOV, V. T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. **Bioinformatics**, v. 17, n. 7, p. 664–668, 2001. ISSN 13674803. Citation on page 129.

MANGUL, S.; MOSQUEIRO, T.; DUONG, D.; MITCHELL, K.; SARWAL, V.; HILL, B.; BRITO, J.; LITTMAN, R.; STATZ, B.; LAM, A.; DAYAMA, G.; GRIENEISEN, L.; MARTIN, L.; FLINT, J.; ESKIN, E.; BLEKHMAN, R. A comprehensive analysis of the usability and archival stability of omics computational tools and resources. **bioRxiv**, v. 119, p. 452532, 11 2018. ISSN 0141-8130. Available: <https://www.biorxiv.org/content/early/2018/10/25/452532.full.pdf+html>. Citation on page 129.

MARTÍNKOVÁ, N.; NOVÁ, P.; SABLINA, O. V.; GRAPHODATSKY, A. S.; ZIMA, J. Karyotypic relationships of the Tatra vole (Microtus tatricus). **Folia Zoologica**, v. 53, n. 3, p. 279–284, 2004. ISSN 01397893. Citation on page 66.

MARTINS, R. M.; ANDERY, G. F.; HEBERLE, H.; PAULOVICH, F. V.; LOPES, A. D. A.; PEDRINI, H.; MINGHIM, R. Multidimensional projections for visual analysis of social networks. **Journal of Computer Science and Technology**, Springer US, v. 27, n. 4, p. 791–810, 7 2012. ISSN 10009000. Available: <http://link.springer.com/10.1007/s11390-012-1265-5>. Citation on page 106.

MCENTYRE, J. Linking up with entrez. **Trends in Genetics**, v. 14, n. 1, p. 39–40, 1 1998. ISSN 01689525. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0168952597013255>. Citation on page 108.

MEINSHAUSEN, N.; BÜHLMANN, P. Stability selection. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 72, n. 4, p. 417–473, 7 2010. ISSN 13697412. Available: <http://doi.wiley.com/10.1111/j.1467-9868.2010.00740.x>. Citation on page 65.

MONNÉ, M. L. M. A.; MONNÉ, M. L. M. A. Synopsis of the Neotropical genus Lepturgantes Gilmour (Coleoptera: Cerambycidae) with description of a new species. **Zootaxa**, v. 8, n. 1876, p. 60–68, 2008. ISSN 11755326. Available: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-460>. Citation on page 129.

MOORE, R. G.; BROWN, A. K.; MILLER, M. C.; SKATES, S.; ALLARD, W. J.; VERCH, T.; STEINHOFF, M.; MESSERLIAN, G.; DISILVESTRO, P.; GRANAI, C. O.; BAST, R. C. The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass. **Gynecologic Oncology**, v. 108, n. 2, p. 402–408, 2 2008. ISSN 00908258. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0090825807008542>. Citation on page 54.

MORRISON, J. L.; BREITLING, R.; HIGHAM, D. J.; GILBERT, D. R. GeneRank: using search engine technology for the analysis of microarray experiments. **BMC Bioinformatics**, v. 6, n. 1, p. 233, 2005. ISSN 1471-2105. Available: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-233>. Citation on page 52.

MOSAIC. **GO network annotation and partition in Cytoscape**. 2012. Available: <http://nrnb.org/tools/mosaic/>. Citation on page 118.

MUELLER, L. A.; DEHMER, M.; EMMERT-STREIB, F. Comparing biological networks: A survey on graph classifying techniques. In: **Systems Biology: Integrative Biology and Simulation Tools**. Dordrecht: Springer Netherlands, 2013. p. 43–63. ISBN 9789400768031. Available: <http://link.springer.com/10.1007/978-94-007-6803-1_2>. Citation on page 106.

MUNZNER, T.; GUIMBRETIÈRE, F.; TASIRAN, S.; ZHANG, L.; ZHOU, Y. TreeJuxtaposer. **ACM Transactions on Graphics**, v. 22, n. 3, p. 453, 2003. ISSN 07300301. Available: <http://portal.acm.org/citation.cfm?doid=882262.882291>. Citation on page 129.

