Universitat Autònoma de Barcelona

PREDICTION OF HOST-PATHOGEN PROTEIN-PROTEIN INTERACTIONS: AN INTEGRATIVE MODEL

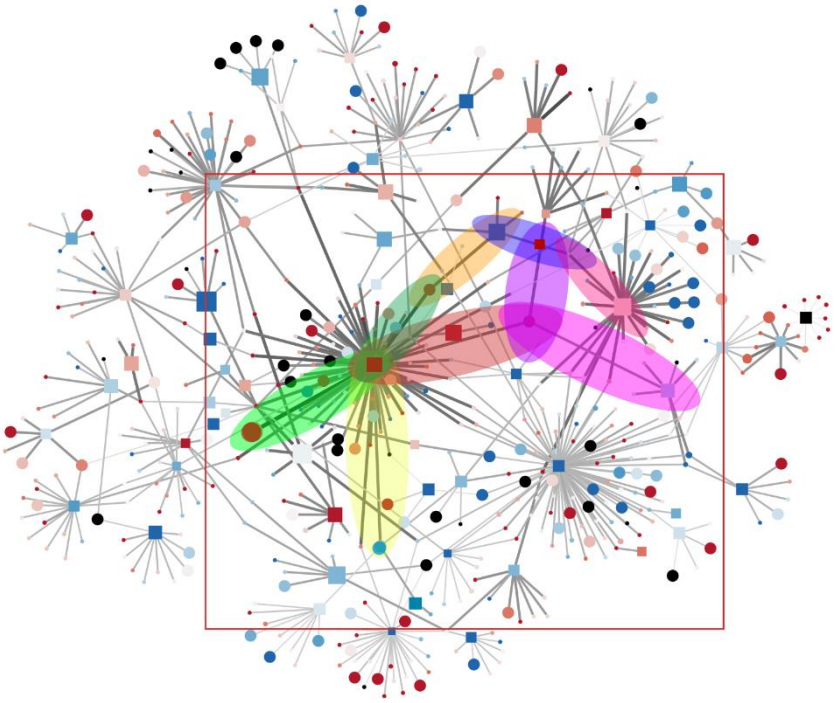
Javier Macho Rendón

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons: http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons: http://es.creativecommons.org/blog/licencias/

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license: https://creativecommons.org/licenses/?lang=en

PREDICTION OF HOST-PATHOGEN PROTEIN-PROTEIN INTERACTIONS: AN INTEGRATIVE MODEL



Javier Macho Rendón

Doctorate Program in Bioinformatics

Departament de Bioquímica i Biologia Molecular

Universitat Autònoma de Barcelona

Thesis supervisor: Dr. Marc Torrent Burgas

Academic tutor: Dr. Xavier Daura Ribera

A thesis submitted for the degree of

Doctor in Bioinformatics

September 2021

Summary

The use and misuse of antibiotics against pathogenic bacteria is accelerating the appearance of multi-drug resistant bacteria and it is becoming a serious concern in public health worldwide. In order to fight antibacterial resistance, it is crucial to get a better understanding of the mechanisms underlying infection, which occur to a great extent due to the interaction between host and pathogen proteins. In this sense, it is necessary to improve the functional annotation and characterization of these proteins and their interactions. With this aim, we have created two databases, BacFITBase and DualSeqDB, which gather information on the importance of bacterial genes and the gene expression changes that occur both in the pathogen and in the host during the infection process, respectively. Similarly, we have developed HPIPred, a host-pathogen protein-protein interaction prediction system, based on the numerical encoding of physicochemical properties of amino acids, capable of integrating phenotypic information to guide the prediction process. The combination of these tools could be useful as a guidance in the development of new drugs against antibacterial resistance.

Resumen

El uso descontrolado de antibióticos contra bacterias patógenas está acelerando la aparición de bacterias resistentes a múltiples fármacos y se está convirtiendo en un grave problema en materia de salud pública mundial. Para combatir esta resistencia bacteriana, es fundamental obtener un mayor conocimiento de los mecanismos de infección, que se producen en gran medida por la interacción entre proteínas del patógeno y proteínas del organismo huésped. En este sentido, es necesaria una mejora en la anotación funcional y caracterización de estas proteínas y de sus interacciones. Con este objetivo, hemos creado las bases de datos BacFITBase y DualSeqDB, que recopilan información sobre la importancia de genes bacterianos y sobre los cambios de expresión génica que se producen en el patógeno y en el huésped durante el proceso infeccioso, respectivamente. Del mismo modo, hemos desarrollado HPIPred, un sistema de predicción de interacciones proteína-proteína entre huésped y patógeno, basado en la codificación numérica de propiedades físico-química de los aminoácidos, capaz de integrar información fenotípica para guiar el proceso de predicción. La combinación de estas herramientas podría servir como guía en el desarrollo de nuevos fármacos contra la resistencia bacteriana.

Table of contents

List of Figures

List of Tables

1. INTRODUCTION

1.1. Bacterial infections and antibiotic treatment failure

1.1.1. Infection and bacterial pathogens

Pathogens are microorganisms, such as bacteria and viruses, that can cause disease. Similarly, pathogenesis is the process by which a disease develops, normally involving exposure of the host to the pathogen, which adheres and colonizes the host, grows within it and ultimately triggers the infection process (1).

Among all bacterial species, only a relatively low number of them are considered pathogenic and causative agents of disease. In this regard, virulence is associated to pathogenicity. The degree of virulence is related directly to the ability of the bacteria to colonize the host and evade the immune response (1,2). Therefore, pathogenic bacteria can be classified as primary or opportunistic pathogens, depending on the susceptibility of the host to infection and the virulence of the bacteria. Primary pathogens can naturally cause disease in any individual, regardless of the host's immunologic and physiologic conditions, whereas opportunistic pathogens are those that can become pathogenic due to an impairment of the host's immune system but are not pathogenic under normal circumstances.

1.1.2. Mechanisms of bacterial infection

Pathogenic bacteria produce a pletora of molecules, known as virulence factors, that subvert host cellular processes to promote infection (3). Bacterial secretion systems are responsible for the delivery of these virulence factors to the host cells, and can be classified as effectors or toxins (4,5). In this sense, toxins are delivered to the cell to damage and irreversible disrupt cellular homeostasis,

whereas effectors are generally translocated to the host cytoplasm through more specialized secretion systems, and their activity is often more subtle and geared towards the modulation of host cellular functions to the pathogen's own benefit. The combined action of toxins and effectors allow the pathogen to control numerous mechanisms related to virulence and host immune system, such as adhesion, encapsulation, ubiquitination of host proteins, inhibition of host cell apoptosis, disruption of host cytoskeleton, membrane trafficking and cell signaling, among others (**Figure 1**) (6–11).

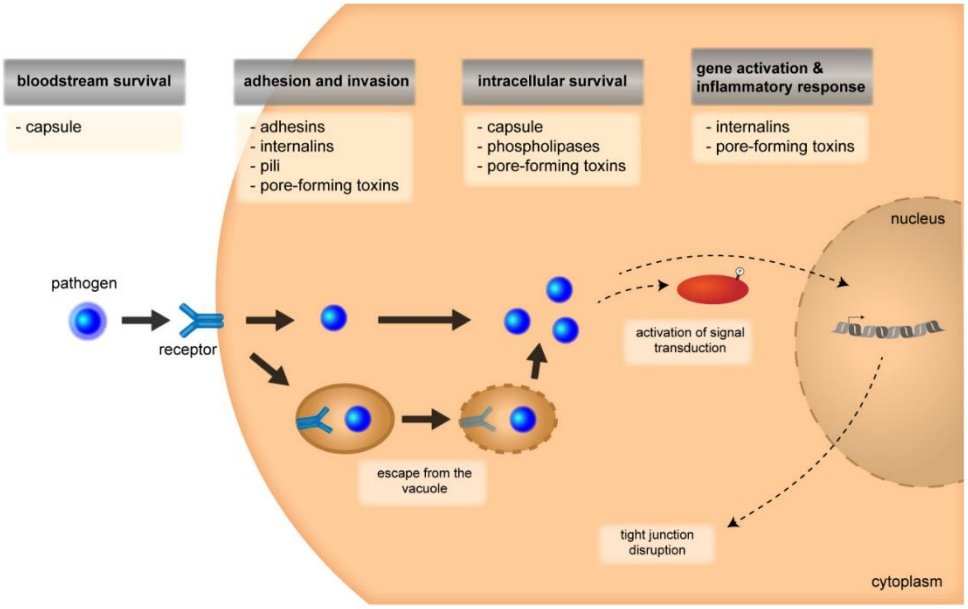


Figure 1. Virulence factors involved in the pathogenesis during bacterial meningitis. Taken from: "Virulence Factors of Meningitis-Causing Bacteria: Enabling Brain Entry across the Blood-Brain Barrier ". Herold R et al. 2019. Int J Mol Sci. 20(21):E5393. Licensed under CC by 4.0: https://creativecommons.org/licenses/by/4.0/

1.1.3. Antibiotic treatments and derived problematics

At the beginning of the 20th Century, life expectancy at birth was 47 years in the developed countries, and infectious diseases, such as cholera, smallpox, tuberculosis, pneumonia, etc., accounted for high morbidity and mortality worldwide (12). Since the discovery of penicillin in 1928 by Alexander Fleming, antibiotics have revolutionized the treatment of infectious diseases, leading to important medical advances, and saving countless lives: while the death rate from these diseases dropped from 0.8% to 0.0036%, life expectancy rose to 78 years (13). Unfortunately, as Flemming predicted nearly a century ago, microorganisms can adapt to the use of antibiotics and develop resistance to them.

1.1.4. Antibiotics: mechanisms of action

Antibiotic treatments are still the most widely used therapeutic agents to fight bacterial pathogens, not only in acute and chronic infections, but also with prophylactic purposes. Antibiotics function in different ways. Some antibiotics have a bactericidal function, killing the bacteria by inhibiting the synthesis of their cell wall, interfering with essential bacterial enzymes. Others are bacteriostatic and hinder bacterial growth by inhibiting protein synthesis and DNA replication or by interfering with other mechanisms of bacterial cellular metabolism (**Figure 2**) (14,15). Based on their target specificity, antibiotics can be categorized as broad-spectrum antibiotics, capable of targeting a wide range of Gram-positive and Gram-negative bacteria, and narrow-spectrum antibiotics, which are able to inhibit or kill specific bacterial species. The choice of antibiotic depends, mainly, on the type of bacteria responsible for the infection. Hence, broad-spectrum antibiotics are the preferred choice when the bacterial agent is unknown or if the infection is suspected to be caused by multiple bacterial species. However, this strategy is not absent of risks, as an incorrect identification of the pathogens can cause toxicity in the host organism, damage the host microbiota or favor the appearance of bacterial resistance (16).

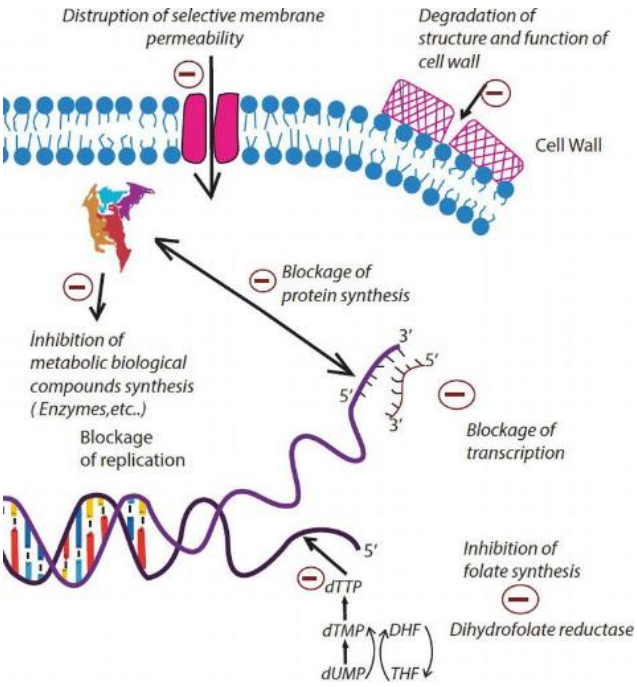


Figure 2. Targets of antimicrobials. Taken from "Introductory Chapter: The Action Mechanisms of Antibiotics and Antibiotic Resistance, Antimicrobials, Antibiotic Resistance, Antibiofilm Strategies and Activity Methods," Sahra Kırmusaoğlu, Nesrin Gareayaghi and Bekir S. Kocazeybek. 2019. IntechOpen, DOI: 10.5772/intechopen.85211. Available from: https://www.intechopen.com/chapters/65914. Licensed under CC by 3.0: https://creativecommons.org/licenses/by/3.0/

1.1.5. Multi-drug resistance in bacteria

Even though bacterial resistance is a naturally occurring phenomena in which genetic mutations allow bacteria to acquire resistance mechanisms, the use and misuse of broad-spectrum antibiotics for the treatment of all kinds of bacterial infections in healthcare has increased the selective pressure against these pathogens. In turn, this pressure has favored the appearance of multi-drug resistance (MDR), a term that refers to the ability to resist simultaneously to different antibiotics. Multi-drug resistant bacteria accomplish this by means of different resistance mechanisms, such as producing enzymes that inactivate or destroy the antibiotic molecule, reducing their cell wall permeability, increasing active efflux to pump the antibiotic out or interfering with the target site of the antibiotic by decreasing its binding affinity (**Figure 3**) (17,18).

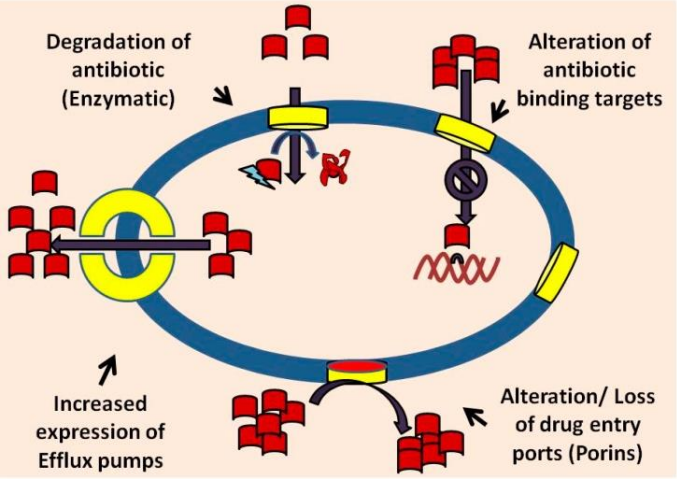


Figure 3. Bacterial antibiotic resistance mechanisms. Taken from "Multidrug efflux pumps from Enterobacteriaceae, Vibrio cholerae and Staphylococcus aureus bacterial food pathogens". Jody L Andersen et at. 2015. Int J Environ Res Public Health 12(2): 1487-547. Licensed under CC by 4.0: https://creativecommons.org/licenses/by/4.0/

Due to the limited amount of effective antibiotics, the World Health Organization (WHO) has catalogued the fight against multi-drug resistant bacteria as one of the most urgent challenges of public health worldwide. In order to promote the research of MDR and the need for the development

of new antibiotics, the WHO has published a list of antibiotic-resistant "priority pathogens", among which *Pseudomonas*, *Acinetobacter* and *Enterobacteriaceae* stand as the most critical group to be dealt with, especially due to the threat that they pose in hospitals, nursing homes and among patients in intensive care units (19). It is estimated that MDR is responsible for 33,000 deaths and costs nearly €1 billion to the European health care systems yearly (20). Similarly, it has been reported that more than 2.8 million antibiotic-resistant infections and more than 35,000 associated deaths occur in the U.S. each year, with associated costs that go as high as $4.6 billion (21). The WHO has warned that common infections could become lethal in the near future. Estimates are that therapeutic coverage will be insufficient within 10 years, where antibiotics will no longer be effective against certain infectious diseases (22).

1.1.6. Limitations of classical drug screening

The lack of scientific research and investment to deal with the proliferation of antibiotic resistant bacteria has fueled the innovation gap in antibiotic discovery, as the majority of antibiotics brought to the market in the past years have been derivatives of already existing drugs (13,23,24). Another major concern is that we may be reaching a point where the number of classical antibiotics is almost complete. This should not come as a surprise: the analysis of 20,000 experimental protein-ligand complex structures available in the Protein Data Bank (PDB) revealed that the number of representative pockets is approaching a plateau, which suggests that the structural space of ligand-binding pockets is nearly completely described (25–27).

During the 1990s, the rise of genome sequencing allowed the identification of 'essential' bacterial targets which are vital for bacterial infection, growth and proliferation, as possible new targets for drug screening. This strategy, however, has had limited to no success in delivering effective novel antibiotics (28). Pathogenic bacteria need the host environment to grow and develop, thus the concept

of ‘essentiality’ for pathogenic bacteria needs to be extended to take into account the biological context of infection, i.e., including those genes that are essential for the pathogen during the host infection. The emergence of bacterial resistance, together with the failure to discover new antibiotics by classical drug screening methods, highlights the need to rethink the antibiotic discovery paradigm, so that screening approaches can identify new targets based on the interactions between the pathogen and the host rather than the pathogen alone (29).

1.2. Importance of protein-protein interactions in host-pathogen infections

The majority of biological processes in the cell are carried out by proteins, which rarely act alone, as their functions tend to be highly regulated. In this sense, protein-protein interactions not only mediate most cellular functions occurring in an organism, but also the complex interplay between a pathogen and its natural host during infection (30).

1.2.1. Protein-protein interactions

Protein-protein interactions (PPIs) are mediated by physical contacts of high specificity between two or more proteins. PPIs are involved in a multitude of functions such as electron transfer, signal transduction, membrane transport or cell metabolism, among others (31,32). Furthermore, PPIs can be transient, when the interaction is reversible and produces short time effects like signal transduction, or permanent, when the interaction is stable and the complexes carry out functional roles. PPIs have been studied from diverse perspectives such as biochemistry, molecular dynamics, quantum chemistry, etc., which has allowed the reconstruction of large ensemble of PPIs occurring within an organism, usually referred to as protein interactome.

1.2.2. Methods for protein-protein interaction identification

Diverse methods have been developed to identify PPIs and reconstruct protein interactomes to gain insight into the processes that occur within the cell and its environment. High-throughput experimental techniques are the most widely used for large-scale PPI detection (**Table 1**), including in vitro assays like affinity chromatography (33), Tandem affinity purification – mass spectrometry (TAP-MS) (34), coimmunoprecipitation (35), X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR spectroscopy) (36), among others, and in vivo assays such as Yeast 2 hybrid (Y2H) screenings (**Figure 4**) (37–39). Nonetheless, experimental techniques have their downsides in terms of required time, cost and reliability, as these methods often fail at identifying weak interactions, their applicability depends on how well assay protocols are optimized for different organisms and they suffer from low-specificity, producing high rates of false positives and false negatives.

