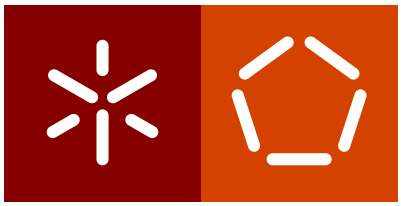
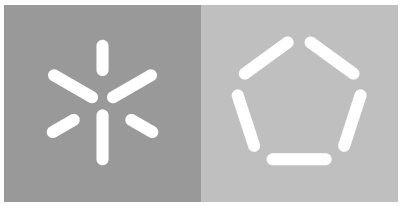


Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Filipe Silva Correia

HIV-TB-Host Protein Interaction Network

November 2018



Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Filipe Silva Correia

HIV-TB-Host Protein Interaction Network

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Nuno Miguel Sampaio Osório
Miguel Francisco de Almeida Pereira da Rocha

November 2018

ACKNOWLEDGEMENTS / AGRADECIMENTOS

Em primeiro lugar, quero agradecer aos meus pais, que me permitiram chegar a este momento. Foram os ideais que me transmitiram que me fizeram a pessoa que sou hoje. Todos os objetivos que alcancei foram em grande parte possíveis graças ao seu apoio, carinho e suor. Aos meus irmãos, por sempre cuidarem do irmão mais novo. A voçês um enorme obrigado por sempre me apoiarem em todas as decisões que tomei e nunca me faltarem em nada.

Queria também expressar o meu agradecimento ao meu orientador Nuno Osório. Obrigado por tudo, pelo apoio, pela disponibilidade, pela motivação e por me ter recebido no seu grupo de investigação. É sem dúvida um exemplo de liderança, foi um prazer trabalhar este ano sobre a sua orientação. Ao meu co-orientador professor Miguel Rocha, que não só durante este ano, mas ao longo de todo o mestrado sempre mostrou total disponibilidade em ajudar. Foi sem dúvida o professor que mais teve impacto durante o meu percurso no Mestrado em Bioinformática, sendo um exemplo de organização e profissionalismo. À Ana Pereira, que apesar de não ser minha orientadora esteve sem dúvida a esse nível. Obrigado Ana por estares sempre disposta a ajudar qualquer que fosse o motivo, e principalmente obrigado por seres uma ótima amiga. Não podia também deixar de referir os restantes elementos do grupo evoBiomed. Obrigado Carlos, Pedro, Isaac e Deisy. Que as cervejas no Pão de Forma continuem por muitos anos.

Finalmente, queria agradecer aos meus amigos por fazerem parte deste 5 anos, desde Vila Real a Braga, que foram sem dúvida os melhores da minha vida. Dizem que se uma amizade durar mais de 7 anos, a probabilidade de durar para a toda a vida é muito grande. Tenho a certeza que acabamos de baixar esse número para 5 anos. Foram sempre um ponto estável durante estes anos e nunca me deixaram ir abaixo nem desmotivar. Que continuemos sempre assim.

Um enorme obrigado a TODOS!

ABSTRACT

Mycobacterium tuberculosis (Mtb) and the *Human Immunodeficiency Virus 1 (HIV-1)* are responsible for the development of *Tuberculosis (TB)* and *Acquired Immunodeficiency Syndrome (AIDS)*, respectively. These are the infectious diseases with the highest mortality rates in the world and Mtb/HIV-1 co-infection further aggravates the severity and burden of both diseases. Additional information on how these pathogens interact with the host and gain reciprocal advantages in co-infection is fundamental to devise better therapeutic strategies. Therefore, it is essential to identify and study all host-pathogen *Protein-Protein Interactions (PPIs)* not only considering a single pathogen and the host but a network of interactions between the three organisms. The use of computational tools for the analysis of PPIs provides efficient assessment, integration and interpretation of data from vast arrays of experiments. However, despite the numerous interactions that have been reported in the literature, there is currently no single database where PPI information for *Human Immunodeficiency Virus (HIV)*, Mtb and human proteins is integrated and can be efficiently accessed.

In this thesis, we aim at providing an integrated and informative network of all known interactions between HIV-1, Mtb and human proteins. To this end, the *Syndemic Protein Interaction NETwork (SPINET)* was designed focusing on the development of a database for data storage and a web tool for data analysis and visualization. The approach presented in this thesis includes the design and implementation of a relational database, the collection of 350,653 HIV-1, Mtb and Human PPIs from multiple sources and the development of a web interface allowing the search, analysis and visualization of the *Protein-Protein Interaction Networks (PPINs)* formed between these organisms.

This work can represent a valuable resource for the scientific community, providing valuable insights on the study of TB, AIDS and Mtb/HIV-1 co-infection. The implementation of SPINET allowed, for the first time, the identification of 81 human proteins that have been experimentally validated to interact directly with Mtb and HIV proteins and also to analyze the inter-species networks formed by these proteins. Interestingly, it was highlighted that, although none of the 81 identified proteins have known inhibitors, they directly interact with other human proteins for which inhibitors have been produced. This list includes tumor necrosis factor inhibitors, that have been used to treat AIDS and are known to promote TB. Other inhibitors that have not been used in the context of these diseases and are potential candidates to be evaluated has host-directed therapies were also identified. This opens new pathways for research toward better control of these deadly diseases.

RESUMO

Mycobacterium tuberculosis (Mtb) e o Vírus da Imunodeficiência Humana 1 (HIV-1) são, respetivamente, os agentes responsáveis pelo desenvolvimento da Tuberculose (TB) e do Síndrome da Imunodeficiência Adquirida (AIDS). Estas são as doenças infeciosas com maior taxa de mortalidade no mundo, e a coinfecção Mtb/HIV-1 agrava ainda mais o impacto de ambas as doenças. Consequentemente, informações adicionais sobre como estes patógenos interagem com o hospedeiro e manipulam os seus mecanismos de defesa durante a coinfecção são fundamentais para o desenvolvimento de estratégias terapêuticas mais eficazes. Assim, é essencial identificar e estudar todas as interações proteína-proteína (PPIs) entre hospedeiro-patógeno, não, apenas, considerando um único patógeno e hospedeiro, mas uma rede de interações entre os três organismos. O uso de ferramentas computacionais para a análise de PPIs oferecem mecanismos eficientes para a avaliação, integração e interpretação de elevadas quantidades de dados experimentais. No entanto, apesar das inúmeras interações que têm sido reportadas na literatura, atualmente, não existe um único repositório em que dados de PPIs para HIV, Mtb e hospedeiro humano estejam reunidos e possam ser eficientemente consultados.

Nesta tese, um dos principais objetivos passou por conseguir disponibilizar uma rede integrada e informativa de todas as interações conhecidas entre proteínas de HIV-1, Mtb e humanas. Neste sentido, o Syndemic Protein Interaction NETwork (SPINET) foi planeado tendo como foco o desenvolvimento de uma base de dados e de uma ferramenta web para armazenamento, análise e visualização de dados de PPIs. A abordagem apresentada nesta tese inclui o design e a implementação de uma base de dados relacional, onde foi possível armazenar 350,653 PPIs experimentalmente validadas entre proteínas humanas, de HIV-1 e Mtb. Estas interações foram recolhidas de múltiplos repositórios e de literatura científica. Adicionalmente, uma interface web que permite a consulta, análise e visualização das redes de PPIs formadas por estes organismos foi desenvolvida.

O trabalho apresentado nesta tese pode representar um recurso relevante para a comunidade científica, fornecendo informações valiosas no estudo da TB, AIDS e da coinfecção Mtb/HIV-1. Com a implementação do SPINET, foi possível pela primeira vez, identificar 81 proteínas humanas com evidência experimental de interagirem simultaneamente com proteínas de HIV e Mtb. Curiosamente, destacou-se que, embora nenhuma das 81 proteínas identificadas tenha inibidores conhecidos, estas interagem diretamente com outras proteínas humanas sobre as quais diversos inibidores foram produzidos. Esta lista inclui inibidores do fator de necrose tumoral, que têm sido usados para tratar a AIDS, e

que são conhecidos por aumentar o risco de desenvolver TB. Outros inibidores que nunca foram usados no contexto destas doenças e que podem representar potenciais candidatos a serem avaliados como terapias direcionadas ao hospedeiro foram também identificados. Este trabalho abre, assim, novos caminhos e possibilidades para a investigação, no sentido de melhor controlar estas doenças mortais.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

A

AD Activation Domain.
AIDS Acquired Immunodeficiency Syndrome.
ANN Artificial Neural Networks.
API Application Programming Interface.
APID Agile Protein Interaction DataAnalyser.
AP-MS Affinity Purification and Mass Spectrometry.
ART Antiretroviral Therapy.

B

BC Betweenness Centrality.
BFS Breadth-First Search.
BioGRID Biological General Repository for Interaction Datasets.
BP Biological Process.

C

CC Closeness Centrality.
CC Cellular Compartment.
CSS Cascading Style Sheets.

D

DAVID Database for Annotation, Visualization and Integrated Discovery.
DBD DNA Binding Domain.
DBMS Database Management System.
DC Degree Centrality.
DFS Depth-First Search.
DIP Database of Interacting Proteins.
DT Decision Tree.

E

EC Eigenvector Centrality.
EM School of Medicine.

G

GO Gene Ontology.

H

HIV Human Immunodeficiency Virus.
HIV-1 HID Human Immunodeficiency Virus Type 1 (HIV-1), Human Interaction Database.
HIV-1 Human Immunodeficiency Virus 1.
HPRD Human Protein Reference Database.
HTML HyperText Markup Language.
HTTPS Hyper Text Transfer Protocol Secure.

I

ICVS Life and Health Science Research Institute.
IG1 Interaction Generality Measurement.
IGRA Interferon Gamma Release Assay.
IMEx International Molecular Exchange Consortium.
IntAct Molecular Interaction Database.
iPPI-DB Inhibitors of Protein-Protein Interaction Database.
IRAP Interaction Reliability by Alternative Path.

J

JS JavaScript.

K

KNN K-Nearest Neighbour.

M

MDR-TB Multidrug-Resistant Tuberculosis.
MF Molecular Function.
MIMIX Minimum Information about a Molecular Interaction eXperiment.
MINT Molecular Interaction Database.
ML Machine Learning.
MS Mass Spectrometry.
Mtb *Mycobacterium tuberculosis*.

N

NB Naive Bayesian.

NER Name Entity Recognition.
NLM National Library of Medicine.

P

PINV Protein Interaction Network Visualizer.
PPI Protein-Protein Interaction.
PPIN Protein-Protein Interaction Network.
PSI-MI Protein Standards Initiative for Molecular Interactions.

R

RefSeq Reference Sequence.
REST Representational State Transfer.
RF Random Forest.

S

SDLC Software Development Life Cycle.
SPINET Syndemic Protein Interaction NETwork.
SQL Structured Query Language.
STRING Search Tool for the Retrieval of Interacting Genes/Proteins.
SVM Support Vector Machine.

T

TB Tuberculosis.
TF Transcription Factor.
TIMBAL TIMBAL.
TNF Tumor Necrosis Factor.
TST Mantoux Tuberculin Skin Test.

U

UniHI Unified Human Interactome.
UniProt Universal Protein Resource.

W

WHO World Health Organization.

Y

Y2H Teast Two-Hibrid.

1

INTRODUCTION

1.1 MOTIVATION

The emergence of high-throughput and computational techniques for screening and prediction of PPIs are currently providing massive amounts of data. The analysis of this type of data has the potential to bring important advances in a wide set of fields. However, no single study has the power to provide a complete information on all interactions of an organism interactome, so data from different studies need to be integrated. The integration of PPI data is not a trivial task, not only for scientists without computational background but also for computational scientists. PPI data is spread among multiple repositories, in disparate formats and with different nomenclatures, making its combination very difficult and time-consuming. The overlap of information is very low, making the use of information from a single source very incomplete and biased. Another huge limitation is the fact that high-throughput methods generate high rates of false positives and false negatives. The combination of data from multiple sources, generated by multiple studies with different methods and the combination of other types of biological data can provide a huge enhancement in the construction of more complete and reliable networks. Therefore, leading to the potential discovery of novel, more accurate and personalized diagnostic and treatment techniques.

PPINs are often modeled as graphs, where proteins are represented as nodes and interactions as connections between nodes. This representation allows PPI data to be analysed using graph-theoretical methods. In fact, most studies made on PPI data have been using graph-based algorithms to analyze PPI data [1, 2, 3]. Visualization of graphs has been also deeply explored over the last decades, not only in the context of PPINs but also in other research fields [4, 5, 6]. In recent years, a lot of platforms focusing on the storage, visualization and analysis of PPI data have been implemented to overcome the above-mentioned issues. However, there is none or little data standardization among these platforms.

In general, the availability of PPI data for an organism varies depending on the knowledge and size of the organism proteome and also on the capacity to handle the organism in the laboratory. For example, in the case of Mtb, these challenges include, among others,

its slow growth in vitro and the necessity of using a biosafety level 3 laboratory. Many organisms secrete virulence effector proteins into host cells, where they interact with host proteins to modulate the host mechanisms of defense. However, these dynamic interactions between host and pathogen have not been fully understood. Thus, the construction of inter-species PPINs is of extreme importance to study these complex interactions. This is especially important and challenging for Mtb and HIV-1 that are obligate intracellular human parasites.

However, despite the numerous interactions that have been reported in databases or in scientific literature between proteins of HIV or Mtb with human proteins there is currently no single database where this information can be efficiently accessed. Thus, this work focuses on the development of a computational platform for PPI data storage, visualization and analysis. Additionally, another objective of this thesis is to perform multiple analyses of the collected data.

1.2 GOALS

The overall objective of this study was to provide researchers with a concise but informative network of all known interactions between HIV-1/Mtb and human proteins.

The specific aims of this thesis were:

1. Review some of the major PPI databases, visualization and analysis tools;
2. Collect, process and standardize PPI data from multiple repositories and from the literature;
3. Develop and implement a relational database to store the collected data and to provide a standardized and structured data format to the scientific community;
4. Develop a web tool to visualize and analyze the collected data;
5. Perform multiple data analysis.

1.3 STRUCTURE OF THE DOCUMENT

This document is organized in the following way:

Chapter 2 - Tuberculosis and Human Immunodeficiency Virus

Introduction to the TB and HIV diseases and Mtb/HIV-1 coinfection.

Chapter 3 - Protein-Protein Interactions

Description of the current state of the art in the PPI field followed by the presentation and description of useful bioinformatics tools and databases for its storage, analysis and visualization. Biological data integration limitations and challenges are also addressed in this chapter.

Chapter 4 - Methodology and Implementation

Here, the methodological steps of the work developed in this thesis are outlined and detailed. The main topics addressed were the SPINET database requirements, design, implementation, testing and maintenance. Data collection, curation, visualization and analysis are also addressed.

Chapter 5 - Results

The main results generated in this thesis are presented in this section. These included data statistics, network measures and the analysis of the PPI sub-network formed by human proteins that interact with both Mtb and HIV proteins.

Chapter 6 - Discussion

Discussion of some of the major results presented in this thesis. Emphasis was given to the discussion of the main characteristics of the SPINET platform and to the discussion of the analysis made on the multiple networks.

Chapter 7 - Conclusion

Main conclusions of the work done during this thesis, followed by a description of possible improvements and future work.

2

TUBERCULOSIS AND HUMAN IMMUNODEFICIENCY VIRUS

TB is by itself an intricate public health threat. Furthermore, TB is one of the major mortality causes in HIV-positive individuals. HIV weakens the immune system, increasing the risk of developing TB. In fact, latent TB is more likely to advance to active TB after coinfection with HIV [7]. The diagnostic and treatment of coinfected people is challenging and depends on the individual circumstances of each patient. The *World Health Organization (WHO)* has implemented several initiatives to reduce the global burden of Mtb-HIV coinfection including measures to achieve better surveillance, improved diagnosis and the implementation of TB preventive therapy in HIV-positive individuals with latent TB.

2.1 TUBERCULOSIS

2.1.1 *Epidemiology*

TB is the ninth leading cause of death worldwide and the leading cause of a single infectious agent. In 2017, the WHO reported 6.4 million new cases of TB, up to 6.3 million in 2016, equivalent to 64% of the estimated incidence of 10 million [8]. TB was responsible for about 1.6 million deaths in 2017, a decline from the 1.7 million reported in 2016 [8]. Despite the decline in the number of TB cases and deaths, the number of *Multidrug-Resistant Tuberculosis (MDR-TB)* is still high. In 2017, 558 0000 new cases with resistance to rifampicin were reported. Rifampicin is the most effective first-line drug to treat TB and 82% of the rifampicin cases were MDR-TB. Most of the estimated number of incident cases in 2017 occurred in Southeast Asia (44%), Africa (25%) and the Western Pacific Region (18%) [8]. The top five countries, with 56% of the estimated cases, were (in descending order) India, Indonesia, China, the Philippines and Pakistan [8]. Figure 1 shows the estimated TB incidence in 2017, for countries with at least 100 000 incident cases.

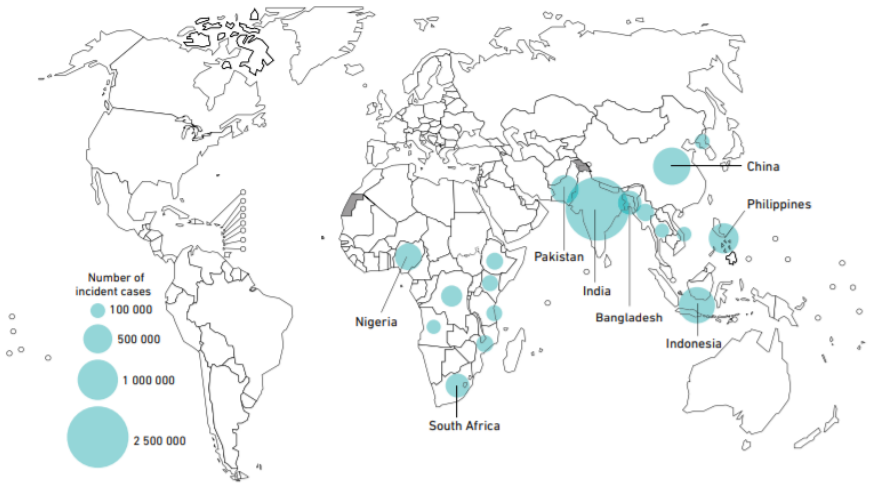


Figure 1: **Estimated TB incidence in 2017, for countries with at least 100 000 incident cases [8].**

2.1.2 *Pathophisiology*

TB is an ancient infectious disease caused by the bacterium Mtb. Mtb is spread by small airborne droplets, expelled by people with active TB. These minuscule droplets can remain airborne for minutes to hours and once inhaled, the droplets settle throughout the airways. When in contact with the lungs, Mtb may cause infection of the respiratory system. In some patients, Mtb is also able to disseminate to other organs and cause extrapulmonary TB.

In most cases, if the contaminated droplets are trapped in the upper parts of the airways where the mucus-secreting goblet cells exist, the infection is prevented [9]. When the infected droplets bypass the mucociliary system and reach the alveoli, an innate immune response is triggered and alveolar macrophages quickly surround and ingest the bacteria trying to prevent infection [10]. Regardless of whether the infection is controlled or progresses, the mycobacteria continue to multiply. In this stage, macrophages produce proteolytic enzymes and cytokines in an attempt to degrade the bacteria [9]. The microorganism continues to grow until reaching sufficient numbers to overcome the cell-mediated immune response. For people with intact cell-mediated immunity, the next response of the organism is the formation of granulomas around the Mtb organisms. In weaker immune systems, the wall formed by the granulomas loses integrity and the bacilli are able to escape and spread in the lungs and eventually also to other organs [9].

TB may develop differently in each patient depending on several host, pathogen and extrinsic factors [11]. The stages of TB disease include latency, primary disease, primary progressive disease and extrapulmonary disease. In the latent stage, the mycobacteria persist in the body but no symptoms occur. In the early infection, the immune system fights the

infection and patients may have fever, paratracheal lymphadenopathy or dyspnea. In this stage, the infection may or may not advance to active disease. In the early primary progressive stage, when the immune system does not control the initial infection, the inflammation of tissues continues. Patients often have fever, fatigue, weight loss and a nonproductive cough develops. In the late primary progressive stage cough becomes productive, patients experience progressive weight loss, rales and anemia. When the disease reaches the extrapulmonary stage, the infection migrates to other locations of the body. The most serious location is the central nervous system, where infection may result in meningitis or space-occupying tuberculomas.

2.1.3 *Diagnosis and Treatment*

The standard method of diagnosis of TB is the examination of a sputum smear for the presence of the Mtb. This approach sometimes requires the detection of the bacilli in a growth solid media for 3 to 6 weeks since Mtb grows slowly. Another option is a method that uses high-performance liquid chromatography to isolate and differentiate cell wall mycolic acids coming from the bacteria in 4 to 14 days can be used [12]. For individuals suspected of having MDR-TB or HIV-associated TB, the WHO recommends that Xpert MTB/RIF be used as an initial diagnosis. The Xpert MTB/RIF assay simultaneously detects Mtb and resistance to rifampin in less than two hours. With this assay, it is possible to purify, concentrate, amplify (by PCR) and identify targeted nucleic acid sequences in the Mtb genome providing real-time detection [13]. Another widely used diagnostic test is the *Mantoux Tuberculin Skin Test (TST)*. This test is performed by injecting 0.1 ml of tuberculin purified protein derivative commonly into the inner surface of the forearm. When correctly administrated a pale elevation of the skin (6-10 mm) should be identified. The skin test reaction should be read between 48 and 72 hours after administration. The interpretation of this test depends mostly on two factors: (1) the measurement of the skin elevation and (2) the person risk of being infected with TB. This test is technically difficult to administer and read often resulting in misleading readings if the tester has insufficient skill [14]. The *Interferon Gamma Release Assay (IGRA)* test is also widely used and consists on a blood test to diagnosis active and latent TB infection. This test is based on the fact that white blood cells from people that have been infected with Mtb will release interferon-gamma when mixed with antigens derived from Mtb. However, despite the simplicity of this test, it does not help to differentiate latent TB infection from TB disease [15].

The standard treatment regimen for TB infected patients includes an induction phase consisting of isoniazid, rifampin, pyrazinamide and either ethambutol or streptomycin as protection against unrecognized resistance to one of the tree core drugs. Once susceptibility to the three core drugs has been confirmed, ethambutol (or streptomycin) can be discontin-

ued. After 2 months of therapy, pyrazinamide can be stopped. Isoniazid and rifampin are continued for 4 more months [16].

Patients in treatment should undergo sputum analysis weekly until sputum conversion is documented. The standard 6-month treatment regimen is a long course of treatment compared with other infectious diseases. The prolonged regimen poses many difficulties including the management of drug toxicity. Thus, it is important to periodically test patients for toxicity [16].

2.2 HUMAN IMMUNODEFICIENCY VIRUS

2.2.1 *Epidemiology*

HIV is the second leading cause of death from a single infectious agent right below TB. In 2017, between 670 000 and 1.3 million people died, which represents a considerable fall since the 2004 peak of deaths, where about 1.9 million HIV related deaths were declared [17]. In 2017, 1.8 million new HIV infections were reported, adding up to a total of 36.9 million people living with the disease [18]. The majority of HIV-positive individuals are located in sub-Saharan Africa, with an estimated 19.5 million infected people, 53% of the estimated incidence worldwide [18]. Globally, HIV-related deaths dropped by 34% and new HIV infections by 47% since the peak in 1996 [18].

2.2.2 *Pathophisiology*

HIV is a lentivirus of the *retroviridae* class of viruses that cause the HIV infection and over time AIDS. There are two types of HIV, HIV-1 and HIV-2. HIV-1 is most virulent and infective than HIV-2 and it is responsible for the majority of infections worldwide. The virus is transmitted through the contact of contaminated body fluids such as semen or blood. The transmission of HIV can also occur between mother and child before or during birth and by breastfeeding. Sexual transmission is the most common way that HIV is passed from person to person, followed by needle sharing between intravenous drug users.

The HIV infection develops in three stages. The earliest stage, the acute HIV infection, develops within 2 to 4 weeks after infection. During this time, commonly people have fever, headaches and rash. In this stage, the virus multiplies rapidly and spreads throughout the body destroying the infectious-fighting $CD4$ cells of the immune system [19]. The level of HIV in the bloodstream is very high in this stage, increasing the risk of transmission. In the second stage, the chronic HIV infection, HIV continues to multiply in the body but at very low levels. People with chronic HIV infection may not have any symptoms, but they can transmit the virus to others. The final and most severe stage of HIV infection is the

AIDS. In this stage, the immunological system is highly debilitated, which potentiates the co-infection by other pathogens, including Mtb.

2.2.3 *Diagnosis an Treatment*

In the early stages of the disease, infected patients may manifest nonspecific symptoms such as fever, lymphadenopathy and myalgia. However, these symptoms usually resolve spontaneously after a period of days or weeks. In addition, these symptoms are also characteristics of other viral infections and consequently, the HIV infection is often missed when specific diagnostic methods are not used [20]. In the chronic state, HIV infected individuals may manifest persistent lymphadenopathy, weight loss, fever, peripheral neuropathy and dementia.

HIV diagnosis is performed by recurring to blood or body fluid collection, which is posteriorly analyzed by enzyme immunoassays, nucleic acid amplification tests or western blot analysis. In the fourth-generation enzyme immunoassays, recombinant antigens capture anti-HIV antibodies, immunoglobulin G and immunoglobulin M using antihuman antibodies plus direct determination of p24Ag, allowing recognition of HIV infection prior to seroconversion. The nucleic acid amplification tests can detect the virus by recurring to specific amplification primers, which is followed by RNA detection by using labeled probes. In the western blot assay, the selected antibodies bind to fixed HIV proteins, which then create a pattern that can be read as positive, negative or indeterminate.

Improvements in diagnosis technologies are crucial in preventing the spread of the disease. Some of the available diagnostic assays lack in sensitivity and specificity or do not detect early HIV infection or AIDS, contributing to the spread of the disease [21].

HIV is an incurable disease. Treatment against HIV infection is called *Antiretroviral Therapy (ART)*. ART reduces plasma viral load allowing the normalization of $CD4^{+}T$ cell levels. However, if the treatment is stopped a resurgence of virus counts occurs. The conventional ART consists of the administration of three or more drugs to target different steps in the HIV replication cycle. ART does not cure HIV, however, it helps people to live longer, healthier and also reduce the risk of transmission.

2.3 HIV/MTB COINFECTION

2.3.1 *Epidemiology*

HIV and Mtb potentiate each other, accelerating the deterioration of the host immunological function [7]. In fact, the risk of developing active TB is estimated to be between 16-27 times greater in HIV-positive than in HIV-negative individuals [8]. TB remains the leading cause

of death among people living with HIV [7]. In 2016, there were 464 633 reported cases of HIV-positive TB [8]. About 0.3 million HIV-positive people died from TB, almost half of the total number of HIV related deaths [7]. From the estimated 10.4 million people that fell ill with TB in 2006, 10% were people living with HIV (74% in Africa) [8].