MURTHY, S. K. Automatic construction of decision trees from data: A multi-disciplinary survey. **Data Mining and Knowledge Discovery**, Kluwer Academic Publishers, Hingham, MA, USA, v. 2, n. 4, p. 345–389, 12 1998. ISSN 13845810. Available: <http://dx.doi.org/10.1023/A:1009744630224>. Citation on page 57.

NAGASAKI, M.; LI, C. Cell Illustrator Online 4 . 0 : A Platform for Systems Biology. **Systems Biology**, IOS Press, v. 10, n. i, p. 4–5, 2010. ISSN 1386-6338. Citations on pages 107 and 116.

OSBORNE, J. D.; FLATOW, J.; HOLKO, M.; LIN, S. M.; KIBBE, W. A.; ZHU, L. J.; DANILA, M. I.; FENG, G.; CHISHOLM, R. L. Annotating the human genome with Disease Ontology. **BMC Genomics**, v. 10, n. SUPPL. 1, p. S6, 2009. ISSN 14712164. Available: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2709267&tool=pmcentrez&rendertype=abstract>. Citation on page 52.

OSHIRO, T. M.; PEREZ, P. S.; BARANAUSKAS, J. A. How many trees in a random forest? In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [s.n.], 2012. v. 7376 LNAI, p. 154–168. ISBN 9783642315367. Available: <http://link.springer.com/10.1007/978-3-642-31537-4_13>. Citation on page 66.

PACZESNY, S. Discovery and validation of graft-versus-host disease biomarkers. **Blood**, v. 121, n. 4, p. 585–594, 1 2013. ISSN 00064971. Available: <http://www.bloodjournal.org/cgi/doi/10.1182/blood-2012-08-355990>. Citation on page 28.

PADUANO, F.; FORBES, A. G. Extended LineSets: a visualization technique for the interactive inspection of biological pathways. **BMC Proceedings**, BioMed Central Ltd, v. 9, n. Suppl 6, p. S4, 2015. ISSN 17536561. Available: <http://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-9-S6-S4>. Citation on page 106.

PAGEL, M. Inferring the historical patterns of biological evolution. **Nature**, v. 401, n. 6756, p. 877–884, 10 1999. ISSN 00280836. Available: <http://www.nature.com/articles/44766>. Citation on page 125.

PAN, W.; LIN, J.; LE, C. T. A mixture model approach to detecting differentially expressed genes with microarray data. **Functional and Integrative Genomics**, v. 3, n. 3, p. 117–124, 7 2003. ISSN 1438793X. Available: <http://link.springer.com/10.1007/s10142-003-0085-7>. Citation on page 54.

PANUCCIO, M.; BARBOUTIS, C.; CHIATANTE, G.; EVANGELIDIS, A.; AGOSTINI, N. Pushed by increasing air temperature and tailwind speed: Weather selectivity of raptors migrating across the Aegean Sea. **Ornis Fennica**, v. 93, n. 3, p. 159–171, 3 2016. ISSN 00305685. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0020025510004147http://linkinghub.elsevier.com/retrieve/pii/S0020025513007020http://ieeexplore.ieee.org/document/6295786/>. Citation on page 106.

PAVLIDIS, P.; QIN, J.; ARANGO, V.; MANN, J. J.; SIBILLE, E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. **Neurochemical Research**, v. 29, n. 6, p. 1213–1222, 2004. ISSN 03643190. Available: <http://dx.doi.org/10.1023/B:NERE.0000023608.29741.45>. Citation on page 56.

PHAM, T. V.; PIERSMA, S. R.; WARMOES, M.; JIMENEZ, C. R. On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. **Bioinformatics**, v. 26, n. 3, p. 363–369, 2 2009. ISSN 14602059. Available: <http://bioinformatics.oxfordjournals.org/content/26/3/363.short>. Citations on pages 29, 31, and 32.