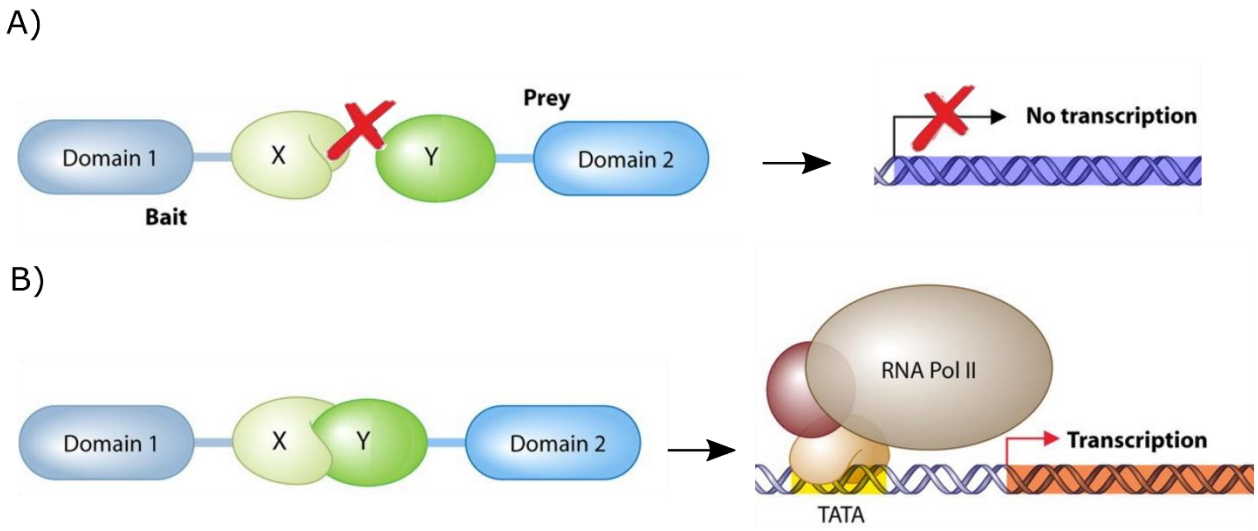


Figure 4. Identification of protein-protein interactions by Yeast two-hybrid technique. Two proteins of interest (X and Y) are each fused to a fixed protein domain, the binding domain and the activation domain, and forming the bait and the prey, respectively. A) in the absence of an interaction between the proteins, the domains remain distance, preventing the detection of any output. B) If the two proteins interact, the bait can recruit the prey to a specific location in the cell, stimulating a detectable output by recruiting RNA Polimerase II and triggering the transcription of a reporter gene. Adapted from "Diversity in Genetic In Vivo Methods for Protein-Protein Interaction Studies: from the Yeast Two-Hybrid System to the Mammalian Split-Luciferase System". Stynen B., et al. 2012. Microbiol Mol Biol Rev. 1;76(2):331–82. Permission to reuse the image is granted under the License agreement available at: https://marketplace.copyright.com/rs-ui-web/mp/license/79dc90d3-93e2-4b95-93ef-875a1c6636bb/b647baca-0694-4647-8c67-e35698cd86e6

Table 1. Experimental methods used for the detection of protein-protein interactions

Experimental method	Basic principle	Advantages	Limitations
Yeast two-hybrid (Y2H)	The binding domain (BD) and the activation domain (AD) of a transcription factor are fused, respectively, to the couple of proteins (A and B) from which the interaction will be studied. Bait (X-BD) and prey (Y-AD) chimeric proteins are co-expressed in yeast cells where, if both proteins interact, the transcription factor is reconstituted, activating the expression of a reporter gene.	- Low cost for protein purification and antibody development. - In vivo technique in eukaryotic cells. - No need for large amounts of highly purified proteins or antibodies.	- Activation of reporter gene may be due to other mechanisms, leading to false positives. - Bait proteins may become toxic. - Risk of incorrect folding or absence of complex protein modifications in yeast.
Affinity chromatography	The protein of interest is immobilized to an insoluble matrix which is then incubated with a solution that contains putative binding partners. After washing away unbound material, the binding partners are eluted and detected by chromatography.	- Can detect very weak interactions in proteins. - Highly responsive. - All the sample proteins are evenly tested for interaction.	- High specificity among proteins causes false positives. - High amounts of sample are lost during elution steps.
Tandem affinity purification-mass spectrometry (TAP-MS)	The protein of interest is fused to a tandem-affinity purification tag (TAP-tag), which is then expressed in cells and used as bait to purify protein complexes through a two-step purification process. These highly purified protein complexes can be then analyzed by mass spectrometry.	- Yields high purity proteins. - Allows for the identification of even very low abundant protein complexes.	- Low exposure of the tag to the affinity beads may alter the results. - The tag may affect levels of protein expression. - Specialized equipment required.
Co-immunoprecipitation	The protein of interest (antigen) is targeted and captured by an antibody, allowing to indirectly pull unknown proteins that form a complex with the protein of interest.	- Isolation of protein complexes in their natural state. - The antigen and the protein complex have similar concentrations in the cell.	- Need for polyclonal antibodies. - Does not guarantee that all the proteins in the complex interact. - Lower sensitivity than other methods like affinity chromatography.
X-ray crystallography	High resolution microscopy technique that can identify the exact three-dimensional position of the molecular and atomic structures of protein-protein complexes.	- High-resolution approach. - Can resolve enzyme conformational changes.	- Specialized equipment required. - Proteins need to be crystalized. - Weak PPIs may be lost
NMR spectroscopy	Ensemble of techniques based on the determination of the protein complex at the atomic level by making use of the nuclear magnetic properties of the atoms.	- High-resolution approach. - Can detect weak PPIs. - Is compatible with crystallized proteins. - Can be combined with functional assays.	- Not well suited for the study of large protein complexes. - Relatively insensitive to the amount of sample needed for good-quality data.

1.2.3. Protein-protein interaction databases

Even though the coverage of the PPI space is still in its infancy, the reduced cost of high-throughput assays has allowed the detection of a considerable amount of PPIs so far. With the aim of gathering and unifying the available data, numerous PPI databases have been developed over the past few years. Text mining of scientific literature is the main resource to systematically categorize individual PPIs, although huge curation efforts are made to manually compile data on PPIs. These databases can be categorized into three main groups: pathway databases such as KEGG (40–42) and Reactome (43), databases that collect experimental data like BioGRID (44) and IntAct (45), and databases like STRING (46) or GeneMANIA (47), where no manual curation is involved but computational predictions of PPIs are included. Recently, large-scale experiments have been performed to determine molecular interactions between human and bacteria, which has increased the availability of PPI data for host-pathogen systems. This has permitted the creation of curated databases of bacterial PPIs such as PATRIC (48), curated databases of host-pathogen PPIs such as HPIDB (49) and PHISTO (50), or related resources like PHI-base (51), which catalogs experimentally validated virulence and pathogenicity not only from bacteria but from other types of pathogens that infect eukaryotes. Even though these advances in data compilation are promising, most relevant data related to host-pathogen interactions is still buried in the literature and hinders progress in infection disease research.

1.2.4. PPIs as drug targets for bacterial infections

During the last decades, most of the drug discovery approaches to fight antibacterial resistance have focused mainly on targeting unique proteins from the pathogen (28). Given the rapid proliferation of MDR bacteria and the failure of classical antibacterial treatments, the study of host-pathogen PPI networks can offer promising insights for the discovery of next-generation antibacterial drug targets (30). The analysis of these PPI networks allows the identification of essential and highly connected proteins within the context of infection. In this sense, genome-wide gene deletion studies indicate that

deleting highly connected proteins (also referred to as hubs) from its interactome is more likely to be lethal for the organism than just removing lowly connected proteins, a phenomenon known as centrality-lethality rule (52). This correlation between network connectivity and essentiality suggests that targeting highly connected proteins in the pathogen PPI network and their interacting partners in the host PPI network may be a promising strategy to find new targets for drug discovery.

1.2.5. Druggability of host-pathogen PPIs

There is a considerable number of small molecule PPI modulators in clinical trials targeting cancer, autoimmune disorders or viral infections as immune suppression agents. However, targeting PPIs as an antibiotic drug discovery strategy remains a relatively unexplored territory by comparison (53). Many pathological processes induced by bacteria are dependent on PPIs, which gives host-pathogen interactions a great potential to shed light on pathogen biology and virulence pathways (**Figure 5**).

In general, using small molecules as PPI modulators is more challenging than the classical screening approaches that target protein-ligand binding pockets (54–57). In contrast with the estimated diversity of 1000 structural pockets (26), the PPI space is thought to be much larger, i.e., a new set of ~10,000 types of PPIs would be available for screening purposes. However, PPIs have large surface contact areas (1,500-3,000 $Å^2$) and small molecules only cover around a fraction of those contact areas (300-1,000 $Å^2$), so they are not well suited to screening such spaces. In addition, PPIs are generally flat and often lack grooves and pockets, which are the preferred targets for small molecules (58). Due to the aforementioned, researchers are starting to consider alternative PPI modulators such as peptides and recombinant proteins, which would allow the exploration of larger interaction surfaces (59).

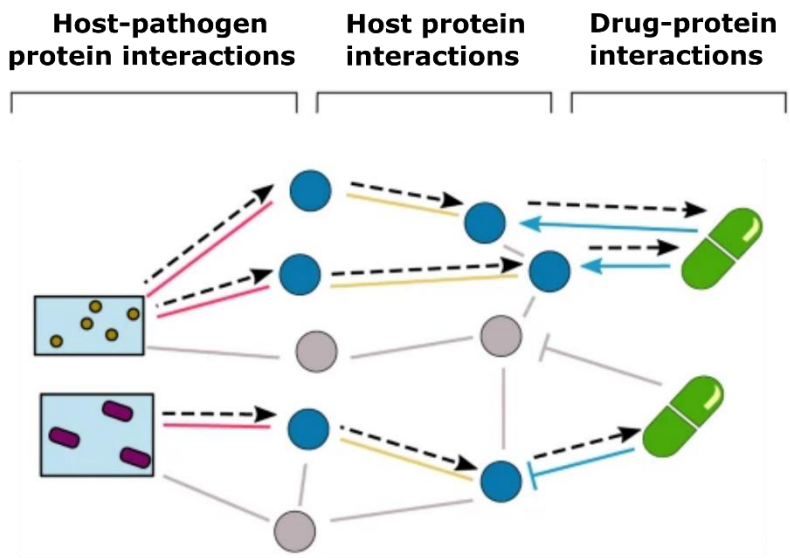


Figure 5. Host-pathogen protein interactome exploration for drug development. Adapted from "Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing". Sepideh Sadegh et al. 2020. Nature Communications volume 11, Article number: 3518. Licensed under CC by 4.0: https://creativecommons.org/licenses/by/4.0/

1.3. Computational prediction of host-pathogen PPIs and limitations of these techniques

The cost in time and resources of experimental assays makes it unfeasible to test and validate all possible host-pathogen PPIs by these procedures alone (60). For instance, the human proteome consists of more than 20,000 validated proteins, which paired with the thousands of proteins from the pathogen, accounts for millions of PPIs to test experimentally. In this sense, there exists several computational strategies that predict host-pathogen PPIs, and their combination with experimental techniques may be a fresh breeding ground for the identification of promising novel PPI drug targets. The most widely used in silico methods for the prediction of these PPIs rely on sequence homology, domain and motif interactions, structure or machine learning techniques (**Figure 6**).

1.3.1. Sequence homology-based methods

Homologous proteins usually share similar functions and three-dimensional structures, so the identification of homologous proteins to a newly determined protein is a method used to infer biological functions for a new protein. This method has been adapted to the identification of PPIs

under the assumption that homologous proteins may share similar interaction partners and functions. Therefore, the homology-based method follows the idea that the interaction between two proteins within an organism may be conserved in a related one and these conserved interactions are called interologs (61). The process that allows the identification of inter-species PPIs involves the obtention of the template PPI pair, followed by the identification of homolog proteins for the host and the pathogen with respect to the template. This method is one of the most extensively used methods for the prediction of host-pathogen PPIs (62–64). However, it is not able to make inferences about interactions between specie-specific families of genes.

1.3.2. Domain and motif interaction-based methods

Domains are the building blocks that determine the structure and function of proteins, and perform specific functions allowing proteins to interact with different types of molecules (65). Most of the PPIs are facilitated by domain-motif interaction by binding domains in a protein to short linear motifs in interacting partners. These interactions are normally involved in key cellular processes, requiring a very tight regulation (66). Not all pathogen systems are appropriate for applying the mentioned domain-based approaches, since domains and the related information are not available for all pathogens.

1.3.3. Structure-based methods

In this method, it is considered that two proteins with similar structures to a known PPI are likely interacting in a structurally similar way. The method generally starts with a set of pathogen and host proteins, followed by sequence matching protocols which are used to determine the similarities between the host or pathogen proteins with known protein partners. Sequence similarity score may be used as a statistical potential assessment when structure information is not available. The last step consists of filtering the set of potential interactions, and it is generally performed using the biological

contexts of protein. The main drawback of this method is that finding high similarity between pathogen proteins and proteins with known structure is far from guaranteed for all pathogen proteins. A number of studies are based on these structural similarities and use known PPIs to identify similar interacting pairs within pathogen and host protein (67,68).

1.3.4. Machine learning techniques

The application of machine learning approaches for the prediction host-pathogen PPIs is a well-accepted idea. These techniques use available PPIs as features for the training and classification of interacting and non-interacting PPIs. Both supervised and semi-supervised methods have been extensively used. The most extensively technique is the Random Forest (RF) algorithm, a classifier algorithm made up of decision trees. Individual trees in the training phase are built by random feature vector sampled from a dataset independently. A subset of the variables is selected at random and individual classification trees are raised for every node in a tree. In order to group a new object, the input vector is set up for each of the trees in the forest. Based on the largest vote, a class is allocated to the object. RF can also classify features based on importance and can also be used to recover missing data. Random forest and Decision trees are widely used in computational biology for the classification of biological data (69), especially for the prediction of PPIs (70).

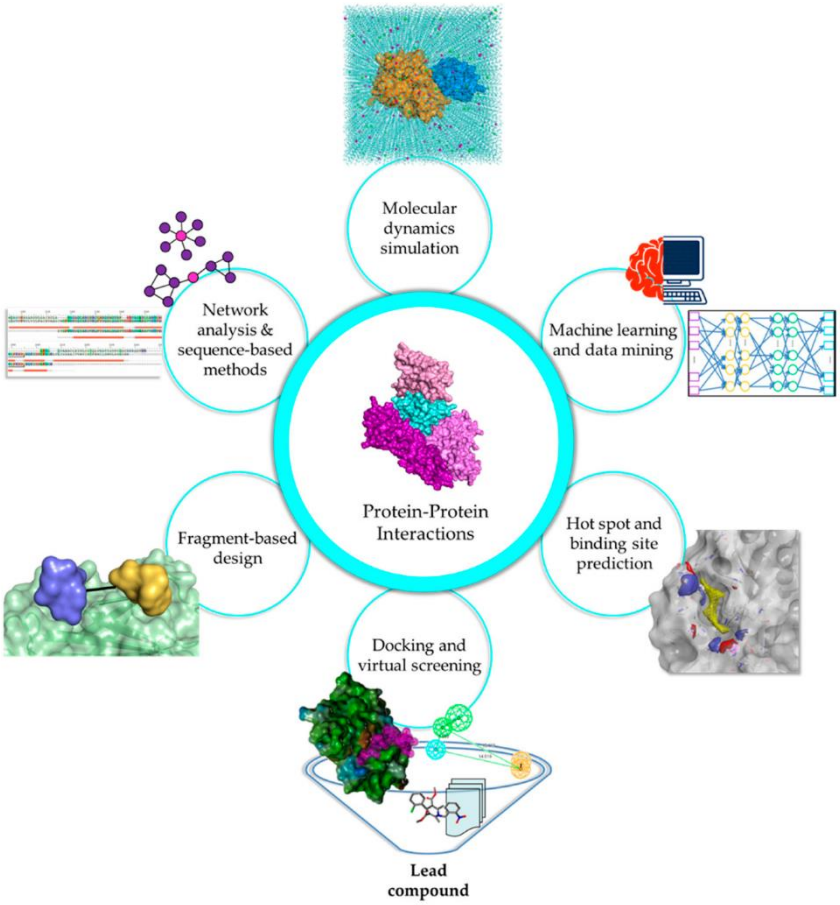


Figure 6. Computational methods for the prediction of host-pathogen PPIs. Taken from "Evolution of In Silico Strategies for Protein-Protein Interaction Drug discovery". Stephani Joy Y. Macalino, et al. 2018. Molecules. 23(8): 1963. Licensed under CC by 4.0: https://creativecommons.org/licenses/by/4.0/

1.3.5. Limitations and future prospects of host-pathogen PPI predictions

As seen, the discussed approaches perform relatively well at predicting PPIs and each of them have their strengths and weaknesses. However, all these methods suffer from the drawback of over-predicting the amount of interaction partners (71). Due to the importance of correctly identifying host-pathogen PPIs for the development of PPI modulators, the challenge now would be to reduce the number of wrongly predicted interactions in a biologically meaningful way to get a more realistic picture of the protein interaction networks during the infection process.

In this sense, the integration of the already existing prediction methods with phenotypic information relevant in the infection process, such as gene expression, gene essentiality, or gene ontology may help improve the performance of current prediction analyses. The advancements in high-throughput sequencing technology, coupled with downstream computational techniques, have allowed the generation of a vast amount of data relevant in the infection process. For instance, the analysis of Dual RNA-Seq experiments makes it possible to quantify gene expression levels of intracellular pathogens and their hosts simultaneously, which helps identify genes that change their expression during the progression of the infection. Similarly, transposon mutagenesis assays have driven the characterization of bacterial genes that play a relevant role in infection. Nevertheless, a lot of these data are disseminated through literature or published experiments with massive amount of heterogeneous data that are not easy to access or interpret, so there is an urgent need to collate and integrate these data in order to facilitate the study of the infection processes.

2. OBJECTIVES

It is necessary to develop tools that help to cope with the shortage of existing validated protein-protein interaction partners in the context of infection in the hope that they will help accelerate the screening process of drug discovery against resistant bacteria.

In this thesis, we address some of the limitations presented regarding the functional annotation of bacterial and host proteins in infection, as well as the problematic of host-pathogen protein-protein interaction predictions and suggest methods to tackle these limitations. Specifically, we propose the development of platforms and databases that compile, harmonize and allow easy access to data that depict bacterial proteins relevant for the infection (chapter 1) as well as gene transcriptional changes in both pathogenic bacteria and their natural host upon infection (chapter 2). Additionally, we propose the development of a computational tool for the prediction of host-pathogen PPIs that integrates phenotypic data to help in the decision making of host-pathogen PPI candidates for follow-up experiments from a systemic point of view (chapter 3).

3. CHAPTER 1. BacFITBase: a database to assess the relevance of bacterial genes during host infection

3.1. Abstract

Bacterial infections have been on the rise world-wide in recent years and have a considerable impact on human well-being in terms of attributable deaths and disability-adjusted life years. Yet many mechanisms underlying bacterial pathogenesis are still poorly understood. Here we introduce the BacFITBase database for the systematic characterization of bacterial proteins relevant for host infection aimed to enable the identification of new antibiotic targets. BacFITBase is manually curated and contains more than 90,000 entries with information on the contribution of individual genes to bacterial fitness under *in vivo* infection conditions in a range of host species. The data were collected from 15 different studies in which transposon mutagenesis was performed, including top-priority pathogens such as *Acinetobacter baumannii* and *Campylobacter jejuni*, for both of which increasing antibiotic resistance has been reported. Overall, BacFITBase includes information on 15 pathogenic bacteria and 5 host vertebrates across 10 different tissues.