2.3.2 *Pathogenesis*

TB and HIV infection impact the pathogenesis of each other. HIV infection alters the course of TB infection, increasing substantially the risk of developing active TB. On the other hand, it is known that TB increases the levels of HIV replication, propagation and genetic diversity [22]. $CD4^+T$ cells are central in controlling the TB infection, thus the decreasing number of this cells during HIV infection contributes to the increased risk of developing TB or the reactivation of latent TB. HIV also leads to a dysfunctional granuloma formation, making the immune system unable to combat Mtb. Full knowledge of the pathogenesis of interaction between TB and HIV infection is not yet achieved. Nevertheless, it is clear that treatment of HIV with ART reduces the risk of TB.

2.3.3 *Diagnosis and Treatment*

HIV-Mtb coinfection presents several challenges related to diagnosis and treatment. HIV-positive patients infected with TB often lack symptoms associated with TB, leading to missed cases. At higher cell levels of $CD4^+T$ cells, it is more common to manifest pulmonary TB. Extrapulmonary and miliary TB becomes more prevalent at lower $CD4^+T$ cell counts. Furthermore, granuloma formation is disrupted as $CD4^+T$ cell counts decline and cavitary lesions may not be seen.

The treatment of HIV-Mtb co-infected patients may be difficult, as drug interactions between ART and TB chemotherapy may cause some unfavorable results. Co-toxicity and increased side effects deteriorate even more the patient health. Coinfected patients are advised to only start the TB treatment 2-8 weeks after ART, depending on the degree of immunodeficiency [23].

Coinfection increases the risk of misleading diagnosis. The detection of TB in HIV-positive patients is also more difficult. Several studies have been conducted to improve the screening of TB in HIV infected patients [24, 25]. Some of these studies revealed good results with a sensitivity of about 79% in HIV-positive people using the Xpert MTB/RIF diagnosis assay [26]. However, there is still a need for efficient methods for diagnosing TB in HIV positive individuals.

3

PROTEIN-PROTEIN INTERACTIONS

3.1 POTEIN-PROTEIN INTERACTION DATA

Since the sequencing of the first human genome carried out by the Human Genome Project and Celera, omics acquired a crucial importance in further theoretical and practical advances within the field of genetics [27]. Proteomics succinctly is the systematic study of the functions, structures and interactions of proteins with the aim of providing detailed descriptions of the structure, function and control of biological systems in health and disease [28].

A particular field of proteomics focuses on the nature and role of interactions between proteins. Most biological processes including transcriptional activation/repression, immunity, metabolism, signaling cascades and biochemical pathways are mediated through protein interactions. Hence, there is a need to understand the chaotic network that forms these processes in order to achieve a better understanding of human diseases and therefore devise better therapeutics [28, 29, 30].

PPIs are commonly understood as physical contacts with molecular docking between proteins in a living organism in vivo [31]. PPIs play diverse biological roles and differ based on the composition, affinity and lifetime of the associations [28]. Biological factors such as cell type, cell cycle phase and state, development state, protein modifications, the presence of cofactors and presence of other binding partners also influence PPIs [28].

PPIs can be categorized based on their structural and functional characteristics. Based on their composition PPIs can be classified as homo- or hetero-oligomeric. If a PPI occurs between identical chains, it is classified as homo-oligomeric, whereas if it takes place among non-identical chains is classified as hetero-oligomeric [30]. PPIs can also be classified as obligate or non-obligate based on their stability. In an obligate PPI, the monomers are not stable structures on their own, whereas the components of a non-obligate interaction can exist independently [30]. As a measure of their persistence PPIs can be classified as transient or permanent. Permanent PPIs are usually very stable and irreversible. Transient interactions are formed and broken easily and can be classified as strong or weak. Weak transient interactions are characterized by a micromolar dissociation constant and lifetimes

of seconds [32]. Strong transient interactions may last longer, have a dissociation constant in the nanomolar range and can shift from an unbound/weakly bound to a strongly bound state when triggered by, for example, ligand binding [32].

It has been shown that mapping PPIs can provide valuable insights into protein function and molecular mechanisms of cellular processes by facilitating the modeling of functional pathways [29, 33, 34].

In recent years, large amounts of PPI data have been generated by high-throughput experimental methods, such as two-hybrid systems, mass spectrometry and protein chip technologies. Analyze such amounts of data presents itself as a challenge to experimental investigations [28]. Thus computational methods for PPI networks analysis has become a necessary tool for understanding the function of uncharacterized proteins and achieve better knowledge of biological processes.

3.2 Experimental Approaches to the Generation of Protein Interaction Data

Proteins seldom act alone but rather execute their functions through interactions with other proteins. As protein and their interactions are fundamental for most biological processes, it is essential to discover and characterize these protein interactions in order to understand the molecular mechanisms of underlying biological processes [35]. This chapter is intended to provide an overview of the most used experimental methods to generate PPI data.

In the beginning, PPIs data was generated via intensive small-scale experiments, yielding a small and limited number of PPIs and consequently providing a data set of limited size. These data, presented in individual research papers, can be considered to be fairly reliable since it was subject to stringent controls and evaluation in the peer-review process. However, the development of high-throughput approaches such as the Y2H system, MS and protein microarrays, allowed the generation of vast arrays of PPI data. These high-throughput methods generate a high amount of false positives. Thus, validation of the data generated by these methods is recommended.

This section will focus on experimental approaches to generate PPIs, reviewing some of the most important and most used techniques to generate PPI data.

3.2.1 *The Y2H System*

Perhaps the most common approach to the detection of interacting proteins *in vivo* is the Y2H system, also known as ”interaction trap”. The Y2H system is a molecular-genetic approach that facilitates the study of PPIs and was firstly described by Fields and Song [36]. This procedure is carried out by screening a protein of interest against a set of potential

protein partners inside the yeast nucleus. The term ”two-hybrid” derives from the two ”hybrid” proteins created by the fusion of each one with a *DNA Binding Domain (DBD)* and an *Activation Domain (AD)* of a *Transcription Factor (TF)*. The protein fused to the DBD is denominated as ”bait” and the protein fused to the AD as the ”prey” [37]. After the interaction between the bait and the prey, the DBD and AD are brought into sufficient proximity to switch on the reporter gene resulting in gene transcription[28]. This process starts a growth or color reaction that can be detected in specific media. Figure 2 shows a schematic version of the Y2H system.

In large-scale screenings, it is common to screen multiple baits against a library of preys. This method is denominated library screening approach. Another commonly used approach in large-scale screening is the matrix approach, in which an array of defined preys is substituted for the library [38].

The Y2H method has several limitations and therefore in some cases it is not possible to detect if there is an interaction between the target proteins. Such limitations include the fact that some proteins do not interact in the yeast nucleus, that many interactions are triggered by post-translational modifications not available in yeast and that some proteins are toxic to yeast [39]. However, despite the fact that Y2H present various limitations, it was rapidly accepted by the scientific community which led to dozens of thousands of publications, remaining the most popular method when it comes to discovering novel protein interactions.

3.2.2 *MS Aproaches*

Another approach to PPIs detection uses quantitative MS to reveal the composition of entire protein complexes. One of the most common applications relies on the combination of *Affinity Purification and Mass Spectrometry (AP-MS)*. In this technique, on the affinity purification step, proteins of interest are expressed in-frame with an epitope tag, which is then used as an affinity handle to purify the tagged protein (the bait) along with its interacting partners (the preys). Then the bait proteins are systematically precipitated along with any associated proteins onto an affinity column. In an optional step, the purified protein complex can be resolved by one-dimensional SDS-PAGE, so that proteins are separated by their mass. Then the protein bands are separated by protein size and posteriorly the protein bands are digested with trypsin. Finally, the component proteins are detected by MS and analyzed with bioinformatic tools. Figure 3 shows a schematic version of the AP-MS experiment.

The advances in MS approaches made over the past few years now enable the discovery of PPIs within protein complexes on a proteomic scale. The knowledge gathered from such projects may then be applied on drug target discovery, validation pipelines, to elucidate

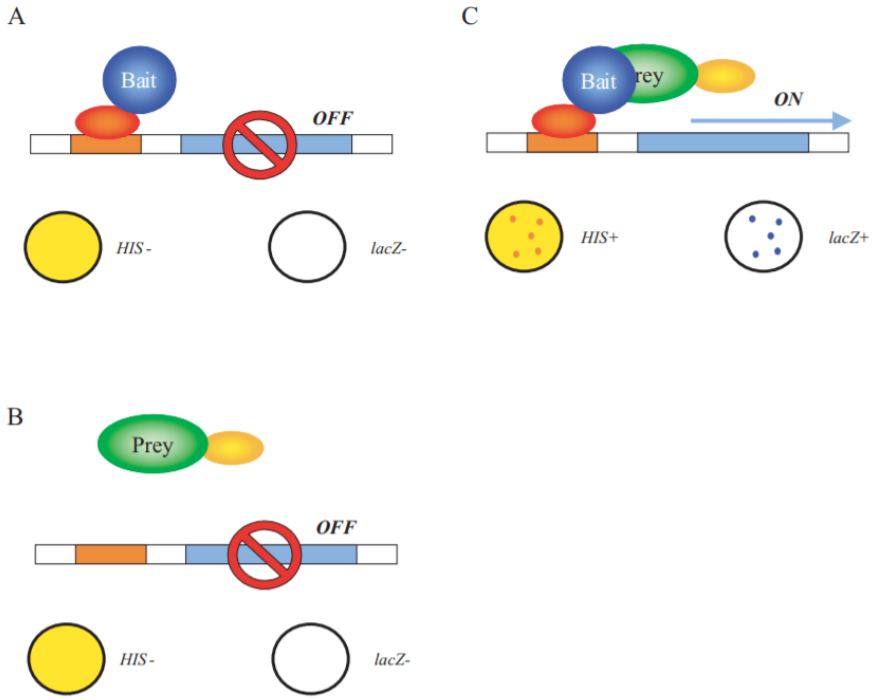


Figure 2: **The yeast two-hybrid system [38].** (**A**) A bait is expressed as a fusion to a DBD. The DBD-bait binds to the operator sequences present in the promoter region upstream of the reporter gene but does not activate its transcription since the DBD-bait does not contain an activation domain. (**B**) A prey is expressed as a fusion to an AD. The AD-prey fusion has the capability to activate transcription but because it is not actively targeted to the promoter it does not activate transcription of the reporter gene. (**C**) The interaction between bait and prey targets the AD-prey fusion protein to the promoter, thereby reconstituting an active transcription factor. The hybrid transcription factor is bound to the promoter upstream of the reporter gene and therefore activates transcription. The readout of the activated reporter gene is measured either as growth on selective medium or in a color reaction (lacZ). Only the DBD-bait or the AD-prey on its own do not allow growth on selective medium (HIS-) and do not display blue staining in a color assay (lacZ-), whereas an interacting DBD-bait and AD-prey display growth (HIS+) and blue color (lacZ+).

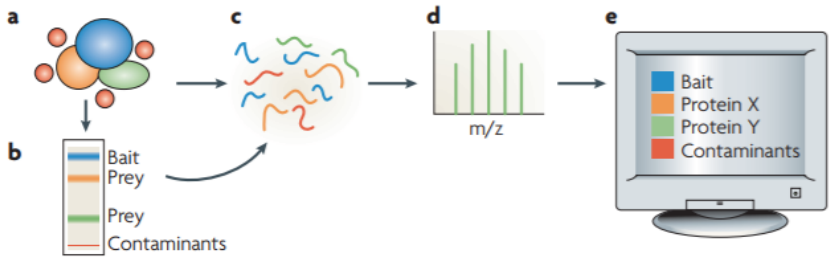


Figure 3: **A general overview of an affinity purification and mass spectrometry experiment [40].** (**a**) The protein of interest (often epitope-tagged, blue) is purified from a cell together with its binding partners (orange and green). Contaminants (red) can also be present. (**b**) In an optional step, proteins in the complex can be separated by SDS-PAGE. (**c**) Proteins are subjected to proteolysis (usually with trypsin). (**d**) Mass spectrometry (MS) analysis of peptides. In most cases, this involves peptide separation by reversed-phase liquid chromatography followed by two MS events: in the first scan, the mass/charge ratio (m/z) of the intact peptide is measured. The most abundant peptides are then specifically selected and subjected to fragmentation, yielding a tandem MS (MS/MS) spectrum (a simplified MS/MS scan is shown for one of the peptides). (**e**) Database searching and statistical software are used to interpret the MS data to yield a list of proteins that were present in the initial sample, including the tagged protein, its interacting partners and contaminants.

systematically pathways and the functional context in which proteins operate in a variety of organisms and cell types [41].

3.2.3 *Protein Microarrays*

Protein microarrays, also known as protein chips, is a high-throughput technology that allows a fast, straightforward and efficient screening of protein-protein interactions alongside with a vast amount of other information during a single assay [42]. The first applications of microarrays were centered on DNA-related applications. To retrieve information about proteins a similar approach was developed, protein microarrays. The key advantage of microarrays is the use of nonporous solid surfaces, such a glass, that allow precise deposition of molecules in a highly dense and ordered way. Protein microarrays allow the examination of protein expression levels and the acquisition of quantitative and qualitative information about proteins of interest [28].

A protein microarray consists of a miniaturized and parallel assay system that contain small amounts of purified proteins in a high-dense and ordered format. Protein microarrays are commonly prepared by immobilizing proteins onto a microscope surface. After the proteins immobilization, they can be probed for a variety of functions including for PPIs inference (functional protein microarrays). Finally, the resulting signals are measured

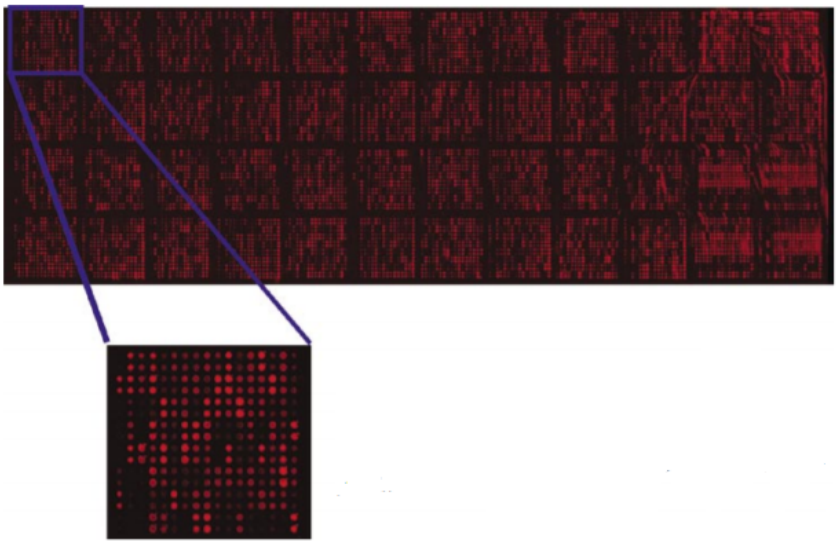


Figure 4: **A typical microarray image [43].** Each spot corresponds to one of the organism's thousands of proteins. The intensity of the dot indicates the amount of protein present. An enlarged image of one of the 48 blocks is depicted below the protein chip.

by detecting fluorescent or radio.isotope labels [43]. Figure 4 shows a typical microarray image.

An advantage of the generation of PPIs in an array format is that the conditions of the experiment can be controlled by the investigator. This includes not only factors such as pH, temperature, ionic strength and the presence or absence of cofactors but also the modification states of the proteins under investigation. Another important advantage of protein microarrays is that thousands of proteins can be spotted on a single slide, enabling the generation of multiple PPIs with minimal sample consumption. In addition, hundreds or even thousands of copies of an array can be fabricated in parallel, enabling the same proteins to be probed repeatedly with many different molecules under many different conditions. These features make microarray technology well suited to PPIs prediction [44]. However, the correct interpretation of data obtained from protein microarrays to create extensive networks of PPIs remains a considerable challenge.

3.3 Computational Methods for the Prediction of Protein-Protein Interactions

Experimental techniques are applied to generate PPI data. However, these techniques are time and money expensive and require some expertise. In addition, there are some discrepancies between the data generated by different experimental techniques [45]. Computa-

tional approaches can offer an alternative to these experimental techniques. Computational techniques have been applied to the generation, indexing, validation, analysis and extrapolation of PPI data [46, 47, 48]. This section will focus on the computational prediction of PPI, reviewing a number of techniques including genome-scale, sequence-based, structure-based, learning-based and network topology-based approaches.

3.3.1 *Genome-Scale Approaches*

The sequencing of complete genomes for various organisms has enabled the prediction of PPIs at a genomic scale. Genomic-scale approaches typically perform a comparison of gene sequences across genomes and are often justified on the basis of the correlated evolutionary mechanisms of genes [28].

One of these approaches is the gene neighborhood conservation. The main idea of gene neighboring is that related genes are located close to each other in the genome and proteins encoded by these genes may physically interact with each other [49]. This method searches if two genes are neighbors in organism X also if their orthologs in organism Y are also neighbors. This may imply that protein-protein interactions impose evolutionary constraints to keep genes together [50]. Although simplicity makes this method very attractive, it may produce some false negative results because distantly located genes may not be recognized as interactors even if they encode proteins that interact with each other [49].

Protein phylogenetic profiles are also used to predict PPIs. The phylogenetic profile of a protein is a binary vector that represents the presence or absence of a protein across a set of organisms. The main idea of this method is that functionally related genes remain together across organisms. This method is an "upgrade" of the gene neighboring method which can detect some interactions that gene neighboring may fail to detect [49]. However, this method presents some drawbacks such as the fact that it can only be used with complete genomes, it can not be applied to essential proteins and the number and distribution of the target genomes can influence the results drastically [49]. This method is experimentally validated as it has been shown that proteins with similar phylogenetic profiles are likely to be functionally linked and to interact with each other [51, 52].

The gene fusion method is also commonly used to predict PPIs. This method predicts a functional link between two proteins when they are separated in one organism and fused in one protein in another organism [50]. A major advantage of this method is its reliability since gene fusion events are very informative about the functional relationship. On the other hand, fusion events are not abundant, especially in prokaryotes [49].

3.3.2 *Sequence-Based Approaches*

Another approach to PPI prediction is based on information regarding sequential homology. This approach was introduced by Mattews et al. [53] and is based on the concept that an interaction in one species can be used to predict an interaction in another species [28]. This approach consists of a systematic search of interologs, potential orthologs of known interacting proteins partners, to identify potentially conserved interactions in different organisms.

In 2001, Wojcik and Schachter [54] proposed a sequence-based prediction approach that takes into account the domain profiles of proteins. The domain information of each interacting protein in one species may help predict interactions in another species, as interactions generally occur between protein domains. In this method, PPI data for a source organism is transformed into a domain interaction map to construct a domain profiles with the multiple alignments of the domain sequences for each cluster. Two domain clusters are connected if the number of interactions between them has a value above a threshold. Finally, each domain cluster is mapped to a similar set of proteins in a target organism. The predicted PPIs are then based on the connectivity between domain clusters.

3.3.3 *Structure-Based Approaches*

Structure-based approaches allow more detailed analysis of protein interactions than the genome-scale and sequence-based approaches. Structural approaches can determine, not only whether two proteins interact, but also the physical characteristics of the interaction, and residues at the protein interface which mediate the interaction [55]. A classic structure-based approach for detecting PPIs is the docking method which detects PPIs by predicting the structure of docked protein complexes. The detection of docked proteins is performed firstly by developing a scoring function that can discriminate between correctly and incorrectly docked orientations and then a search method is applied to identify correctly docked orientations with reasonable reliability [28] The algorithm applied in this method, searches for protein complexes by treating proteins as rigid bodies and generates a list of possible docked complexes. These complexes are scored based on the energy of their association, i.e. the evaluation of statistical potentials, electrostatics and hydrogen bonding. Optionally it is possible to introduce flexibility through side-chains rearrangements.

In 2002, Aloy and Russel [56] shown that protein complexes with known three-dimensional structures offer better conditions to identify reliable PPIs. This can be achieved by modeling putative interactions upon three-dimensional protein complexes and then determining the compatibility of the proposed interaction and the complexes. This compati-

bility is measured by empirical potentials using molar-fraction random state models based on the observed tendency of residues to persist on protein surfaces.

Alternatively, Aytuna et al. [57] proposed an algorithm that takes into account the similarities in interface surfaces. This algorithm starts with a set of structurally known protein interfaces and searches for pairs of proteins having similar residues. Proteins with similar residues have more probability of interacting with each other.

3.3.4 *Network Topology-Based Approaches*

PPINs can be useful resources to predict new or to identify the reliability of protein interactions. Topological features of PPINs can provide valuable information on this task. For example, Goldberg and Roth [58] proposed the use of clustering coefficients based on neighborhood cohesiveness. The main idea of this method is that two proteins are more likely to interact if they share many interacting neighbors, thus having mutual clustering coefficients. This property is very common in small-world networks, which is the case of PPINs.

Based on the idea that interactions involving proteins that have many interacting partners are likely to be false positives and that highly interconnected sets of interactions or interactions forming a loop are likely to be true positives, Saito et al. [59] proposed an *Interaction Generality Measurement (IG1)*. This measurement is defined as the number of proteins that directly interact with a target protein pair, as reduced by the number of proteins interacting with more than one protein. This is a local measurement as only considers the direct neighbors of a protein. However, in a subsequent work [60] the authors extended this measured to consider the topological properties beyond the direct neighbors of a protein. Other methods based in the IG1 was also proposed by other authors. For example, Chen et al. [43] presented the *Interaction Reliability by Alternative Path (IRAP)* approach to measure the reliability of an interaction in terms of the alternative path. This approach uses the reversed and normalized IG1 as initial edge weights to reflect the local reliability of each interaction in a PPIN.

Probabilistic weighted interaction models are also commonly used to estimate the probability that a pair of proteins interact directly and stable [61, 62, 63].

3.3.5 *Learning-Based Approaches*

Various ML techniques are recognized as useful and reliable methods for the prediction of PPIs. Given a set of known interactions, an ML system can be trained to recognize interactions based on specific biological features. The prediction of PPIs can be defined as a classification problem. Therefore, multiple ML techniques can be applied to determine

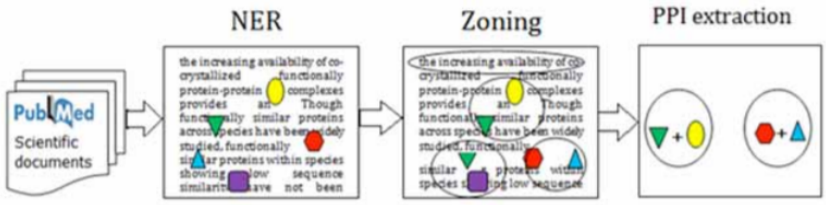


Figure 5: **A schematic version of text mining approaches for protein-protein interaction prediction [49].** In general, each literature mining system consists of three steps: Named Entity Recognition or NER step, it does the identification task of protein. Zoning step, in this step the text split into basic building blocks and extract sentences from the text. PPI step that uses various algorithms to infer protein-protein interaction.

whether a pair of proteins are interacting or noninteracting. However, sometimes datasets are not balanced, i.e. there are much more noninteracting than interacting PPIs, and may be noisy and contain missing values [64]. Therefore, the selection of an adequated classification technique is a challenging and important task.

The most common ML techniques include *Support Vector Machine (SVM)*, *K-Nearest Neighbour (KNN)*, *Naive Bayesian (NB)*, *Artificial Neural Networks (ANN)*, *Decision Tree (DT)* and *Random Forest (RF)*. A brief summary of the main features of these techniques is presented in Table 1.

Text mining techniques can also be applied to extract useful knowledge from large data sources and therefore to predict PPI. In general, literature mining systems are performed in three steps. The *Name Entity Recognition (NER)* step, consists of the identification of protein names. The second step, the zoning, the text is split into basic building blocks and sentences are extracted from the text. Finally, the protein-protein interaction extraction, where multiple algorithms are used to infer a possible PPI. Text mining approaches that use machine learning may not be as reliable as manually curated data, but the fast growth of publications in this field can make these methods a valuable resource. Figure 5 shows a schematic version of text mining approaches for PPI prediction.

3.4 Computational Analysis of Protein-Protein Interactions

PPINs can be described as complex systems of proteins formed by biochemical events and electrostatic forces that serves multiple biological functions. At a computational level, a PPIN is commonly represented mathematically by a graph G = (V, E) consisting of nodes (V) and edges (E). In a graph, proteins are represented as nodes and interactions as edges. Two proteins that interact with each other are represented as adjacent nodes and are connected by an edge.

Table 1: A summary of ML approaches for PPI prediction. Advantages and limitations for each approach are presented along with a reference to studies where they have been applied.

Classifier	Description	Reference
NB	• Simple and easy to interpret. • Ability to handle diverse heterogeneous data. • Copes well with missing values. • Assumes conditional independence between datasets. • Performance deteriorates when dependencies between features exist.	[65], [66]
KNN	• Simple to understand. • Requires no training. • The computational cost and memory requirement grows rapidly with increasing feature vectors dimension.	[67], [68]
SVM	• Can handle non-linear separable datasets. • Copes well with high-dimensional data. • It is very powerful. • The parameters can greatly affect the results.	[69], [70]
RF and DT	• Can handle missing values. • Copes well with high-dimensional data. • Can integrate diverse heterogeneous data. • The patterns in data can be easily explained.	[71], [72]
ANN	• Good generalization capabilities. • Ability to recognize complex patterns.	[73], [74]

Mining the organization of PPINs can yield a variety of insights [75]. For example, the assignment of putative roles to uncharacterized proteins as neighboring proteins in a graph are generally considered to share functions ("guilt by association") [28]. PPINs can also provide details about the steps within a pathway or help to characterize the relations between proteins of multi-molecular complexes [75, 76, 77]. In addition, densely connected subgraphs in the network are likely to form protein complexes that function as a unit [28]. Other characteristics can also be explored as for example the topological features and centralities of the network to enhance our understanding of the biological system.

In general, although graphs are well-known structures, the computational analysis of PPINs is challenging, with these major difficulties being commonly encountered:

- The high number and heterogeneity of nodes and edges complicate the network visualization [5].
- Some protein interactions are not reliable. Experimental methods for interactomics such as Y2H assays produce high amounts of false positives (physical interactions detected in the screening method that are not reproducible in an independent system) and false negatives (protein-protein interactions undetected by the screening method) [78].
- Some proteins can have several different functions and consequently belong to multiple functional groups. Classic clustering approaches usually produce disjoint clusters, allowing a protein to belong to only one cluster, therefore impeding a realistic assignment of multifunctional proteins to clusters [79].
- Two proteins from different functional groups frequently interact with each other increasing the topological complexity of the PPIN, posing difficulties to the detection of unambiguous partitions [28].