PISANI, D.; COTTON, J. A.; MCINERNEY, J. O. Supertrees disentangle the chimerical origin of eukaryotic genomes. **Molecular Biology and Evolution**, v. 24, n. 8, p. 1752–1760, 2007. ISSN 07374038. Citation on page 126.

PRATT, G. Molecular aspects of multiple myeloma. **Journal of Clinical Pathology - Molecular Pathology**, v. 55, n. 5, p. 273–283, 10 2002. ISSN 13668714. Available: <http://mp.bmj.com/cgi/doi/10.1136/mp.55.5.273>. Citation on page 100.

PUNIN, J.; KRISHNAMOORTHY, M. XGMML (eXtensible Graph Markup and Modeling Language) 1.0 draft specification. 2001. Available: <http://www.cs.rpi.edu/~puninj/XGMML/draft-xgmml.html>. Citation on page 108.

PUNTMANN, V. O. How-to guide on biomarkers: Biomarker definitions, validation and applications with examples from cardiovascular disease. **Postgraduate Medical Journal**, v. 85, n. 1008, p. 538–545, 10 2009. ISSN 00325473. Available: <http://pmj.bmj.com/cgi/doi/10.1136/pgmj.2008.073759>. Citation on page 52.

QUINN, J.; FISHER, P. W.; CAPOCASALE, R. J.; ACHUTHANANDAM, R.; KAM, M.; BUGELSKI, P. J.; HREBIEN, L. A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow. **Cytometry Part A**, v. 71, n. 8, p. 612–624, 8 2007. ISSN 1552-4922. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=824819>. Citation on page 26.

RAPAPORT, F.; ZINOVYEV, A.; DUTREIX, M.; BARILLOT, E.; VERT, J. P. Classification of microarray data using gene networks. **BMC Bioinformatics**, v. 8, n. 1, p. 35, 2007. ISSN 14712105. Available: <http://www.ncbi.nlm.nih.gov/pubmed/17270037>. Citation on page 59.

REDFIELD, R. J. Do bacteria have sex? **Nature Reviews Genetics**, v. 2, n. 8, p. 634–639, 8 2001. ISSN 14710064. Available: <http://www.nature.com/articles/35084593>. Citation on page 126.

REVIEWS, C. T. I. **Discover Biology: Biology, Biology**. Cram101, 2016. ISBN 9781478416135. Available: <https://books.google.com.br/books?id=nI8xl9i_9ugC>. Citation on page 126.

REX, D. K.; BOLAND, C. R.; DOMINITZ, J. A.; GIARDIELLO, F. M.; JOHNSON, D. A.; KALTENBACH, T.; LEVIN, T. R.; LIEBERMAN, D.; ROBERTSON, D. J. Colorectal Cancer Screening: Recommendations for Physicians and Patients from the U.S. Multi-Society Task Force on Colorectal Cancer. **American Journal of Gastroenterology**, v. 112, n. 7, p. 1016–1030, 7 2017. ISSN 15720241. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0016510717318059>. Citation on page 24.

ROBINSON, M. M.; PH, D.; HARTLEY, A. G. J.; NUTTING, C.; PH, D.; POWELL, N.; PH, D.; BOOZ, H. A.; ROBINSON, M. M.; SMITH, A. F.; SC, M.; HALL, P.; PH, D.; DUNN, J.; PH, D. PET-CT surveillance versus neck dissection in advanced head and neck cancer. **British dental journal**, v. 220, n. 9, p. 449, 5 2016. ISSN 14765373. Available: <http://www.nature.com/articles/sj.bdj.2016.327>. Citation on page 40.

ROBINSON, O.; DYLUS, D.; DESSIMOZ, C. Phylo.io: Interactive viewing and comparison of large phylogenetic trees on the web. **Mol. Biol. Evol.**, v. 33, p. 2163–2166, 2016. Citation on page 129.