3.2. Introduction

The development of new antimicrobial therapies relies heavily on our understanding of the mechanisms of bacterial infection. Bacterial proteins are responsible for rewiring a myriad of biochemical processes essential for the efficient propagation of the pathogen (72,73). Recently, our group showed that bacterial fitness *in vivo* does not correlate with data from *in vitro* studies (52). This is a major drawback for antimicrobial target discovery as many *in vitro* false negatives are disregarded for further testing. Therefore, it is crucial to understand how bacterial infection develops *in vivo* and which bacterial genes are required to infect a host.

BacFITBase is a manually curated database of bacterial genes that collates *in vivo* information on their relevance during host infection, as measured by transposon mutagenesis. Transposon mutagenesis experiments allow the measurement of fitness values for individual genes, allowing us to assess which genes are fundamental to infect a specific host organism (74). To address the contribution of a bacterial gene to infection, its fitness is measured through genome-wide transposon mutagenesis coupled with next-generation sequencing (Tn-seq) (75) . Briefly, mutations targeting virtually all genes in the bacterial genome are generated by random insertion of transposons. Afterwards, these mutants with randomly inactivated genes are grown in culture medium (input pool), inoculated in a host organism, and finally recovered after infection (output pool). Genomic DNA from the input and output pool is extracted, and transposon insertion site junctions are amplified and quantified by next-generation sequencing.

BacFITBase provides a common framework and easy access to fitness data for individual bacterial genes during infection. At present, it covers the infection processes of 15 pathogenic bacteria in 5 model vertebrates across 10 different tissues (**Table 2**). A tutorial section provides a detailed step-

by-step description of how to search and browse our data. This resource is available at www.tartaglialab.com/bacfitbase to help the research community in the systematic characterization of bacterial proteins involved in host infection.

3.3. Methods

3.3.1. Fitness scores and z-scores

Data were collected from publicly available transposon mutagenesis experiments containing either raw input/output read counts or fitness scores for all mutant genes available (**Table 2**) (76–90). The fitness score of a gene is calculated as the ratio of the normalized frequencies of input/output read counts. Reads for each transposon insertion site are normalized to the total number of reads obtained from a sample according to the following equation:

$$F^a = \frac{C_i^a / N_i}{C_0^a / N_0}$$

where F^a is the fitness score for gene a, C_0^a and C_i^a are the number of reads for a gene a before and after infection and N_0 and N_i are the number of total reads before and after infection, respectively. In order to normalize the fitness scores to allow comparison between different studies, the corresponding z-scores were calculated for each individual experiment (**Figure 7**). To assess the significance of a bacterial gene's fitness impact, p-values were calculated using a two-tailed one-sample Student's t-test on the distribution of fitness scores within each study provided that raw data was available. Otherwise, the p-values reported in the original study are shown.

Table 2. List of all studies included in the BacFITBase.

Pathogen	Host organism	Tissue	Reference
Salmonella enterica Serovar Typhimurium ST4/74	*Bos taurus (cow)*	Ileal mucosa	(76)
	Sus scrofa (pig)	Colonic mucosa	
	Gallus gallus (chicken)	Cecum	
Haemophilus influenzae Rd KW20	*Mus musculus (mouse)*	Lung	(77)
Streptococcus pyogenes M1 5448	*Mus musculus (mouse)*	Skin	(86)
Porphyromonas gingivalis ATCC 33277	*Mus musculus (mouse)*	Skin	(87)
Escherichia coli CFT073	*Mus musculus (mouse)*	Spleen	(88)
Mycobacterium avium subsp. Paratuberculosis K10	*Mus musculus (mouse)*	Spleen	(89)
Escherichia coli M12	*Mus musculus (mouse)*	Spleen and mammary gland	(90)
Escherichia coli O157:H7	*Bos taurus (cow)*	Feces	(78)
Vibrio cholerae O1 biovar El Tor str. N16961	*Oryctolagus cuniculus (rabbit)*	Small intestine	(79)
Campylobacter jejuni subsp. jejuni 81-176	*Mus musculus (mouse)*	Cecum	(80)
Klebsiella pneumoniae subsp. pneumoniae pneumoniae ATCC 43816 KPPR1	*Mus musculus (mouse)*	Lung	(81)
Acinetobacter baumannii ATCC 17978	*Mus musculus (mouse)*	Lung	(82)
Salmonella enterica Serovar Typhimurium SL1344	*Mus musculus (mouse)*	Liver and spleen	(83)
Serratia marcescens Strain UMH9	*Mus musculus (mouse)*	Spleen	(84)
Vibrio parahaemolyticus RIMD 2210633	*Oryctolagus cuniculus (rabbit)*	Small intestine	(85)

3.3.2. Technical aspects

BacFITBase was built using PHP on an Apache web server with a MySQL database backend. BacFITBase stores no user data, except for the anonymous caching of BLAST search results for a given sequence in order to greatly speed up repeated searches. The open-source Bootstrap library was used to allow display on devices of any screen size, including mobile devices. Several icons were included from Font Awesome and the Noun Project, and a number of JavaScript libraries are used for table export and sorting.

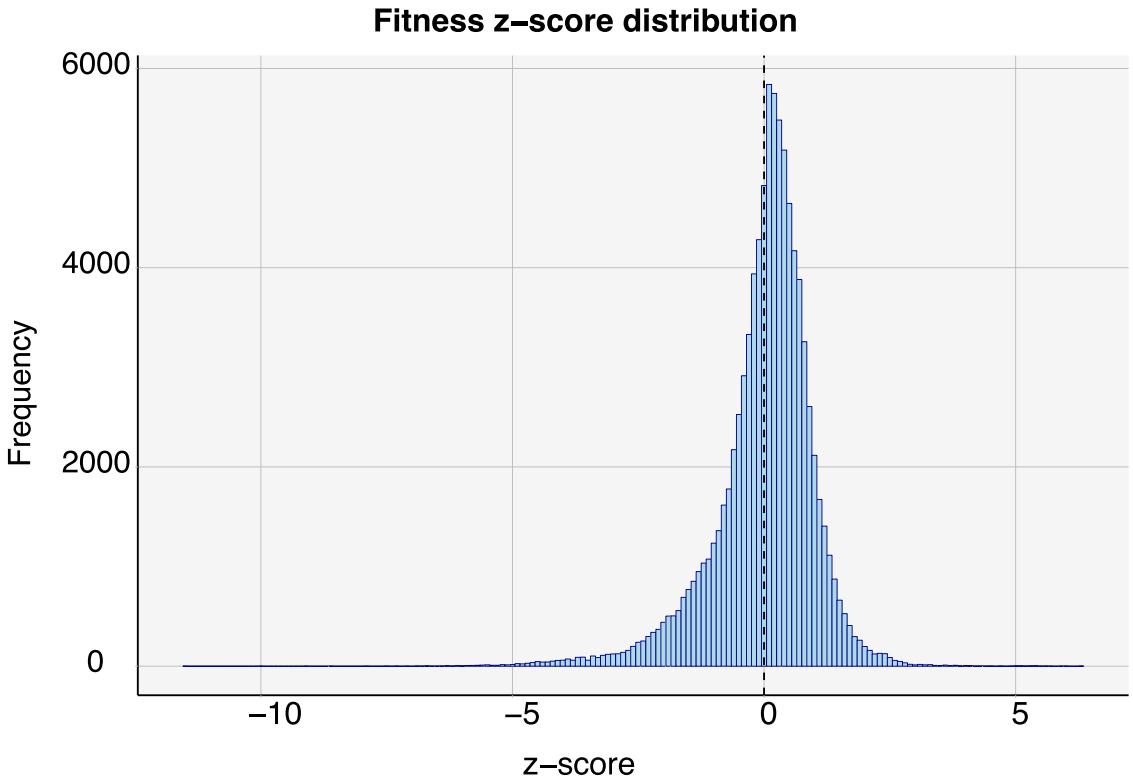


Figure 7. Histogram representing the z-score distribution among all studies included in the BacFITBase database. A fitness z-score < 0 indicates that a given mutation is more detrimental than the average mutation during infection, while a raw fitness score < 1 indicates that a given mutant is underrepresented in the output pool. As we can observe in the histogram, the extreme values in the fitness z-score distribution are shifted towards negative values. Underrepresented instances give strong evidence that this particular gene is relevant in the infection process, as transposon insertion is typically associated with a loss of function phenotype.

3.3.3. BLAST search

The NCBI BLAST suite version 2.9.0+ (March 2019) (91) is used to search by sequence similarity. The BLASTP program is used for amino acid sequences, and BLASTX for nucleic acid (coding) sequences. BLAST search results are cached for each unique sequence, which means that re-running a search using the same sequence will yield results nearly instantaneously. As on all other pages, results from the BLAST search page can be linked to and shared with other researchers using the "*Link to these results*" link at the bottom of the page. For sequences above a URL length of 2,000 characters this link uses a sequence hash identifying the cached sequence, rather than the sequence itself.

3.3.4. Protein visualisation (ProViz)

For UniProt proteins, a protein visualization is automatically generated and displayed by ProViz (92). ProViz is an interactive exploration tool for investigating the structural, functional and evolutionary features of proteins and is likely to be particularly helpful for analyzing uncharacterized proteins.

3.4. Using BacFITBase

BacFITBase consists of a text search function to find specific pathogenic bacterial genes, a BLAST search function to find bacterial genes similar to a protein or nucleic acid sequence of interest, a Browse function to quickly identify genes of high fitness impact during infection, and a Tutorial section to get started quickly by following a step-by-step guide.

3.4.1. Searching for a gene or protein

To search for a gene or protein, users just need to type its name or identifier in the Search tab and press the "*Search*" button. Any of the following options are available: gene symbols, gene locus identifiers, NCBI protein identifiers, UniProt protein accessions, or a free-text search in the gene product's description. The Search function also supports smart partial matches, so free-text terms can be used (e.g., "ribonuclease").

3.4.2. Searching within specific hosts and pathogens

Both pathogen and/or host species can be specified in the drop-down menus on the Search page. If no search term is given, this will result in a complete list of genes for these species, similar to the Browse view (described below).

3.4.3. Search results

After searching, the search results page will display a list of any bacterial genes matching the search term and species selected (**Figure 8**). The column matched by the search term is highlighted in green (if a search term was provided). For each pathogen species and gene, the search results page already shows a preview of the lowest fitness z-score across all available hosts, tissues, and post-infection time points, and the corresponding p-value.

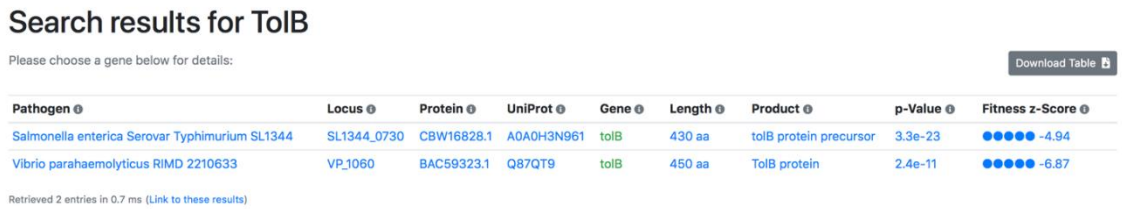


Figure 8. Search results in BacFITBase. This page displays a list of any bacterial genes matching the search term and the host and pathogen species that were selected. The search results page displays the gene locus and NCBI protein identifier, the UniProt accessions code and the gene symbol (Please see BacFITBase description section for further information). This preview also shows the gene product description and its length, together with the fitness z-score and the corresponding p-value. In this example, we show the case of TolB, a periplasmic component of the Tol-Pal complex that plays a central role in the maintenance of cell wall integrity. Although this complex is not essential in vitro, mutants defective in the TolB-Pal complex have a reduced infectivity. The results in BacFITBase show that, upon deletion of TolB, the infection fitness of the mutant is strongly decreased.

3.4.4. Tables on BacFITBase: Sorting, downloading, and linking to results

To sort any table on BacFITBase as desired, users can simply click on any of the column headers. All tables can be downloaded as a comma-separated CSV file for import into spreadsheet software such as Microsoft Excel or Apple Numbers using the "*Download Table*" button in the top right corner. An appropriate, readable file name is automatically generated. Any results pages can also be linked to and shared with other researchers by right-clicking and copying the "*Link to these results*" link at the bottom of a page or table.

3.4.5. Detailed view of infection fitness scores for a gene

After selecting a gene of interest, a view will open with all the infection fitness information available for this gene (**Figure 9**). The heading of this page provides information on the selected protein: protein and pathogen identifiers, sequence length, gene name and UniProt identifier.

In the table, all available experimental data are shown: Tissue name and ontology of the host organism, time after infection, transposon insertion site on the pathogen's main chromosome (if reported by the original study), fitness data during infection including the raw fitness score, normalized fitness z-score, and p-value, and the reference to the original paper where the data was published. A brief description on the meaning of the raw score, normalized z-score and p-value is also available as mouse-over explanation on the column headers.

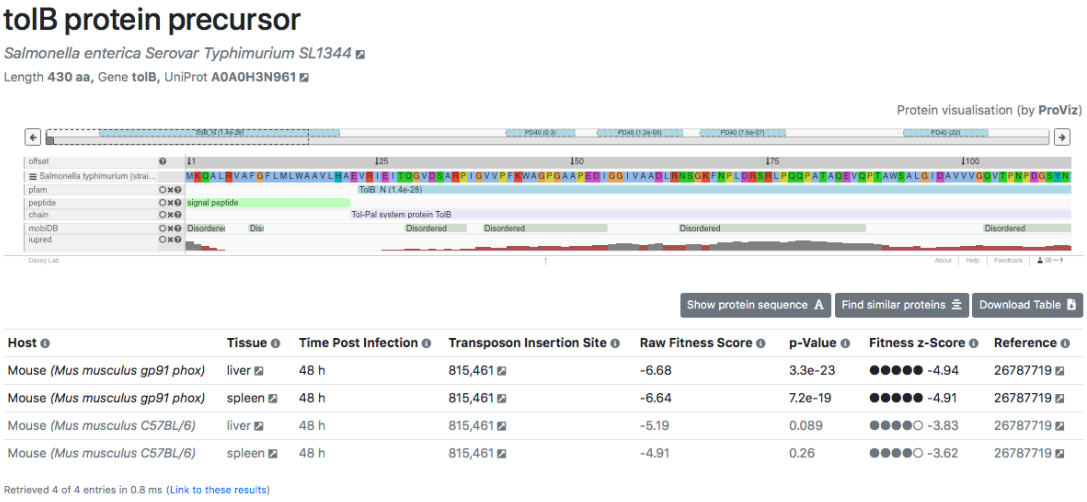


Figure 9. Detailed view of infection fitness scores for a gene. This page displays all the infection fitness information available for a bacterial gene, along with a ProViz visualization of the protein's sequence and structural features. The detailed view shows the fitness raw values and z-scores for each entry together with all the details of the experiment: hosts, tissues, transposon insertion sites, post-infection time points, and the corresponding p-value. The reference to the original paper where the data was published is also included in the last column of the table with the corresponding PubMed link.

For any protein in UniProt, a protein visualization is automatically provided by ProViz, an interactive exploration tool for investigating the structural, functional and evolutionary features of proteins,

including Pfam domains (93) and transmembrane regions. This is particularly useful for uncharacterized proteins, which account for a large fraction of bacterial proteins (94). Alternatively, the protein's FASTA sequence can be displayed by pressing the "*Show protein sequence*" button, along with a "*Copy*" link in the top right corner to copy and paste the protein's sequence into other research tools, or into the BacFITBase BLAST Search to search for similar proteins. Similar proteins can be searched via BLAST using the "*Find similar proteins*" button. This allows rapid assessment of the fitness impact of a group of similar proteins across all pathogens in BacFITBase.

3.4.6. BLAST Search

The BLAST Search tab provides a search by sequence similarity. When the protein of interest is not in our database, the user may search for similar proteins using BLAST sequence alignment. Finding a similar protein with low z-score (and low p-value) is a strong indication that the query sequence may be relevant for infection.

To search for similar proteins in our database using BLAST, the user can paste in the protein or coding sequence of interest in FASTA format and then press the Search button. Both protein and coding sequences can be used, but the proper format (protein or coding sequence) must be specified in the drop-down menu next to the Search button (as illustrated by the examples provided on the BLAST search page).

3.4.7. BLAST Results

When the BLAST alignment is ready (usually within a second or less), a search results page will open displaying alignment performance together with a complete description of the identified hits. This includes output columns from BLAST, such as the percentage of sequence identity between query and target in the successfully aligned region (“Identity”), the total number of amino acids that were

successfully aligned between query and target (“Aligned”), and the required size of a sequence database in which the current match could be found just by chance (the “bit score”). The E-value is the expected number of false positive matches given the size of the search database used. Matches with 100% sequence identity in the aligned region are highlighted in green. The meaning of the Pathogen, Locus, Protein, Gene, Product, p-value, and Fitness z-Score columns can be found in the Browse Tab section below, or via the mouse-over information symbols in the top row of any table.

3.4.8. Browsing the entire database (the Browse Tab)

The Browse tab provides an overview of all entries in the BacFITBase database. A pathogenic species of interest can be chosen in the selection element at the top. This table is sorted by significance and fitness z-score, which means that bacterial genes with a high and significant fitness impact during infection are listed first. Arrows next to each field link out to useful external databases, including the NCBI Taxonomy database, a comprehensive taxonomic database (“Pathogen”), the Ensembl Bacteria database (“Locus”), which provides genome annotation for many bacterial species, and the NCBI Protein database (“Protein”), which provides protein sequences and information. Additionally, the “UniProt Accession” and “Gene Symbol” columns link out to the UniProt Knowledgebase (95), which provides comprehensive protein annotation.

3.4.9. Downloading the entire database (the Download Tab)

The entire BacFITBase database is available for download on the Download page. Currently, BacFITBase v1 is available, and will be upgraded with new data as they become available.

3.5. Discussion

Infectious diseases are caused by microorganisms known as pathogens, which have the capacity to enter, colonize, and grow within a host, causing infection and damage. The use of antibiotics to treat bacterial infections has undoubtedly been one of the most important advances in healthcare, saving millions of lives since their discovery and widespread use (20,96).

Antibiotic development has mainly focused on the identification of 'essential' genes and proteins in bacteria whose inhibition is lethal under *in vitro* conditions (i.e., bacterial growth in culture). However, pathogenic bacteria do not grow alone but in a complex host environment. Thus, we need to revise the definition of what 'essential' means so that it includes the biological context of infection, i.e., which genes are essential for the pathogen during host infection.