A common approach to understand and characterize complex networks is the exploration of the topological features of such networks [80, 81, 82]. These features include small-world properties, scale-free degree distributions and hierarchical modularity.

3.4.1 *Topological Features of Protein-Protein Interaction Networks*

In 1957, Rapoport [83] studied perhaps the simplest useful model of a network, the random graph. In this model, a fixed number n of vertices are randomly connected by undirected edges to create a network. Networks of different types can be distinguished by their degree distributions. The degree distribution P(k), a concept introduced by Barabasi and Oltvai [84], quantifies the probability that a selected node in a network will have exactly k links. Random networks follow a Poisson distribution. However, real-world networks are not

like random graphs. Real networks can be classified by their degree distribution, PPINs typically follow a power-law degree distribution $P(k) \sim k^{-\gamma}$. These networks are known as scale-free networks and are characterized by having few nodes with many edges and many nodes with few edges [85]. Biologically this means that most proteins participate in only a few interactions, while a small set of hubs, highly connected nodes, participate in dozens of interactions. Another characteristic of PPINs is that hubs rarely directly link to each other [86].

PPINs are also characterized by a property called "small-world effect", which states that any pair of vertices can be connected through a short path of a few links [87]. Although random networks show the small-world property, scale-free networks have path length much more small and are called "ultra-small" [88]. This short path length indicates that local perturbations in metabolite concentrations could affect an entire network very quickly [28].

Multiple centrality measures can be used as an evaluation of the importance of components in a network. Four of the most representative centrality measures are the *Degree Centrality (DC)*, *Closeness Centrality (CC)*, *Betweenness Centrality (BC)*, and *Eigenvector Centrality (EC)*.

- DC of a node measures the number of edges of the node. This measure can be helpful to identify nodes with high importance in the network. However, this measure does not take into account the rest of the network and the importance given to this measure strongly depends on the network size. For example, in a study conducted by Joeng et al, the authors found that proteins with high DC (hubs) are three times more likely to be essential than proteins with only a small number of links [89].

- BC of a node measures the number of shortest paths in a network that pass through the node. Proteins that have high BC participate in many more interactions than others, thus these proteins are more likely to be essential to the organism survival than proteins with fewer interactions [90].

- CC of a node measures how close the node is to other nodes in the network. The distance between two nodes is defined as the length of the shortest path between them. The smaller the sum of the distance of the node to all other nodes, the higher its closeness is [91].

- EC measures the centrality of a node taking into account that each neighbor will have a different weight in the centrality of the node. This notion is known as "prestige" in social networks. A node is considered more central if it is connected to many central nodes. Thus, this measure takes into account not only the quantity but also the quality of node connections [91].

The unique topological features that characterize each PPIN play an important role in the computational analysis of these networks providing crucial information for inferring functional properties of the network.

3.4.2 *Modularity Analysis*

Complex networks can usually be decomposed to several highly inter-connected sub-units. The identification of these sub-units is very important as they may help to discover the unknown function of these functional modules [92]. A functional module in a PPIN may represent a set of functionally associated proteins that are involved in a given biological process or function [28]. Several graph-based approaches have been employed to identify functional modules in PPINs. However, due to the presence of unreliable interactions and the fact that real functional modules are overlapping, i.e. one protein may participate in multiple biological processes or functions, these approaches tend to be limited in accuracy [93].

The identification of functional modules in PPINs can be successfully accomplished through the use of cluster analysis. In general, modularity detection can be performed by three primary approaches [94]: (i) divisive (or top-down) techniques, in which initially the entire graph is considered as a cluster and then edges are successively removed to detect other clusters until further division is no longer worthwhile; (ii) agglomerative (or bottom-up) techniques, in which modules are constructed by adding elements to an initial seed, i.e. initially every vertex is in a separate cluster and successively pairs of clusters are merged until the clustering can no longer be improved; (iii) force-directed methods, in which nodes belonging to the same module are considered to be spatially close.

Several studies have been conducted and multiple algorithms aiming at modularity analysis have been developed [95, 96, 97]. However, each algorithm has its own advantages and disadvantages being able to exhibit good and bad performances in different cases. The main challenges for PPINs modularity analysis are: (i) the fact that PPI data yield a significant amount of false positives and miss a high fraction of existing interactions; (ii) clusters may overlap each other and therefore traditional cluster approaches that assign each protein to a single cluster do not suit this problem well; (iii) the recent advances in high-throughput techniques have led to a huge amount of PPI data making the computational clustering difficult; (iv) it is difficult to determine the number and size that each cluster should have.

Topological metrics can also be incorporated into modularity analysis of PPINs in order to be able to retrieve more accurate conclusions. For example, bridging nodes in PPINs serve as the connecting nodes between protein modules. Therefore, the removal of the bridging nodes yields a set of components disconnected from the network. Thus, using this

measure can be an excellent preprocessing procedure to estimate the number and location of modules in a network.

3.5 APPLICATIONS

As stated before, the analysis of PPI and PPIN can provide valuable insights into the cellular organization, processes and functions. There are many applications following this analysis. Some of the main applications in which the discovery and exploration of PPI can be applied are described below:

- **Identifying protein and protein complexes and predicting their functions.** The most basic application of PPINs is the prediction of protein functions. This prediction generally relies on the fact that interacting proteins may belong to at least one common functional class, and thus knowledge of the function of a subset of the proteins involved in the network may lead to an accurate prediction of the function of the remaining subset of uncharacterized proteins [98]. Asthana et al. [61] proposed the use of semantic similarity and semantic interactivity to measure the reliability of PPIs based on *Gene Ontology (GO)* annotations to construct weighted PPINs. Then flow-based modularization algorithms are used to identify overlapping modules in the weighted PPINs.

- **Essential protein identification.** Essential proteins are indispensable to maintain normal function of life activities in living organisms. These proteins play important roles in the studies of pathology, synthetic biology and drug development [99]. Topological analysis of PPINs can be used to systematically assess the biological importance of bridging and other nodes in a PPIN. For example, by integrating network topology with gene expression profiles, Li et al. [100] and Tang et al. [101] exploited the Pearson correlation coefficient and edge clustering coefficient to weight PPIs and predicted essential proteins on the weighted PPINs.

- **Discovering signal transduction pathways.** Many approaches have been proposed to discover signal transduction pathways from PPINs. Shlomi et al. [102] developed an algorithm, QPath, that given a query pathway and a network of interest, searches the network for homologous pathways, allowing both insertions and deletions of proteins in the identified pathways. The identified homologous pathways are scored according to their variation in terms of the protein insertions and deletions against the query pathway, the sequence similarity between their constituent proteins and the query proteins and the reliability of their constituent interactions. Gitter et al. [103] proposed an alternative method to discover signal transduction pathways. Taking into account that pathways are directed and PPIs are undirected, they developed three algorithms

based on either weighted Boolean satisfiability solvers or probabilistic assignments to formalize the orientation problem in protein interaction graphs as an optimization problem. To construct weighted PPINs, they used two weighted schemes. The first weighting scheme is to increase the weights for those interactions that are supported by multiple databases. The second one is based on the type of experiments used to detect the interactions. The discovered paths may match several known signaling pathways and suggest new mechanisms that are not currently present in signaling databases.

- **Disease gene prioritization.** The analysis of PPINs could provide biological insights into disease mechanisms. Chen et al. [104] used the PPI data and confidence scores inferred from *Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)* [105] to discover new genes related to brain development. By applying a shortest path algorithm on a weighted graph, they identified new candidate genes falling in at least one of the shortest paths connecting two known genes that are related to brain development. I. To identify potential drug target proteins, Li et al. [106] applied graph theory to the human PPIN, in which proteins were weighted by descriptors of protein primary structure and PPIs were weighted by the confidence scores. Proteins with higher weights were considered to be stronger potential drug targets.

3.6 BIOLOGICAL DATA INTEGRATION

Over the last years, the rapid technological advances have led to a exponential growth of biological data. Therefore data management assumes a key step in present biological projects ensuring proper data sharing, integration and annotation. Numerous databases containing biological knowledge are available online and are daily assessed by scientists all over the world. However, the use of different databases to achieve a better understanding of any biological question is not a straightforward task. The information spread across these repositories is stored in multiple and disparate formats and nomenclatures [107]. These issues lead to an incomplete use of all the knowledge made available as scientists most of the times only use information of one source since analyses based on multiple sources are more difficult and subjected to errors. For example, the same protein might have different identifiers in each repository and the lack of mapping between these identifiers make the integration of this information almost impossible [108].

Data integration frameworks can be classified into two major categories, eager and lazy [109]. In the eager approach, the data are structured into a global schema and stored in a single data repository. In the lazy approach, data remains on different sources and are integrated on demand based on a global schema used to map the data between sources or are made available through hyperlinks to the original sources. Both approaches have their

advantages and disadvantages and the choice between them depends on multiple aspects such as the amount of data, their availability and the quality of existing infrastructures.

In the case of PPI data integration, there are two main aspects to take into account. First is collecting data from as many sources as possible in order to create a unified interaction network. The second is to combine the PPI data with other types of biological data.

3.6.1 *Protein-Protein Interaction Data Integration from Multiple Sources*

With the emergence of high-throughput methods for PPI detection multiple public repositories to store these data were created. However, the data either directly submitted to these repositories or manually curated from the literature is stored in multiple and disparate formats and with different and ambiguous protein identifiers. These problems prevent the proper unification of these data and, consequently, a greater coverage of the interactome space.

In recent years, the *International Molecular Exchange Consortium (IMEx)* [110] is promoting a standard format, the *Protein Standards Initiative for Molecular Interactions (PSI-MI)* data exchange format, in an attempt to uniform the way interactions are formatted and codified. This data format was adopted by some PPI databases including *Molecular Interaction Database (MINT)*, *Database of Interacting Proteins (DIP)* and *Molecular Interaction Database (IntAct)*. However, despite the efforts to adopt this data format most databases do not follow it in a uniform manner or did not adopt it at all. Thus, using all available PPI data in an integrated way is still not trivial for most computational biologists.

Despite the challenges, it is essential to perform PPI data integration in order to obtain high confidence networks. Interactions present in multiple sources and detected in multiple and by multiple detection methods are more reliable than those detected by just one experiment.

3.6.2 *Combining Protein Interaction Data With Other Types of Biological Data*

The combination of PPI data with other types of biological data is a key aspect in the analysis of the networks formed by these interactions. Not only it will increase the reliability of the interactions but will also provide valuable additional knowledge that interaction data did not contain.

High-throughput methods for detecting PPI generate high rates of false-positives and false-negatives. One way of verifying the reliability of an interaction could be through the use of information contained in databases like *Universal Protein Resource (UniProt)* [111]. Proteins that are never colocalized are not likely to interact with each other. Shared phe-

notypes, correlated expression and shared GO terms can also support the reliability of an interaction.

The use of known inhibitors targeting PPI could also bring valuable insights to the analysis of the networks formed by these interactions. Studying how these inhibitors perturb PPINs can be very useful in understanding disease mechanisms. However, targeting PPIs has been extremely challenging to convert into therapeutics with only a few compounds following for trials over the last years [112]. Some databases like TIMBAL [113] and *Inhibitors of Protein-Protein Interaction Database (iPPI-DB)* [114] were implemented over the last few years and store data on small-molecule inhibitors of PPIs. These databases can be a useful source of information to use together with PPI and other biological data. Other types of biological data like gene expression data can also be used to identify essential proteins [115], identify protein complexes and functional modules [100, 116] and discover disease signaling pathways and regulatory networks [117]. This kind of information provides valuable biological insights that otherwise cannot be achieved with the use of PPI data alone.

3.7 Public Protein-Protein Interaction Databases

In the last years, high-throughput technologies have generated massive amounts of PPI data of various organisms. These data are currently stored in several databases. The majority of the data contained in these databases are manually curated from literature data-mining. However, there is none or little data standardization among these databases, with each presenting different data structure, format and mode of description. Currently, more than 100 PPI related repositories are available online [110]. Some of the major open PPI databases will be described as follows:

- The ***Biological General Repository for Interaction Datasets (BioGRID)*** [118] is an open-access database that focuses on the manual curation of experimental validated genetic and protein interactions that are reported in peer-reviewed biomedical publications. At this moment (December 2017), it comprises more than 1.4 million curated interactions derived from over 57 thousand publications. This database represents interaction records for 66 model organisms, including humans, with a recent emphasis on central biological processes and specific human diseases. All data contained in this database can be consulted in https://thebiogrid.org/.

- **MINT** [119] stores information about molecular interactions extracted from experimental works published in peer-reviewed journals. Genetic and computational inferred interactions are not included in this database. All data contained in MINT are manually curated and can be freely accessed online at http://dip.mbi.ucla.edu/dip/. In addition, MINT includes a separated database of human protein interaction data,

the HomoMINT. Currently (December 2017), MINT comprises more than 27 thousand curated interactions from 255 organisms derived from over 1964 publications.

- **DIP** [120] combines data from a variety of sources to create a single and consistent set of PPI. The data stored in the DIP database were curated, both, manually and using computational approaches. At this moment (December 2017), DIP comprises more than 81 thousand curated interactions from 834 organisms derived from over 8233 publications. This database also provides tools that allow users to analyze, visualize and integrate their own experimental data with the information about PPIs available in the database. The data stored at DIP can be consulted online at http://dip.mbi.ucla.edu/dip/.

- **IntAct** [121] is a molecular interaction database for modeling, storing and analyzing molecular interaction data. The data stored in IntAct is either curated from the literature or from direct data depositions. Currently (December 2017), IntAct comprises more than 794 thousand interactions from over 20047 publications. Recently, the MINT and IntAct databases have merged their efforts to make optimal use of limited resources and maximize the curation output. All data curated by MINT curators have been incorporated into the IntAct database. Both IntAct and MINT data are freely available at http://www.ebi.ac.uk/intact.

- The ***Human Protein Reference Database (HPRD)*** [122] is a database of curated proteomic information that provides a collection of human PPIs linked to protein features such as protein function, post-transcriptional modifications, enzyme-substrate relationships, subcellular localization, protein isoforms and domain architectures. At this moment (December 2007), HPRD comprises more than 41 thousand PPIs. Data contained in this database is manually extracted from the literature and can be consulted at http://www.hprd.org/.

- The ***Human Immunodeficiency Virus Type 1 (HIV-1), Human Interaction Database (HIV-1 HID)*** [123] available through the *National Library of Medicine (NLM)* at `https://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/` aims to provide scientists in the field of HIV/AIDS research a detailed summary of all known interactions of HIV-1 and the human host. The data is collected of published reports of two types of interactions, protein interactions and human gene knock-downs that affect virus replication and infectivity. This database has been designed to retrieve PPIs without restrictions and includes *Reference Sequence (RefSeq)* protein accession numbers, National Center for Biotechnology Information Gene identification numbers, brief descriptions of the interactions, searchable keywords for interactions and PubMed identification numbers (PMIDs) of journal articles

describing the interactions. A total of 4 006 human genes are described participating in 14 102 interactions.

3.8 Resources for Protein-Protein Interaction Networks Visualization and Analysis

Scientists in the domain of Bioinformatics that study PPINs have to deal with huge amounts of data. They have to rely on comprehensive data from web resources. Getting an overview of the complex networks formed by PPIs is crucial to understanding living systems. Visualization supports this complex task. In the last years, bioinformatics has come up with a lot of tools that support the analysis of such complex networks. They provide multiple functionalities to layout and query the network, to visually integrate the network with expression profiles, phenotypes and other molecular states and to link the network to databases of functional annotations among others. Some of the most promising PPI web resources [124] will be described as follows:

- Cytoscape [125] is an open source Java-based bioinformatics software project for visualizing biomolecular interaction networks and integrating it with high-throughput expression data and other molecular states. It has an intuitive user interface and several features such as filters, attribute browser and so on. Cytoscape has a high degree of customization through the addition of multiple external plug-ins. An important feature is that Cytoscape is able to manage and visualize nested networks and thus create network hierarchies. Cytoscape supports and can export files in multiple formats. Furthermore, Cytoscape is able to link the network to databases of functional annotations such as the GO.

- Cytoscape.js [4] provides a *JavaScript (JS) Application Programming Interface (API)* to enable software developers to integrate graphs into their data models and web user interfaces. With this tool, it is possible to render PPINs in an interactive way on a server. Cytoscape.js provides interesting features such as the incorporated graph algorithms such as connectivity search, shortest path, minimum spanning tree, minimum cut, ranking and centrality measures. In terms of performance, cytoscape.js can render thousands of graph elements on average hardware.

- *Agile Protein Interaction DataAnalyser (APID)* [126] is a Java-based resource that allows the visualization of PPIs as a graph. APID allows queries of several input proteins. The visualization is dynamic and provides options for zooming, filtering and limit details. It also allows the addition of other proteins to the previously generated graph. However, it lacks visual clustering, highlighting certain nodes and edges and associa-

tions to diseases are not available. The data can be exported in a tabular format and the graph can be stored as an image. Import possibilities are very limited.

- *Protein Interaction Network Visualizer (PINV)* [127] provides PPINs visualization as graphs in browsers having Javascript installed and activated. The use of BioJS and D3 framework to create an HTML5 application offers a wide range of possibilities for visual analysis online. However, it presents some performance limitations for large and dense graphs. The user interface is very intuitive. The visualization is dynamic and provides options for zooming, filtering, highlighting, coloring and even uploading expression data. PINV also provides circular layouts, heatmaps and simple table views. The graphs can be exported both graphically and as text tables.

- STRING [105] is an interactive network viewer that only requires a web browser with the Flash plugin. The query interface is simple and includes data from several databases for multiple organisms. Graphs are rendered dynamically and offer multiple possibilities. STRING provides a variation of four different designs, namely confidence, evidence, action and interactive view. It is also possible to apply filters and control features, zooming and scaling functionality. The nodes and edges are colored. Node colors represent direct associations. Edge colors are mapped to types of evidence and line thickness represents confidence. An important feature that STRING offers is functional options for clustering and enrichment. Finally, the graphs can be exported as several file formats, both as graphics and as text.

- *Unified Human Interactome (UniHI)* [128] is a Java-based resource that makes use of the Cytoscape Web for the network visualization. The user interface is simple and intuitive. This resource is capable of dealing with large graphs. However, the graph does not include any visual details. UniHI makes use of clustering and enrichment functions and includes common control features such as zoom, repositioning and scaling. Details of a protein are provided in a separated window by clicking on a node. Information on the target proteins is extracted from the KEGG database. The information provided on pathological associations are scarce or nonexistent. Export options include text files, png and pdf.

4

METHODOLOGY AND IMPLEMENTATION

The development of the tools described in this thesis was devised to be divided into three main tasks. The SPINET database development, the data collection and curation and finally the data analysis and visualization tool. Figure 6 shows the global architecture of the SPINET web server. This thesis took part in the development of tasks 1, 2 and 3. Task 1 consisted in the collection of PPI data from major web repositories and from the literature and on the creation of multiple parsers to convert the data into a single and structured data format. In task 2 the SPINET database was designed and implemented to follow the proposed requirements. Task 3 consisted of the analysis of the data collected and on the creation of a visualization tool.

4.1 SPINET DATABASE

An essential aspect of the development of a database is the subdivision of the development process in multiple phases focusing on different aspects. The collection of these phases, called the *Software Development Life Cycle (SDLC)*, assumes a relevant importance in the organization of this process. The development of the SPINET database was carried out following an SDLC, in particular, the waterfall model. The phases involved in this model include:

1. Requirements Analysis
2. Design
3. Implementation
4. Testing
5. Maintenance

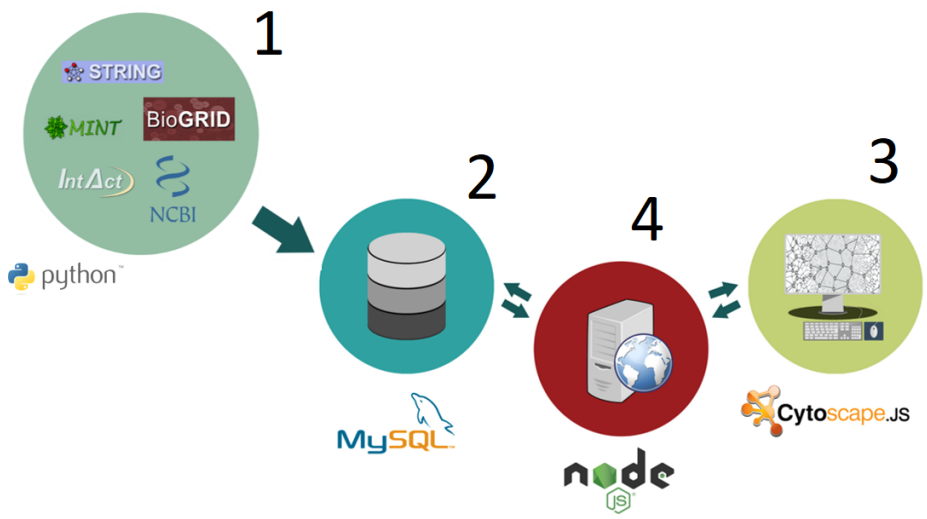


Figure 6: SPINET architecture.

4.1.1 *Requirements Analysis*

Due to the high heterogeneity of PPI data stored in multiple repositories available online, selecting which information is essential and which information holds redundancies or unimportant details becomes even more important but difficult task. In the requirements analysis phase, the specifications of what the database should be able to provide are gathered and analyzed. In this phase, some of the specified requirements that SPINET database should be able to answer were:

- Study PPINs online;
- Be "user-friendly" to users without computational background;
- Integrate data from the primary PPI databases and from the literature;
- Map different nomenclatures (the same biological object (e.g. a protein) might be identified with a different name in each repository);
- Provide a unique data format, allowing to uniformly use all available information;
- Provide more reliable and informative networks by combining PPI data with other types of biological data;
- Filter interactions by the number of experimental detection methods;
- Filter predicted and experimentally validated interactions;

- Filter interactions by score;
- Filter interactions by organism;
- Link the interaction with the original reference/source;
- Link proteins/interactions with known inhibitors.

4.1.2 *Database Design*

The database design phase was divided into three main steps:

- Conceptual model design;
- Logical model design;
- Physical model design.

Conceptual model

Using an entity-relationship representation, the conceptual model was implemented to create a high-level overview of the database. This data model was developed based on the requirements specifications gathered in the requirements analysis step and was independent of any physical consideration. This model is focused on defining important entities and the relationships between them. Figure 7 shows the conceptual schema of the SPINET database.

Logical model

In general, the logical data model is considered the implementation of the conceptual data model. The logical data model consists of data entities, keys, attributes and relationships between the entities. This model is focused on defining the data as much as possible, regardless of how it is to be implemented. The steps followed in the implementation of the logical model were:

1. Specify primary keys for all entities;
2. Define the relationships between different entities;
3. Define the attributes of all entities;
4. Resolve many-to-many relationships;
5. Normalization.

Figure 8 shows the logical schema of the SPINET database.

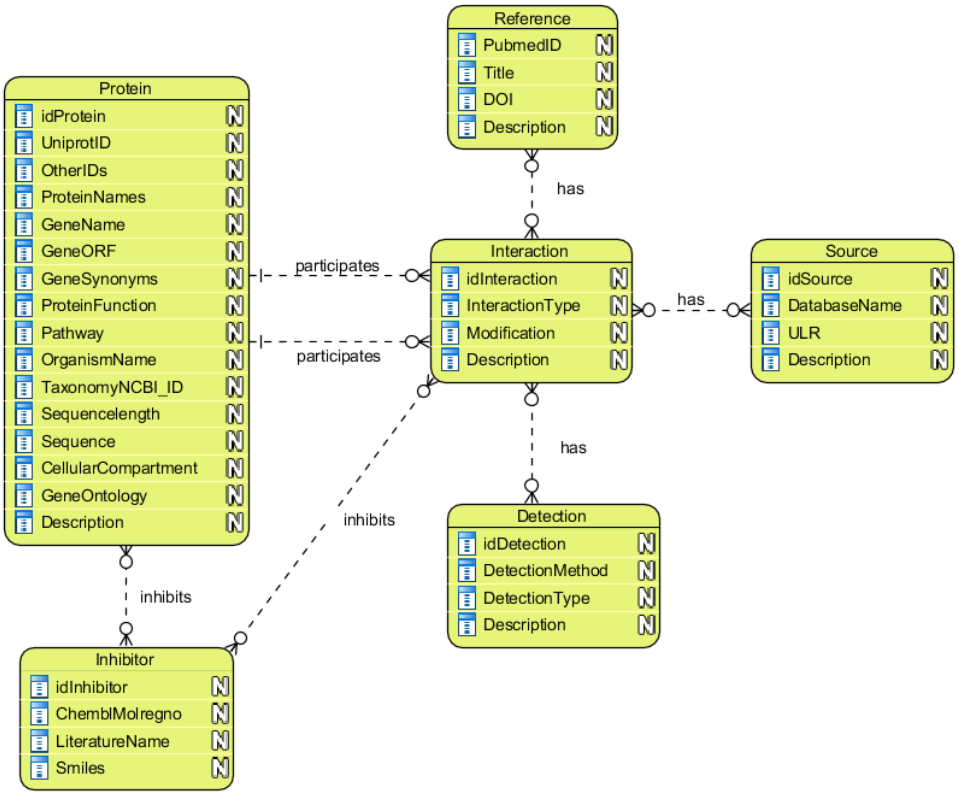


Figure 7: SPINET database conceptual model.

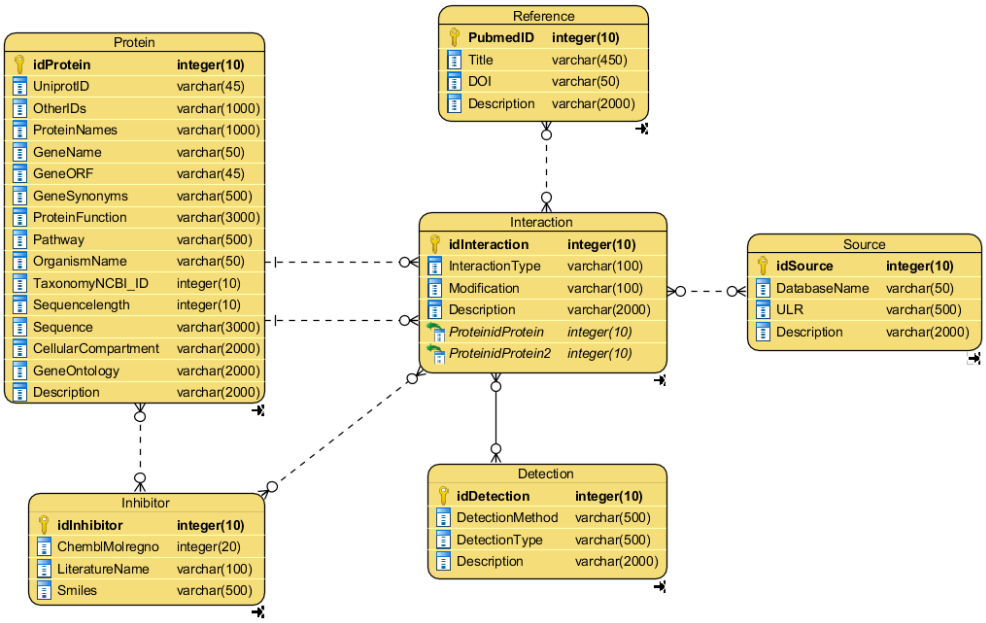


Figure 8: SPINET database logical model.