ROSENBERG, S. A.; BRANCH, S.; MEDICINE, N.; INSTITUTES, N.; MISTELI, T. **New approach to immunotherapy leads to complete response in breast cancer patient unresponsive to other treatments**. 2018. 1–2 p. Available: <https://www.cancer.gov/news-events/press-releases/2018/immunotherapy-targets-breast-cancer-case-report>. Citation on page 25.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, v. 4, n. 4, p. 406–425, 7 1987. ISSN 1537-1719. Available: <http://mbe.oxfordjournals.org/content/4/4/406.abstracthttps://academic.oup.com/mbe/article/4/4/406/1029664/The-neighborjoining-method-a-new-method-for>. Citation on page 127.

SALDANHA, A. J. Java Treeview - Extensible visualization of microarray data. **Bioinformatics**, v. 20, n. 17, p. 3246–3248, 11 2004. ISSN 13674803. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15180930>. Citation on page 129.

SAMET, H. The Quadtree and Related Hierarchical Data Structures. ACM, v. 16, n. 2, p. 187–260. ISSN 0360-0300. Citation on page 111.

SANAVIA, T.; AIOLLI, F.; MARTINO, G. D. S.; BISOGNIN, A.; CAMILLO, B. D. Improving biomarker list stability by integration of biological knowledge in the learning process. **BMC Bioinformatics**, BioMed Central Ltd, v. 13, n. SUPPL.4, p. S22, 2012. ISSN 14712105. Available: <http://www.biomedcentral.com/1471-2105/13/S4/S22>. Citations on pages 51 and 59.

SANTAMARÍA, R.; THERÓN, R. Treevolution: Visual analysis of phylogenetic trees. **Bioinformatics**, v. 25, n. 15, p. 1970–1971, 8 2009. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/content/25/15/1970.abstracthttp://www.ncbi.nlm.nih.gov/pubmed/19470585>. Citation on page 129.

SCHUHMACHER, A. **Software Visualization via Hierarchic Graphs**. Phd Thesis (PhD Thesis) — Karlsruhe Institute of Technology, 2015. Citation on page 106.

SCHULZ, H. J. Treevis.net: A tree visualization reference. **IEEE Computer Graphics and Applications**, v. 31, n. 6, p. 11–15, 11 2011. ISSN 02721716. Available: <http://ieeexplore.ieee.org/document/6056510/>. Citation on page 128.

Scikit-learn developers. **sklearn.preprocessing.MinMaxScaler - scikit-learn 0.20.0 documentation**. 2017. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Citations on pages 65, 71, and 76.

SHAH, R. D.; SAMWORTH, R. J. Variable selection with error control: another look at stability selection. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 75, n. 1, p. 55–80, 1 2013. ISSN 13697412. Available: <http://doi.wiley.com/10.1111/j.1467-9868.2011.01034.x>. Citation on page 65.

SHANNON, P.; MARKIEL, A.; OZIER, O.; BALIGA, N. S.; WANG, J. T.; RAMAGE, D.; AMIN, N.; SCHWIKOWSKI, B.; IDEKER, T. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. **Genome Research**, Cold Spring Harbor Lab, v. 13, n. 11, p. 2498–2504, 11 2003. ISSN 10889051. Available: <http://www.ncbi.nlm.nih.gov/pubmed/14597658>. Citation on page 106.

SHNEIDERMAN, B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: **The Craft of Information Visualization**. IEEE Comput. Soc. Press, 2003. p. 364–371. ISBN 0-8186-7508-X. ISSN 1049-2615. Available: <http://linkinghub.elsevier.com/retrieve/pii/B9781558609150500469>. Citation on page 122.

SILVA, S. D. da; HIER, M.; MLYNAREK, A.; KOWALSKI, L. P.; ALAOUI-JAMALI, M. A. Recurrent oral cancer: Current and emerging therapeutic approaches. **Frontiers in Pharmacology**, v. 3 JUL, n. July, p. 1–7, 2012. ISSN 16639812. Available: <http://journal.frontiersin.org/article/10.3389/fphar.2012.00149/abstract>. Citation on page 40.