For a given bacterium, the *in vivo* fitness cost of deleting a single gene is correlated with the number of interactions with host proteins (52). Therefore, proteins with high impact on pathogen fitness during infection may cause extensive rewiring of the host interactome. These observations indicate that infectious diseases are only properly understood in the context of the host-pathogen interactions. Towards this end, a promising approach is to systematically characterize proteins involved in host infection. The possibility of accessing fitness data from disparate sources quickly and easily should accelerate the identification of new proteins involved in life-threatening infectious diseases.

In this context, BacFITBase constitutes a valuable resource to systematically classify bacterial proteins relevant for host cell invasion and infection. BacFITBase will facilitate the task of identifying target proteins and interspecies complexes that will help us to understand the mechanisms of infection, and the design of new antimicrobial molecules aimed to interfere with the formation of

such complexes. In the next several years, BacFITBase is expected to grow continuously, becoming an even more comprehensive repository of bacterial proteins which could be important targets in the fight against infectious diseases.

3.6. Availability

BacFITBase is freely available at www.tartaglialab.com/bacfitbase

4. CHAPTER 2. DualSeqDB: The host-pathogen dual RNA sequencing database for infection processes

4.1. Abstract

Despite antibiotic resistance is a matter of growing concern worldwide, the bacterial mechanisms of pathogenesis remain underexplored, restraining our ability to develop new antimicrobials. The rise of high-throughput sequencing technology has made available a massive amount of transcriptomic data that could help elucidate the mechanisms underlying bacterial infection. Here we introduce the DualSeqDB database, a resource that helps the identification of gene transcriptional changes in both pathogenic bacteria and their natural hosts upon infection. DualSeqDB comprises nearly 300,000 entries from eight different studies, with information on bacterial and host differential gene expression under *in vivo* or *in vitro* conditions. Expression data were calculated entirely from raw data and analyzed through a standardized pipeline to ensure consistency between different experiments. It includes information on seven different strains of pathogenic bacteria and a variety of cell types and tissues in *Homo sapiens*, *Mus musculus* and *Macaca fascicularis* at different time-points. We envisage that DualSeqDB can help the research community in the systematic characterization of genes involved in host infection and help the development and tailoring of new molecules against infectious diseases. DualSeqDB is freely available at http://www.tartaglialab.com/dualseq.

4.2. Introduction

During infection, pathogens trigger the expression of unique genes that ensure their survival and allow replicating within the host. In turn, the host activates complex mechanisms to recognize and kill pathogens. Hence, the simultaneous detection of host and pathogen transcripts during the infection process can provide deeper insights into the host–pathogen interaction than those detected from the host or pathogen in isolation. The term "dual RNA-seq" refers to the process of simultaneously analysing RNA-seq data of a pathogenic bacteria and the infected host (97). Dual RNA-seq has become a leading approach to uncover the intricate relationship between pathogen and host interactions allowing researchers to identify 'molecular phenotypes' that would otherwise remain undetected (98–100).

In a typical dual RNA-seq experiment, either animals are inoculated with a defined load of bacteria (*in vivo*) or relevant cell culture models are inoculated with bacteria at a defined multiplicity of infection (*in vitro*). After inoculation, samples are taken over time to determine the time response. At each time point, infected cells are lysed, RNA is isolated and the cDNA library is prepared and sequenced using high-throughput sequencing technologies, which generates large amounts of data. RNA-seq data of mock-infected host cells and initial bacterial cultures are used as control conditions for expression analysis. Dual RNA-seq experiments have several technical difficulties, including the different nature and content of RNA between bacteria and eukaryotic cells, the larger proportion of RNA from eukaryotic cells and the need to account for the prevalence of rRNA transcripts and variable infection rates (97,101). Usually, such limitations can solved using high-depth sequencing, pathogen and host rRNA depletion, and enrichment of samples for infected host cells by fluorescence-activated cell sorting (102).

Dual RNA-seq is a mixture of host and pathogen transcripts where different RNA samples may contain variable proportions of pathogen to host reads (103,104). These transcripts can be sorted into the corresponding organisms by different computational strategies (97,98) and the accuracy of

differential expression values depends on the analytical method used. To circumvent these biases, we need a standard pipeline to compare data from different sources.

Despite the increasing availability of raw sequencing data from dual RNA-seq experiments, the existence of multiple analysis pipelines may hinder the comparison between datasets. Dual RNA-seq pipelines are very sensitive to software selection and parameter definition. This lack of standardization motivated the creation of DualSeqDB, a user-friendly platform to search for changes in gene expression levels during infection at both pathogen and host levels. To build this database, we analyzed raw sequencing data from heterogeneous dual RNA-seq studies using a well-defined pipeline, to generate comparable gene expression data. This setup allows DualSeqDB to compare across multiple species and experimental conditions.

4.3. Methods

4.3.1. Processing sequencing data

To build DualSeqDB, we reprocessed raw data from available studies (**Table 3**) and used a well-defined pipeline to provide robust and homogeneous information in our database (**Figure 10**). To this end, we selected only dual RNA-seq studies containing at least two biological replicates and only when data were available for infected and control conditions for both pathogen and host (98–100,105–109). For each study, genome and annotation files were downloaded from the NCBI Reference Sequence Database (RefSeq) (110). Bacterial and eukaryotic genome indices were created with Bowtie2 (111) and HISAT2 (112), respectively. HISAT2 can take into account alternative splicing of genes and was used for eukaryotic genome indexing. For each biological replicate, raw sequencing reads in FastQ format were trimmed with Trimmomatic (113) to remove adapter content. During this process, reads that are <36 bases long are dropped from the analysis. Afterward, surviving reads were mapped to host genome index with HISAT2. Mapped reads were stored as BAM files, and unmapped

reads were kept in a separate FastQ file. FeatureCounts (114), together with the host annotation file, was used for gene counting, and a matrix of read counts was generated where each row represents an annotated gene and each column represents a different condition or biological replicate. Unmapped reads from the previous mapping step were then mapped back to the bacterial genome index with Bowtie2, and a matrix of read counts was produced similarly by using the bacterial annotation file and FeatureCounts. HISAT2 and Bowtie2 are run with default parameters as a way to simplify and standardize criteria when analyzing data coming from heterogeneous sources. Finally, to calculate gene expression changes in treated against control conditions, differential expression analysis was performed separately for the bacterial and the host matrices by using the DESeq2 R Package (115). For this, the Wald test was used under the null hypothesis that there is no differential expression between the control and the treated samples. The estimated gene expression change value [measured in $\log_2$ fold change (FC)] and its associated *P*-value were generated for each annotated gene with detected reads in at least one condition (**Figures 10 and 11**). *P*-values were corrected for multiple testing using the Benjamini–Hochberg method. All additional information such as bacterial ID, host ID, time point, experimental condition (*in vivo*/*in vitro*) and cell type/tissue was added to each gene to create the final format as displayed in DualSeqDB.

Table 3. List of dual RNA-Seq studies included in DualSeqDB.

Pathogen	Host Organism	Tissue/cell-type	Condition	GEO code	Reference
Streptococcus pyogenes	*Macaca fascicularis*	Skeletal muscle tissue	*In vivo*	GSE144100	(2)
Salmonella enterica Serovar Typhimurium SL1344	*Homo sapiens*	Hella-S3 cells	*In vitro*	GSE60144	(3)
Salmonella enterica Serovar Typhimurium SL1344	*Homo sapiens*	Endothelial cells Epithelial cells Monocytic cells NK cells	*In vitro*	GSE136717	(4)
Yersinia pseudotuberculosis IP 32953	*Mus musculus*	Lymphoid tissue	*In vivo*	PRJEB14242 (ENA)	(5)
Pseudomonas aeruginosa PA01	*Mus musculus*	Lung tissue	*In vivo*	SRP090213 (SRA)	(6)
Haemophilus ducreyi 35000HP	*Homo sapiens*	Skin tissue	*In vivo*	GSE130901	(7)
Mycobacterium tuberculosis ATCC 35733	*Homo sapiens*	THP-1 cells	*In vitro*	PRJEB6552 (ENA)	(8)
Streptococcus pneumoniae D39	*Homo sapiens*	A549 cells	*In vitro*	GSE79595	(9)

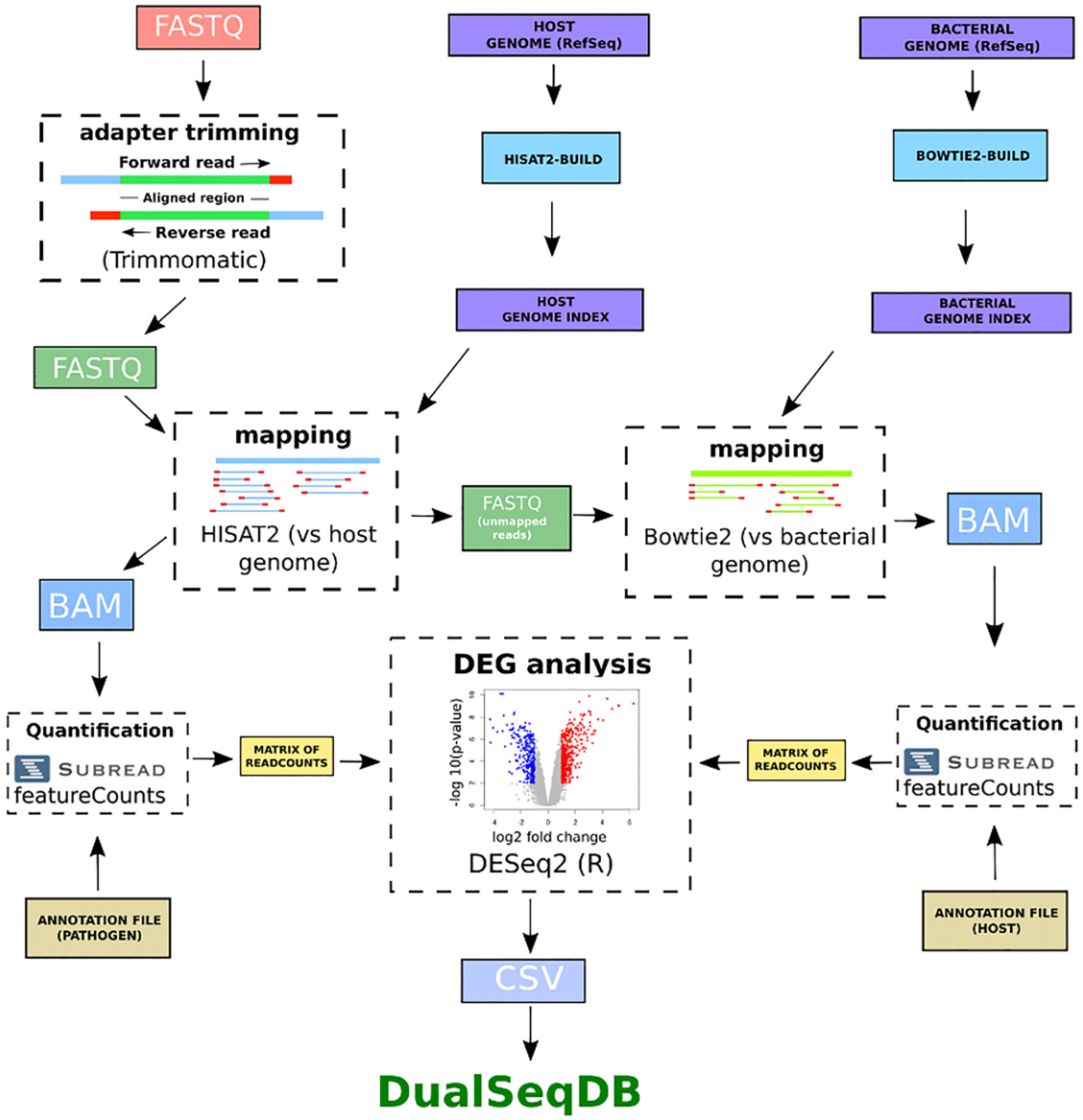


Figure 10. Pipeline used to re-process raw sequencing data from dual RNA-Sequencing studies. Raw sequencing data was downloaded from the GEO repository in FastQ format and adapter sequences were removed with Trimmomatic. Pathogen and host genomes and annotations were downloaded to build genome indices. Trimmed FastQ files were then mapped to the host index genome with HISAT2 and the unmapped reads were subsequently mapped to the pathogen index genome using Bowtie2. From this point onwards, pathogen and host reads were analyzed in parallel: mapped reads were quantified with FeatureCounts and their respective annotation files, creating a matrix of readcounts; this matrix of readcounts containing control and treated samples is then used as input for the DESeq2 R package to perform a differential expression analysis. The differential gene expression change (measured as $\log_2$ fold change) and corresponding p-value (Benjamini-Hochberg correction for multiple testing) was calculated using DESEq2.

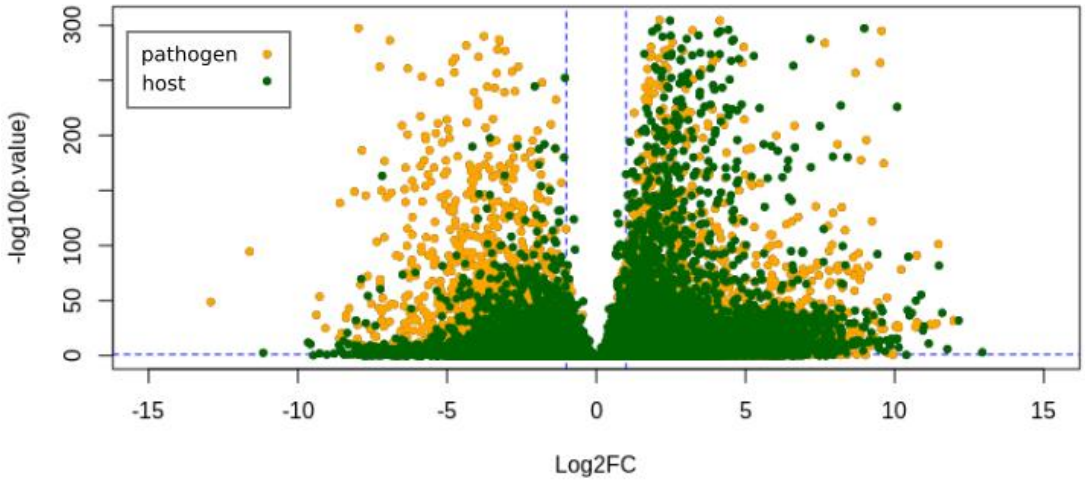


Figure 11. Visualization of overall statistical significance (p-value) and magnitude of change (log2FC) of all entries in DualSeqDB. Gene expression changes were considered significant when log2FC > 2 (upregulated) or log2FC < -2 (downregulated) and p-values < 0.05 (dashed blue lines). Pathogen genes are labelled in yellow and host genes in green.

4.3.2. Technical aspects

DualSeqDB was built using PHP on an Apache web server with a MySQL database backend. Sequence identifiers and cross-references were obtained from UniProt and the NCBI RefSeq, Gene and Genome resources (116). DualSeqDB stores no user data, except for the anonymous caching of BLAST search results for a given sequence in order to greatly speed up repeated searches. The open-source Bootstrap library was used to allow display on devices of any screen size, including mobile devices. Several icons were included from Font Awesome and the Noun Project, and a number of JavaScript libraries are used for table export and sorting.

4.3.3. BLAST search

The NCBI BLAST suite version 2.9.0+ (March 2019) (91) is used to search by sequence similarity. The BLASTP program is used for amino acid sequences, and BLASTX for nucleic acid (coding) sequences. BLAST search results are cached for each unique sequence, which means that re-running a search using the same sequence will yield results nearly instantaneously. As on all other pages, results from the BLAST search page can be linked to and shared with other researchers using the 'Link to these results' link at the bottom of the page. For sequences above a URL length of 2000 characters this link uses a sequence hash identifying the cached sequence, rather than the sequence itself.

4.4. Using DualSeqDB

DualSeqDB consists of several elements: (i) a text search function to find specific eukaryotic and bacterial genes, (ii) a BLAST search function to find genes similar to a protein or nucleic acid sequence of interest, (iii) a Browse function to quickly identify genes up and down-regulated during infection, and (iv) a Tutorial section to get started quickly by following a step-by-step guide. DualSeqDB relies on JavaScript, users will need to enable this in their web browser for full functionality.

4.4.1. Search function

To search for a gene or protein, users simply need to type its name or identifier. Any of the following options are available: gene symbols, gene locus identifiers, NCBI protein identifiers, UniProt protein accessions or a free-text search in the gene product's description (**Figure 12**). To search within a particular host and/or pathogen, users can select the pathogen and/or host name in the drop-down

menu. If no gene or protein name is given, the output will display a complete list of genes, similar to the Browse view (described below).

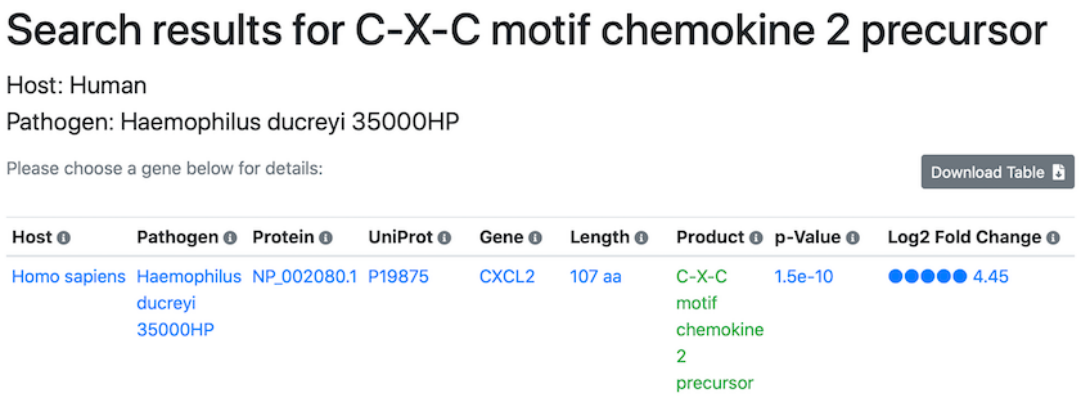


Figure 12. Search results in DualSeqDB. The search results page displays a list of any host or bacterial genes matching the search term. It also displays information on the infected host species and its associated pathogenic bacteria, the NCBI protein identifier, the UniProt protein accession code and the gene symbol of the gene for which the expression change was measured. This preview section also shows a description of the gene product and its length, together with the expression change value (measured as log_2 fold change) and the corresponding p-value. In this example, we show the case of CXCL2, a chemoattractant chemokine with pro-inflammatory function, involved in many immune responses, such as cancer metastasis, wound healing or angiogenesis. The results collected in DualSeqDB show that, upon infection of skin tissue human cells by the pathogen Haemophilus ducreyi, the human gene CXCL2 increases its expression levels, as indicated by a log_2 fold change above 4.

4.4.2. Tables on DualSeqDB: sorting, downloading and linking to results

After selecting a gene of interest, a view will open with all the infection information available for the corresponding gene (**Figure 13**). The heading of this page provides information on the selected protein: protein and host/pathogen name, length, gene name and UniProt ID. In the table, all available experimental data are listed: tissue of the host organism, tissue condition (whether the experiment was carried out *in vivo* or *in vitro*), time after infection, differential expression gene data, including the log_2 fold-change and the associated p-value, a note giving information on the growth conditions of control bacteria (including temperature and growth phase, whenever specified in its study, otherwise it is shown as “none”) and the reference to the original paper where the data was published.