Physical model

The final step in the design of the database is the design of the physical data model. This model represents how the database will be built in the *Database Management System (DBMS)*. In this model, the data type of each attribute is specified and constraints are defined. Figure 9 shows the physical schema of the SPINET database.

The design complexity increases from the conceptual to the logical to the physical model. Therefore, in summary, with the conceptual model we aim to understand at a high level what are the different entities in our data and how they relate to one another, then we move to the logical model to understand the details of our data without worrying about how they will be implemented and finally the physical model to know exactly how to implement our data model in the DBMS of our choice. All the database schemas were created using the Visual Paradigm software.

4.1.3 *Database Implementation*

For the implementation of the SPINET database, the MySQL Workbench DBMS was used. This is a free, robust, and flexible relational DBMS that effectively removes the need of costly unsupportable informatics overhead associated with other systems such as Oracle or DB2.

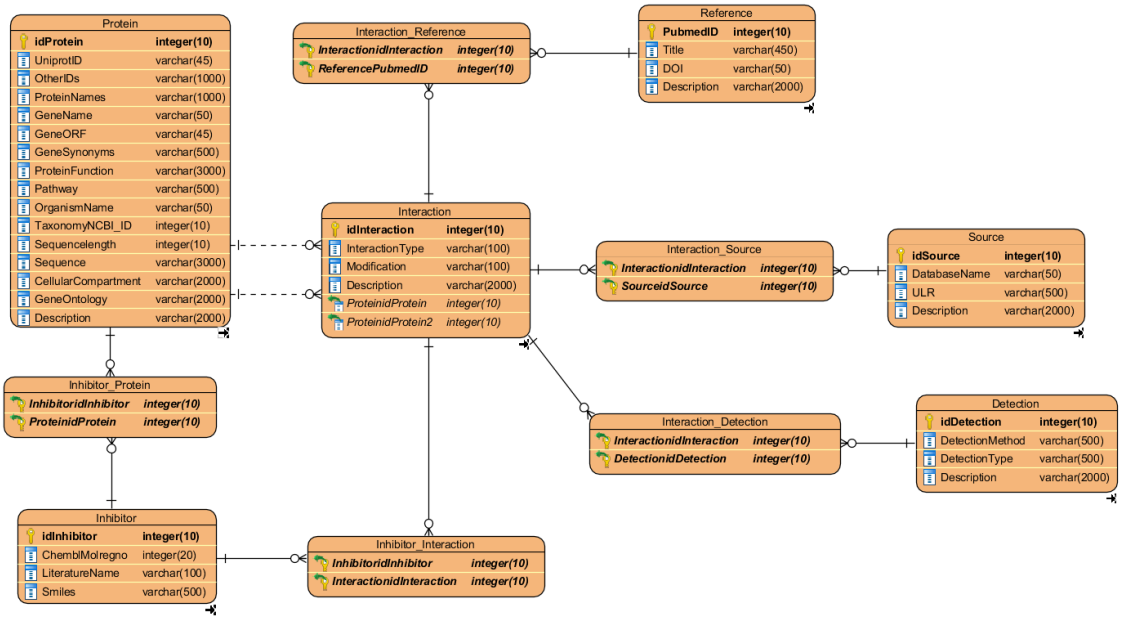


Figure 9: SPINET database physical model.

To avoid redundancy and dependency of data we made sure that the database obeys to the database normal forms. These normal forms ensure that are no repeating fields and groups (first normal form), there are no partial dependencies between records (second normal form) and that there are no transitive dependencies between records (third normal form). The current schema of the database consists of 9 tables and is shown in Figure 10.

Finally, the database was imported into phpMyAdmin allowing us to manage it in an easier and more efficient way from anywhere. Figure 11 shows the implemented database on phpMyAdmin.

4.1.4 *Database Testing*

Regarding database testing, multiple tasks were performed to ensure proper database functionality, data integrity and data mapping.

Some of the tasks performed in the database testing were:

- Check for any incorrect data;
- Try to upload inconsistent data to some tables and see if any failure occurs;
- Try to upload repeated information;
- Insert child data before inserting the respective parent's data;

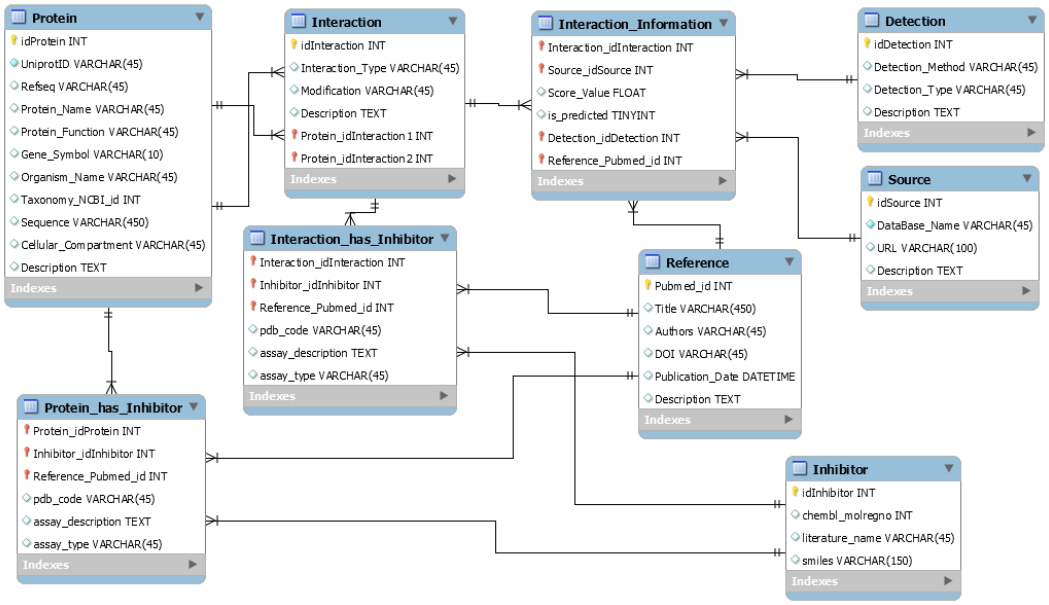


Figure 10: SPINET database schema.

Tabela	Acções	Registos	Tipo	Agrupamento (Collation)	Tamanho	Suspenso
Detection	Procurar Estrutura Pesquisar Insere Limpa Elimina	207	InnoDB	utf8_general_ci	32 KB	-
Inhibitor	Procurar Estrutura Pesquisar Insere Limpa Elimina	7,735	InnoDB	utf8_general_ci	1.6 MB	-
Interaction	Procurar Estrutura Pesquisar Insere Limpa Elimina	~4,343,452	InnoDB	utf8_general_ci	594 MB	-
Interaction_has_Inhibitor	Procurar Estrutura Pesquisar Insere Limpa Elimina	0	InnoDB	utf8_general_ci	48 KB	-
Interaction_Information	Procurar Estrutura Pesquisar Insere Limpa Elimina	~1,565,475	InnoDB	utf8_general_ci	479.5 MB	-
Protein	Procurar Estrutura Pesquisar Insere Limpa Elimina	27,863	InnoDB	utf8_general_ci	57.4 MB	-
Protein_has_Inhibitor	Procurar Estrutura Pesquisar Insere Limpa Elimina	11,903	InnoDB	utf8_general_ci	4.6 MB	-
Reference	Procurar Estrutura Pesquisar Insere Limpa Elimina	47,166	InnoDB	utf8_general_ci	15 MB	-
Source	Procurar Estrutura Pesquisar Insere Limpa Elimina	10	InnoDB	utf8_general_ci	48 KB	-
9 tabela(s)	Soma	94,884	InnoDB	utf8_general_ci	1.1 GB	0 Bytes

Figure 11: Implemented database on phpMyAdmin.

- Try to delete a record referenced by the data in another table;
- Check whether the references for foreign keys are valid;
- Check whether the field is mandatory while allowing NULL values;
- Check whether the length of each field is of sufficient size.

Besides the above-mentioned tasks, multiple *Structured Query Language (SQL)* queries were performed to ensure data integrity from the database to the user.

4.1.5 *Database Maintenance*

Maintaining and upgrading the database is a never-ending phase. As soon as new requirements arise, the SDLC will be again restarted and almost all of the above steps will have to be remade.

During this phase, some minor changes have been done, such as the alteration of the name of some attributes, data types and constraints.

4.2 DATA COLLECTION AND CURATION

The first step in performing PPIN analysis is, naturally, building the network. Nowadays, several databases store huge amounts of PPI data. The combination of the data stored in these databases allied with information from other biological databases can be used to build more reliable and informative networks. At this moment, the data used in this project was collected from multiple sources (APID, BioGRID, DIP, HPRD, IntAct, MINT, HIV-1 HID, STRING, UniProt, TIMBAL and from the literature [129, 130]). However, the data stored in these repositories exhibits multiple and distinct formats, increasing the difficulty of integrating these data into a uniform format. Moreover, the majority of PPI databases adopts different nomenclatures: the same protein might be identified with a different name in each database.

Taking into account the abovementioned issues, using the Python [131] language multiple parsers were implemented in order to uniformize the data from the different sources. In general, all the parsers perform the following tasks:

- Data filtering: remove interactions where at least one interactor does not belong to the interest organisms (*Homo sapiens*, *Human Immunodeficiency Virus Type 1* and *Mycobacterium tuberculosis H37Rv*); Remove proteins without uniprotID; Remove inconsistent data; Remove repeated data.
- Protein id mapping and enrichment: using the UniProt Retrieve/ID mapping tool `https://www.uniprot.org/uploadlists/`, convert the database internal identifiers to

UniProt identifiers. Before downloading the identifiers list, UniProt allows retrieving additional information about each protein. In our case we are interested in retrieving the following information: original identifier, UniProt identifier, RefSeq, protein names, gene names (primary), gene names (ORF), gene names (synonym), protein function, protein pathway, organism name, organism taxonomy, sequence length, sequence, subcellular location and gene ontology. This method allows not only to obtain a uniform data format but also to enrich data with valuable biological insights.

- Check if the proteins, interaction, source, detection method and pubmedID are already in the SPINET database. In the negative case, this information is written in multiple files for further upload to the database. This allows avoiding data redundancies.

Regarding literature and inhibitors data, similar parsers were implemented adapted to each case. Additionally, using the easyPubmed package [132], an R [133] script was developed to efficiently search and retrieve scientific publication records from PubMed. Figures 12 and 13 show the SPINET data warehouse and parsers architectures respectively.

4.3 DATA VISUALIZATION AND ANALYSIS

Having the database implemented and populated, the next step was to be able to benefit from the huge amount of data without being overwhelmed by it. The evaluation made on the already available resources for PPI visualization and analysis allowed to conclude that none of them fulfills all of our requirements. Thus, there was a need for a new visualization and analysis framework that will support features such the combination of PPI and inhibitors data, the presence of both inter and intra-species interactions and the inclusion of different algorithms.

4.3.1 *Data Analysis*

With the data collected, and resorting to the Python package NetworkX [134] it was possible to create 6 different networks with different properties. A network formed with all proteins, only with human, Mtb or HIV proteins, with both human and Mtb proteins and finally with human and HIV proteins.

Topological analysis of the networks

To analyze each of the generated networks, multiple measures were calculated:

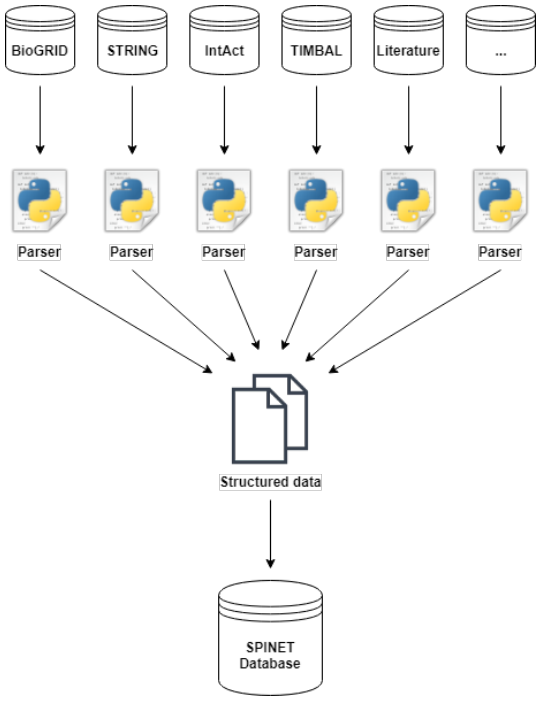


Figure 12: SPINET - Data warehouse.

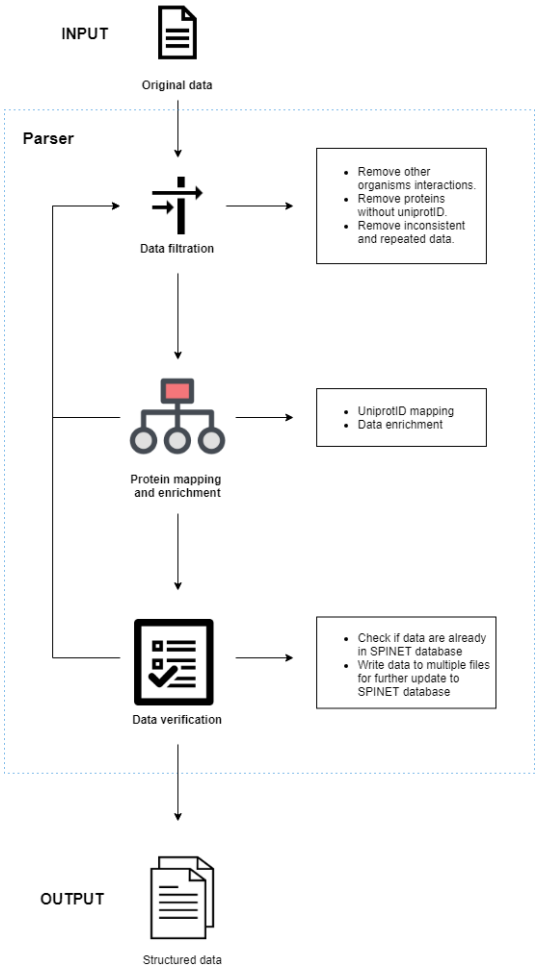


Figure 13: Parsers architecture.

- Network density: the density of a graph G is the ratio of the number of edges and the number of possible edges. Thus, $density(G) = \frac{2m}{n(n-1)}$, where n is the number of nodes and m is the number of edges in G.
- Network diameter: the diameter of a graph G is the maximum eccentricity among the nodes of G. Thus, $diameter(G) = max\{e(v) : v\ in\ N(G)\}$. The eccentricity e of a node v is the maximum distance from v to any node w. That is, $e(v) = max\{d(v,w) : w\ in\ N(G)\}$.
- Network radius: the radius of a graph G is the minimum eccentricity among the nodes of G. Therefore, $radius(G) = min\{e(v) : v\ in\ N(G)\}$.
- Network center: the center of a graph G is the set of vertices of eccentricity equal to the radius. Hence, $center(G) = \{v\ in\ N(G) : e(v) = radius(G)\}$.

- Average path length (l): in a graph G, the average path length is the average shortest path between every pair of nodes. Thus, $l = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, w_j)$, where n is the number of nodes and $d(v_i, w_j)$ is the length of the shortest path between nodes i and j.
- Clustering coefficient: the clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. In practice, it is the fraction of triangles (three connected nodes) that actually exist over all possible triangles in its neighborhood. The average clustering coefficient for the whole graph is the average of the local values C_i, $C = \frac{1}{n} \sum_{i=1}^{n} C_i$, where $C_1 = \frac{number\ of\ triangles\ connected\ to\ node\ i}{number\ of\ triples\ centered\ around\ node\ i}$.
- Scale-free property: a scale-free network is typified by the presence of hubs whose degree greatly exceeds the average. The presence of hubs will give the degree distribution a long tail fitting a power-law degree distribution ($P(k) \sim k^{-\gamma}$). To determine whether each network exhibits the scale-free property the python powerlaw package [135] was used. This package was used to calculate the parameter alpha, to plot the probability density function, the cumulative distribution function ($p(X < x)$) and the complementary cumulative distribution function ($p(X \geq x)$, also known as the survival function). We used log-log axes since a typical histogram on linear axes is not helpful for visualizing heavy-tailed distributions.
- Small-world property: in a small-world network the distance between any pair of nodes is relatively small when compared to random networks. In these networks the typical distance L between two randomly chosen nodes grows proportionally to the logarithm of the number of nodes n in the network, that is: $L \propto \log n$. To determine whether each network exhibits the small-world property the average-path length was calculated and then compared with the value of the logarithm of the number of nodes.

Centrality measures

In a network, especially in biological networks like PPINs, every node has a different importance. For example, a node with a high degree is more probable to be an essential node in the network compared to a node with a lower degree. The importance of a node can be mathematically determined through the calculation of many centrality measures. In the analysis of each node four major centralities measures were used, as follows:

- DC: this measure assigns an importance score based purely on the number of edges held by each node. In this case, to be able to compare this measure across different networks, we calculated the normalized degree for each node n_i as follows: $DC(n_i) = \frac{1}{N-1} e_i$, where e_i is the number of edges held by the node and $N-1$ corresponds to the maximum possible number of edges that a node can have (normalization).

- CC: corresponds to the sum of the length of the shortest paths between the node and all other nodes in the graph. The node with the highest closeness centrality has on average, the shortest distance to all the other nodes. Mathematically, the CC is calculated as follows: $CC(n_i) = \frac{N-1}{\sum_{n_j \in G} d(n_i, n_j)}$, where $N-1$ corresponds to the minimum distance to all the other $N-1$ nodes (normalization) and $\sum_{n_j \in G} d(n_i, n_j)$ to the total distance of n_i to all other nodes in G.

- BC: measures the fraction of shortest paths between every pair of nodes in the graph that passes through the node. This measure shows which nodes act as "bridges" between nodes in a network and can be mathematically calculated by the following formula: $BC(n_i) = \frac{\sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}}{(N-1)(N-2)/2}$, where $\sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}$ corresponds to the fraction of shortest paths that pass through the node n_i and $(N-1)(N-2)/2$ is the total number of pairs of nodes (normalization).

- EC: like DC it measures the influence of a node based on its number of edges. However, EC also takes into account how "well" connected a node is by measuring the number connections of their neighbors, and so on through the network. By calculating the extended connections of a node, EC can identify nodes with influence over the whole network, not just to those directly connected to it. $EC(n_i) = \frac{1}{\lambda} \sum_j A_{ij} EC(n_j)$, where λ corresponds to the leading eigenvalue of the matrix A (this measure should satisfy $Ax = \lambda x$, where A is the adjacency matrix of the graph G with eigenvalue λ. By virtue of the Perron-Frobenius theorem, there is a unique and positive solution if λ is the largest eigenvalue associated with the eigenvector of the adjacency matrix A [136]) and $\sum_j A_{ij} EC(n_j)$ is the sum of the centralities of the node n_i neighbors.

4.3.2 *SPINET Visualization and Analysis Tool*

To overcome some of the research gaps and to fulfill the established requirements, the SPINET visualization and analysis tool will try to accomplish, among others, the following tasks:

- Study PPINs in an interactive way;
- Be "user-friendly" to users without computational background;
- Generate appealing networks in different layouts;
- Generate networks with both inter and intra-species PPIs;
- Apply multiple filters;

- Integrate information of known inhibitors of both proteins and interactions;
- Perform multiple centrality measures;
- Perform multiple clustering algorithms;
- Make use of other useful algorithms.

Data Visualization

To be able to visualize the collected data, we decided to develop a simple tool to visualize and analyze sub-networks in an interactive and appealing way. Towards this aim, the SPINET analysis and visualization tool was written using the JS programming language, the *Cascading Style Sheets (CSS)* style sheet language, the *HyperText Markup Language (HTML)* language and the JS library cytoscape.js. The tool holds multiple menus and buttons with the following functionalities:

- Layout menu:

With the SPINET visualization tool it is possible to use three different layouts to specify how the networks are positioned in the viewport:

1. Force-directed layout: through the use of force-directed graph drawing algorithms it is possible to position the nodes of the networks so that all the edges are of more or less equal length and there are as few crossing edges as possible (Figure 14).
2. Concentric by centrality layout: in a concentric layout, the higher the centrality value of a node (in our case the CC), the closer the node will be to the center of the graph.
3. Hierarchy by centrality layout: this layout is very similar to the concentric layout. The higher the centrality value of a node (in our case the CC), the higher the position the node will occupy in the graph.

- Centrality measures menu:

With this tool, it is also possible to calculate the above-mentioned centrality measures DC, CC, BC and EC. These measures can be consulted by clicking on a node (this shows the node and the 1st-grade neighbors) and the measures will be displayed in a box. If we want these measures to be calculated to all nodes, the measures need to be selected from the topological analysis menu and then a list with all nodes and respective value will be displayed. The two alternatives are shown in figures 15 and 16.

- Clustering menu:

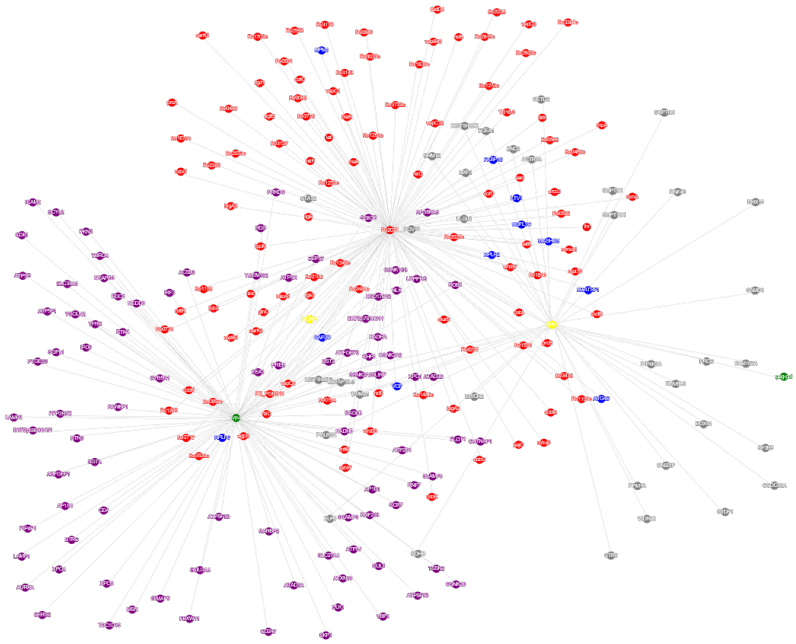


Figure 14: Force-directed layout.

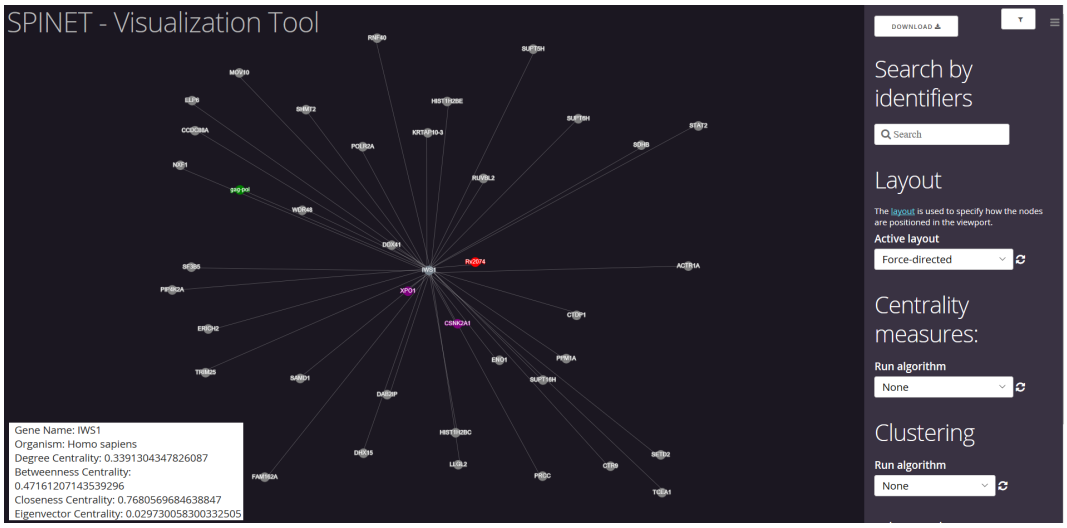


Figure 15: Centrality measures of the protein IWS1 (UniProtKB - Q96ST2).

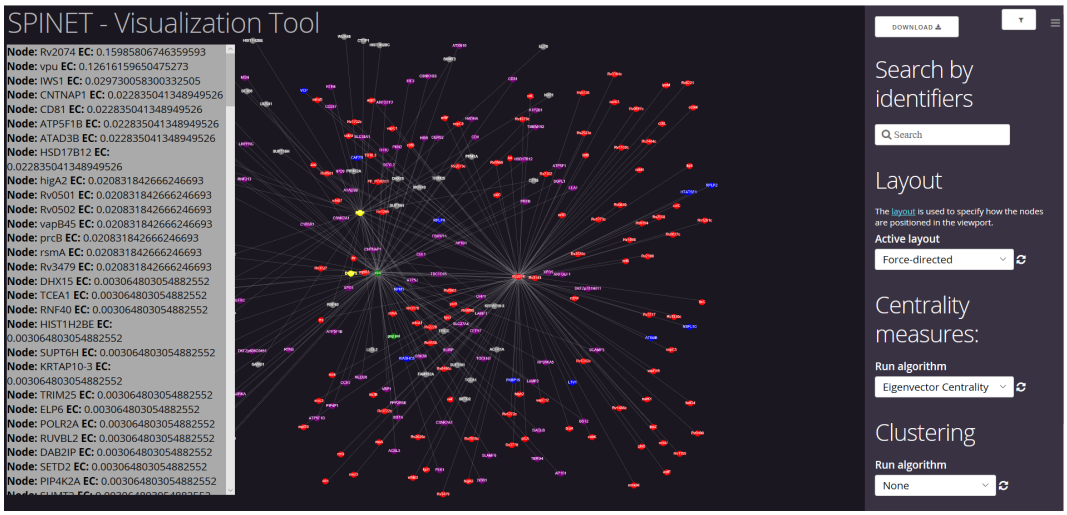


Figure 16: Eigenvector centrality for all proteins in the network.