Springer Nature Publishing AG. **Transcriptomics - Latest research and news | Nature**. 2019. Available: <https://www.nature.com/subjects/transcriptomics>. Citation on page 26.

STRIMBU, K.; TAVEL, J. A. What are biomarkers? **Current Opinion in HIV and AIDS**, v. 5, n. 6, p. 463–466, 11 2010. ISSN 1746630X. Available: <http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=01222929-201011000-00003>. Citation on page 24.

SUGAR, C. **Force Directed Edge Bundling (FDEB) in Javascript**. 2015. Available: <https://github.com/upphiminn/d3.ForceBundle>. Citation on page 111.

SYMEONIDIS, A.; TOLLIS, I. G. to Visualize Phylogenetic Trees. **Exchange Organizational Behavior Teaching Journal**, Springer, p. 283–293, 2005. Citation on page 129.

TAYLOR, I. W.; LINDING, R.; WARDE-FARLEY, D.; LIU, Y.; PESQUITA, C.; FARIA, D.; BULL, S.; PAWSON, T.; MORRIS, Q.; WRANA, J. L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. **Nature Biotechnology**, v. 27, n. 2, p. 199–204, 2 2009. ISSN 10870156. Available: <http://www.nature.com/doifinder/10.1038/nbt.1522>. Citation on page 59.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. **Nature Genetics**, v. 25, n. may, p. 25–29, 2000. ISSN 1061-4036. Citation on page 43.

The International Agency for Research on Cancer. **Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018**. [S.l.], 2018. Available: <https://www.iarc.fr/wp-content/uploads/2018/09/pr263{_}E.pdf>. Citation on page 23.

Thomas Huijskens. **Stability selection – stability-selection 0.1.0 documentation**. 2018. Available: <https://thuijskens.github.io/stability-selection/docs/stability_selection.html>. Citation on page 65.

THURSTON, R. C.; HERNANDEZ, J.; RIO, J. M. D.; TORRE, F. D. L. Support Vector Machines to improve physiologic hot flash measures: Application to the ambulatory setting. **Psychophysiology**, v. 48, n. 7, p. 1015–1021, 7 2011. ISSN 14698986. Available: <http://www.ncbi.nlm.nih.gov/pubmed/21143609>. Citations on pages 29, 32, 33, and 59.

TIAN, L. L.; GREENBERG, S. A.; KONG, S. W.; ALTSCHULER, J.; KOHANE, I. S.; PARK, P. J. Discovering statistically significant pathways in expression profiling studies. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 38, p. 13544–9, 2005. ISSN 0027-8424. Available: <http://www.pnas.org/cgi/doi/10.1073/pnas.0506577102http://www.ncbi.nlm.nih.gov/pubmed/16174746>. Citation on page 56.

TIBSHIRANI, R.; HASTIE, T.; NARASIMHAN, B.; CHU, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. **Proceedings of the National Academy of Sciences**, v. 99, n. 10, p. 6567–6572, 5 2002. ISSN 0027-8424. Available: <http://www.pnas.org/cgi/doi/10.1073/pnas.082099299>. Citations on pages 29, 32, and 33.

____. Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. **Statistical Science**, v. 18, n. 1, p. 104–117, 2 2003. ISSN 0883-4237. Available: <http://projecteuclid.org/euclid.ss/1056397488>. Citation on page 59.

TLSTY, T. D.; COUSSENS, L. M. Tumor Stroma and Regulation of Cancer Development. **Annual Review of Pathology: Mechanisms of Disease**, v. 1, n. 1, p. 119–150, 2006. ISSN 1553-4006. Available: <http://www.annualreviews.org/doi/10.1146/annurev.pathol.1.110304.100224>. Citation on page 41.

TODUA, F.; GAGUA, R.; MAGLAKELIDZE, M.; MAGLAKELIDZE, D. Cancer incidence and mortality - Major patterns in GLOBOCAN 2012, worldwide and Georgia. **Bulletin of the Georgian National Academy of Sciences**, v. 9, n. 1, p. 168–173, 2015. ISSN 01321447. Citation on page 40.