A brief description on the meaning of $\log_2$ fold-change and p-value is also available as mouse-over explanation on the column headers. For any proteins in UniProt, a protein visualisation is automatically provided by ProViz from the Davey lab. Proteins larger than 5,000 amino acids are not displayed due to display speed limitations. ProViz is an interactive exploration tool that allows inspection of the structural, functional and evolutionary features of proteins, including Pfam domains and transmembrane regions. This tool is particularly useful for unknown and uncharacterised proteins.

Alternatively, the protein's FASTA sequence can be displayed by pressing the "Show protein sequence" button, along with a "Copy" link in the top right corner to copy and paste the protein's sequence into other research tools, or into the DualSeqDB BLAST Search to search for similar proteins. You can also immediately search for similar proteins via BLAST (see below for more details) by pressing the "Find similar proteins" button.

To sort the table as desired, users can select any of the column headers. The current table can be downloaded as a comma-separated CSV file for export into spreadsheet software such as Microsoft Excel using the "Download Table" button in the top right. An appropriate readable file name is automatically generated. The results can also be linked to and shared with other researchers by right-clicking and copying the "Link to these results" link at the bottom of the page.

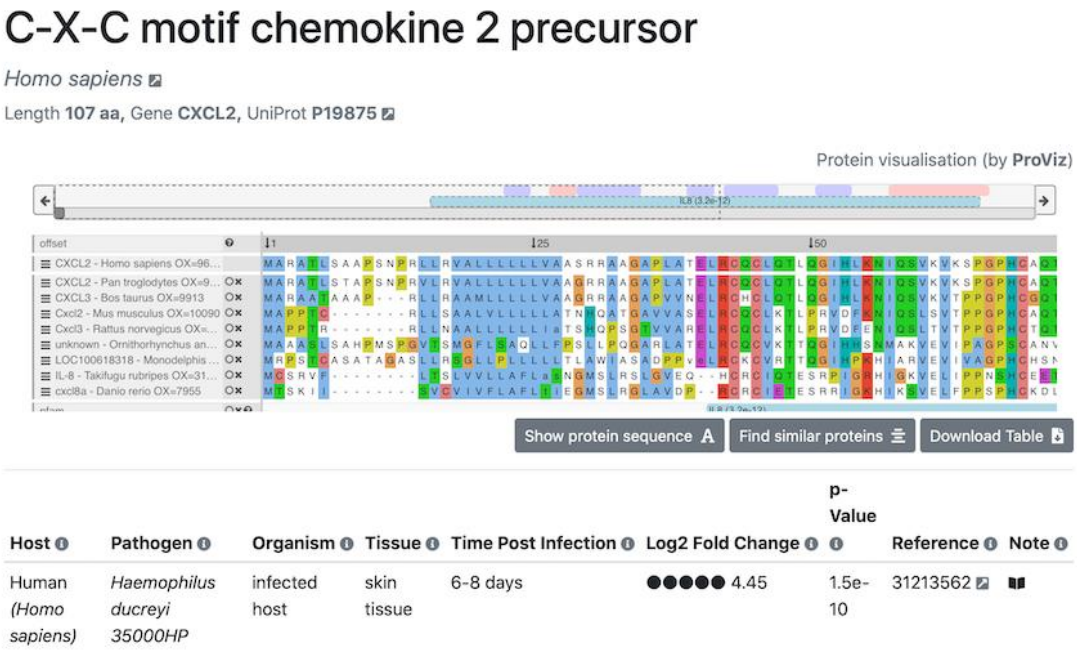


Figure 13. Detailed view of gene expression changes. The detailed view page displays all the information available for a host or bacterial gene, along with a ProViz visualization of the protein's sequence and structural features, showing sequence conservation with similar proteins. It also shows the log_2 fold change and the associated p-value for each entry, together with all the details of the experiment: host name, pathogen name, organism (indicating whether the measured gene belongs to the host or the pathogen), cell type/tissues, post-infection time points, as well as the PMID reference with a PubMed link to the original study, and a note column, specifying the bacterial growth conditions if listed in the original study.

4.4.3. BLAST search

The BLAST Search tab provides a search by sequence similarity. When the protein of interest is not in our database, the user may search for similar proteins using BLAST sequence alignment. Finding a similar protein with a high variation in log_2 fold-change and low p-value is a strong indication that the query sequence may be relevant during infection.

To search for similar proteins in our database using BLAST, protein or coding sequences in FASTA format can be used and has to be properly identified in the drop-down menu. When the BLAST alignment is ready, a search results page will open with the following information:

1. *Identity:* The percentage of sequence identity between query and target in the successfully aligned region.
2. *Aligned:* The total number of amino acids that were successfully aligned between query and target.
3. *Bit score:* The required size of a sequence database in which the current match could be found just by chance. The bit score is a $\log_2$-scaled and normalized raw score, meaning that each increase by one double the required database size.
4. *E-value:* The number of expected hits of similar quality (score) that could be found in the BLAST sequence database just by chance.

The meaning of the Host, Pathogen, Locus, Protein, Gene, Product, p-value, and $\log_2$ fold-change columns can be found in the Browse Tab section below, or via the mouse-over information symbols in the top row of any table. By default, BLAST matches with the highest Bit scores are shown first and matches with 100% sequence identity will be highlighted in green. Tables can also be sorted as desired using the column headers. As for all tables, the results can be downloaded as a comma-separated CSV file for export into spreadsheet software such as Microsoft Excel using the "Download Table" button in the top right corner. An appropriate readable file name is automatically generated. The results can also be linked to and shared with other researchers by right-clicking and copying the "Link to these results" link at the bottom of the results table.

4.4.4. Browsing the entire database

The Browse tab provides an overview of all entries in the DualSeqDB database. A pathogenic species or a host of interest can be chosen in the selection element at the top. This table is sorted by significance and $\log_2$ fold-change. It displays pathogen/host genes with a high and significant change in expression during infection at the top, followed by insignificant genes by decreasing $\log_2$ fold-

change in absolute value. Genes with very little expression changes and high p-value are listed at the very end of the table. Arrows next to each field provide links to useful external databases:

1. *Pathogen/Host:* Links out to the NCBI Taxonomy database, a comprehensive taxonomic database.
2. *Locus:* Links out to the Ensembl database, which provides genome annotation for all species included in the database.
3. *Protein:* Links out to the NCBI Protein database, which provides protein sequences and information.
4. *UniProt Accession and Gene Symbol:* Links out to the UniProt Knowledgebase, which provides comprehensive protein annotation.

Users can select the Locus, Protein, UniProt Accession or Gene Symbol entries to view details for the given protein in the external databases. This information is also available as a mouse-over explanation in the Browse tab.

4.4.5. Downloading the Entire Database

To download the entire DualSeqDB database for local analysis, please click the link available under the Download tab. Currently, DualSeqDB v1 is available, and will be upgraded with new data as they become available.

4.5. Discussion

The development of new antimicrobial therapies heavily relies on our knowledge of the mechanisms of bacterial infection (52,117,118). Therefore, it is crucial to understand how bacterial infection develops and which bacterial genes are required to infect a host. The use of high-throughput sequencing technologies has unveiled new levels of complexity in the transcriptomic response of pathogens and hosts during infection. In the last few years, dual RNA-seq has become the leading approach to uncover the intricate relationship between pathogen and host interactions. Hence, dual

RNA-seq could be used to define host–pathogen interactions or identify potential biomarkers of infection (119). At present, dual RNA-seq data are disseminated in multiple locations and incompatible formats and are therefore not accessible to the scientific community without specialized tools and knowledge.

DualSeqDB intends to be a valuable central resource for the systematic identification of proteins that are crucial for successful infection, aimed to understand how the bacterial and host transcriptomes change and interact during infection. In this context, we envisage that DualSeqDB will facilitate the finding of interspecies relationships between pathogen and host and will help us to uncover new mechanisms of infection. The analysis of the results included in DualSeqDB may inspire the design of new therapeutic interventions aimed to prevent the spread of infection. Given the current *momentum* of sequencing technologies in research and clinics, we expect that our database will grow continuously and become a comprehensive repository that will help us in the fight against infectious diseases.

4.6. Availability

To download the entire DualSeqDB database for local analysis, please click the link available under the Download tab at http://www.tartaglialab.com/dualseq. Currently, DualSeqDB v1 is available, and will be upgraded with new data as they become available.

5. CHAPTER 3. HPIPred: The Host-Pathogen Interactome Prediction tool

5.1. Abstract

Protein-protein interactions are involved in most of the cellular processes. Their correct identification is especially important in the context of pathogenic infections, as they can help get a better understanding of the infection process. The experimental methods commonly used for the detection of host-pathogen protein-protein interactions have their limitations due to cost and large-scale constraints. To circumvent these limitations, computational methods are nowadays used for the prediction of these protein interactions to support experimental data, although they generally suffer from high rates of false positive predictions. With the purpose to tackle this problematic, we have created HPIPred, a host-pathogen protein-protein interaction prediction tool based on the numerical encoding of different physicochemical properties of the amino acids, which also integrates phenotypic data related to the infection process to filter the results in a biologically meaningful way. By using the *Homo sapiens* and the *Pseudomonas aeruginosa PAO1* proteomes as input to our prediction tool, we generated a predicted host-pathogen of 763 interactions showing a highly connected network topology. We hope that our predictive model can be used by researchers to prioritize candidate protein-protein interactions as possible targets for the development of new antibacterial drugs.

5.2. Introduction

During the course of infection, pathogen proteins play a crucial role in the re-wiring of multiple biochemical processes occurring in the host, which ultimately allow for the progression of the infection (120). As a counterpart, hosts make use of their protein machinery to trigger defense mechanisms against the pathogen (121). Therefore, it is vital to characterize host-pathogen protein-protein interactions (PPIs) in order to gain better insight into the biological processes that rule pathogen infection and replication (53). Several detection methods, such as yeast two-hybrid (37–39), pull-down assays (122) or coimmunoprecipitation (35) are commonly used for the identification of novel PPIs, but it is estimated that only a small fraction of the PPI database has been characterized so far. With the purpose to tackle the scarcity of validated PPIs, it is becoming increasingly common to develop in silico methods that aid in the prediction of PPIs.

In spite of the fact that there are many protein-protein interaction predictors available, such as homology-based (61), domain and motif interaction-based (65), structure-based (67) and machine learning methods (69,70), the lack of validated datasets and the problem of class imbalance causes that the predicting potential of these algorithms is still far from optimal. With the purpose to address the aforementioned, we have developed a predictive algorithm of host-pathogen protein-protein interactions that determines protein similarity based on cross-correlation between numerical representations of the proteins by using physicochemical properties of their amino acids. In order to improve robustness, the predictions from single models are then combined into a consensus interactome, which is finally integrated with phenotypic data collected in biological databases related with infection processes, allowing us to give a ranked score to each interaction.

5.3. Methods

5.3.1. Data collection and dataset construction

5.3.1.1. Positive dataset

Host-pathogen protein-protein interactions (PPIs) were obtained from PHISTO (50), a bioinformatics Web resource which enables access to the most updated and experimentally validated host-pathogen protein-protein data. A total of 9,237 intra-species PPIs between bacteria and *Homo sapiens* were used as the initial dataset of positives. This dataset contains interaction data described between *Homo sapiens* and 95 different bacterial strains, with more than 90% of these entries belonging to *Homo sapiens – Yersinia pestis* interactions (4,069), *Homo sapiens – Bacillus anthracis* interactions (3,053) and *Homo sapiens – Francisella tularensis* interactions (1,348). We applied a length filtering criterion to remove PPIs containing any protein shorter than 100 amino acids or longer than 2,000 amino acids, obtaining a final dataset of 7,423 PPIs. This dataset represents the interactome of 3,327 human proteins against 2,496 bacterial proteins.

5.3.1.2. Synthetic negative dataset

Random-sequence protein libraries were created with the 20-amino acid alphabet using a gamma distribution to fit the observed protein-length distribution in eukaryotic and bacterial organisms. Fixed average protein lengths of 472 and 319 amino acids were used for the creation of the eukaryotic and bacterial proteomes, respectively (123). The chosen sizes for the proteomes were 20,000 and 3,000 proteins, similar to the sizes of the human proteome and bacterial proteome, respectively. Afterwards, a length filtering criterion was used to remove proteins shorter than 100 or longer than 2,000 amino acids, giving a total of 2,598 and 18,669 surviving proteins for the bacteria and the host, respectively.

5.3.1.3. Negative dataset for model validation

The *Homo sapiens* proteome was downloaded from UniProt (proteome ref: UP000005640) (95) and used as the host fraction of proteins. The combined proteomes of *Yersinia pestis* (proteome ref: UP000000815), *Francisella tularensis* (proteome ref: UP000001174) and *Bacillus anthracis* (proteome ref: UP000000594) were also downloaded from UniProt and used as the pathogen fraction of proteins. Non-interacting pairs of proteins were created by randomly pairing proteins from the host and the pathogen fractions, discarding those pairs that were also part of the positive PPI network, and applying the length filtering criteria previously described, to finally obtain a dataset of 7,421 entries, which were divided into the host (2,734) and bacterial (2,102) fraction of unique proteins.

5.3.1.4. Query proteome datasets

Homo sapiens and *Pseudomonas aeruginosa PAO1* (proteome ref: UP000002438) proteomes were used as host and pathogen query, respectively. After applying our length filtering criteria previously described, host and bacteria proteome were composed by 19,192 and 1,314 proteins, respectively.

5.3.2. Prediction of protein-protein interactions (single model)

The main steps involved in our prediction algorithm are the transformation of amino acid sequences to numerical sequences, the calculation of similarity scores between the numerical sequences from the query datasets and the positive datasets, and the prediction of putative PPIs based on their similarity scores to known PPIs (**Figure 15**).

5.3.2.1. Numerical encoding of protein sequences

Each protein sequence from the positive, negative and query datasets was transformed into a numerical sequence by using a physicochemical property of the amino acids, which allowed to treat

the amino acid sequence as a numerical signal (**Figure 15 A**). These physicochemical properties are experimentally measured and represented as numerical indices for each amino acid as included in Aaindex (124–127). In order to represent the forces that contribute to the binding interaction between proteins, the numerical encoding of the protein sequences was performed individually for five different physicochemical properties. Specifically, we used alpha-helix indices (GEIM800101) and beta-strand indices (GEIM800105) to represent structure and ultimately hydrogen bonding, hydrophobicity index (ZIMJ680101) to represent hydrophobic effect and isoelectric point (ZIMJ680104) and electron-ion interaction potential values (COSI940101) indices to account for electrostatic forces. A 0-1 normalization step was applied to each descriptor to scale the values to the same range, allow direct comparison between and avoid biases. After numerical encoding, a moving average with a sliding window of 9 positions was used to smooth the data and represent each amino acid's numerical value as a measure of itself and its near environment.

5.3.2.2. Assessing protein similarity by cross-correlation

To determine similarity in the physicochemical profiles of proteins, we calculated the cross-correlation coefficients (CCCs) between the query dataset and the positive dataset by performing one-vs-all pairwise comparisons, that is, each query protein was individually tested against all proteins in the positive dataset, separately for the host and the pathogen proteins (**Figure 15 B**). The maximum lag at which the CCCs were calculated was set to a fixed value of 200 (**Figure 14**). The highest CCC obtained for each pairwise comparison was assigned as a measure of similarity between the two proteins being compared, creating a database of similar proteins. Each entry in the database represented a pairwise comparison and included the highest CCC and the length of the query protein.

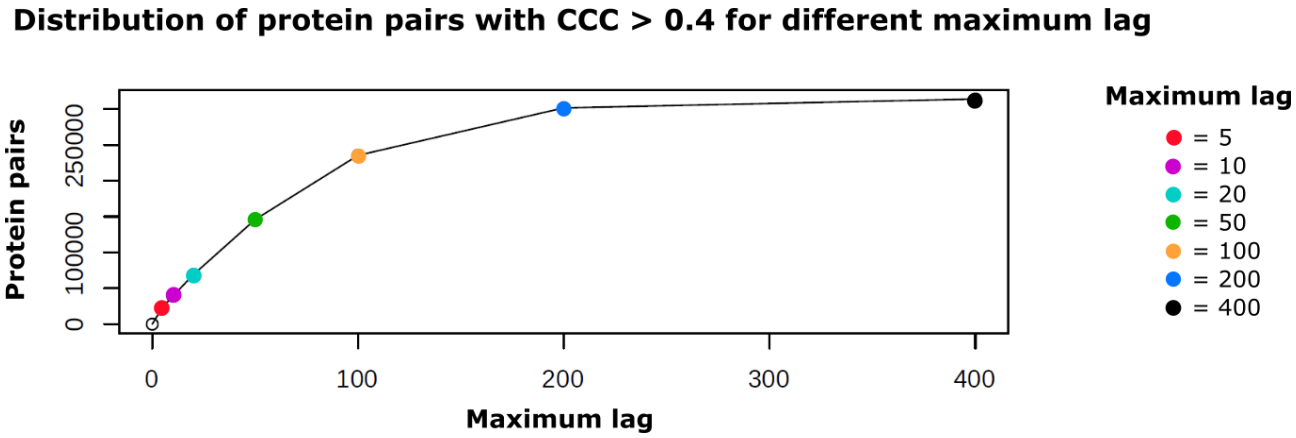


Figure 14. Determination of the optimal maximum lag value for the calculation of CCCs. In order to determine the optimal maximum lag value, the hydrophobicity index was used to calculate the CCCs between the Homo sapiens query proteins and the positive dataset for different maximum lag values, namely 5, 10, 20, 50, 100, 200 and 400. After plotting the number of protein pairs with a CCC > 0.4 for each maximum lag value tested, the values adjusted to a logarithmic function and the number of surviving protein pairs did not increase substantially after a maximum lag max of 200.

5.3.2.3. Filtering low-scored proteins (Filtering step I)

To remove low-scored proteins, the database of similar proteins obtained in the previous step was filtered by removing all the pairwise comparisons with an associated CCC lower than 0.4 (**Figure 15 C.1**).

5.3.2.4. Using a synthetic negative dataset for filtering (Filtering step II)

We introduced a second filtering criterion by making use of the synthetic dataset of proteins. We evaluated the synthetic dataset against the positive dataset and obtained the CCCs for all proteins in the set, respectively for the host and pathogen proteins. As the synthetic set contains only random-sequence proteins, none of them should be considered as part of putative interacting pairs, irrespective of whether they passed the 0.4 threshold. The cross-correlation values of the synthetic dataset were plotted against their corresponding protein lengths. As the length of the synthetic proteins increased,

their associated CCCs decreased linearly, a correlation that was also observed with the query datasets. Hence, we manually determined the slope and intercept points of parallel linear equations of the type:

$$Y = aX + b$$

that represented such negative linear correlation between protein length and CCC for the synthetic dataset, so that only 0.1%, 0.01%, 0.001% and 0.0001% of the data points fell above the equations. The determination of these linear equations was carried out and averaged over the five different physicochemical properties used, namely hydrophobicity, isoelectric point, alpha-helix and beta-sheet indices and electron-ion interaction potential (**Figure 15 C.2**). We calculated a threshold value for each pairwise comparison in the database of similar proteins by substituting the X variable in the linear equation with the length of the query protein. Afterwards, we discarded those pairwise comparisons whose threshold value was higher than its associated CCC (**Figure 15 C.3**), to obtain a filtered database of similar proteins, repeating this process for each parallel linear equation.