K-means, K-medoids and Markov Cluster clustering algorithms can be performed through the clustering menu of the SPINET visualization tool. By running any of these algorithms a random color will be assigned to each cluster. These algorithms were implemented resorting to the cytoscape.js extensions cytoscape-k-means.js and cytoscape-markov-cluster.js [137].

1. K-means Algorithm: this is one of the simplest clustering algorithms. The goal is to find groups in the data, with the K variable representing the number of groups. The algorithm works iteratively to assign each node to one of the K groups based on the features that are provided. Initially, K initial random centroids are chosen, then every node is assigned to the closest centroid. After this, K new centroids are computed by averaging the examples in each cluster. Finally, if the centroids do not change the algorithm stops, otherwise the previous steps are repeated until the centroids' composition does not change. To run this algorithm in this tool, it is required as input the number of clusters (k), the distance metric (Euclidean, Manhattan and Max) and the maximum number of iterations that the algorithm should do.

2. K-medoids Algorithm: the k-medoids algorithm is very similar to the k-means algorithm, the major difference between them is the manner in which the cluster centers are initialized. In k-medoids, the cluster centers (medoids) are random nodes from the graph. To run this algorithm the inputs are the same as the ones required by the K-means algorithm.

3. Markov Cluster Algorithm: this is a fast and scalable cluster algorithm for graphs based on simulation of (stochastic) flow in graphs. This algorithm tries to find highly

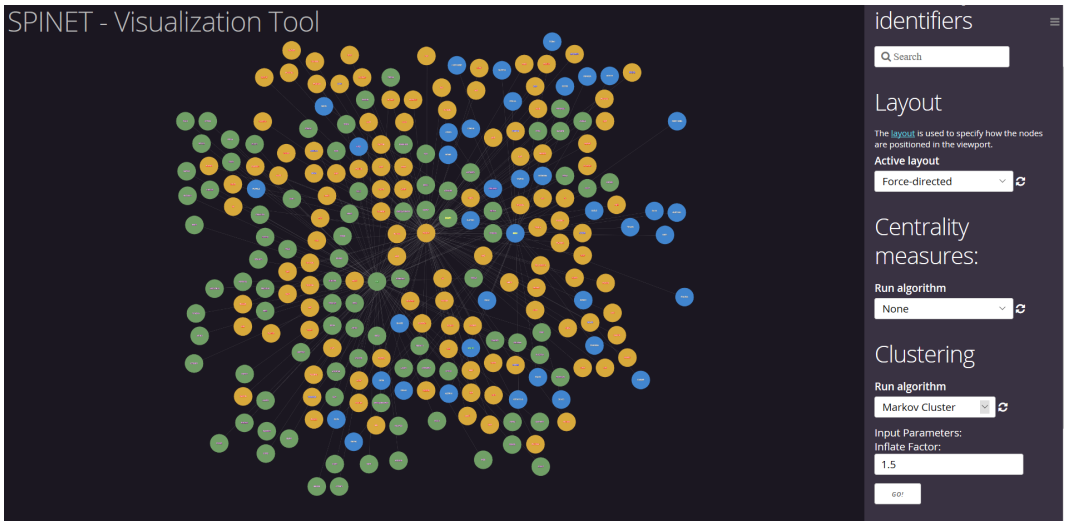


Figure 17: Markov Cluster algorithm layout.

interconnected regions (clusters) in the network by doing random walks upon the graph to discover where the flow tends to gather, and therefore, where clusters are. The algorithm was discovered by Stijn van Dongenat at the Centre for Mathematics and Computer Science in the Netherlands [138]. The only input that this algorithm requires is the inflate factor, that influences the size of the clusters. The higher the inflate factor, the smaller the clusters. Figure 17 shows how the Markov Cluster results are displayed in the SPINET visualization tool.

- Other algorithms menu:

Within the SPINET visualization tool, it is also possible to make use of three different search algorithms. The A* Search, *Breadth-First Search (BFS)* and *Depth-First Search (DFS)* algorithms. All these algorithms are widely used in pathfinding and graph traversal.

1. A* Search Algorithm: this algorithm is used in this tool to find the shortest path between two nodes in the network. It takes as input the start protein and the target protein and returns the path and distance between the two nodes. It also shows in the network the path, highlighting the edges of the path between the two proteins. Figure 18 how this algorithm is presented to the user. In this case, the path between the proteins NPM1 (UniProtKB - P06748) and RANBP1 (UniProtKB - P43487) is shown (NPM1 $\rightarrow$ Rv2074 $\rightarrow$ DCAF1 $\rightarrow$ vpu $\rightarrow$ RANBP1).

2. BFS: this algorithm visits all nodes in the network. First, it starts at an arbitrary node and then explores all the neighbor nodes at the present depth and only then moving to the nodes at the next depth level.

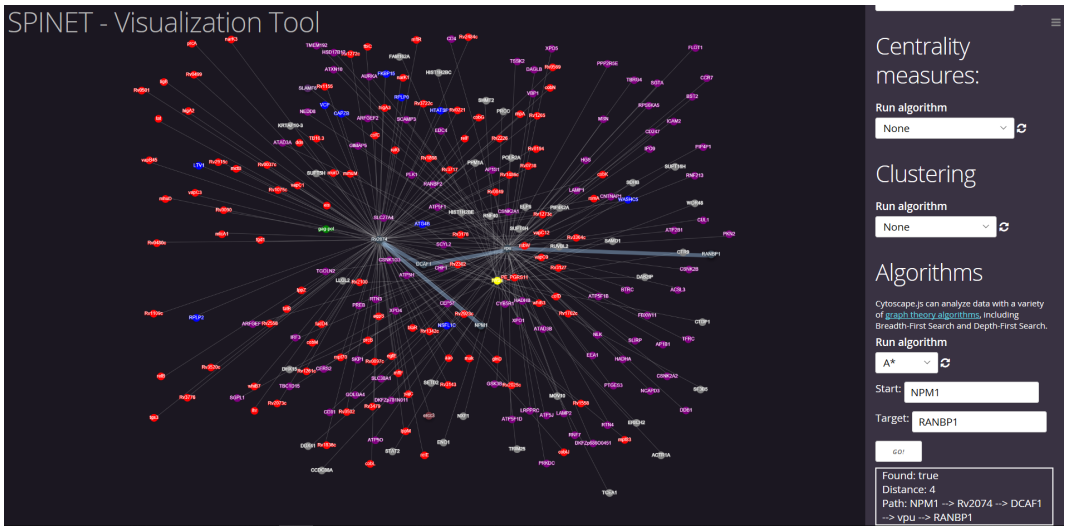


Figure 18: A* Search algorithm output layout.

3. DFS: this algorithm also visits all the nodes in the network but explores the network as far as possible along each branch before backtracking, repeating this procedure until all nodes are visited.

- Other features:

This tool also presents features such as the filtering of proteins by organism, the search of proteins by identifiers showing the protein and the first-grade neighbors and allows the exportation of the network as a png image file.

5

RESULTS

5.1 DATA STATISTICS

As already mentioned, at this moment, the data integrated into the SPINET database was retrieved from 11 different repositories and from 2 scientific articles. The integrated set of interactions consists of 4,737,904 interactions between 27,863 different proteins. The interactions can be divided into predicted or experimentally validated. Only the 350,653 unique experimentally validated interactions between 22,171 proteins were used to produce all the results included in this dissertation. In addition, the SPINET database also includes information about 7,735 inhibitors targeting 60 different human proteins.

5.1.1 *Interactions Distribution*

The experimentally validated interactions were detected using 177 different experimental methods (grouped by PSI-MI). As shown in Table 2, we can see that as expected the species with the largest number of interactions is *Homo sapiens* with 342551 interactions. Table 3 also shows the number of interactions between the three species and the respective number of proteins. The number of interactions between human and HIV proteins goes according to what was expected with 2916 interactions between the two organisms. On the other hand, it was expected that the number of interactions between human and Mtb proteins would be higher, especially if we take into account that 186 out of the 196 interactions were described in a single source ([130]). Of the 186 interactions, only 12 can be found in another source, which goes according to the fact that the mechanisms by which Mtb disrupts the host immune response are still very poorly understood. This also shows that recent studies are still not stored in the major PPI data repositories, reflecting the fact that these platforms are not regularly updated.

Most interactions were found just in one or two sources (Table 4) and were normally detected also by one or two detection method (Table 5). The interactions overlap between the different sources is very low with almost 60% being only found in one or two of the

Table 2: Number of proteins and interactions per species. (*This number is different from the total number of interactions (350653) because interactions between organisms were counted in both organisms.)

Species	Number of Interactions	Number of Proteins
Homo sapiens	342551	17788
HIV-1	2930	10
Mtb H37Rv	8284	3983
Total	353765*	27863

Table 3: Number of interactions between species and respective number of proteins.

	Interactions			Proteins		
	Homo sapiens	*Mtb H37Rv*	*HIV-1*	*Homo sapiens*	*Mtb H37Rv*	*HIV-1*
Homo sapiens	339439	196	2916	17655	192	1776
Mtb H37Rv	196	8008	-	39	3947	-
HIV-1	2916	-	14	10	-	8

multiple sources, reinforcing the need for an integration of this type of data into a single and structured source to build more reliable and informative PPINs.

It was observed that the majority of the interactions were detected using high-throughput methods. Affinity methods were the most used experimental assay to the screening of PPIs being used 482,337 times (one interaction can be detected by multiple types of experimental assays). Table 6 shows the number times that each one of the top 5 most used PPI detection methods was used.

5.1.2 *Inhibitors Data*

Regarding inhibitors data, as already mentioned the SPINET database include data on 7,735 inhibitors targetting 60 different human proteins. The integrins are the proteins that are targeted by the higher number of inhibitors with the ITGB3 (UniprotKB - P05106) protein alone being targeted by 2,191 different inhibitors. Table 7 shows the 6 proteins that are targeted by a higher number of inhibitors.

On the other hand, the inhibitors that target the higher number of proteins are the compounds with the following ChEMBL IDs: CHEMBL369635, CHEMBL88478, and CHEMBL173552 targeting 7 different proteins each. Table 8 shows 5 inhibitors that target a higher number of proteins.

Table 4: Distribution of interactions across different sources.

Number of Sources	Number of Interactions
1	196056
2	122509
3	23056
4	4711
5	877
6	122

Table 5: Distribution of interactions in SPINET database across different detection methods.

Number of Detection Methods	Number of Interactions
1	106725
2	156137
3	46961
...	...
15	90
16	47
...	...
41	1

Table 6: Number times that each one of the top 5 most used PPI detection methods was used.

Detection Method	PSI-MI	Number of Occurrences
Affinity Chromatography Technology	MI:0004	482337
Two Hybrid	MI:0018	83956
Anti Tag Coimmunoprecipitation	MI:0007	43088
Anti Bait Coimmunoprecipitation	MI:0006	24909
Enzymatic Study	MI:0415	18249

Table 7: Top 6 proteins targeted by the higher number of inhibitors.

UniprotID	Gene Name	Number of inhibitors
P05106	ITGB3	2191
P08514	ITGA4	1307
P13612	ITGA2B	1308
P05556	ITGB1	1266
P62942	FKBP1A	693
Q00987	MDM2	639

Table 8: Five inhibitors that target a higher number of proteins.

Compound ID	PubChem CID	Chemical Name	Number of Proteins
CHEMBL369635	9851886	-	7
CHEMBL88478	9961766	-	7
CHEMBL173552	11799915	-	7
CHEMBL429876	176873	Cilengitide	6
CHEMBL2332367	10196873	Cyclo(-RGDfK)	6

5.2 NETWORK MEASURES

5.2.1 *Properties of the Integrated Protein-Protein Interaction Network*

After the data being collected and stored in the SPINET database, the next step was to create and analyze the networks formed by these interactions. Multiple parameters were calculated for each one of the six different networks. In table 9 it is possible to verify the properties of each network.

As expected, the network with all proteins has the higher diameter and radius value. However, the network formed only by Mtb proteins presents a considerably higher density value (excluding the HIV network that has a 0.5 density value due to the low number of proteins), this goes according to the fact that despite having a lower number of proteins the Mtb network has a high number of interactions. The Mtb network also has the highest clustering coefficient, meaning that this network has a higher tendency to contain groups of nodes that are densely connected internally.

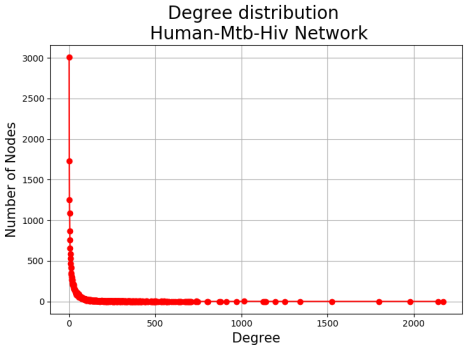
Regarding the scale-free property, the powerlaw python package was used to fit the data to a power law distribution. The gamma value of each network can be consulted in table 9. Despite the fact that the usual gamma value for scale-free networks is between 2 and 3 [139], which is the case of the human-Mtb, human-HIV and human networks, with the graphic analysis made on the human-Mtb-HIV we can conclude that this network can be considered to fit a scale-free network. Figure 19 shows the graphic analysis of the human network. As we can see the degree distribution follows the power-law distribution with many nodes having few edges and few nodes (hubs) having a high number of edges. The PDF, CDF and CCDF graphics also seem to fit a power law distribution. In the case of the Mtb network, the results were inconclusive both in the statistical analysis and in the graphical analysis (Table 9 and Figure 20). In the case of the HIV network, due to the small number of interactions, the analysis was not significant.

All networks can be considered small-world having a low average path length. The human network has the lowest average path length of 3.12 (excluding the HIV network), meaning that the distance between two random proteins is of 2.23 edges, in a mean.

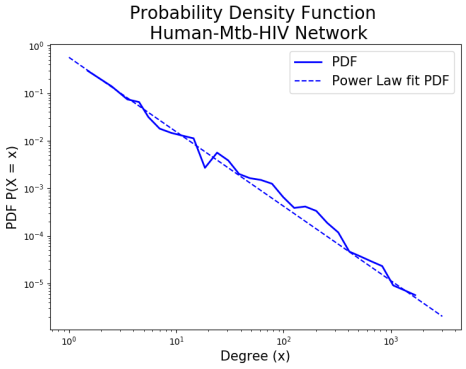
5.2.2 *Centrality measures*

The four previous referred centrality measures, DC, BC, CC and EC were calculated for each one of the six different networks. In the next subsections, the results for the proteins with higher centralities measures are shown.

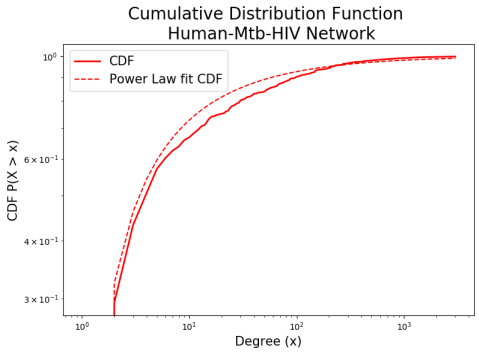
Figure 19: Scale-free analysis of the Human-Mtb-HIV PPIN.



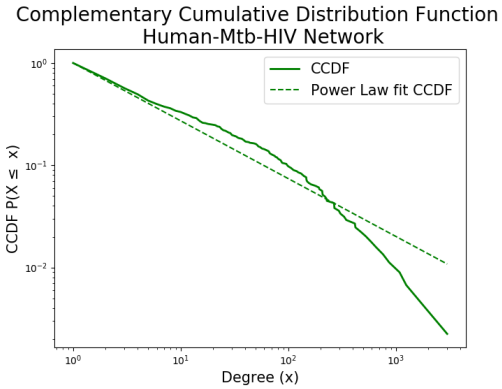
(a) Degree Distribution



(b) PDF

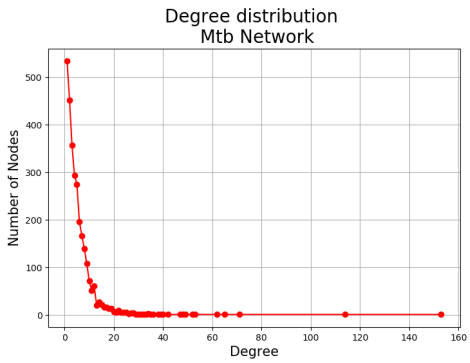


(c) CDF

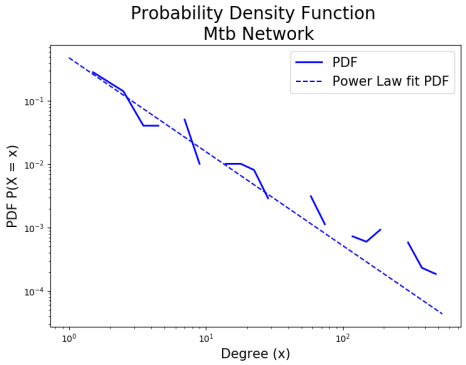


(d) CCDF

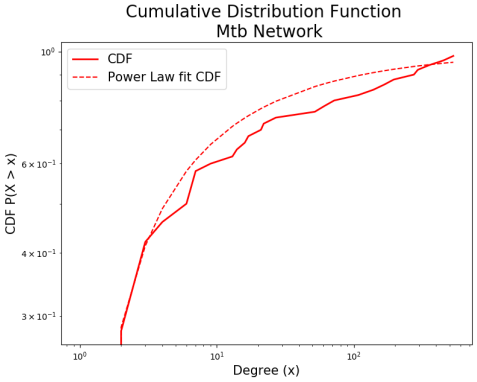
Figure 20: Scale-free analysis of the Mtb PPIN.



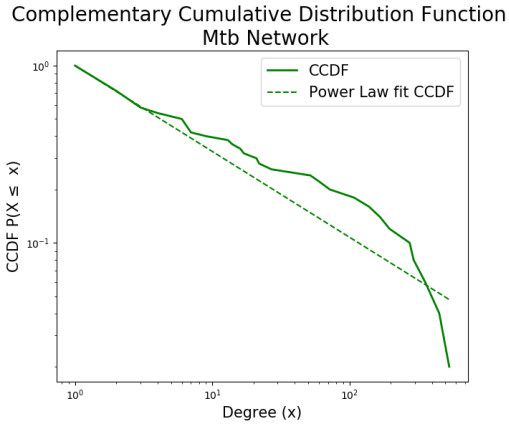
(a) Degree Distribution



(b) PDF



(c) CDF



(d) CCDF

Table 9: Networks properties.

Network	Human-Mtb-HIV	Human-Mtb	Human-HIV	Human	Mtb	HIV
Diameter	12	12	8	8	10	4
Radius	7	7	5	5	6	2
Density	0.0014	0.0014	0.0018	0.0019	0.0019	0.5
Number of Central Proteins	91	91	1038	1033	72	2
Central Proteins	DDX1 PE10 RAB1B	espC RAB1B recD	LRRC59 NFRKB GLP1R	PPP2R2B XBP1 SRPK1	esxB trpG coaX	gag nef
Average path Length	3.75	3.75	3.12	3.12	4.31	1.95
Average Clustering Coefficient	0.101	0.101	0.116	0.115	0.008	0.229
Scale-Free	Yes	Yes	Yes	Yes	inconclusive	n.s.
Gamma	1.57	2.16	2.51	2.87	1.48	3.79
Small-World	Yes	Yes	Yes	Yes	Yes	Yes

Human-Mtb-HIV Network

In the network formed by the three organisms, as expected, the proteins with higher centrality measures are predominantly human proteins. The first non-human protein to appear is the HIV protein gag polyprotein occupying the 72nd place with a DC of 0.0247, BC of 0.0058, CC of 0.3622 and EC of 0.0441. In the first 200 places are also the envelope glycoprotein gp160 (UniProtKB: P04578), tat (UniProtKB: P04608), gag-pol polyprotein (UniProtKB: P04585) and nef (UniProtKB: P04601) proteins. In the case of Mtb proteins, the first to appear is the Mtb nucleoid-associated protein EspR (UniProtKB: P9WJB7) which occupies the 1486th place with a DC of 5.8638e-04, BC of 0.0349, CC of 0.3084 and EC of 7.124e-4. The proteins F420H(2)-dependent biliverdin reductase Rv2074 (UniProtKB: P9WLL7) and the probable conserved lipoprotein LpqN (UniProtKB: O53780) are also between the 2000 proteins with higher centrality measures. Table 10 shows the 10 genes with higher centrality measures in this network. In the network formed only by human proteins, the proteins with higher centralities remain the same, with their centrality values suffering minor changes. The low number of HIV interactions/proteins does not cause remarkable changes, with the centrality values remaining similar in the human-HIV network.

Mtb Network

In the network formed with only Mtb proteins, the proteins with higher centrality measures are shown in table 11. However, despite the relevance of these proteins in this network, when compared with its importance in the network formed by the three organisms, we noticed that other proteins (e.g. lpqN (UniProtKB: O53780), esxA (UniProtKB: P9WNK7) and espB (UniProtKB: P9WJD9) overcome these ones in terms of centrality measures. For example, the protein 4-hydroxy-tetrahydrodipicolinate reductase (gene name: dapB; UniProtKB: P9WP23) which is the protein with higher centrality measures in the Mtb networks, in the Human-Mtb-HIV network is only the 20th Mtb protein with higher centrality measures.

5.3 Human Proteins Interacting With Both Organisms

In the network formed by all interactions, 81 human proteins that interact with both Mtb and HIV proteins were identified. Using the python package goatools [140], a GO enrichment analysis of these genes was performed, the p-value of each GO term was generated using a false rate discovery correction with Benjamini/Hochberg (non-negative) with a p-value cutoff of 0.05. As result, a list of 46 enriched terms (7 *Molecular Functions (MFs)*, 18 *Biological Processs (BPs)* and 21 *Cellular Compartments (CCs)*) was obtained. Figure 21a shows the 7 enriched MFs. Almost every enriched term is related to binding activity with only the protein serine/threonine phosphatase activity term being related to other function. In

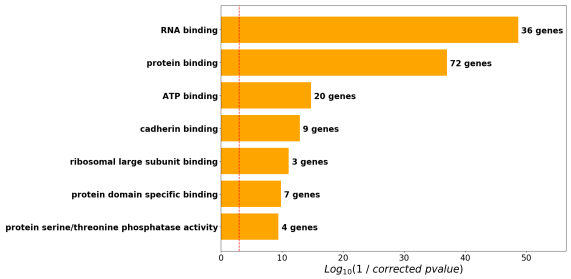
Table 10: Top 10 genes with higher centrality measures in the network formed by the three organisms.

Gene	Description	DC	BC	CC	EC
TRIM25	Functions as a ubiquitin E3 ligase and as an ISG15 E3 ligase. Involved in innate immune defense against viruses.	0.0965	0.0475	0.4056	0.1041
NTRK1	Receptor tyrosine kinase involved in the development and the maturation of the central and peripheral nervous systems.	0.0965	0.0256	0.4028	0.1310
APP	Functions as a cell surface receptor and performs physiological functions on the surface of neurons relevant to neurite growth, neuronal adhesion and axonogenesis.	0.0979	0.0626	0.4036	0.0722
JUN	Transcription factor that recognizes and binds to the enhancer heptamer motif 5′-TGA[CG]TCA-3′.	0.0688	0.0149	0.4000	0.1273
ELAVL1	RNA-binding protein that binds to the 3′-UTR region of mRNAs and increases their stability.	0.0810	0.0425	0.4007	0.0766
CUL3	Mediates the ubiquitination and subsequent proteasomal degradation of target proteins.	0.0540	0.0105	0.3891	0.1042
EGFR	Activates several signaling cascades to convert extracellular cues into appropriate cellular responses	0.0605	0.0216	0.3949	0.0786
TP53	Acts as a tumor suppressor in many tumor types.	0.0515	0.0154	0.3928	0.0889
UBC	Polyubiquitin precursor..	0.0508	0.0148	0.3925	0.0784
XPO1	Receptor for the leucine-rich nuclear export signal (NES).	0.0564	0.0141	0.3857	0.0727

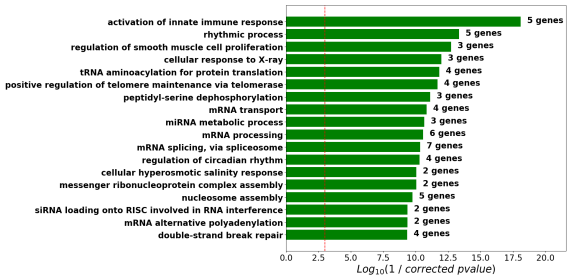
Table 11: Top 10 genes with higher centrality measures in the Mtb network.

Gene	Description	DC	BC	CC	EC
dapB	Catalyzes the conversion of (HTPA) to tetrahydrodipicolinate.	0.0524	0.1282	0.3512	0.5273
whiB3	Maintains intracellular redox homeostasis by regulating catabolic metabolism and polyketide biosynthesis	0.0391	0.0833	0.3367	0.2495
PPE22	Uncharacterized PPE family protein PPE22. Was identified as a high-confidence drug target.	0.0182	0.0342	0.3146	0.0782
wbbL	Involved in the biosynthesis of the mAGP complex, an essential component of the mycobacterial cell wall.	0.0212	0.0282	0.3072	0.0787
rsmI	Catalyzes the 2'-O-methylation of the ribose of cytidine 1402 (C1402) in 16S rRNA.	0.0243	0.0367	0.3092	0.0566
Rv2669	Involved in the regulation of response to oxidative stress.	0.0223	0.0272	0.2951	0.0754
coaX	Catalyzes the phosphorylation of pantothenate, the first step in CoA biosynthesis.	0.0144	0.0204	0.3058	0.0792
tsaE	Involved in the tRNA threonylcarbamoyladenosine modification.	0.0137	0.0195	0.3041	0.0792
mmsA	Involved in the methylmalonate-semialdehyde dehydrogenase (acylating) activity	0.0123	0.0181	0.3062	0.0673
Rv0027	Involved in protein secretion.	0.0123	0.0141	0.2981	0.0779

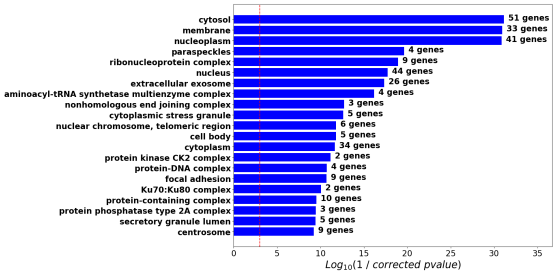
the case of the BP enrichment analysis, Figure 21b shows the 18 enriched terms. It should be noted that BPs like activation of innate immune response and positive regulation of telomere maintenance via telomerase were enriched. The most enriched CCs were cytosol, membrane and nucleoplasm. All enriched CCs terms are given in Figure 21c.