TURLEY, S. J.; CREMASCO, V.; ASTARITA, J. L. Immunological hallmarks of stromal cells in the tumour microenvironment. **Nature Reviews Immunology**, Nature Publishing Group, v. 15, n. 11, p. 669–682, 2015. ISSN 14741741. Available: <http://dx.doi.org/10.1038/nri3902>. Citation on page 42.

TUSHER, V. G.; TIBSHIRANI, R.; CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. **Proceedings of the National Academy of Sciences**, v. 98, n. 9, p. 5116–5121, 4 2001. ISSN 0027-8424. Available: <http://www.pnas.org/cgi/doi/10.1073/pnas.091062498>. Citations on pages 54, 55, and 59.

UHLEN, M.; OKSVOLD, P.; FAGERBERG, L.; LUNDBERG, E.; JONASSON, K.; FORSBERG, M.; ZWAHLEN, M.; KAMPF, C.; WESTER, K.; HOBER, S.; WERNERUS, H.; BJÖRLING, L.; PONTEN, F. Towards a knowledge-based Human Protein Atlas. **Nature Biotechnology**, v. 28, n. 12, p. 1248–1250, 12 2010. ISSN 10870156. Available: <http://www.nature.com/doifinder/10.1038/nbt1210-1248>. Citations on pages 44 and 113.

VASILIJEVIC, A.; JAMBROSIC, K.; VUKIC, Z. Teleoperated path following and trajectory tracking of unmanned vehicles using spatial auditory guidance system. **Applied Acoustics**, v. 129, n. 4598, p. 72–85, 5 2018. ISSN 1872910X. Available: <http://www.sciencemag.org/cgi/doi/10.1126/science.220.4598.671>. Citation on page 57.

VAUGHAN, T. G. IcyTree: Rapid browser-based visualization for phylogenetic trees and networks. **Bioinformatics**, v. 33, n. 15, p. 2392–2394, 2017. ISSN 14602059. Citation on page 129.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. **Pattern Recognition**, v. 44, n. 2, p. 330–349, 2011. ISSN 00313203. Available: <http://www.sciencedirect.com/science/article/pii/S0031320310003973>. Citation on page 57.

W3C. **5.1 Access-Control-Allow-Origin Response Header**. 2014. Available: <https://www.w3.org/TR/cors/>. Citation on page 112.

WAAL, I. van der. Are we able to reduce the mortality and morbidity of oral cancer; some considerations. **Medicina Oral, Patologia Oral y Cirugia Bucal**, v. 18, n. 1, p. 33–37, 2013. ISSN 16984447. Citation on page 40.

WANG, L.; ZHU, J.; ZOU, H. Hybrid huberized support vector machines for microarray classification and gene selection. **Bioinformatics**, v. 24, n. 3, p. 412–419, 2 2008. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm579>. Citations on pages 55 and 59.

WANG, S. L.; LI, X. L.; FANG, J. Finding minimum gene subsets with heuristic breadth-first search algorithm for robust tumor classification. **BMC Bioinformatics**, BMC Bioinformatics, v. 13, n. 1, p. 178, 2012. ISSN 14712105. Available: <http://www.biomedcentral.com/1471-2105/13/178>. Citations on pages 52, 54, and 55.

WANG, Y.; KLIJN, J. G.; ZHANG, Y.; SIEUWERTS, A. M.; LOOK, M. P.; YANG, F.; TALANTOV, D.; TIMMERMANS, M.; GELDER, M. E. M.-V.; YU, J.; JATKOE, T.; BERNS, E. M.; ATKINS, D.; FOEKENS, J. A. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. **Lancet**, v. 365, n. 9460, p. 671–679, 2 2005. ISSN 01406736. Available: <http://www.ncbi.nlm.nih.gov/pubmed/15721472>. Citations on pages 51 and 57.

WANG, Y.; MAKEDON, F. S.; FORD, J. C.; PEARLMAN, J. HykGene: A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. **Bioinformatics**, v. 21, n. 8, p. 1530–1537, 4 2005. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti192>. Citation on page 55.