5.3.2.5. Prediction of protein-protein interactions

To predict putative PPIs, each PPI from the positive dataset was collated with the filtered databases of similar proteins in the following way: the host protein in the positive PPI was searched against all the pairwise comparisons in the database of similar proteins of the host and whenever a match was found, the query protein was kept; the same procedure was repeated for the pathogen protein in the positive PPI, and any pair of host-pathogen query proteins obtained through this search process was considered a putative PPI due to its similarity with the positive PPI (**Figure 15 D.1**). We performed this search sequentially with all the PPIs in the positive dataset to obtain a predicted interactome of PPIs (**Figure 15 D.2**).

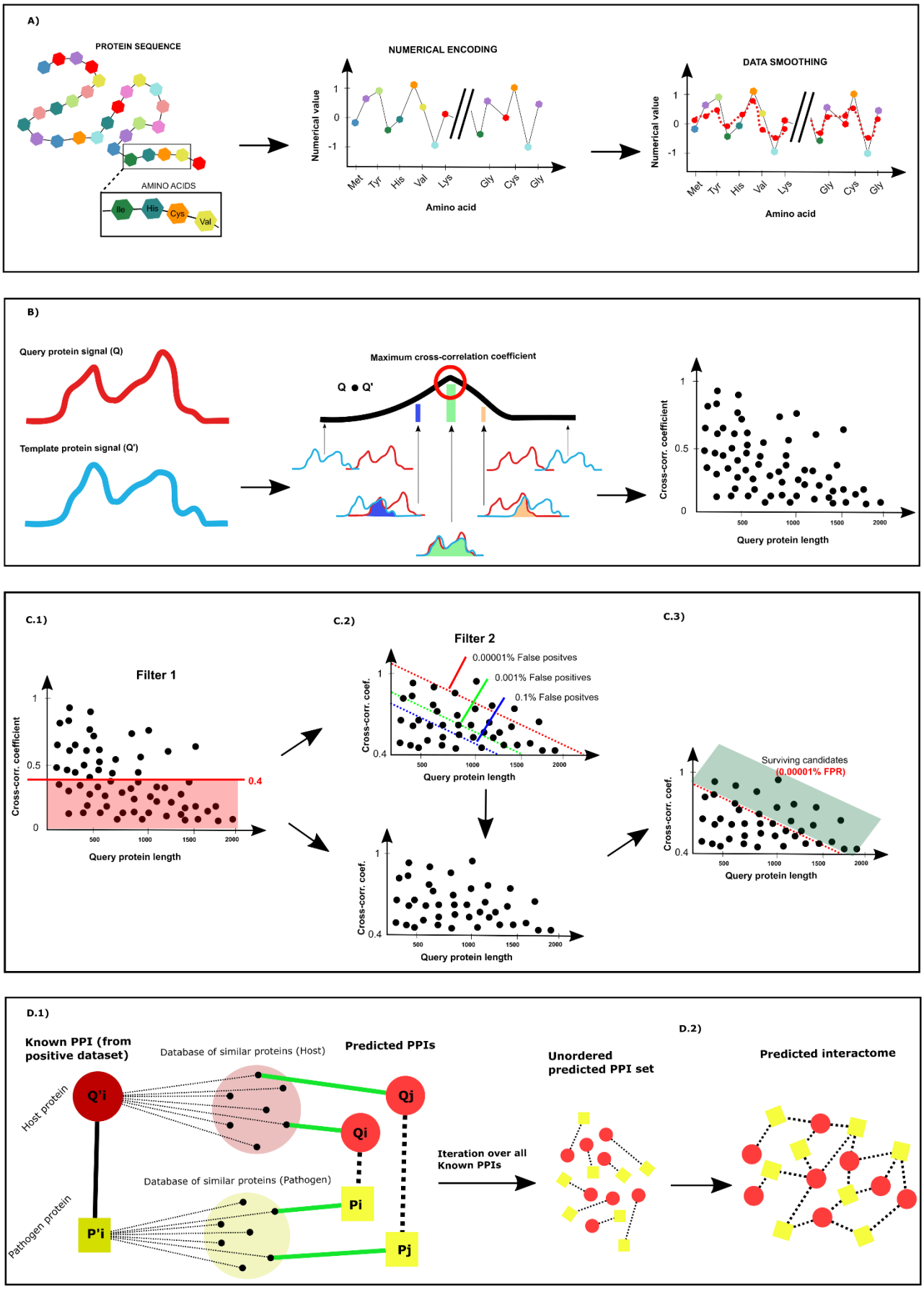


Figure 15. Pipeline of the prediction algorithm of PPIs for a single model. A) Numerical encoding of proteins. B) Assessing protein similarity by cross-correlation. C) Filtering steps. D) Prediction of protein-protein interactions.

5.3.3. Prediction of a consensus interactome

5.3.3.1. Model combination

Individual host-pathogen interactomes were predicted for the five different physicochemical properties previously described. These five interactomes were combined to create a consensus interactome, such that it included any PPI that had been predicted by at least three individual models (**Figure 16 A**).

5.3.3.2. Phenotypic scoring of predicted PPIs

In order to add additional layers of information to the predicted interactions, we used the proteins in these interactions as queries to run protein BLAST (91) against several databases that contain information about the infection phenotype, namely BacFITbase, DualSeqDB and PHI-base (51,118,128) (**Figure 16 B**).

5.3.3.2.1. Sequence alignment against BacFITbase

We downloaded the file including all the entries from BacFITBase v1.0, a database that contains data on bacterial fitness, which represents the importance of each bacterial gene in the infection process, measured through transposon mutagenesis. Afterwards, we performed protein sequence alignment between each query protein from the pathogen proteins in our predicted PPIs and this database, discarding all reported hits with a percent of identity ≤ 40% and an Expect value (e-value) ≥ 10^{-50}. We then filtered out those whose fitness score in BacFITBase had an associated p-value ≥ 0.05. In order to assign an average fitness score to a query protein, we calculated the mean fitness score for all the surviving hits of that query, as well as a mean standard deviation score. Those queries with no surviving hits were assigned “NA”. Finally, under the assumption that the lowest fitness scores are

the most relevant in infection, we performed 0-1 normalization, assigning a value of 1 to the lowest fitness score reported, and a value of 0 to the highest.

5.3.3.2.2. Sequence alignment against DualSeqDB

We downloaded the file containing all the entries from DualSeqDB, a database that contains data of bacterial and host genes and their expression changes in different bacterial infection models, represented as $\log_2$ fold change as measured by dual RNA-Seq experiments. Following the same criteria and parameters used with the BacFITBase database, we performed protein sequence alignment between each query protein and the filtered DualSeqDB, respectively for the bacterial and the host fractions, discarding those hits with a percent of identity $\leq 40\%$ and an e-value $\geq 10^{-50}$. We then filtered out those hits with an associated p-value ≥ 0.05 and for each query protein we averaged over the $\log_2$ fold changes of all surviving hits to assign a representative $\log_2$ fold change and standard deviation scores, or "NA" if not hit passed the threshold. Standard 0-1 normalization was performed at the end.

5.3.3.2.3. Sequence alignment against PHI-base

We downloaded the file including all entries in the PHI-base, a dataset that contains information on pathogen genes and assigns each of them a "mutant phenotype" depending on how their mutation affects the organism's pathogenicity. In some cases, the same gene can have more than one entry as it may have been measured in a different infection setting. We filtered out those entries referring to pathogens which do not belong to the bacterial kingdom. Afterwards, we only kept entries with mutant phenotype tags that matched "unaffected pathogenicity", "loss of pathogenicity", "reduced virulence", "lethal" or "increased virulence (hypervirulence)", and transformed them into numeric

values, 0, 0.5 or 1 in the following way: “lethal” = 1, “loss of pathogenicity” = 1, “reduced virulence” = 0.5, “increased virulence (hypervirulence)” = 0.5, “unaffected pathogenicity” = 0. In order to assign a unique numerical value to the genes which had several discrepant phenotypes associated, only the most abundant case was chosen as the representative value, filtering out the rest of the genes for which this requirement was not fulfilled. We then performed sequence alignment between the query entries from the pathogen in our predicted PPIs and the surviving genes in PHI-base in the same way described for the previous cases, obtaining an average PHI-base score and a mean standard deviation score for each query protein, and “NAs” for those queries with no hits found.

5.3.3.3. Betweenness centrality of host proteins

As indicated by the centrality-lethality rule (52,129), those proteins that are important for network connectivity and centrality are usually essential proteins for the organism. In this sense, betweenness centrality is a relevant centrality measure, as nodes with high betweenness centrality lie on communication paths and determine network integrity (130,131). As a way to measure the essentiality of the *Homo sapiens* protein interactome within its biological context, we calculated the betweenness centrality. To accomplish it, the host interactome was downloaded from the STRING database (46). We filtered out all PPIs with a confidence score lower than 0.9. We then used igraph R package (132) to build an undirected graph with the surviving PPIs, calculated the node betweenness centrality for all nodes in the graph, each representing a protein from the host proteome, and performed standard 0-1 normalization (**Figure 16 B**).

5.3.3.4. Determination of a ranked score

For each predicted PPI in the combined interactome we compiled all the normalized scores obtained in the previous steps (BacFITBase, DualSeqDB, PHI-base scores for the pathogen proteins and betweeness centrality and DualSeqDB scores for the host proteins) and calculated an average weighted score with a ranging value from 0 to 1 with the following formula:

$$Weighted\ score = \frac{fitness + log2FC\ (pathogen) + PHIbase + log2FC(host) + centrality}{Number\ of\ non - missing\ values}$$

Also, a phenotypic confidence score was calculated with values ranging from 0 to 5. For each PPI to indicate the number of missing values (NAs) in the previous formula, namely a score of 5 showing no missing value and 0 indicating that no value had been reported for that specific PPI. In order to account for both the weighted score and the confidence score, we performed 0-1 normalization to adjust for the number of missing values and obtained a normalized ranked score (**Figure 16 B**):

$$Normalized\ ranked\ score = \frac{weighted\ score\ \cdot confidence\ score}{\max(weighted\ score\ \cdot\ confidence\ score)}$$

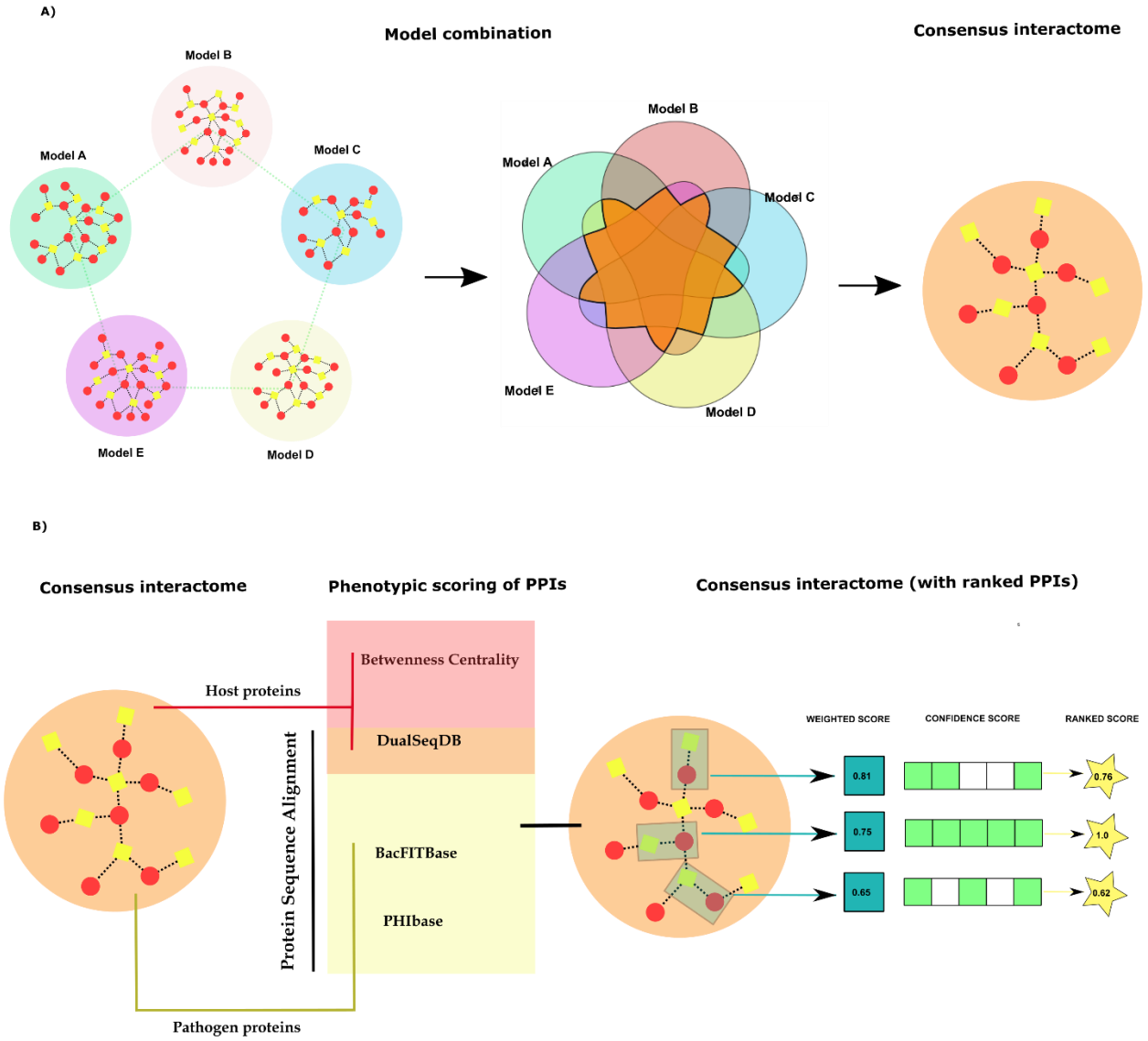


Figure 16. Model combination and calculation of ranked scores. A) Model combination. B) Determination of ranked scores for the PPIs in the consensus interactome.

5.3.4. Validation

In order to validate the performance of our algorithm, each individual PPI from the positive dataset was taken out of the predictive models and used as query input (leave-one-out cross validation). All the PPIs recovered in the combined interactome generated were considered as True Positives (TP), while the rest were considered as False Negatives (FN). Subsequently, we passed each of the non-interacting pairs of proteins as input to our predictive algorithm. In this case, any non-interacting pair recovered in the combined interactome was treated as a False Positive (FP), whereas the remaining ones were treated as True Positives (TP).

5.3.5. Software implementation

Our predictive algorithm has been wrapped up in a command-line tool that displays dialog boxes and allows the user to choose the host and pathogen organisms to be used as queries or introduce custom files, select among more than 400 different physicochemical descriptors and set up the desired false positive rate and percentage of agreeing models to reconstruct a consensus interactome. We also pre-calculated protein similarity between the positive dataset and five model organisms that can be used as hosts, namely *Homo sapiens*, *Mus musculus*, *Dario rerio*, *Caenorhabditis elegans* and *Drosophila melanogaster*, for five different physicochemical descriptors, specifically hydrophobicity, isoelectric point, alpha-helix indices, beta-helix indices and electron-interaction potential. Finally, the user can download and integrate these data into the software in order to speed up calculations related to any of the mentioned model organisms.

5.4. Results

5.4.1. Validation

The results of the model validation, summarized in **Table 4**, showed that there was a relatively low recovery rate of positive PPIs, ranging from ~2% (150/7423) with the most permissive filter to ~1% (85/7423) when applying the most restrictive one, whereas the wrongly predicted positives were nearly zero in most conditions. There was a mild improvement in the recovery rate of positive PPIs when using a more permissive filter at the expense of increasing the computational cost and the run time, due to the higher number of calculations that needed to be performed. Nonetheless, the low efficiency of the method goes in accordance with the predictive power reported by other protein-protein prediction algorithms which suffer mainly from using incomplete datasets of PPIs: due to the fact that only a small fraction of the search space of PPIs has been validated experimentally models don't perform well at generalizing, because the penalty of removing a known PPI from the positive dataset is very costly.

Table 4. Model evaluation results. CI = number combined interactomes. TP = True Positives. FP = False Positives. TN = True negatives. FN = False Negatives.

FPR	**0.1**			**0.01**			**0.001**			**0.0001**			**0.00001**		
# CI	**3**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**	**3**	**4**	**5**
TP	150	136	120	136	121	107	129	118	97	111	105	91	108	100	85
FP	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0
TN	7,419	7,421	7,421	7,420	7,421	7,421	7,421	7,421	7,421	7,421	7,421	7,421	7,421	7,421	7,421
FN	7,273	7,287	7,303	7,287	7,302	7,316	7,294	7,305	7,326	7,312	7,318	7,332	7,315	7,323	7,338

5.4.2. Prediction of the host-pathogen interactome

Due to the size of the positive datasets and the input query datasets, the algorithm tested ~64,000,000 and ~3,300,000 pairwise comparisons to determine the maximum cross-correlation coefficient for the host and pathogen fractions, respectively, repeating this process for each physicochemical property. After discarding those protein pairs with a cross-correlation score lower that 0.4 (Filter I), we obtained ~5,500,000 and ~470,000 protein pairs, on average for each 5 physicochemical properties tested. The filtering step involving the negative dataset (Filter II) allowed to further reduce the amount of surviving protein pairs in the host fraction to ~105,000 (FPR = 0.1%), ~21,000 (FPR = 0.01%), ~8,000 (FPR = 0.001%), ~4,900 (FPR = 0.0001%) and ~4,400 (FPR = 0.00001%) for the host. As for the pathogen fraction, the surviving protein pairs were also greatly reduced to ~3,500 combinations (FPR = 0.1%), ~620 (FPR = 0.01%), ~250 (FPR = 0.001%), ~175 (FPR = 0.0001%) and ~130 (FPR = 0.00001).

The surviving host-host and pathogen-pathogen protein pairs from the previous filtering step were used to predict host-pathogen interactomes for *Homo sapiens* and *Pseudomonas PAO1* at 5 different FPR (0.1%, 0.01 %, 0.001 %, 0.0001% and 0.00001%) by using each of the five physicochemical descriptors mentioned (GEIM800101, GEIM800105, ZIMJ680101, ZIMJ680104, COSI940101), as well as their combined model. The whole search space of putative PPIs that can be predicted, calculated by multiplying the filtered proteome sizes of *Homo sapiens* and *Pseudomonas aeruginosa PAO1,* add up to ~25,000,000 possible PPIs. However, the results of the different models run by our predictive algorithm, summarized in terms of interactome sizes (**Table 5**), showed how the filtering step where different FPRs are used had an impact on narrowing down the search space of PPIs by several orders of magnitude. For instance, in the case of the single models we went from generating interactomes of a few hundred thousand PPIs when applying the most permissive filter (0.1%) to predict interactomes of just several hundred PPIs when the most restrictive filter (0.00001%) was used. Furthermore, it can be observed the effect of the combination of models to further decrease the

number of predicted PPIs, as indicated by the reduction by two orders of magnitude of the interactome predicted by the most permissive filter. In the same way, we can notice how, as the filter criteria becomes more restrictive, the sizes of the interactomes predicted by individual models are more similar to the combined model, indicating the filtering power of the combination of models and suggesting that upon using our most restrictive filtering criteria, the single models become almost as robust as the combined models. Moreover, the prediction of interactomes of relatively small sizes can be seen as an advantage for the downstream analysis, in terms of network visualization or gene ontology enrichment analysis.