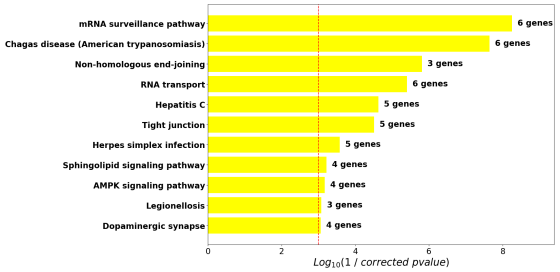
(a) Molecular function enrichment.



(b) Biological process enrichment



(c) Cellular compartment enrichment.



(d) Pathway enrichment.

Figure 21: Functional annotation analysis of the 81 human proteins that interact with both Mtb and HIV proteins.

In addition, a pathway enrichment analysis was performed using the *Database for Annotation, Visualization and Integrated Discovery (DAVID)* functional analysis tool. The p-value of each GO term was generated using a Benjamini correction with a p-value cutoff of 0.05. Figure 21d shows the 11 pathways that were significantly enriched. The most enriched pathway was the mRNA surveillance pathway. Several other pathways involved in infectious diseases were also enriched. For example, the Chagas disease (American trypanosomiasis), Hepatitis C, Herpes simplex infection and legionellosis pathways.

To explore the interactive relationships between the 81 human proteins that interact with both Mtb and HIV proteins and his first-degree human proteins that have known inhibitors, we performed a PPIN analysis based on the data contained in the SPINET database. We obtained a network of 175 proteins (9 HIV-1, 26 Mtb and 140 human proteins) and 60 nodes representing the number of inhibitors that each protein has. The Integrin alpha-4 (gene name: ITGA4; UniProtKB: P13612) gene was the most highly connected, interacting with 43 other human proteins. This protein does not interact directly with Mtb or HIV proteins but interacts with 33 other human proteins that interact with proteins of these two organisms. The Myc proto-oncogene protein (gene name: MYC; UniProtKB: P01106),

Exportin-1 (gene name: XPO1; UniProtKB: O14980) and Polyubiquitin-C (gene name: UBC; UniProtKB: P0CG48) also interact with a high number of proteins in this sub-network, interacting with respectively 43, 43 and 38 different proteins. Among these ones, the only one that interacts with both Mtb and HIV proteins is the Polyubiquitin-C interacting with the Mtb protein-tyrosine-phosphatase ptpA and with the HIV vpr, vif, gag, rev and tat proteins. Figure 22 shows the graphic aspect of the generated sub-network. This network was visualized by the SPINET Visualization Tool. In Figure 23, it is also possible to consult the number of inhibitors of each protein.

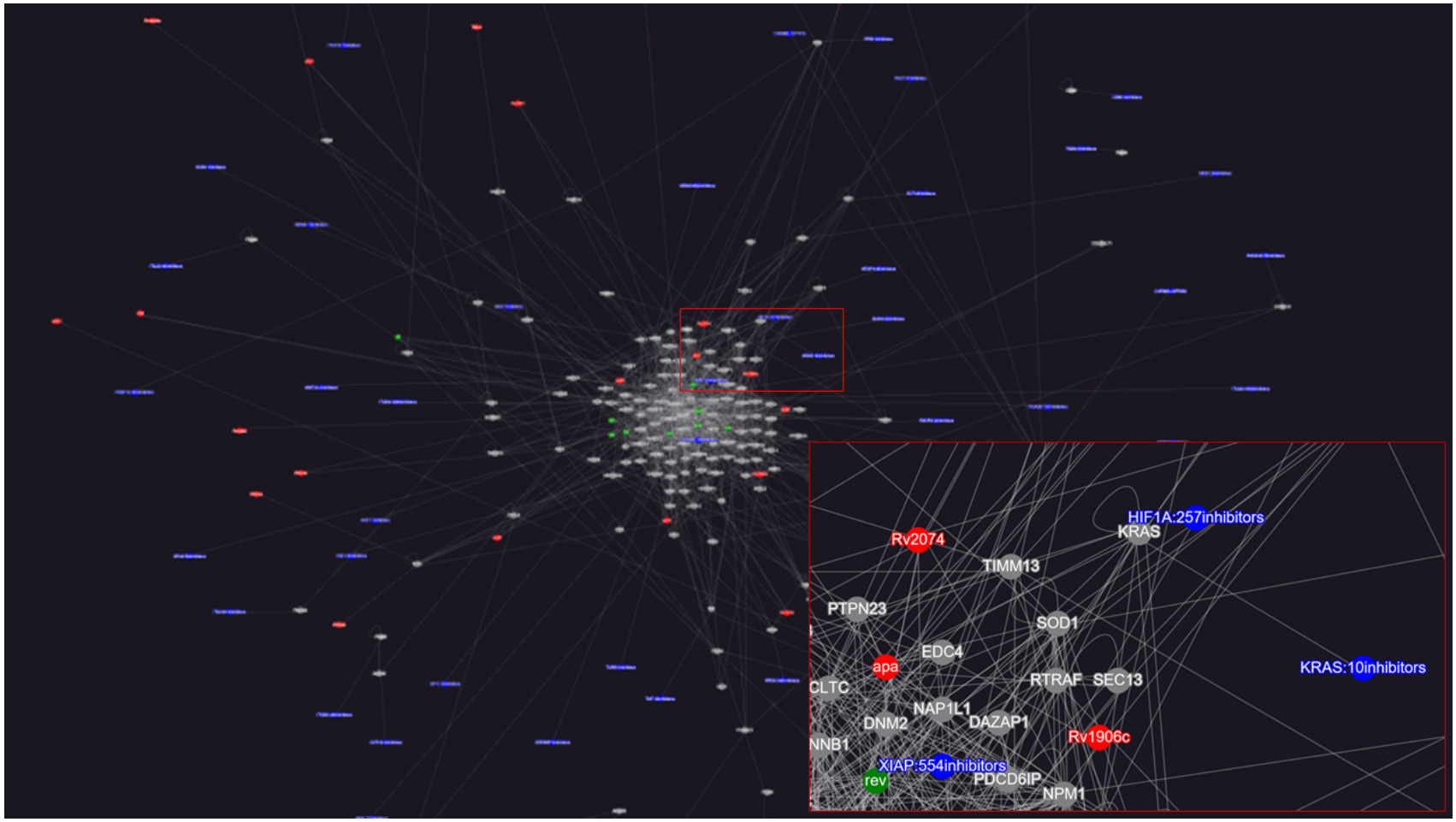


Figure 22: Protein-protein interaction network of the 87 human proteins (grey) and respective interactions (red: Mtb proteins, green: HIV-1 proteins and blue: number of inhibitors).

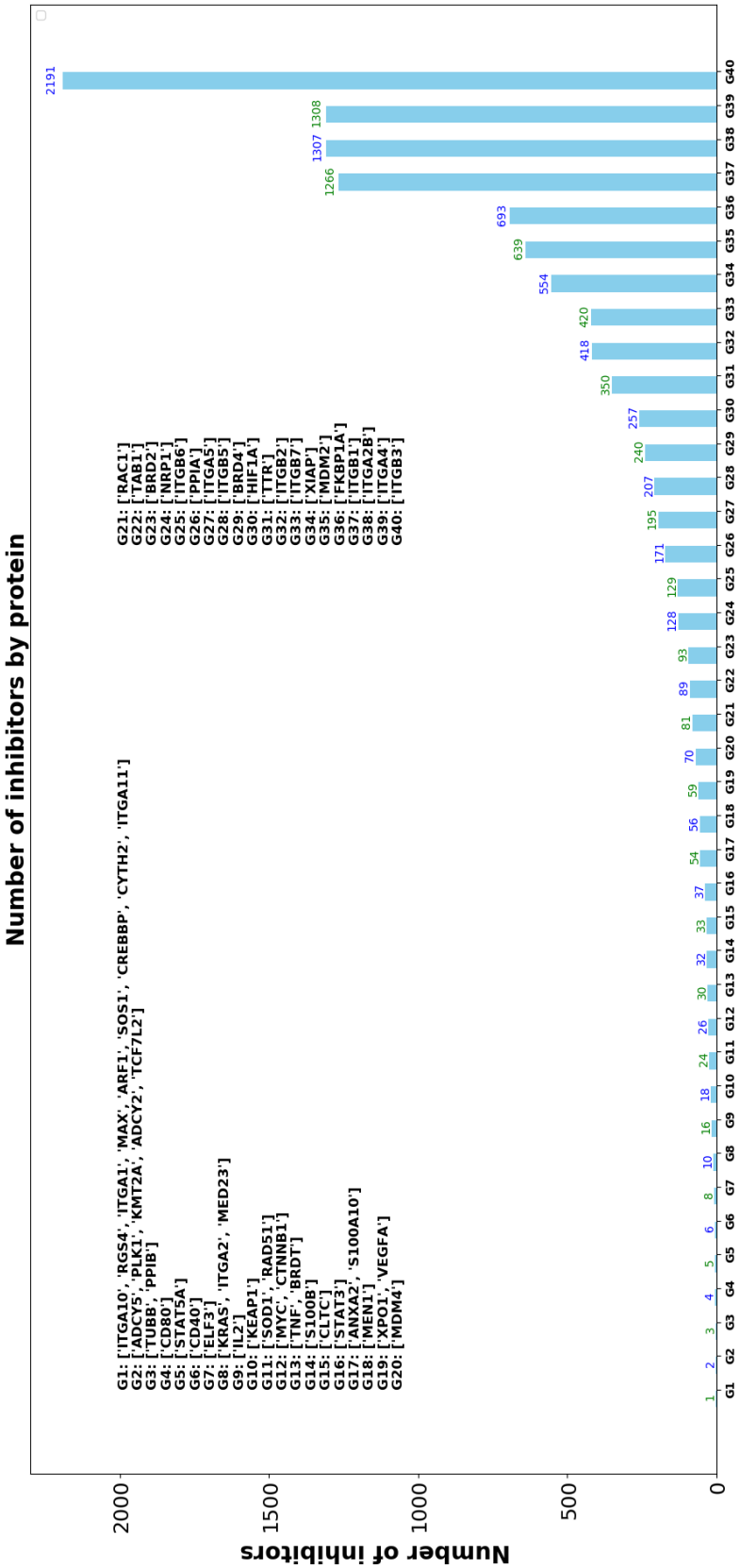


Figure 23: Number of inhibitors by protein.

6

DISCUSSION

In this thesis, we presented SPINET, a framework for PPI data integration, storage, analysis and visualization. The SPINET was created to address the current issues present in the use, manipulation and analysis of PPI data. Nowadays, PPI data are spread among multiple repositories that contain different information stored with different nomenclatures. The SPINET database will serve as a data warehouse storing data contained in the main PPI data repositories in a uniform and structured way, making it a viable primary source for the consultation of HIV-1/Mtb and Human PPIs. This tool can also be used as an initial resource for the analysis and visualization of sub-networks formed by the user proteins of interest.

Data sharing, integration and annotation are key aspects among the scientific community. However, the efficient use of the data highly depends on the existence and adoption of standards, shared formats and nomenclatures. In the case of PPI data, some efforts were made to overcome these issues. These efforts began about ten years ago with the creation of a common file format for representing PPI data, the *Minimum Information about a Molecular Interaction eXperiment (MIMIX)* [141]. This file format consisted of a list of information that had to be supplied when describing an experimental molecular interaction in a journal article. More recently, the IMEx consortium proposed a standard format for PPIs, the PSI-MI. The guidelines introduced by this initiative allowed the creation of a new exchanging format with a set of controlled vocabularies for different types of information (e.g. established names for types of detection methods). However, despite the efforts of some of the main PPI data repositories to adhere to these initiatives, each repository did it in a different way. Thus, the use of data from multiple sources is usually made in a biased way. SPINET is one of the few PPI platforms where all data from the main repositories can be found using a wide variety of identifiers for a set of proteins of interest. Furthermore, SPINET can provide additional information about the proteins involved in the generated networks as well as information on known protein inhibitors. This last feature brings additional value to this platform allowing not only the consultation of PPI data but also the possibility to study potential drug target to treat these diseases. Additionally, the SPINET is equipped with an

analysis and visualization tool that allows an initial analysis of sub-networks queried by the user.

The data stored in the SPINET database was collected from 11 different repositories and from 2 scientific articles, making it one of the largest (if not the largest) repository of human, Mtb and HIV-1 PPI data. Protein data was retrieved from UniProt, inhibitor data from *TIMBAL (TIMBAL)* and PPI data from IntAct, MINT, STRING, HPRD, APID, DIP, BioGRID, HIV-1 HID and from the works of Yi Wang et al and Bennett H. Penn et al [129, 130]. These platforms constitute some of the major data repositories available for these types of data. We collected a total of 4,737,904 interactions between 27,863 different proteins including 350,653 experimentally validated interactions. To maximize the scientific validity of the results, the work produced in this thesis was focused on experimentally validated interactions. As expected the majority of these data corresponded to human interactions. However, a considerable number of interaction involving Mtb or HIV proteins were found. The number of known interaction between human and Mtb proteins is small, with only 196 interactions reported. This fact is related to the challenges and limitations of working with Mtb in the laboratory. These include, among others, its slow growth in vitro and the necessity of using a biosafety level 3 laboratory. In fact, only 10 interactions between these two organisms were found in other databases. The remaining 186 interactions were retrieved from recent literature, highlighting the fact that these platforms do not frequently integrate newly generated data in their databases. The low number of interactions between these two organisms also reflects the lack of knowledge on how human innate immune system responds to Mtb infection, largely because of the difficulties in studying lung-specific immunity in humans [142]. Regarding inhibitors data, we found information on 7,735 inhibitors targeting 60 different human proteins. The four most targeted proteins are all integrins, with some of them having a role in microbial infection. For example, the Integrin beta-3 (gene name: ITGB3; UniProtKB: P05106), is involved in HIV-1 infection by interacting with the extracellular viral Tat protein, enhancing angiogenesis in Kaposi's sarcoma lesions [143]. Integrin beta 3 has also shown to be a contributor to macrophage-related inflammation [144]. Thus it is conceivable that its inhibition might be useful as a host directed therapy to treat high inflammation levels often associated with severe active TB. This kind of information can be deeply explored and potentially provide valuable insights for targeting specific proteins in specific diseases using already launched and approved drugs.

In this thesis, we presented a statistical analysis of the data collected from the multiple sources and then stored in the SPINET database. We verified that most of the interactions are only available in one or two databases, reinforcing the need for platforms like SPINET to unify all known interactions. Regarding experimental PPI detection methods, most of the interactions were also detected by one or two experimental methods. These experimental detection methods are almost entirely constituted by high-throughput methods. A big

problem associated with high-throughput methods is that they generate high rates of false positives. In this aspect, SPINET can offer a set of filters that aim to produce PPINs with a higher reliability. It is possible to filter predicted from experimentally validated interactions, filter by number of sources, number of experimental detection methods, score (if provided in the study), study (pubmed id), by original database among many others. In addition, to improve the confidence of a interaction it is also possible to access information about the proteins involved in the interaction. For example, it is possible to consult information about the cellular compartment of each protein. Proteins that are never co-located are not probable to interact with each other.

With the data already stored and validated, it was possible to create six different networks. The Human-Mtb-HIV, Human-Mtb, Human-HIV, Human, Mtb and HIV networks. Multiple network measures were calculated for each one of the networks and the results were given in Table 9. Due to the low number of HIV interactions compared with human and Mtb interactions, this organism does not considerably change the properties of the other networks. Excluding the Mtb and HIV networks that resulted in inconclusive and not significant results in the scale-free analysis, the other networks all follow a scale-free organization as their degree distribution followed a power law distribution. Previous studies made on human [145] and Mtb [146] PPINs also showed that these networks follow the scale-free property. Despite the fact that our analysis on the Mtb network resulted in a γ value slightly outside the common range ($2<\gamma<3$) and the graphical analysis turned out to be inconclusive, we can conclude that perhaps the fact that our analysis, unlike most published studies, does not use computationally generated interactions, and consequently leading to a smaller number of interactions, can be the key factor to that differences. The average clustering coefficient, along with the average shortest path are strong factors to determine if a network exhibits the small-world property. Although the average clustering coefficients of our networks appear to be small, they are in most cases, much higher than average clustering coefficients of random graphs with the same number of nodes and edges [147]. This reflects the fact that PPINs have a higher tendency to form clusters. In addition, our networks have a shorter average path length than expected when compared with random networks with the same dimensions. This means that any two proteins are separated by fewer interactions than expected regardless of the size of the network. All of our networks have an average path length lower than 6, reflecting the popularised "six degrees of separation" theory used in social networks [148]. With these two characteristics, we can conclude that our networks fit the small-world property having small path lengths between nodes.

As shown in this thesis and in other studies, PPINs normally manifest the scale-free and small-world properties. But biologically speaking, what does that mean? In a PPIN, the small-world property means that there is great connectivity between proteins. Naturally,

this level of connectivity has important biological consequences, allowing an efficient and quick flow of information within the network [149]. However, if the network is so highly connected, why don't perturbations in a single protein affect the entire network? This characteristic can be explained by the scale-free property. In scale-free networks, the majority of proteins participate in only a few interactions, while few proteins (hubs) participate in many interactions in the network. This characteristic grants PPINs a high stability because if a failure occurs at a random protein, the likelihood that a hub would be affected is very small [84]. However, if some major hubs are targeted, the network can be turned into a set of disconnected networks. This fact goes in accordance with the fact that hubs are enriched with essential/lethal proteins. Thus, the study of these highly connected proteins assumes high importance. In this thesis, to determine the importance of each protein in the respective network, some of the major centralities measures (DC, BC, CC and EC) were calculated for the generated networks.

Centrality gives an estimation of how important a protein is for the connectivity of the network. Multiple centrality measures can be calculated to identify important proteins in different contexts. Hubs can be important for the stability of the networks, however, sometimes proteins that act as bridges between protein complexes or proteins that lay in the shortest paths between multiple pairs of proteins can be of higher importance for the network. For example proteins with high BC sometimes represent proteins that lie on communication paths and can control information flow. Proteins with high BC normally represent important proteins in signaling pathways and can represent potential targets for drugs. Tables 10 and 11 show the top 10 genes with higher centrality values (mean of the four values) for the Human-Mtb-HIV and Mtb networks respectively. In the first network, between the top 10 proteins with higher centrality, the only one that interacts with both Mtb and HIV proteins is the Polyubiquitin-C (gene name: UBC; UniProtKB: P0CG48) interacting with the Mtb protein-tyrosine-phosphatase ptpA and with the HIV vpr, vif, gag, rev and tat proteins. Ubiquitination has been associated with protein degradation, DNA repair, cell cycle regulation, kinase modification, endocytosis, and regulation of other cell signaling pathways. Regarding Mtb and HIV infection, it is known that UBC protein participates in the budding, maturation and assembly of HIV virion, however, their role in the Mtb infection is still unknown. Other protein like TP53, JUN, APP, ELAVL1 and XPO1 also interacts with HIV proteins. In the case of the Mtb network, the proteins that have higher centralities in this network are not the same when compared with the ones that have higher values in the Human-Mtb network. This means that some proteins assume a higher relevance when in contact with the host in comparison with their relevance to the Mtb network. For example, the nucleoid-associated protein EspR (UniProtKB: P9WJB7) that only appear in the 86th place in terms of centrality in the Mtb network is one of the most relevant Mtb protein when in contact with the host. This makes total sense as this a is a relevant protein in controlling

the virulence of Mtb by specifically regulating expression of the exported EspA protein, which is required for ESX-1 secretion system to function [150]. Thus, the study of central proteins in individual and in combined networks provided by the SPINET platform stands out as an interesting resource for identifying and study proteins that could go unnoticed in the organism PPIN but stand out when including proteins from other organisms. This is especially relevant for Mtb and HIV-1 that are obligate intracellular human parasites.

The primary goal of the platform developed throughout this thesis is to, as much as possible, provide a complete and reliable study of the PPINs formed by Human, Mtb and HIV PPIs and their relationships. Taking advantage of SPINET, we were able to list, for the first time, 81 human proteins that interact with both Mtb and HIV proteins. To characterize this list of proteins a GO enrichment analysis was performed and 57 significantly enriched terms were highlighted. The most significantly enriched BP was the activation of innate immune response. This BP was enriched by 5 proteins, the X-ray repair cross-complementing protein 5 (gene name: XRCC5; UniprotKB: P13010), X-ray repair cross-complementing protein 6 (gene name: XRCC6; UniprotKB: P12956), Non-POU domain-containing octamer-binding protein (gene name: NONO; UniprotKB: Q15233), DNA-dependent protein kinase catalytic subunit (gene name: PRKDC; UniprotKB: P78527) and Splicing factor, proline and glutamine-rich (gene name: SFPQ; UniprotKB: P23246). All these proteins play an important role in the regulation and function of the cGAS-STING pathway of cytosolic DNA sensing. This pathway has a major role in the mechanisms by which the immune system detects pathogens. In particular, all the referred proteins have a role in the regulation of DNA virus-mediated innate immune response by assembling into the HDP-RNP complex, a complex that serves as a platform for IRF3 phosphorylation and subsequent innate immune response activation through the cGAS-STING pathway [151]. The positive regulation of telomere maintenance via telomerase was also enriched. This BP is related to the HIV infection [152, 153]. Telomere activity in HIV infection is complicated because telomerase possesses a reverse transcriptase that shares homology with HIV reverse transcriptase, suggesting that enzyme processivity is a limiting factor for telomere maintenance and potentially leading to further shortening telomere length and potentially mimicking immunosenescence [154]. Regarding CC enrichment, the most significantly enriched term was the cytosol with 51 proteins being located at this compartment at some moment in the cell. In the pathway enrichment analysis, pathways involved in other infectious diseases were enriched including diseases like Chagas disease, hepatitis C, herpes simplex infection and legionellosis. This suggests that some of these genes and encoding proteins have central roles in response to several infectious diseases.

Among the 81 identified human proteins that interact with both Mtb and HIV proteins, none of them has known inhibitors. However, several of these proteins interact directly with other human proteins that have known inhibitors. In order to better perceive and ana-

lyze the interactions between these proteins, using the SPINET Visualization and Analysis tool, a sub-network was created with the human proteins that interact with both Mtb and HIV proteins, the respective Mtb and HIV proteins and the 60 human proteins with known inhibitors (Figure 22). In this network, the human protein with the higher number of interactions was the Integrin alpha-4 (gene name: ITGA4; UniprotKB: P13612) interacting with 42 other human proteins. The Alanine and proline-rich secreted protein Apa (gene name: apa; UniProtKB - P9WIR7) interacts with 11 human proteins being the Mtb protein with most interactions in this sub-network. The HIV proteins were the Gag polyprotein (gene name: gag; UniProtKB: P14349) interacting with 34 human proteins. By the analysis of this network, it was possible to identify multiple proteins with known inhibitors that interact directly with human proteins that interact with both pathogens. For example, the *Tumor Necrosis Factor (TNF)* (gene name: TNF; UniprotKB: P01375), that has 30 known inhibitors, does not interact directly with any Mtb or HIV protein, however, interacts with three human proteins (Polyubiquitin-C (gene name:UBC; UniProtKB: P0CG48), F-box/WD repeat-containing protein 11 (gene name: FBXW11; UniProtKB: Q9UKB1) and the heterogeneous nuclear ribonucleoprotein A1 (gene name: HNRNPA1; UniProtKB: P09651)) that interact with proteins of both pathogens. TNF inhibitors are commonly used to treat multiple inflammatory diseases, including rheumatoid arthritis, the seronegative spondyloarthropathies, psoriasis, and inflammatory bowel disease [155]. Interestingly, in the case of TB and AIDS/HIV, the use of TNF inhibitors induce different effects on both diseases. For example, the use of infliximab, a TNF inhibitor, is associated with an increased risk of developing active TB [156, 157]. Instead, in the case of AIDS/HIV, some studies showed that no significant clinical adverse effects were associated with the treatment of HIV-positive patients with anti-TNF therapy [158, 159]. The identification of proteins like this one could offer an excellent opportunity in the discovery of target proteins for the reuse of known drugs used for different purposes and in the identification of potential side effects in the treatment of patients with AIDS/HIV and or TB.

SPINET's approach to data integration gives a high coverage of all information on PPI data available for the Human, Mtb and HIV organisms while maintaining a good reliability and integrity of the data. This platform can be used as a start point for comprehensive types of analysis as the information gathered in it offers a high quantity of raw data willing to be mined. The storage of all these data, not only PPI data but also other types of biological data in one platform stands as an attractive source of more reliable and complete data, without the hard and time-consuming task of data and nomenclature mapping required from the use of data from multiple sources. SPINET also provides a set of tools that allow an initial analysis and visualization of the relevant data. Thus, the SPINET platform stands out as one of the most comprehensive PPI data repository storing the largest dataset of PPIs for the human, Mtb and HIV organisms in the world.