WANG, Y.; THILMONY, R.; GU, Y. Q. NetVenn: an integrated network analysis web platform for gene lists. **Nucleic Acids Research**, v. 42, n. W1, p. W161–W166, 7 2014. ISSN 0305-1048. Available: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku331>. Citation on page 114.

WANG, Y.; WANG, J.; HUANG, Y. MicroRNAs as new biomarkers for human papilloma virus related head and neck cancers. **Cancer Biomarkers**, v. 15, n. 3, p. 213–218, 2015. ISSN 18758592. Citation on page 41.

WEILAND, S. Aging According to Biography. In: **Gerontologist**. IEEE, 1989. v. 29, n. 2, p. 191–194. ISBN 9781479956692. ISSN 17585341. Available: <http://www.sciencedirect.com/science/article/pii/S1046202315001528>. Citations on pages 51 and 57.

WELINDER, C.; JIRSTRÖM, K.; LEHN, S.; NODIN, B.; MARKO-VARGA, G.; BLIXT, O.; DANIELSSON, L.; JANSSON, B. Intra-tumour IgA1 is common in cancer and is correlated with poor prognosis in bladder cancer. **Heliyon**, Elsevier Ltd, v. 2, n. 8, p. e00143, 8 2016. ISSN 24058440. Available: <http://dx.doi.org/10.1016/j.heliyon.2016.e00143https://linkinghub.elsevier.com/retrieve/pii/S2405844016303693>. Citation on page 100.

WHIDDEN, C.; ZEH, N.; BEIKO, R. G. Supertrees based on the subtree prune-and-regraft distance. **Systematic Biology**, v. 63, n. 4, p. 566–581, 7 2014. ISSN 1076836X. Available: <https://academic.oup.com/sysbio/article/63/4/566/2848417>. Citations on pages 127, 128, 129, 131, 132, 133, and 135.

WHO, W. H. O. **Biomarkers In Risk Assessment: Validity And Validation. International Programme on Chemical Safety (IPCS), Environmental Health Criteria 222**. [S.l.]: World Health Organization, 2001. 144 p. ISBN 924572221. Citation on page 27.

Wikimedia Commons. **File:Felidae phylogeny (eng).png**. 2017. Available: <https://commons.wikimedia.org/wiki/File:Felidae_phylogeny_(eng).png>. Citation on page 127.

WINCK, F. V.; RIBEIRO, A. C. P.; DOMINGUES, R. R.; LING, L. Y.; RIAÑO-PACHÓN, D. M.; RIVERA, C.; BRANDÃO, T. B.; GOUVEA, A. F.; SANTOS-SILVA, A. R.; COLETTA, R. D.; LEME, A. F. Insights into immune responses in oral cancer through proteomic analysis of saliva and salivary extracellular vesicles. **Scientific Reports**, Nature Publishing Group, v. 5, n. November, p. 1–13, 2015. ISSN 20452322. Available: <http://dx.doi.org/10.1038/srep16305>. Citation on page 41.

XIONG, M.; FANG, X.; ZHAO, J. Biomarker {Identification} by {Feature} {Wrappers}. **Genome Research**, v. 11, n. 11, p. 1878–1887, 11 2001. ISSN 1088-9051. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC311150/>. Citation on page 55.

YAN, X.; DENG, M.; FUNG, W. K.; QIAN, M. Detecting differentially expressed genes by relative entropy. **Journal of Theoretical Biology**, v. 234, n. 3, p. 395–402, 6 2005. ISSN 00225193. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0022519304005995>. Citation on page 54.