Table 5. PPI sizes of the predicted interactomes by individual and combined models, at different FPRs.

	PREDICTED PPIs					
Model / FPR	**ZIM101**	**ZIM104**	**GEIM101**	**GEIM105**	**COSI940101**	**COMB PPI**
0.1%	142,089	386,529	551,043	143,493	225,456	**2,594**
0.01%	8,469	15,155	22,907	7,082	15,119	**1,661**
0.001%	1,853	2,322	3,491	1,485	3,717	**1,104**
0.0001%	1,035	1,085	1,439	968	1,779	**894**
0.00001%	814	915	915	763	1,146	**763**

5.4.2. Analysis of the *Homo sapiens - Pseudomonas aeruginosa PAO1* predicted interactome

We chose to further analyze the predicted interactome generated after using the most restrictive filtering criteria (FPR = 0.00001%) because of its adequate proteome size (763 PPIs) and also due to the fact that the interactomes predicted by the most restrictive FPRs are subsets of the most permissive models. After calculating the node betweenness centrality of the proteins in the *Homo sapiens* proteome and performing sequence alignment against DualSeqDB, BacFITBase and PHI-base, we generated a ranked score for each PPI. This, in turn, allowed us to prioritize the PPIs not only

according to protein similarity based on physicochemical properties, but also on network topology and biological properties related to the infection process. Afterwards, we visualized the predicted interactome in Cytoscape (**Figure 17**) (133) by representing it as a bipartite graph (depicted in figure 3), where host proteins (host nodes) are only connected to pathogen proteins (pathogen nodes) and vice versa. The ranked score of each predicted PPI was represented by the thickness and the color intensity of the edges connecting the PPIs in a proportional way, that is, the closer the ranked score of a PPI was to 1, the thicker its edge and the more intense its color, meaning that this PPI scored good on average on the biological databases it was compared to. Furthermore, we colored the nodes to represent changes in expression, derived from the BLAST search against DualSeqDB, in a range from blue (downregulated) to red (upregulated). Some of the top scoring PPIs (**Table 6**) have been highlighted in the Cytoscape network as a way to show how highly ranked PPIs, which represent protein pairs with inferred biological relevance in infection, seem to also be important for network integrity and connectivity in our predicted interactome.

At the same time, we filtered out all PPIs with a ranked normalized score lower or equal than 0.6, separated the proteins that integrated these PPIs into a pathogen fraction (16 proteins) and a host fraction (128 proteins) of unique proteins, respectively, which were then used to perform gene ontology enrichment analysis with DAVID functional annotation tool (**Figure 18**) (134,135). The results showed that the host fraction of proteins was enriched in biological processes related to immune and inflammatory response, such as regulation of NF-KappaB activity, cellular iron homeostasis and actin filament bundle assembly. As for the pathogen fraction, we observed an enrichment in biological terms related to amino acid and nucleotide biosynthesis, as well as folate biosynthesis, which are required for bacterial proliferation; these biosynthetic routes have been used as molecular targets for the development of antibacterials and bacterial resistance against these has also been described.

Table 6. PPIs with high ranked score which greatly interconnect the Homo sapiens – Pseudomonas aeruginosa PAO1 predicted interactome.

Query Pathogen	Query Host	Template Pathogen	Template Host	Supporting predictions	Weighted score	Confidence	Ranked Score
Q9HXN2	Q99459	Q5NEC0	Q99459	5	0.704	5	1.00
Q9HXN2	P46379	Q8ZCQ2	P46379	5	0.737	4	0.819
P50587	P46379	P46379	P46379	4	0.664	4	0.726
Q9HXN2	Q9Y6X8	Q5NEC0	Q9Y6X8	5	0.631	4	0.685
Q9I6M5	Q9Y6X8	Q81ZE2	Q9Y6X8	3	0.621	4	0.672
P48247	Q9NQB0	Q81LD0	Q9NQB0	3	0.472	5	0.633
Q9HXN2	Q9GZM7	Q8ZCQ2	Q9GZM7	5	0.583	4	0.624
Q9I6E0	Q9Y6X8	Q8ZAB3	Q9Y6X8	4	0.538	4	0.566
P50587	Q9NQB0	Q8ZJP7	Q9NQB0	4	0.424	5	0.556

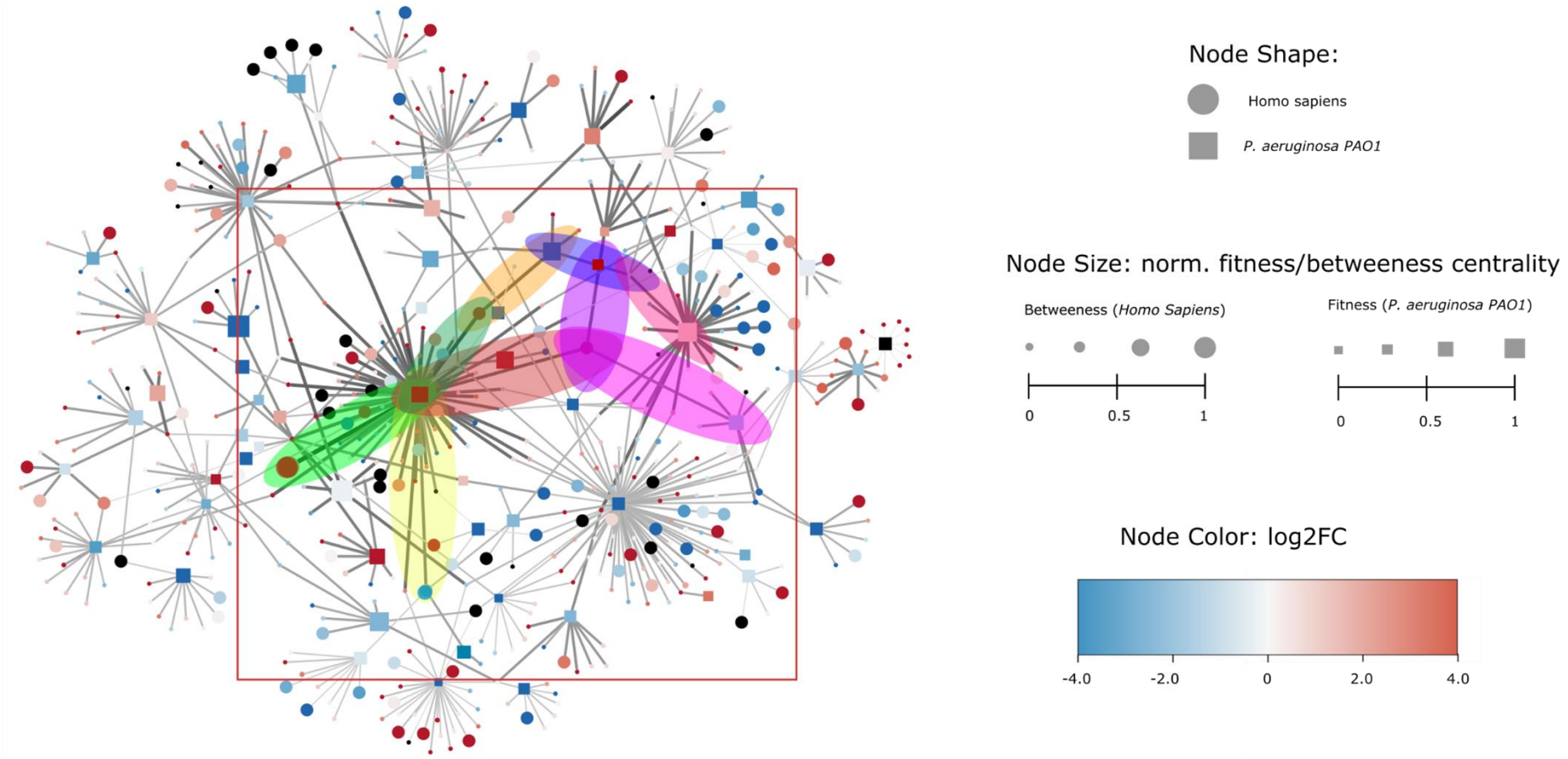


Figure 17. Network representation of the Homo sapiens - Pseudomonas aeruginosa PAO1 interactome predicted by the combined models. Nodes represent proteins and edges represent predicted interactions between proteins. Host and pathogen proteins are represented by circles and squares, respectively. Nodes are colored according to the normalized expression changes (computationally derived from DualSeqDB) or black in case of missing information. Node sizes are proportionally to normalized betweenness centrality and fitness for the host and pathogen proteins, respectively. Edge size and width corresponds to the PPI ranked score (0-1 scale). Some of the PPIs with the highest final scores have been highlighted with colors according to the color code in Table 2, in order to show how the highest ranked PPIs from our predictive algorithm allow to reconstruct a highly connected subnetwork.

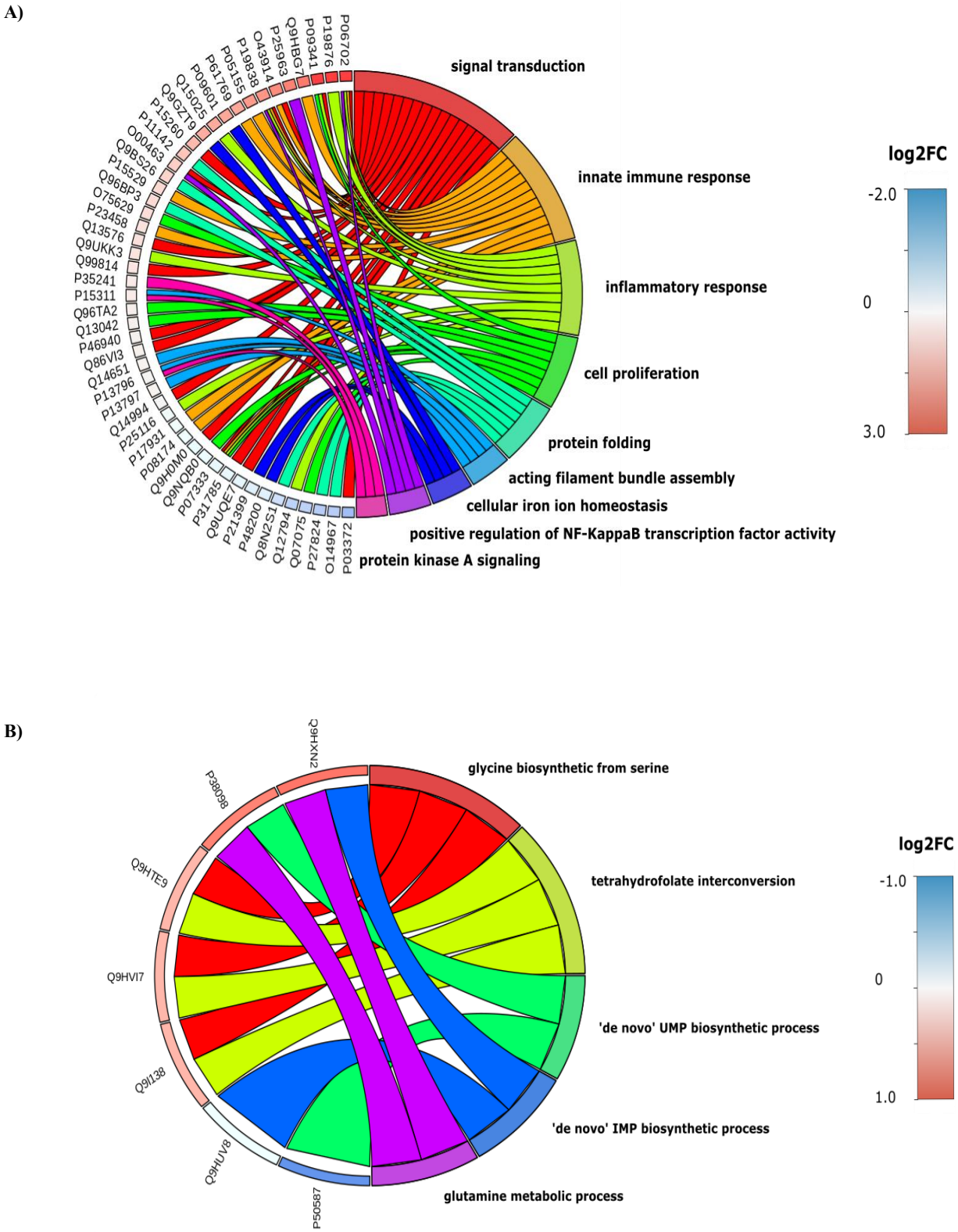


Figure 18. Gene Ontology enrichment analysis of the proteins from the highest scoring PPIs predicted. Biological processes (BP) depicted. Log Fold changes assigned from sequence similarity to log2 fold changes from DualSeqDB. A) Host proteins. B) Pathogen proteins.

5.4.3. Benchmarking

We compared HPIPred with one publicly available predictive software called BIPS (Biana Interolog Prediction Server) (136), which predicts putative PPIs based on interolog information found in different protein-protein interaction databases. We used *Homo sapiens* (proteome ref: UP000005640) and *Pseudomonas PAO1* (proteome ref: UP000002438) proteomes as query inputs to make the results directly comparable to the consensus interactome that we previously showed. We ran the software with default parameters and obtained very few predictions, so we relaxed the filtering criteria involving identity similarity to 40%. BIPS predictive tool generated an interactome consisting of 963 PPIs, compared to the 763 PPIs predicted by our algorithm. The results from BIPS and our own algorithm were then compared by creating an intersection of the predicted interactomes, which revealed that both methods shared a total of 262 common PPIs (**Figure 19**). Afterwards, we represented these common PPIs as a network in Cytoscape (**Figure 20**), where we observed that, in general, the shared PPIs maintained a certain degree of network connectivity and the proteins involved presented a high score in terms of betweenness centrality and fitness for the host and the pathogen, respectively.

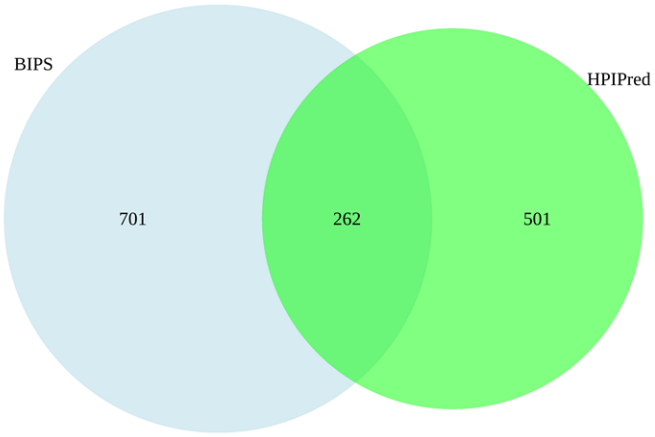


Figure 19. Venn diagram showing the number of predicted PPIs shared by BIPS and HPIPred.

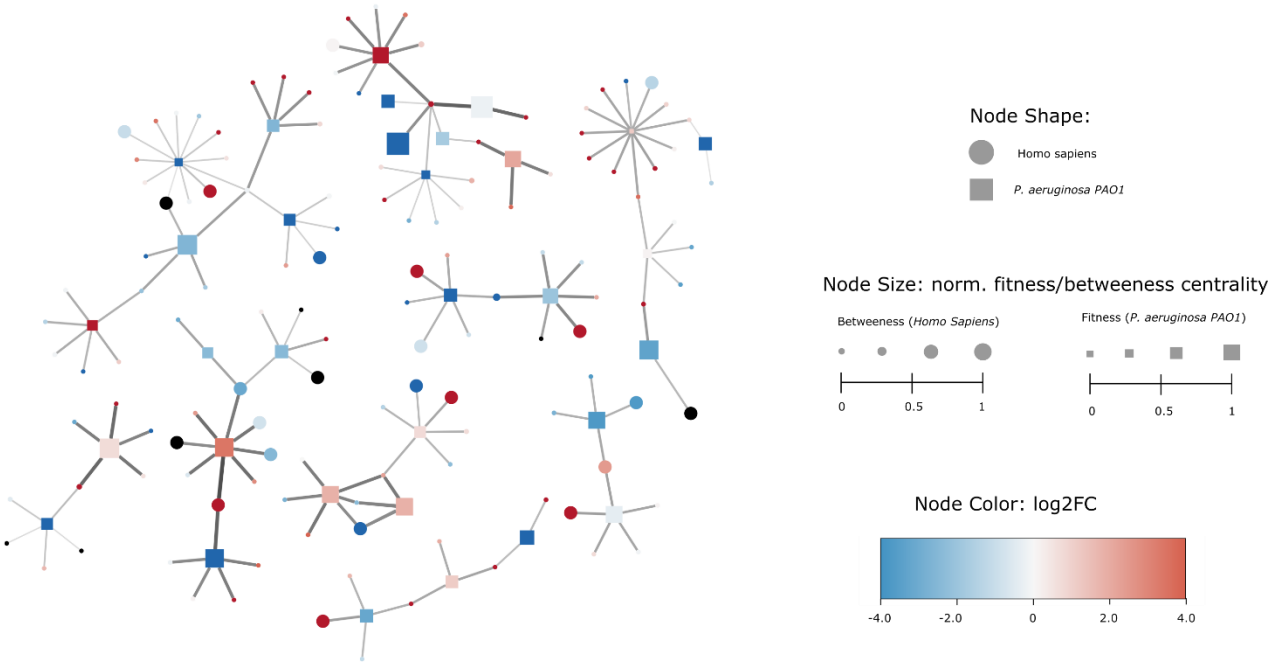


Figure 20. Network representation of the common PPIs by BIPS and HPIPred.

5.5. Discussion

Infectious diseases are a growing health concern worldwide due to the rise of multi-drug resistant bacteria. These pathogenic bacteria cause prolonged hospitalization, higher medical treatment expenses and an increase in the mortality rate. In this sense, protein interactions between pathogenic bacteria and their natural hosts play a key role in the infection mechanism and a thorough understanding of their complex interplay is required for the development of new antibiotics. Numerous experimental techniques, such as yeast two-hybrid, pulldown assays or co-immunoprecipitation, are currently used for the detection of these interactions, which are collected in existing databases through literature mining or manual curation. However, experimental techniques are time-consuming, costly and suffer from low-specificity, making it unfeasible to evaluate all possible protein-protein interactions.