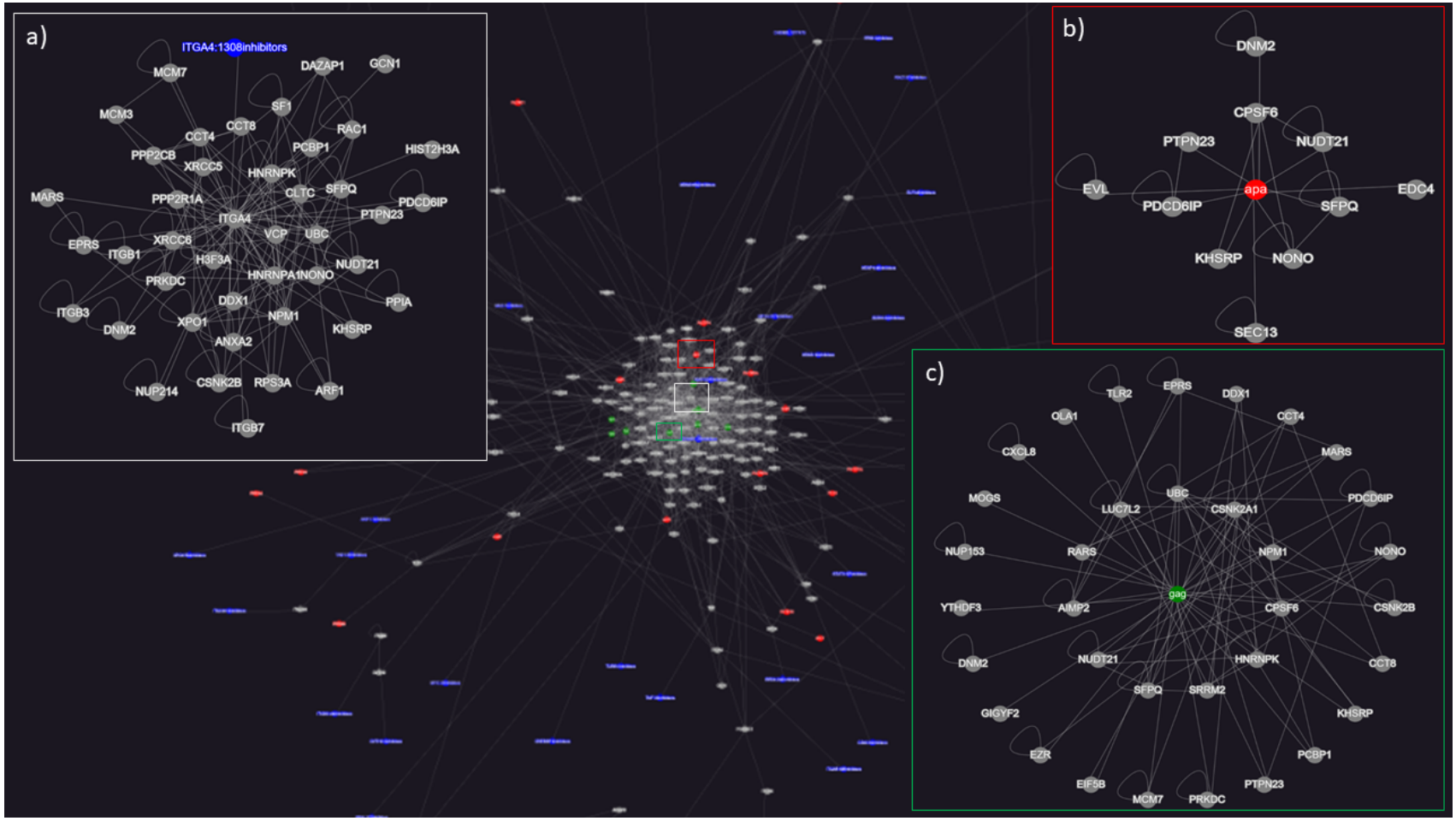


Figure 24: Sub-networks for the proteins: a) human protein ITGA4 ; b) Mtb protein Apa; c) HIV protein gag.

7

CONCLUSION

7.1 CONCLUSIONS

Our approach to data integration is based on a data warehousing technology that aggregates data from the major PPI data repositories as well as from other databases containing different types of biological data, so that it can facilitate high-level analysis, summarization of information, and extraction of new knowledge hidden in the data. Therefore, SPINET provides, in a single platform, a complete set of data that is spread over multiple sources without the hard and time-consuming task of dealing with the huge amount of available data, the heterogeneity of the data, and the different capabilities of the sources. The low data overlap found between the different databases allied with the fact that the majority of these databases do not frequently update their platforms with data generated in recent studies reinforces the need for initiatives like SPINET.

The presented platform has many differentiating factors that distinguish it from the existing PPI platforms. SPINET enables the use of both intra and inter-species interactions allowing the identification of proteins that in the organism network may go unnoticed but in the network involving multiple organisms can have relevant functions. The use of inhibitors data also provides an important resource in the study of potential drug targets. These targets often occupy a central position in the PPIN and can be easily identified using the SPINET analysis and visualization tool. In the global network formed by the three organisms, 81 human proteins were identified, and despite the fact that none of them have known inhibitors, these proteins interact with many other human proteins that are targeted by multiple inhibitors. The study of the networks formed by these proteins allied with information on inhibitors could provide valuable insights in the discovering of new drug targets using already launched and approved drugs in the fight against the TB and AIDS diseases.

The major objectives of this thesis were the data collection and curation, development of a relational database, and the development of an analysis and visualization tool. In summary, the SPINET platform stands out as an intuitive tool that bridges the gap between experi-

mental and computational biologists providing an easy PPI data accessibility, discovery, re-use, preservation and, especially, sharing for the Human, Mtb and HIV organisms.

Finally, the tools developed during the course of this thesis are part of a larger project planned around the SPINET platform held by the evoBiomed group at the *Life and Health Science Research Institute (ICVS)*, *School of Medicine (EM)*, University of Minho.

7.2 PROSPECT FOR FUTURE WORK

There are a considerable number of ongoing and future projects that will be implemented to improve and exploit the SPINET capabilities. The objective is to increase the usefulness of the platform and to extract knowledge from the data contained in it.

7.2.1 *Continue to develop SPINET and add new features to it*

The next step is to continue to develop the SPINET platform by turning it into a dynamic web server. Thus, through a common web browser, any person may request, visualize and analyze any data contained in the SPINET database.

Some new features to add to the SPINET platform are already being planned:

- Creation of a web form so that PPI data from individual studies can be submitted to the SPINET database.
- Creation of a module to display the networks in 3D. The use of a 3D layout adds more available space, making it easier to display and navigate through larger structures.
- Development of a SPINET *Representational State Transfer (REST)* Service to provide access to the interaction data in the SPINET database over *Hyper Text Transfer Protocol Secure (HTTPS)*, programmatically or in a browser.

7.2.2 *Keep adding more data*

It is of primordial importance to keep the data in the SPINET database updated. Data generated in recent studies bring important value to the already huge amount of data stored in the SPINET database as the majority of these data is still not integrated into some or none of the major public PPI repositories.

It is also important to keep adding data from other inhibitors databases like PubChem [160] and DrugBank [161], into the SPINET database.

7.2.3 *Data mining*

The quantity of valuable insights hidden in the networks formed by PPIs is endless and willing to be mined. Multiple types of analysis can be performed to extract the maximum information possible from it. Multiple studies were already performed with smaller networks [162, 163, 164] and with the data stored in the SPINET database, they can be done on a larger scale and with a higher reliability.

The first planned analysis is focused on finding possible human targets by using computational and statisctical analysis of the networks formed by the three organisms of interest. Hereafter, in a wet lab context test inhibitors for that human proteins, focusing on host-directed therapy, and evaluate possible differences on the course of infection.

BIBLIOGRAPHY

[1] Matthias E. Futschik, Gautam Chaurasia, and Hanspeter Herzel. Comparison of human protein–protein interaction maps. *Bioinformatics*, 23(5):605–611, jan 2007. doi: 10.1093/bioinformatics/btl683. URL https://doi.org/10.1093/bioinformatics/btl683.

[2] Wolfgang Huber, Vincent J Carey, Li Long, Seth Falcon, and Robert Gentleman. Graphs in molecular biology. *BMC Bioinformatics*, 8(Suppl 6):S8, 2007. doi: 10.1186/1471-2105-8-s6-s8. URL https://doi.org/10.1186/1471-2105-8-s6-s8.

[3] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, may 2002. doi: 10.1038/nature750. URL https://doi.org/10.1038/nature750.

[4] Max Franz, Christian T. Lopes, Gerardo Huck, Yue Dong, Onur Sumer, and Gary D. Bader. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, page btv557, sep 2015. doi: 10.1093/bioinformatics/btv557. URL https://doi.org/10.1093/bioinformatics/btv557.

[5] Giuseppe Agapito, Pietro Guzzi, and Mario Cannataro. Visualization of protein interaction networks: problems and solutions. *BMC Bioinformatics*, 14(Suppl 1):S1, 2013. doi: 10.1186/1471-2105-14-s1-s1. URL https://doi.org/10.1186/1471-2105-14-s1-s1.

[6] Amadeu S. Campos Filho, Magdala A. Novaes, and Alex S. Gomes. A 3d visualization framework to social network monitoring and analysis. *Computers in Human Behavior*, 49:623–634, aug 2015. doi: 10.1016/j.chb.2015.03.053. URL https://doi.org/10.1016/j.chb.2015.03.053.

[7] Judith Bruchfeld, Margarida Correia-Neves, and Gunilla Källenius. Tuberculosis and HIV coinfection. *Cold Spring Harbor Perspectives in Medicine*, 5(7):a017871, feb 2015. doi: 10.1101/cshperspect.a017871. URL https://doi.org/10.1101/cshperspect.a017871.

[8] World Health Organization. *Global Tuberculosis Report 2018*. Global Tuberculosis Control. World Health Organization, 2018. ISBN 978-92-4-156564-6. URL http://www.who.int/tb/publications/global_report/en/.

[9] N. A. Knechel. Tuberculosis: Pathophysiology, clinical features, and diagnosis. *Critical Care Nurse*, 29(2):34–43, apr 2009. doi: 10.4037/ccn2009968. URL https://doi.org/10.4037/ccn2009968.

[10] K. Sakamoto. The pathology of mycobacterium tuberculosis infection. *Veterinary Pathology*, 49(3):423–439, jan 2012. doi: 10.1177/0300985811429313. URL https://doi.org/10.1177/0300985811429313.

[11] Helder Novais Bastos, Nuno S. Osório, Sebastien Gagneux, Iñaki Comas, and Margarida Saraiva. The troika host–pathogen–extrinsic factors in tuberculosis: Modulating inflammation and clinical outcomes. *Frontiers in Immunology*, 8, jan 2018. doi: 10.3389/fimmu.2017.01948. URL https://doi.org/10.3389/fimmu.2017.01948.

[12] Yon Ju Ryu. Diagnosis of pulmonary tuberculosis: Recent advances and diagnostic algorithms. *Tuberculosis and Respiratory Diseases*, 78(2):64, 2015. doi: 10.4046/trd.2015.78.2.64. URL https://doi.org/10.4046/trd.2015.78.2.64.

[13] Xpert mtb/rif assay for the diagnosis of pulmonary and extrapulmonary tb in adults and children. URL http://www.who.int/tb/publications/xpert-mtb-rif-assay-diagnosis-policy-update/en/.

[14] Surajit Nayak and Basanti Acharjya. Mantoux test and its interpretation. *Indian Dermatology Online Journal*, 3(1):2, 2012. doi: 10.4103/2229-5178.93479. URL https://doi.org/10.4103/2229-5178.93479.

[15] M. Pai, C. M. Denkinger, S. V. Kik, M. X. Rangaka, A. Zwerling, O. Oxlade, J. Z. Metcalfe, A. Cattamanchi, D. W. Dowdy, K. Dheda, and N. Banaei. Gamma interferon release assays for detection of mycobacterium tuberculosis infection. *Clinical Microbiology Reviews*, 27(1):3–20, jan 2014. doi: 10.1128/cmr.00034-13. URL https://doi.org/10.1128/cmr.00034-13.

[16] C. Robert Horsburgh, Clifton E. Barry, and Christoph Lange. Treatment of tuberculosis. *New England Journal of Medicine*, 373(22):2149–2160, nov 2015. doi: 10.1056/nejmra1413919. URL https://doi.org/10.1056/nejmra1413919.

[17] Unaids data 2018, . URL http://www.unaids.org/en/resources/documents/2018/unaids-data-2018.

[18] Fact sheet - latest statistics on the status of the aids epidemic, . URL http://www.unaids.org/en/resources/fact-sheet.

[19] Gary Maartens, Connie Celum, and Sharon R Lewin. HIV infection: epidemiology, pathogenesis, treatment, and prevention. *The Lancet*, 384(9939):258–271, jul 2014. doi:

10.1016/s0140-6736(14)60164-1. URL https://doi.org/10.1016/s0140-6736(14)60164-1.

[20] Eric S Daar, Christopher D Pilcher, and Frederick M Hecht. Clinical presentation and diagnosis of primary HIV-1 infection. *Current Opinion in HIV and AIDS*, 3(1):10–15, jan 2008. doi: 10.1097/coh.0b013e3282f2e295. URL https://doi.org/10.1097/coh.0b013e3282f2e295.

[21] J. K. Cornett and T. J. Kirn. Laboratory diagnosis of HIV in adults: A review of current methods. *Clinical Infectious Diseases*, 57(5):712–718, may 2013. doi: 10.1093/cid/cit281. URL https://doi.org/10.1093/cid/cit281.

[22] Lucy C. K. Bell and Mahdad Noursadeghi. Pathogenesis of HIV-1 and mycobacterium tuberculosis co-infection. *Nature Reviews Microbiology*, 16(2):80–90, nov 2017. doi: 10.1038/nrmicro.2017.128. URL https://doi.org/10.1038/nrmicro.2017.128.

[23] Karen Cohen and Graeme Meintjes. Management of individuals requiring antiretroviral therapy and TB treatment. *Current Opinion in HIV and AIDS*, 5(1):61–69, jan 2010. doi: 10.1097/coh.0b013e3283339309. URL https://doi.org/10.1097/coh.0b013e3283339309.

[24] Stephen D. Lawn, Sophie V. Brooks, Katharina Kranzer, Mark P. Nicol, Andrew Whitelaw, Monica Vogt, Linda-Gail Bekker, and Robin Wood. Screening for HIV-associated tuberculosis and rifampicin resistance before antiretroviral therapy using the xpert MTB/RIF assay: A prospective study. *PLoS Medicine*, 8(7):e1001067, jul 2011. doi: 10.1371/journal.pmed.1001067. URL https://doi.org/10.1371/journal.pmed.1001067.

[25] Colleen F Hanrahan and Annelies Van Rie. Urine antigen test for diagnosis of HIV-associated tuberculosis. *The Lancet Infectious Diseases*, 12(11):826, nov 2012. doi: 10.1016/s1473-3099(12)70219-0. URL https://doi.org/10.1016/s1473-3099(12)70219-0.

[26] Ishani Pathmanathan, Anand Date, William L. Coggin, John Nkengasong, Amy S. Piatek, and Heather Alexander. Rolling out xpert MTB/RIF® for tuberculosis detection in HIV-positive populations: An opportunity for systems strengthening. *African Journal of Laboratory Medicine*, 6(2), mar 2017. doi: 10.4102/ajlm.v6i2.460. URL https://doi.org/10.4102/ajlm.v6i2.460.

[27] Vicki Brower. Proteomics: biology in the post-genomic era. *EMBO reports*, 2(7):558–560, jul 2001. doi: 10.1093/embo-reports/kve144. URL https://doi.org/10.1093/embo-reports/kve144.

[28] Aidong Zhang. *Protein Interaction Networks: Computational Analysis*. Cambridge University Press, New York, NY, USA, 1st edition, 2009. ISBN 0521888956, 9780521888950.

[29] Tuba Sevimoglu and Kazim Yalcin Arga. The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, 11(18):22–27, aug 2014. doi: 10.1016/j.csbj.2014.08.008. URL https://doi.org/10.1016/j.csbj.2014.08.008.

[30] S. E. Acuner Ozbabacan, H. B. Engin, A. Gursoy, and O. Keskin. Transient protein-protein interactions. *Protein Engineering Design and Selection*, 24(9):635–648, jun 2011. doi: 10.1093/protein/gzr025. URL https://doi.org/10.1093/protein/gzr025.

[31] Javier De Las Rivas and Celia Fontanillo. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, 6(6):e1000807, jun 2010. doi: 10.1371/journal.pcbi.1000807. URL https://doi.org/10.1371/journal.pcbi.1000807.

[32] James R. Perkins, Ilhem Diboun, Benoit H. Dessailly, Jon G. Lees, and Christine Orengo. Transient protein-protein interactions: Structural, functional, and network properties. *Structure*, 18(10):1233–1243, oct 2010. doi: 10.1016/j.str.2010.08.007. URL https://doi.org/10.1016/j.str.2010.08.007.

[33] Kristine A Pattin and Jason H Moore. Role for protein-protein interaction databases in human genetics. *Expert Review of Proteomics*, 6(6):647–659, dec 2009. doi: 10.1586/epr.09.86. URL https://doi.org/10.1586/epr.09.86.

[34] Amit V. Pandey, Colin J. Henderson, Yuji Ishii, Michel Kranendonk, Wayne L. Backes, and Ulrich M. Zanger. Editorial: Role of protein-protein interactions in metabolism: Genetics, structure, function. *Frontiers in Pharmacology*, 8, nov 2017. doi: 10.3389/fphar.2017.00881. URL https://doi.org/10.3389/fphar.2017.00881.

[35] Michael E. Cusick, Niels Klitgord, Marc Vidal, and David E. Hill. Interactome: gateway into systems biology. *Human Molecular Genetics*, 14(suppl_2):R171–R181, oct 2005. doi: 10.1093/hmg/ddi335. URL https://doi.org/10.1093/hmg/ddi335.

[36] Stanley Fields and Ok kyu Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, jul 1989. doi: 10.1038/340245a0. URL https://doi.org/10.1038/340245a0.

[37] I. Serebriiskii. Yeast two-hybrid system for studying protein-protein interactions–stage 3: Screen for interacting proteins. *Cold Spring Harbor Protocols*, 2010(5):pdb.prot5431–pdb.prot5431, may 2010. doi: 10.1101/pdb.prot5431. URL https://doi.org/10.1101/pdb.prot5431.

[38] Daniel Auerbach and Igor Stagljar. Yeast two-hybrid protein-protein interaction networks. In *Proteomics and Protein-Protein Interactions*, pages 19–31. Springer US, 2005. doi: 10.1007/0-387-24532-4_2. URL https://doi.org/10.1007/0-387-24532-4_2.

[39] M. Koegl and P. Uetz. Improving yeast two-hybrid screening systems. *Briefings in Functional Genomics and Proteomics*, 6(4):302–312, jan 2008. doi: 10.1093/bfgp/elm035. URL https://doi.org/10.1093/bfgp/elm035.

[40] Anne-Claude Gingras, Matthias Gstaiger, Brian Raught, and Ruedi Aebersold. Analysis of protein complexes using mass spectrometry. *Nature Reviews Molecular Cell Biology*, 8(8):645–654, aug 2007. doi: 10.1038/nrm2208. URL https://doi.org/10.1038/nrm2208.

[41] Andreas Bauer and Bernhard Kuster. Affinity purification-mass spectrometry. *European Journal of Biochemistry*, 270(4):570–578, jan 2003. doi: 10.1046/j.1432-1033.2003.03428.x. URL https://doi.org/10.1046/j.1432-1033.2003.03428.x.

[42] Corry Paul, HeeSool Rho, Johnathan Neiswinger, and Heng Zhu. Characterization of protein-protein interactions using protein microarrays. *Cold Spring Harbor Protocols*, 2016(10):pdb.prot087965, oct 2016. doi: 10.1101/pdb.prot087965. URL https://doi.org/10.1101/pdb.prot087965.

[43] Chien-Sheng Chen and Heng Zhu. Protein microarrays. *BioTechniques*, 40(4):423–429, apr 2006. doi: 10.2144/06404te01. URL https://doi.org/10.2144/06404te01.

[44] Gavin MacBeath. Protein microarrays and proteomics. *Nature Genetics*, 32(Supp):526–532, dec 2002. doi: 10.1038/ng1037. URL https://doi.org/10.1038/ng1037.

[45] Sylvain Pitre, Md Alamgir, James R. Green, Michel Dumontier, Frank Dehne, and Ashkan Golshani. Computational methods for predicting protein-protein interactions. In *Protein - Protein Interaction*, pages 247–267. Springer Berlin Heidelberg, 2008. doi: 10.1007/10_2007_089. URL https://doi.org/10.1007/10_2007_089.

[46] Tobias Ehrenberger, Lewis C. Cantley, and Michael B. Yaffe. Computational prediction of protein-protein interactions. In *Methods in Molecular Biology*, pages 57–75. Springer New York, 2015. doi: 10.1007/978-1-4939-2425-7_4. URL https://doi.org/10.1007/978-1-4939-2425-7_4.

[47] Haidong Wang, Eran Segal, Asa Ben-Hur, Qian-Ru Li, Marc Vidal, and Daphne Koller. InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biology*, 8(9):R192, 2007. doi: 10.1186/gb-2007-8-9-r192. URL https://doi.org/10.1186/gb-2007-8-9-r192.

[48] Jiang Xie, Wu Zhang, Jian Mei, Zhi li Gu, Ji zong Wu, Hui Li, and Lü wen Zhang. A novel computational method for protein-protein interaction networks prediction of alpha-synuclein. *Journal of Shanghai University (English Edition)*, 12(6):501–507, dec 2008. doi: 10.1007/s11741-008-0608-1. URL https://doi.org/10.1007/s11741-008-0608-1.

[49] Javad Zahiri, Joseph Bozorgmehr, and Ali Masoudi-Nejad. Computational prediction of protein-protein interaction networks: Algorithms and resources. *Current Genomics*, 14(6):397–414, sep 2013. doi: 10.2174/1389202911314060004. URL https://doi.org/10.2174/1389202911314060004.

[50] Jingkai Yu and Farshad Fotouhi. Computational approaches for predicting protein-protein interactions: A survey. *Journal of Medical Systems*, 30(1):39–44, feb 2006. doi: 10.1007/s10916-006-7402-3. URL https://doi.org/10.1007/s10916-006-7402-3.

[51] David Eisenberg, Edward M. Marcotte, Ioannis Xenarios, and Todd O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, jun 2000. doi: 10.1038/35015694. URL https://doi.org/10.1038/35015694.

[52] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, apr 1999. doi: 10.1073/pnas.96.8.4285. URL https://doi.org/10.1073/pnas.96.8.4285.

[53] L. R. Matthews. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126, dec 2001. doi: 10.1101/gr.205301. URL https://doi.org/10.1101/gr.205301.

[54] J. Wojcik and V. Schachter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17(Suppl 1):S296–S305, jun 2001. doi: 10.1093/bioinformatics/17.suppl_1.s296. URL https://doi.org/10.1093/bioinformatics/17.suppl_1.s296.

[55] Lucy Skrabanek, Harpreet K. Saini, Gary D. Bader, and Anton J. Enright. Computational prediction of protein-protein interactions. *Molecular Biotechnology*, 38(1):1–17, aug 2007. doi: 10.1007/s12033-007-0069-2. URL https://doi.org/10.1007/s12033-007-0069-2.

[56] P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences*, 99(9):5896–5901, apr 2002. doi: 10.1073/pnas.092147999. URL https://doi.org/10.1073/pnas.092147999.

[57] A. S. Aytuna, A. Gursoy, and O. Keskin. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12):2850–2855, apr 2005. doi: 10.1093/bioinformatics/bti443. URL https://doi.org/10.1093/bioinformatics/bti443.

[58] D. S. Goldberg and F. P. Roth. Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences*, 100(8):4372–4376, apr 2003. doi: 10.1073/pnas.0735871100. URL https://doi.org/10.1073/pnas.0735871100.

[59] Rintaro Saito, Masaru Tomita, Harukazu Suzuki, and Yoshihide Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Genome Informatics*, 13:324–325, 2002. doi: 10.11234/gi1990.13.324. URL https://doi.org/10.11234/gi1990.13.324.

[60] R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19(6):756–763, apr 2003. doi: 10.1093/bioinformatics/btg070. URL https://doi.org/10.1093/bioinformatics/btg070.

[61] S. Asthana. Predicting protein complex membership using probabilistic network reliability. *Genome Research*, 14(6):1170–1175, may 2004. doi: 10.1101/gr.2203804. URL https://doi.org/10.1101/gr.2203804.

[62] S. Letovsky and S. Kasif. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19(Suppl 1):i197–i204, jul 2003. doi: 10.1093/bioinformatics/btg1026. URL https://doi.org/10.1093/bioinformatics/btg1026.

[63] H. N. Chua, W.-K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, apr 2006. doi: 10.1093/bioinformatics/btl145. URL https://doi.org/10.1093/bioinformatics/btl145.

[64] Fiona Browne, Huiru Zheng, Haiying Wang, and Francisco Azuaje. From experimental approaches to computational techniques: A review on the prediction of protein-protein interactions. *Advances in Artificial Intelligence*, 2010:1–15, 2010. doi: 10.1155/2010/924529. URL https://doi.org/10.1155/2010/924529.

[65] R. Jansen. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, oct 2003. doi: 10.1126/science.1087361. URL https://doi.org/10.1126/science.1087361.

[66] L. J. Lu. Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15(7):945–953, jun 2005. doi: 10.1101/gr.3610305. URL https://doi.org/10.1101/gr.3610305.

[67] Fiona Browne, Haiying Wang, Huiru Zheng, and Francisco Azuaje. Supervised statistical and machine learning approaches to inferring pairwise and module-based protein interaction networks. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*. IEEE, oct 2007. doi: 10.1109/bibe.2007.4375748. URL https://doi.org/10.1109/bibe.2007.4375748.

[68] Yanjun Qi, Ziv Bar-Joseph, and Judith Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics*, 63(3):490–500, jan 2006. doi: 10.1002/prot.20865. URL https://doi.org/10.1002/prot.20865.

[69] Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Research*, 36(9):3025–3030, apr 2008. doi: 10.1093/nar/gkn159. URL https://doi.org/10.1093/nar/gkn159.

[70] Siaw Ling Lo, Cong Zhong Cai, Yu Zong Chen, and Maxey C. M. Chung. Effect of training datasets on support vector machine prediction of protein-protein interactions. *PROTEOMICS*, 5(4):876–884, mar 2005. doi: 10.1002/pmic.200401118. URL https://doi.org/10.1002/pmic.200401118.

[71] X.-W. Chen and M. Liu. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400, oct 2005. doi: 10.1093/bioinformatics/bti721. URL https://doi.org/10.1093/bioinformatics/bti721.

[72] Yanjun Qi, Judith Klein-Seetharaman, and Ziv Bar-Joseph. Random forest similarity for protein-protein interaction prediction from multiple sources. In *Biocomputing 2005*. World Scientific, dec 2004. doi: 10.1142/9789812702456_0050. URL https://doi.org/10.1142/9789812702456_0050.

[73] Zhiqiang Ma, Chunguang Zhou, Linying Lu, Yanan Ma, Pingping Sun, and Ying Cui. Predicting protein-protein interactions based on BP neural network. In *2007 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. IEEE, nov 2007. doi: 10.1109/bibmw.2007.4425393. URL https://doi.org/10.1109/bibmw.2007.4425393.

[74] P. Fariselli, A. Zauli, I. Rossi, M. Finelli, P.L. Martelli, and R. Casadio. A neural network method to improve prediction of protein-protein interaction sites in het-

erocomplexes. In *2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718)*. IEEE, 2003. doi: 10.1109/nnsp.2003.1318002. URL https://doi.org/10.1109/nnsp.2003.1318002.

[75] Luke Hakes, John W Pinney, David L Robertson, and Simon C Lovell. Protein-protein interaction networks and biology - what is the connection? *Nature Biotechnology*, 26(1):69–72, jan 2008. doi: 10.1038/nbt0108-69. URL https://doi.org/10.1038/nbt0108-69.