YATES, A.; AKANNI, W.; AMODE, M. R.; BARRELL, D.; BILLIS, K.; CARVALHO-SILVA, D.; CUMMINS, C.; CLAPHAM, P.; FITZGERALD, S.; GIL, L.; GIRÓN, C. G.; GORDON, L.; HOURLIER, T.; HUNT, S. E.; JANACEK, S. H.; JOHNSON, N.; JUETTEMANN, T.; KEENAN, S.; LAVIDAS, I.; MARTIN, F. J.; MAUREL, T.; MCLAREN, W.; MURPHY, D. N.; NAG, R.; NUHN, M.; PARKER, A.; PATRICIO, M.; PIGNATELLI, M.; RAHTZ, M.; RIAT, H. S.; SHEPPARD, D.; TAYLOR, K.; THORMANN, A.; VULLO, A.; WILDER, S. P.; ZADISSA, A.; BIRNEY, E.; HARROW, J.; MUFFATO, M.; PERRY, E.; RUFFIER, M.; SPUDICH, G.; TREVANION, S. J.; CUNNINGHAM, F.; AKEN, B. L.; ZERBINO, D. R.; FLICEK, P. Ensembl 2016. **Nucleic Acids Research**, v. 44, n. D1, p. D710–D716, 1 2016. ISSN 0305-1048. Available: <http://dx.doi.org/10.1093/nar/gkv1157https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1157>. Citation on page 108.

YU, G.; SMITH, D. K.; ZHU, H.; GUAN, Y.; LAM, T. T. Y. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. **Methods in Ecology and Evolution**, v. 8, n. 1, p. 28–36, 2017. ISSN 2041210X. Citation on page 129.

ZHANG, C. **MOSAIC - Figure**. 2012. Available: <http://nrnb.org/tools/mosaic/images/mosaicresults.png>. Citation on page 119.

____. **MOSAIC - Figure**. 2012. Available: <http://nrnb.org/tools/mosaic/images/mosaic-selectnodes.png>. Citation on page 119.

____. **MOSAIC - Figure**. 2012. Available: <http://nrnb.org/tools/mosaic/images/mosaicresults.png>. Citation on page 119.

ZHANG, C.; HANSPERS, K.; KUCHINSKY, A.; SALOMONIS, N.; XU, D.; PICO, A. R. Mosaic: making biological sense of complex networks. **Bioinformatics**, v. 28, n. 14, p. 1943–1944, 7 2012. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts278https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts278>. Citations on pages 106, 107, 116, and 118.

ZHANG, H. H.; AHN, J.; LIN, X.; PARK, C. Gene selection using support vector machines with non-convex penalty. **Bioinformatics**, v. 22, n. 1, p. 88–95, 1 2006. ISSN 13674803. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti736https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti736>. Citations on pages 56 and 59.

ZHANG, L.; FARRELL, J. J.; ZHOU, H.; ELASHOFF, D.; AKIN, D.; PARK, N. H.; CHIA, D.; WONG, D. T. Salivary Transcriptomic Biomarkers for Detection of Resectable Pancreatic Cancer. **Gastroenterology**, Elsevier Inc., v. 138, n. 3, p. 949–957, 3 2010. ISSN 00165085. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0016508509020009http://dx.doi.org/10.1053/j.gastro.2009.11.010>. Citations on pages 54 and 55.

ZHU, Y.; SHEN, X.; PAN, W. Network-based support vector machine for classification of microarray samples. **BMC Bioinformatics**, v. 10, n. SUPPL. 1, p. S21, 2009. ISSN 14712105. Available: <http://www.biomedcentral.com/1471-2105/10/S1/S21>. Citation on page 59.

ZHU, Z.; ONG, Y. S.; DASH, M. Markov blanket-embedded genetic algorithm for gene selection. **Pattern Recognition**, v. 40, n. 11, p. 3236–3248, 11 2007. ISSN 00313203. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0031320307000945>. Citation on page 56.

ZIZAS, R.; SHAMOVICH, D.; KURLAVIČIUS, P.; BELOVA, O.; BRAZAITIS, G. Radio-tracking of capercaillie (tetrao urogallus l.) in north belarus. **Baltic Forestry**, v. 18, n. 2, p. 270–277, 9 2012. ISSN 13921355. Available: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq345>. Citations on pages 52 and 59.


ICMC
USP
SÃO CARLOS