[76] Zoozeal Thakur, , Renu Dharra, Vandana Saini, Ajit Kumar, Promod K. Mehta, , , and and. Insights from the protein-protein interaction network analysis of mycobacterium tuberculosis toxinantitoxin systems. *Bioinformation*, 13(11):380–387, nov 2017. doi: 10.6026/97320630013380. URL https://doi.org/10.6026/97320630013380.

[77] Xiaoke Ma and Lin Gao. Discovering protein complexes in protein interaction networks via exploring the weak ties effect. *BMC Systems Biology*, 6(Suppl 1):S6, 2012. doi: 10.1186/1752-0509-6-s1-s6. URL https://doi.org/10.1186/1752-0509-6-s1-s6.

[78] Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, 10(6):2763–2788, jun 2009. doi: 10.3390/ijms10062763. URL https://doi.org/10.3390/ijms10062763.

[79] Emmanuelle Becker, Benoît Robisson, Charles E. Chapple, Alain Guénoche, and Christine Brun. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28(1):84–90, nov 2011. doi: 10.1093/bioinformatics/btr621. URL https://doi.org/10.1093/bioinformatics/btr621.

[80] Yuri Pritykin and Mona Singh. Simple topological features reflect dynamics and modularity in protein interaction networks. *PLoS Computational Biology*, 9(10):e1003243, oct 2013. doi: 10.1371/journal.pcbi.1003243. URL https://doi.org/10.1371/journal.pcbi.1003243.

[81] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800–2805, sep 2006. doi: 10.1093/bioinformatics/btl467. URL https://doi.org/10.1093/bioinformatics/btl467.

[82] Darren Davis, Ömer Nebil Yaveroğlu, Nol Malod-Dognin, Aleksandar Stojmirovic, and Nataša Pržulj. Topology-function conservation in protein-protein interaction networks. *Bioinformatics*, 31(10):1632–1639, jan 2015. doi: 10.1093/bioinformatics/btv026. URL https://doi.org/10.1093/bioinformatics/btv026.

[83] Anatol Rapoport. Contribution to the theory of random and biased nets. *The Bulletin of Mathematical Biophysics*, 19(4):257–277, dec 1957. doi: 10.1007/bf02478417. URL https://doi.org/10.1007/bf02478417.

[84] Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, feb 2004. doi: 10.1038/nrg1272. URL https://doi.org/10.1038/nrg1272.

[85] J.C. Nacher, M. Hayashida, and T. Akutsu. Emergence of scale-free distribution in protein-protein interaction networks based on random selection of interacting domain pairs. *Biosystems*, 95(2):155–159, feb 2009. doi: 10.1016/j.biosystems.2008.10.002. URL https://doi.org/10.1016/j.biosystems.2008.10.002.

[86] S. Maslov. Specificity and stability in topology of protein networks. *Science*, 296(5569): 910–913, may 2002. doi: 10.1126/science.1065103. URL https://doi.org/10.1126/science.1065103.

[87] A. del Sol, H. Fujihashi, and P. O'Meara. Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 21(8):1311–1315, jan 2005. doi: 10.1093/bioinformatics/bti167. URL https://doi.org/10.1093/bioinformatics/bti167.

[88] Reuven Cohen and Shlomo Havlin. Scale-free networks are ultrasmall. *Physical Review Letters*, 90(5), feb 2003. doi: 10.1103/physrevlett.90.058701. URL https://doi.org/10.1103/physrevlett.90.058701.

[89] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, may 2001. doi: 10.1038/35075138. URL https://doi.org/10.1038/35075138.

[90] Maliackal Poulo Joy, Amy Brock, Donald E. Ingber, and Sui Huang. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, 2005(2):96–103, 2005. doi: 10.1155/jbb.2005.96. URL https://doi.org/10.1155/jbb.2005.96.

[91] A. Ozgur, T. Vu, G. Erkan, and D. R. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13): i277–i285, jun 2008. doi: 10.1093/bioinformatics/btn182. URL https://doi.org/10.1093/bioinformatics/btn182.

[92] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL https://doi.org/10.1088/1742-5468/2008/10/p10008.

[93] Einat Sprinzak, Shmuel Sattath, and Hanah Margalit. How reliable are experimental protein–protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, apr 2003. doi: 10.1016/s0022-2836(03)00239-0. URL https://doi.org/10.1016/s0022-2836(03)00239-0.

[94] Tejaswini Narayanan, Merril Gersten, Shankar Subramaniam, and Ananth Grama. Modularity detection in protein-protein interaction networks. *BMC Research Notes*, 4(1):569, 2011. doi: 10.1186/1756-0500-4-569. URL https://doi.org/10.1186/1756-0500-4-569.

[95] A. D. King, N. Przulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, jun 2004. doi: 10.1093/bioinformatics/bth351. URL https://doi.org/10.1093/bioinformatics/bth351.

[96] F. Luo, Y. Yang, C.-F. Chen, R. Chang, J. Zhou, and R. H. Scheuermann. Modular organization of protein interaction networks. *Bioinformatics*, 23(2):207–214, nov 2006. doi: 10.1093/bioinformatics/btl562. URL https://doi.org/10.1093/bioinformatics/btl562.

[97] Min Li, Jianxin Wang, and Jianer Chen. A fast agglomerate algorithm for mining functional modules in protein interaction networks. In *2008 International Conference on BioMedical Engineering and Informatics*. IEEE, may 2008. doi: 10.1109/bmei.2008.121. URL https://doi.org/10.1109/bmei.2008.121.

[98] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697–700, may 2003. doi: 10.1038/nbt825. URL https://doi.org/10.1038/nbt825.

[99] Xiaoqing Peng, Jianxin Wang, Wei Peng, Fang-Xiang Wu, and Yi Pan. Protein–protein interactions: detection, reliability assessment and applications. *Briefings in Bioinformatics*, page bbw066, jul 2016. doi: 10.1093/bib/bbw066. URL https://doi.org/10.1093/bib/bbw066.

[100] Min Li, Hanhui Zhang, Jian xin Wang, and Yi Pan. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Systems Biology*, 6(1):15, 2012. doi: 10.1186/1752-0509-6-15. URL https://doi.org/10.1186/1752-0509-6-15.

[101] Xiwei Tang, Jianxin Wang, Jiancheng Zhong, and Yi Pan. Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(2):407–418, mar 2014. doi: 10.1109/tcbb.2013.2295318. URL https://doi.org/10.1109/tcbb.2013.2295318.

[102] Tomer Shlomi, Daniel Segal, Eytan Ruppin, and Roded Sharan. Qpath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7(1):199, 2006. doi: 10.1186/1471-2105-7-199. URL https://doi.org/10.1186/1471-2105-7-199.

[103] Anthony Gitter, Judith Klein-Seetharaman, Anupam Gupta, and Ziv Bar-Joseph. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Research*, 39(4):e22–e22, nov 2010. doi: 10.1093/nar/gkq1207. URL https://doi.org/10.1093/nar/gkq1207.

[104] Lei Chen, Chen Chu, Xiangyin Kong, Tao Huang, and Yu-Dong Cai. Discovery of new candidate genes related to brain development using protein interaction information. *PLOS ONE*, 10(1):e0118003, jan 2015. doi: 10.1371/journal.pone.0118003. URL https://doi.org/10.1371/journal.pone.0118003.

[105] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, oct 2016. doi: 10.1093/nar/gkw937. URL https://doi.org/10.1093/nar/gkw937.

[106] Zhan-Chao Li, Wen-Qian Zhong, Zhi-Qing Liu, Meng-Hua Huang, Yun Xie, Zong Dai, and Xiao-Yong Zou. Large-scale identification of potential drug targets based on the topological features of human protein–protein interaction network. *Analytica Chimica Acta*, 871:18–27, apr 2015. doi: 10.1016/j.aca.2015.02.032. URL https://doi.org/10.1016/j.aca.2015.02.032.

[107] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of The Royal Society Interface*, 12(112):20150571, oct 2015. doi: 10.1098/rsif.2015.0571. URL https://doi.org/10.1098/rsif.2015.0571.

[108] Helen Pearson. Biology's name game. *Nature*, 411(6838):631–632, jun 2001. doi: 10.1038/35079694. URL https://doi.org/10.1038/35079694.

[109] Jennifer Widom. Integrating heterogeneous databases: lazy or eager? *ACM Computing Surveys*, 28(4es):91–es, dec 1996. doi: 10.1145/242224.242344. URL https://doi.org/10.1145/242224.242344.

[110] Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti, Fiona S L Brinkman, Gianni Cesareni, Andrew Chatr-aryamontri, Emilie Chautard, Carol Chen, Marine Dumousseau, Johannes Goll, Robert E W Hancock, Linda I Hannick, Igor Jurisica, Jyoti Khadake,

David J Lynn, Usha Mahadevan, Livia Perfetto, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Lukasz Salwinski, Volker Stümpflen, Mike Tyers, Peter Uetz, Ioannis Xenarios, and Henning Hermjakob. Protein interaction data curation: the international molecular exchange (IMEx) consortium. *Nature Methods*, 9(4):345–350, mar 2012. doi: 10.1038/nmeth.1931. URL https://doi.org/10.1038/nmeth.1931.

[111] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, nov 2016. doi: 10.1093/nar/gkw1099. URL https://doi.org/10.1093/nar/gkw1099.

[112] Michelle R. Arkin, Yinyan Tang, and James A. Wells. Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality. *Chemistry & Biology*, 21 (9):1102–1114, sep 2014. doi: 10.1016/j.chembiol.2014.09.001. URL https://doi.org/10.1016/j.chembiol.2014.09.001.

[113] Alicia P. Higueruelo, Harry Jubb, and Tom L. Blundell. TIMBAL v2: update of a database holding small molecules modulating protein–protein interactions. *Database*, 2013, jan 2013. doi: 10.1093/database/bat039. URL https://doi.org/10.1093/database/bat039.

[114] Céline M. Labbé, Mélaine A. Kuenemann, Barbara Zarzycka, Gert Vriend, Gerry A.F. Nicolaes, David Lagorce, Maria A. Miteva, Bruno O. Villoutreix, and Olivier Sperandio. iPPI-DB: an online database of modulators of protein–protein interactions. *Nucleic Acids Research*, 44(D1):D542–D547, oct 2015. doi: 10.1093/nar/gkv982. URL https://doi.org/10.1093/nar/gkv982.

[115] Qianghua Xiao, Jianxin Wang, Xiaoqing Peng, Fang xiang Wu, and Yi Pan. Identifying essential proteins from active PPI networks constructed with dynamic gene expression. *BMC Genomics*, 16(Suppl 3):S1, 2015. doi: 10.1186/1471-2164-16-s3-s1. URL https://doi.org/10.1186/1471-2164-16-s3-s1.

[116] R. Jansen. Relating whole-genome expression data with protein-protein interactions. *Genome Research*, 12(1):37–46, jan 2002. doi: 10.1101/gr.205602. URL https://doi.org/10.1101/gr.205602.

[117] Wei Kong, Jingmao Zhang, Xiaoyang Mou, and Yang Yang. Integrating gene expression and protein interaction data for signaling pathway prediction of alzheimer's disease. *Computational and Mathematical Methods in Medicine*, 2014:1–7, 2014. doi: 10.1155/2014/340758. URL https://doi.org/10.1155/2014/340758.

[118] Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara ODonnell, Sara Oster, Chandra Theesfeld, Adnane

Sellam, and et al. The biogrid interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369D379, Dec 2016. doi: 10.1093/nar/gkw1102. URL http://dx.doi.org/10.1093/nar/gkw1102.

[119] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(D1):D857–D861, nov 2011. doi: 10.1093/nar/gkr930. URL https://doi.org/10.1093/nar/gkr930.

[120] L. Salwinski. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(90001):449D–451, jan 2004. doi: 10.1093/nar/gkh086. URL https://doi.org/10.1093/nar/gkh086.

[121] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, and et al. The mintact project–intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1), 2013. doi: 10.1093/nar/gkt1115.

[122] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. Human protein reference database–2009 update. *Nucleic Acids Research*, 37(Database):D767–D772, jan 2009. doi: 10.1093/nar/gkn892. URL https://doi.org/10.1093/nar/gkn892.

[123] Danso Ako-Adjei, William Fu, Craig Wallin, Kenneth S. Katz, Guangfeng Song, Dakshesh Darji, J. Rodney Brister, Roger G. Ptak, and Kim D. Pruitt. HIV-1, human interaction database: current status and new features. *Nucleic Acids Research*, 43(D1): D566–D570, nov 2014. doi: 10.1093/nar/gku1126. URL https://doi.org/10.1093/nar/gku1126.

[124] Fleur Jeanquartier, Claire Jean-Quartier, and Andreas Holzinger. Integrated web visualizations for protein-protein interaction databases. *BMC Bioinformatics*, 16 (1), jun 2015. doi: 10.1186/s12859-015-0615-z. URL https://doi.org/10.1186/s12859-015-0615-z.

[125] P. Shannon. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, nov 2003. doi: 10.1101/gr.1239303. URL https://doi.org/10.1101/gr.1239303.

[126] C. Prieto and J. De Las Rivas. APID: Agile protein interaction DataAnalyzer. *Nucleic Acids Research*, 34(Web Server):W298–W302, jul 2006. doi: 10.1093/nar/gkl128. URL https://doi.org/10.1093/nar/gkl128.

[127] Gustavo A Salazar, Ayton Meintjes, Gaston K Mazandu, Holifidy A Rapanol, Richard O Akinola, and Nicola J Mulder. A web-based protein interaction network visualizer. *BMC Bioinformatics*, 15(1):129, 2014. doi: 10.1186/1471-2105-15-129. URL https://doi.org/10.1186/1471-2105-15-129.

[128] Ravi Kiran Reddy Kalathur, José Pedro Pinto, Miguel A. Hernández-Prieto, Rui S.R. Machado, Dulce Almeida, Gautam Chaurasia, and Matthias E. Futschik. UniHI 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. *Nucleic Acids Research*, 42(D1):D408–D414, nov 2013. doi: 10.1093/nar/gkt1100. URL https://doi.org/10.1093/nar/gkt1100.

[129] Yi Wang, Tao Cui, Cong Zhang, Min Yang, Yuanxia Huang, Weihui Li, Lei Zhang, Chunhui Gao, Yang He, Yuqing Li, Feng Huang, Jumei Zeng, Cheng Huang, Qiong Yang, Yuxi Tian, Chunchao Zhao, Huanchun Chen, Hua Zhang, and Zheng-Guo He. Global protein-protein interaction network in the human PathogenMycobacterium tuberculosisH37rv. *Journal of Proteome Research*, 9(12):6665–6677, dec 2010. doi: 10.1021/pr100808n. URL https://doi.org/10.1021/pr100808n.

[130] Bennett H. Penn, Zoe Netter, Jeffrey R. Johnson, John Von Dollen, Gwendolyn M. Jang, Tasha Johnson, Yamini M. Ohol, Cyrus Maher, Samantha L. Bell, Kristina Geiger, Xiaotang Du, Alex Choi, Trevor Parry, Mayumi Naramura, Chen Chen, Stefanie Jaeger, Michael Shales, Dan A. Portnoy, Ryan Hernandez, Laurent Coscoy, Jeffery S. Cox, and Nevan J. Krogan. An mtb-human protein-protein interaction map reveals that bacterial LpqN antagonizes CBL, a host ubiquitin ligase that regulates the balance between anti-viral and anti-bacterial responses. oct 2017. doi: 10.1101/202598. URL https://doi.org/10.1101/202598.

[131] Guido Rossum. Python reference manual. Technical report, Amsterdam, The Netherlands, The Netherlands, 1995.

[132] Damiano Fantini. *easyPubMed: Search and Retrieve Scientific Publication Records from PubMed*, 2018. URL https://CRAN.R-project.org/package=easyPubMed. R package version 2.5.

[133] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

[134] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[135] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777, jan 2014. doi: 10.1371/journal.pone.0085777. URL https://doi.org/10.1371/journal.pone.0085777.

[136] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010. ISBN 0199206651, 9780199206650.

[137] Zoe Xi. Extending cytoscape.js: Addition of clustering algorithms, 2016. URL https://gist.github.com/zoexi/ec20705f6669882039b1ef8b05aaa4c1.

[138] A. J. Enright. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, apr 2002. doi: 10.1093/nar/30.7.1575. URL https://doi.org/10.1093/nar/30.7.1575.

[139] R. Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957, nov 2005. doi: 10.1242/jcs.02714. URL https://doi.org/10.1242/jcs.02714.

[140] D. V. Klopfenstein, Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, Jeffrey M. Yunes, Olga Botvinnik, Mark Weigel, Will Dampier, Christophe Dessimoz, Patrick Flick, and Haibao Tang. GOATOOLS: A python library for gene ontology analyses. *Scientific Reports*, 8(1), jul 2018. doi: 10.1038/s41598-018-28948-z. URL https://doi.org/10.1038/s41598-018-28948-z.

[141] Sandra Orchard, Lukasz Salwinski, Samuel Kerrien, Luisa Montecchi-Palazzi, Matthias Oesterheld, Volker Stümpflen, Arnaud Ceol, Andrew Chatr-aryamontri, John Armstrong, Peter Woollard, John J Salama, Susan Moore, Jérôme Wojcik, Gary D Bader, Marc Vidal, Michael E Cusick, Mark Gerstein, Anne-Claude Gavin, Giulio Superti-Furga, Jack Greenblatt, Joel Bader, Peter Uetz, Mike Tyers, Pierre Legrain, Stan Fields, Nicola Mulder, Michael Gilson, Michael Niepmann, Lyle Burgoon, Javier De Las Rivas, Carlos Prieto, Victoria M Perreau, Chris Hogue, Hans-Werner Mewes, Rolf Apweiler, Ioannis Xenarios, David Eisenberg, Gianni Cesareni, and Henning

Hermjakob. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature Biotechnology*, 25(8):894–898, aug 2007. doi: 10.1038/nbt1324. URL https://doi.org/10.1038/nbt1324.

[142] Jonathan Kevin Sia, Maria Georgieva, and Jyothi Rengarajan. Innate immune defenses in human tuberculosis: An overview of the interactions betweenMycobacterium tuberculosisand innate immune cells. *Journal of Immunology Research*, 2015:1–12, 2015. doi: 10.1155/2015/747543. URL https://doi.org/10.1155/2015/747543.

[143] Giovanni Barillari, Cecilia Sgadari, Valeria Fiorelli, Felipe Samaniego, Sandra Colombini, Vittorio Manzari, Andrea Modesti, Bala C. Nair, Aurelio Cafaro, Michael Stürzl, and Barbara Ensoli. The tat protein of human immunodeficiency virus type-1 promotes vascular cell growth and locomotion by engaging the alpha5beta1 and alpha5beta3 integrins and by mobilizing sequestered basic fibroblast growth factor. *Blood*, 94(2):663–672, 1999. ISSN 0006-4971. URL http://www.bloodjournal.org/content/94/2/663.

[144] Alexander S. Antonov, Galina N. Antonova, David H. Munn, Nahid Mivechi, Rudolf Lucas, John D. Catravas, and Alexander D. Verin. $\alpha v\beta 3$ integrin regulates macrophage inflammatory responses via PI3 kinase/akt-dependent NF-κb activation. *Journal of Cellular Physiology*, 226(2):469–476, nov 2010. doi: 10.1002/jcp.22356. URL https://doi.org/10.1002/jcp.22356.

[145] Ashwini Patil, Kengo Kinoshita, and Haruki Nakamura. Hub promiscuity in protein-protein interaction networks. *International Journal of Molecular Sciences*, 11(4):1930–1943, apr 2010. doi: 10.3390/ijms11041930. URL https://doi.org/10.3390/ijms11041930.

[146] Zhi-Ping Liu, Jiguang Wang, Yu-Qing Qiu, Ross KK Leung, Xiang-Sun Zhang, Stephen KW Tsui, and Luonan Chen. Inferring a protein interaction map of mycobacterium tuberculosis based on sequences and interologs. *BMC Bioinformatics*, 13(Suppl 7):S6, 2012. doi: 10.1186/1471-2105-13-s7-s6. URL https://doi.org/10.1186/1471-2105-13-s7-s6.

[147] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, jun 1998. doi: 10.1038/30918. URL https://doi.org/10.1038/30918.

[148] Nicholas Jarman, Erik Steur, Chris Trengove, Ivan Y. Tyukin, and Cees van Leeuwen. Self-organisation of small-world networks by adaptive rewiring in response to graph diffusion. *Scientific Reports*, 7(1), oct 2017. doi: 10.1038/s41598-017-12589-9. URL https://doi.org/10.1038/s41598-017-12589-9.

[149] Aleksandar Stojmirović and Yi-Kuo Yu. Information flow in interaction networks. *Journal of Computational Biology*, 14(8):1115–1143, oct 2007. doi: 10.1089/cmb.2007.0069. URL https://doi.org/10.1089/cmb.2007.0069.

[150] Benjamin Blasco, Jeffrey M. Chen, Ruben Hartkoorn, Claudia Sala, Swapna Uplekar, Jacques Rougemont, Florence Pojer, and Stewart T. Cole. Virulence regulator EspR of mycobacterium tuberculosis is a nucleoid-associated protein. *PLoS Pathogens*, 8(3): e1002621, mar 2012. doi: 10.1371/journal.ppat.1002621. URL https://doi.org/10.1371/journal.ppat.1002621.

[151] Mehdi Morchikh, Alexandra Cribier, Raoul Raffel, Sonia Amraoui, Julien Cau, Dany Severac, Emeric Dubois, Olivier Schwartz, Yamina Bennasser, and Monsef Benkirane. HEXIM1 and NEAT1 long non-coding RNA form a multi-subunit complex that regulates DNA-mediated innate immune response. *Molecular Cell*, 67(3):387–399.e5, aug 2017. doi: 10.1016/j.molcel.2017.06.020. URL https://doi.org/10.1016/j.molcel.2017.06.020.

[152] DeAnna L. Zanet, Anona Thorne, Joel Singer, Evelyn J. Maan, Beheroze Sattha, Armelle Le Campion, Hugo Soudeyns, Neora Pick, Melanie Murray, Deborah M. Money, and Hélène C. F. Côté and. Association between short leukocyte telomere length and HIV infection in a cohort study: No evidence of a relationship with antiretroviral therapy. *Clinical Infectious Diseases*, 58(9):1322–1332, jan 2014. doi: 10.1093/cid/ciu051. URL https://doi.org/10.1093/cid/ciu051.

[153] M. Lichterfeld, D. Mou, T. D. H. Cung, K. L. Williams, M. T. Waring, J. Huang, F. Pereyra, A. Trocha, G. J. Freeman, E. S. Rosenberg, B. D. Walker, and X. G. Yu. Telomerase activity of hiv-1-specific cd8+ t cells: constitutive up-regulation in controllers and selective increase by blockade of PD ligand 1 in progressors. *Blood*, 112(9):3679–3687, aug 2008. doi: 10.1182/blood-2008-01-135442. URL https://doi.org/10.1182/blood-2008-01-135442.

[154] Yun Peng, I.Saira Mian, and Neal F Lue. Analysis of telomerase processivity. *Molecular Cell*, 7(6):1201–1211, jun 2001. doi: 10.1016/s1097-2765(01)00268-4. URL https://doi.org/10.1016/s1097-2765(01)00268-4.

[155] JR Bradley. TNF-mediated inflammatory disease. *The Journal of Pathology*, 214(2):149–160, jan 2008. doi: 10.1002/path.2287. URL https://doi.org/10.1002/path.2287.

[156] Joseph Keane, Sharon Gershon, Robert P. Wise, Elizabeth Mirabile-Levens, John Kasznica, William D. Schwieterman, Jeffrey N. Siegel, and M. Miles Braun. Tuberculosis associated with infliximab, a tumor necrosis factor α-neutralizing agent. *New*

England Journal of Medicine, 345(15):1098–1104, oct 2001. doi: 10.1056/nejmoa011110. URL https://doi.org/10.1056/nejmoa011110.

[157] Juan J. Gómez-Reino, Loreto Carmona, Vicente Rodríguez Valverde, Emilio Martín Mola, and Maria Dolores Montero. Treatment of rheumatoid arthritis with tumor necrosis factor inhibitors may predispose to significant increase in tuberculosis risk: A multicenter active-surveillance report. *Arthritis & Rheumatism*, 48(8):2122–2127, aug 2003. doi: 10.1002/art.11137. URL https://doi.org/10.1002/art.11137.

[158] E J Cepeda, F M Williams, M L Ishimori, M H Weisman, and J D Reveille. The use of anti-tumour necrosis factor therapy in HIV-positive individuals with rheumatic disease. *Annals of the Rheumatic Diseases*, 67(5):710–712, aug 2007. doi: 10.1136/ard.2007.081513. URL https://doi.org/10.1136/ard.2007.081513.

[159] Primal P. Kaur, Virginia C. Chan, and Steven N. Berney. Successful etanercept use in an HIV-positive patient with rheumatoid arthritis. *JCR: Journal of Clinical Rheumatology*, 13(2):79–80, apr 2007. doi: 10.1097/01.rhu.0000260411.75599.39. URL https://doi.org/10.1097/01.rhu.0000260411.75599.39.

[160] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, sep 2015. doi: 10.1093/nar/gkv951. URL https://doi.org/10.1093/nar/gkv951.

[161] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, nov 2017. doi: 10.1093/nar/gkx1037. URL https://doi.org/10.1093/nar/gkx1037.

[162] Chaoran Yu, Hiju Hong, Jiaoyang Lu, Xuan Zhao, Wenjun Hu, Sen Zhang, Yaping Zong, Zhihai Mao, Jianwen Li, Mingliang Wang, Bo Feng, Jing Sun, and Minhua Zheng. Prediction of target genes and pathways associated with cetuximab insensitivity in colorectal cancer. *Technology in Cancer Research & Treatment*, 17: 153303381880690, jan 2018. doi: 10.1177/1533033818806905. URL https://doi.org/10.1177/1533033818806905.

[163] Cun Liu, Lijuan Liu, Chao Zhou, Jing Zhuang, Lu Wang, Yue Sun, and Changgang Sun. Protein–protein interaction networks and different clustering analysis in

burkitt's lymphoma. *Hematology*, 23(7):391–398, nov 2017. doi: 10.1080/10245332.2017.1409947. URL https://doi.org/10.1080/10245332.2017.1409947.

[164] Chen-Ching Lin, Hsueh-Fen Juan, Jen-Tsung Hsiang, Yih-Chii Hwang, Hirotada Mori, and Hsuan-Cheng Huang. Essential core of protein-protein interaction network inEscherichia coli. *Journal of Proteome Research*, 8(4):1925–1931, apr 2009. doi: 10.1021/pr8008786. URL https://doi.org/10.1021/pr8008786.