Recently, computational approaches such as machine learning, homology-based methods or structure-based methods, have allowed the prediction of putative PPIs, complementing experimental techniques. The main caveat of these prediction methods is the high false positive rates that they generate, mainly due to the lack of robust databases of experimentally validated PPIs. This data shortage causes the prediction methods to perform poorly in terms of generalization.

HPIPred has been developed as a tool that could help reduce the false positive rate compared to other methods. To this end, HPIPred predicts putative PPIs through numerical encoding of the proteins based on physicochemical properties of the amino acids and integrates these predictions with biologically meaningful data in infection. These data include information on the *in vivo* relevance of bacterial genes for the infection process, the *in vitro* and *in vivo* gene expression changes occurring in the host and the pathogen, as well as topology information that allows to highlight the importance of central hubs on the host side. By using the *Homo sapiens* and the *Pseudomonas aeruginosa PAO1* proteomes as input to our prediction tool, we generated 763 host-pathogen interactions showing a highly connected network topology. We expect that our prediction tool will allow to get a more realistic picture of host-pathogen interactomes and will help pave the way in the prioritization of PPIs that can be explored as potential targets for the development of new antibacterial drugs.

5.6. Availability

HPIPred tool is available under https://github.com/SysBioUAB/hpi_predictor. Datasets are available under https://zenodo.org/record/4668840#.YPWY3u0zaV4.

6. DISCUSSION

Antibiotic development has been one of the most important advances for the healthcare system in the treatment of infectious diseases caused by pathogens during the last century, having saved millions of lives since their discovery and widespread use. However, the natural ability of these pathogens to acquire resistance and the high selective pressure imposed by the use and misuse of antibiotics, is accelerating the proliferation of multi-drug resistant bacteria. This, in turn, is causing a dangerous increase in diseases produced by these bacteria, which are harder to treat, and are increasing the medical expenses and the hospitalization time derived from their treatments, as well as the number of deaths, as we face the antibiotic shortage in the fight against multi-drug resistance.

This situation reflects the urge to deviate from traditional drug development approaches, which are normally based on small molecules directed to bacterial targets that are considered essential for the development of the bacteria. In order to counter the rise of multi-drug resistance, it is necessary to develop new and innovative antibiotics that address new targets and use new modes of action.

In this respect, we need to redefine the concept of essentiality in the context of infection, and identify those genes that are essential for the pathogen during host infection. For a given bacterium, the *in vivo* fitness cost of deleting a single gene is correlated with the number of interactions with host proteins. Therefore, proteins with high impact on pathogen fitness during infection may cause extensive rewiring of the host interactome.

With the intention to offer an easy access to high-throughput bacterial fitness data from heterogeneous sources, we created BacFITBase, which we believe constitutes a valuable resource for the systematic classification and annotation of bacterial proteins relevant for host cell invasion and infection.

Similarly, we need to characterize the proteins involved in the host defense mechanisms activated upon bacterial infections. In this regard, the advancements in the field of high-throughput sequencing technologies have allowed the development of techniques such as dual RNA-Sequencing, which allows to measure the time course evolution of gene expression changes occurring during *in vitro* and *in vivo* infections in the pathogen and the host simultaneously. To facilitate accessibility to highly disseminated heterogeneous dual RNA-Seq data, we created DualSeqDB as a resource for the systematic identification of bacterial and host transcriptomic changes that can be crucial for the development of the infection process. We hope that BacFITBase and DualSeqDB will serve as powerful tools for the functional annotation of relevant proteins in the context of infection and will help unveil interspecies relationships between pathogen and host.

These observations show that infectious diseases can only be properly understood in the context of the complex interplay that occurs between the host and the pathogen proteins. Thus, it is of vital importance to identify and characterize PPIs between host and pathogens that can be used as targets for the development of novel antibacterial drugs.

In this regard, the use of PPI modulators is a proposed model for the development of new antibacterial drugs, as it may help reduce side effects associated with classical

antibacterial treatments, such as toxicity and damage of host microbiota. Such inhibitors would produce a bacteriostatic effect rather than a bactericidal one. It has been proposed that they could be used as prophylactics to prevent the spread and progression of MDR bacteria, in combination with classical antibiotics that show a synergistic effect, or as a last resource against organisms that are resistant to all available antibiotics. Nevertheless, due to the limitation of experimental techniques to correctly identify host-pathogen PPIs, it is essential to find alternatives that complement detection methods and allow to understand the functions of PPIs, so that they can later be investigated as possible drug targets.

Diverse computational strategies are nowadays used to predict host-pathogen PPIs, such as machine learning, homology-based methods or structure-based methods. However, these prediction methods generally suffer from high rates of false positive predictions as well as limited overlap with experimental validated interactions. In an attempt to reduce false positive predictions, we have developed HPIPred, a host-pathogen protein-protein interaction prediction tool that uses numerical encoding of amino acids based on their physicochemical properties and is able to integrate these predictions with phenotypic data from biological sources such as BacFITBase or DualSeqDB, among others. The integration of these phenotype data is used as a filtering criterion to reduce the number of putative predictions in a biologically meaningful way. By using the *Homo sapiens* and the *Pseudomonas aeruginosa PAO1* proteomes as input to our prediction tool, we generated a predicted host-pathogen of 763 interactions showing a highly connected network topology and a high degree of overlap in the predictions compared BIPS, an interolog prediction tool. We expect that HPIPred will help reduce the PPI candidates to

be tested in follow-up experiments and facilitate the identification of key PPIs that can be further explored as potential targets for PPI modulators.

Finally, even though it is beyond of the scope of this thesis, it is important to highlight how the combination of detection and prediction methods capable of prioritizing key host-pathogen PPIs, coupled with the recent development of protein structure prediction systems as accurate as AlphaFold (137) could entail a paradigm shift in the field of drug development, by speeding up the identification of PPI modulators that can be used as antibacterial drugs against multi-drug resistant bacteria.

7. CONCLUSIONS

In an era of antibiotic resistance, it is crucial to get a better understanding of the molecular mechanisms that control the infection process. To such end, we need to develop tools that allow us to deal with the limitations of the existing methods to detect PPIs in order to accelerate the screening process of drug discovery against resistant bacteria.

For a given bacterium, the *in vivo* fitness cost of deleting a single gene is correlated with the number of interactions with host proteins. Therefore, proteins with high impact on pathogen fitness during infection may cause extensive rewiring of the host interactome. These observations indicate that infectious diseases are only properly understood in the context of the host–pathogen interactions. We believe that the creation of BacFITBase will constitute a valuable resource to systematically classify bacterial proteins relevant for host cell invasion and infection. In the same line, we created DualSeqDB to systematically identify proteins that are crucial for successful infection, with the aim to understand how the bacterial and host transcriptomes change and interact during infection.

As discussed, infectious diseases can only be properly understood in the context of the complex interplay that occurs between the host and the pathogen proteins. Thus, it is of vital importance to identify and characterize PPIs between host and pathogens. Numerous experimental techniques, such as yeast two-hybrid, pulldown assays or co-immunoprecipitation, are currently used for the detection of these interactions, which are collected in existing databases through literature mining or manual curation. However, experimental techniques are time-consuming, costly and suffer from low-specificity, making it unfeasible to evaluate all possible protein-protein interactions. In this context, computational approaches such as machine learning, homology-based methods or

structure-based methods, are allowing the prediction of putative PPIs, in turn helping to complement experimental techniques. However, they are limited by the high rate of false positive that they generate, mainly due to the lack of robust databases of experimentally validated PPIs.

With the intention of addressing this challenge, we developed HPIPred, a tool that combines host-pathogen protein-protein interaction prediction with phenotype data from biological sources such as BacFITBase or DualSeqDB. The integration of these phenotype data is used as a filtering criterion to reduce the number of putative predictions in a biologically meaningful way. We hope that HPIPred predictions tool will allow to get a more realistic picture of host-pathogen interactomes and will help pave the way in the prioritization of PPIs that can be explored as potential targets for the development of new antibacterial drugs.

8. PUBLICATIONS FROM THIS THESIS

The results presented in “Chapter 1. BacFITBase: a database to assess the relevance of bacterial genes during host infection” have been published in Nucleic Acids Research (118).

The results presented in “Chapter 2. DualSeqDB: The host-pathogen dual RNA sequencing database for infection processes” have been published in Nucleic Acids Research (128).

9. REFERENCES

1. Peterson JW. Bacterial Pathogenesis. In: Baron S, editor. Medical Microbiology [Internet]. 4th ed. Galveston (TX): University of Texas Medical Branch at Galveston; 1996 [cited 2021 Aug 15]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK8526/

2. Xiao Y, Cai W. Autophagy and Bacterial Infection. Adv Exp Med Biol. 2020;1207:413–23.

3. Bench-to-bedside review: Bacterial virulence and subversion of host defences [Internet]. [cited 2021 Aug 15]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2646333/

4. Lee VT, Schneewind O. Protein secretion and the pathogenesis of bacterial infections. Genes Dev. 2001 Jul 15;15(14):1725–52.

5. Green ER, Mecsas J. Bacterial Secretion Systems: An Overview. Microbiol Spectr. 2016 Feb;4(1).

6. Kline KA, Fälker S, Dahlberg S, Normark S, Henriques-Normark B. Bacterial adhesins in host-microbe interactions. Cell Host Microbe. 2009 Jun 18;5(6):580–92.

7. Roberts IS. The biochemistry and genetics of capsular polysaccharide production in bacteria. Annu Rev Microbiol. 1996;50:285–315.

8. Komander D, Rape M. The ubiquitin code. Annu Rev Biochem. 2012;81:203–29.

9. Lamkanfi M, Dixit VM. Manipulation of host cell death pathways during microbial infections. Cell Host Microbe. 2010 Jul 22;8(1):44–54.

10. Ham H, Sreelatha A, Orth K. Manipulation of host membranes by bacterial effectors. Nat Rev Microbiol. 2011 Jul 18;9(9):635–46.

11. Goldberg MB. Actin-Based Motility of Intracellular Microbial Pathogens. Microbiol Mol Biol Rev. 2001 Dec;65(4):595–626.

12. Adedeji WA. THE TREASURE CALLED ANTIBIOTICS. Ann Ib Postgrad Med. 2016 Dec;14(2):56–7.

13. National Research Council (US) Committee on New Directions in the Study of Antimicrobial Therapeutics: New Classes of Antimicrobials, National Research Council (US) Committee on New Directions in the Study of Antimicrobial Therapeutics: Immunomodulation. Treating Infectious Diseases in a Microbial World: Report of Two Workshops on Novel Antimicrobial Therapeutics [Internet]. Washington (DC): National Academies Press (US); 2006 [cited 2021 Aug 15]. (The National Academies Collection: Reports funded by National Institutes of Health). Available from: http://www.ncbi.nlm.nih.gov/books/NBK19849/

14. Armstrong GL, Conn LA, Pinner RW. Trends in infectious disease mortality in the United States during the 20th century. JAMA. 1999 Jan 6;281(1):61–6.

15. The ongoing challenge of latent tuberculosis - PubMed [Internet]. [cited 2021 Aug 15]. Available from: https://pubmed.ncbi.nlm.nih.gov/24821923/

16. Origins and Evolution of Antibiotic Resistance [Internet]. [cited 2021 Aug 15]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2937522/

17. Blair JMA, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJV. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol. 2015 Jan;13(1):42–51.

18. Nikaido H. Multidrug resistance in bacteria. Annu Rev Biochem. 2009;78:119–46.

19. Asokan GV, Ramadhan T, Ahmed E, Sanad H. WHO Global Priority Pathogens List: A Bibliometric Analysis of Medline-PubMed for Knowledge Mobilization to Infection Prevention and Control Practices in Bahrain. Oman Med J. 2019 May;34(3):184–93.

20. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. Lancet Infect Dis. 2019;19(1):56–66.

21. Nelson RE, Hatfield KM, Wolford H, Samore MH, Scott RD II, Reddy SC, et al. National Estimates of Healthcare Costs Associated With Multidrug-Resistant Bacterial Infections Among Hospitalized Patients in the United States. Clin Infect Dis. 2021 Jan 15;72(Supplement_1):S17–26.

22. World Health Organization. Antimicrobial resistance: global report on surveillance [Internet]. World Health Organization; 2014 [cited 2021 Aug 15]. xxii, 232 p. Available from: https://apps.who.int/iris/handle/10665/112642

23. Wright PM, Seiple IB, Myers AG. The evolving role of chemical synthesis in antibacterial drug discovery. Angew Chem Int Ed Engl. 2014 Aug 18;53(34):8840–69.

24. Ventola CL. The antibiotic resistance crisis: part 2: management strategies and new agents. P T Peer-Rev J Formul Manag. 2015 May;40(5):344–52.

25. Spellberg B. The future of antibiotics. Crit Care Lond Engl. 2014 Jun 27;18(3):228.

26. Interplay of physics and evolution in the likely origin of protein biochemical function | PNAS [Internet]. [cited 2021 Aug 15]. Available from: https://www.pnas.org/content/110/23/9344

27. Comprehensive Classification and Diversity Assessment of Atomic Contacts in Protein–Small Ligand Interactions | Journal of Chemical Information and Modeling [Internet]. [cited 2021 Aug 15]. Available from: https://pubs.acs.org/doi/abs/10.1021/ci300377f

28. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nat Rev Drug Discov. 2007 Jan;6(1):29–40.

29. Zoraghi R, Reiner NE. Protein interaction networks as starting points to identify novel antimicrobial drug targets. Curr Opin Microbiol. 2013 Oct;16(5):566–72.

30. Schweppe DK, Harding C, Chavez JD, Wu X, Ramage E, Singh PK, et al. Host-microbe protein interactions during bacterial infection. Chem Biol. 2015 Nov 19;22(11):1521–30.

31. Titeca K, Lemmens I, Tavernier J, Eyckerman S. Discovering cellular protein-protein interactions: Technological strategies and opportunities. Mass Spectrom Rev. 2019 Jan;38(1):79–111.

32. Seychell BC, Beck T. Molecular basis for protein-protein interactions. Beilstein J Org Chem. 2021;17:1–10.

33. Muronetz VI, Sholukh M, Korpela T. Use of protein-protein interactions in affinity chromatography. J Biochem Biophys Methods. 2001 Oct 30;49(1–3):29–47.

34. Interaction proteomics: characterization of protein complexes using tandem affinity purification-mass spectrometry - PubMed [Internet]. [cited 2021 Aug 15]. Available from: https://pubmed.ncbi.nlm.nih.gov/20658971/

35. Lin J-S, Lai E-M. Protein–Protein Interactions: Co-Immunoprecipitation. In: Journet L, Cascales E, editors. Bacterial Protein Secretion Systems: Methods and Protocols [Internet]. New York, NY: Springer; 2017 [cited 2021 Aug 15]. p. 211–9. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-7033-9_17

36. Du Z, Lee JK, Fenn S, Tjhen R, Stroud RM, James TL. X-ray crystallographic and NMR studies of protein–protein and protein–nucleic acid interactions involving the KH domains from human poly(C)-binding protein-2. RNA. 2007 Jan 7;13(7):1043–51.

37. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast two-hybrid, a powerful tool for systems biology. Int J Mol Sci. 2009 Jun 18;10(6):2763–88.

38. Fields S. High-throughput two-hybrid analysis. The promise and the peril. FEBS J. 2005 Nov;272(21):5391–9.

39. Legrain P, Selig L. Genome-wide protein interaction maps using two-hybrid systems. FEBS Lett. 2000 Aug 25;480(1):32–6.

40. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021 Jan 8;49(D1):D545–51.

41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.

42. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci Publ Protein Soc. 2019 Nov;28(11):1947–51.

43. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020 Jan 8;48(D1):D498–503.

44. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D535–9.

45. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D358-63.

46. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009 Jan;37(Database issue):D412-416.

47. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010 Jul 1;38(suppl_2):W214–20.

48. Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res. 2020 Jan 8;48(D1):D606–12.

49. Ammari MG, Gresham CR, McCarthy FM, Nanduri B. HPIDB 2.0: a curated database for host–pathogen interactions. Database [Internet]. 2016 Jan 1 [cited 2021 Aug 15];2016(baw103). Available from: https://doi.org/10.1093/database/baw103

50. Durmuş Tekir S, Çakır T, Ardıç E, Sayılırbaş AS, Konuk G, Konuk M, et al. PHISTO: pathogen–host interaction search tool. Bioinformatics. 2013 May 15;29(10):1357–8.

51. Urban M, Cuzick A, Seager J, Wood V, Rutherford K, Venkatesh SY, et al. PHI-base: the pathogen–host interactions database. Nucleic Acids Res. 2020 Jan 8;48(D1):D613–20.

52. Crua Asensio N, Munõz Giner E, De Groot NS, Torrent Burgas M. Centrality in the host-pathogen interactome is associated with pathogen fitness during infection. Nat Commun. 2017;8.

53. Cossar PJ, Lewis PJ, McCluskey A. Protein-protein interactions as antibiotic targets: A medicinal chemistry perspective. Med Res Rev. 2020;40(2):469–94.

54. Zinzalla G, Thurston DE. Targeting protein-protein interactions for therapeutic intervention: a challenge for the future. Future Med Chem. 2009 Apr;1(1):65–93.

55. Arkin MR, Wells JA. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. Nat Rev Drug Discov. 2004 Apr;3(4):301–17.

56. Smith MC, Gestwicki JE. Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity. Expert Rev Mol Med. 2012 Jul 26;14:e16.

57. Higueruelo AP, Jubb H, Blundell TL. Protein-protein interactions as druggable targets: recent technological advances. Curr Opin Pharmacol. 2013 Oct;13(5):791–6.

58. Aloy P, Russell RB. Ten thousand interactions for the molecular biologist. Nat Biotechnol. 2004 Oct;22(10):1317–21.

59. Nevola L, Giralt E. Modulating protein-protein interactions: the potential of peptides. Chem Commun Camb Engl. 2015 Feb 25;51(16):3302–15.

60. Zahiri J, Bozorgmehr JH, Masoudi-Nejad A. Computational Prediction of Protein-Protein Interaction Networks: Algo-rithms and Resources. Curr Genomics. 2013 Sep;14(6):397–414.

61. Murakami Y, Mizuguchi K. Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators. BMC Bioinformatics. 2014 Jun 23;15(1):213.

62. Huang Y-A, You Z-H, Chen X, Chan K, Luo X. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. BMC Bioinformatics. 2016 Apr 26;17(1):184.

63. Zaki N. Protein-Protein Interaction Prediction Using Homology and Inter-domain Linker Region Information. In: Ao S-I, Gelman L, editors. Advances in Electrical Engineering and Computational Science [Internet]. Dordrecht: Springer Netherlands; 2009 [cited 2021 Aug 15]. p. 635–45. (Lecture Notes in Electrical Engineering). Available from: https://doi.org/10.1007/978-90-481-2311-7_54

64. Predicting protein–protein interactions through sequence-based deep learning | Bioinformatics | Oxford Academic [Internet]. [cited 2021 Aug 15]. Available from: https://academic.oup.com/bioinformatics/article/34/17/i802/5093239

65. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008 Apr;18(4):644–52.

66. Akiva E, Friedlander G, Itzhaki Z, Margalit H. A Dynamic View of Domain-Motif Interactions. PLOS Comput Biol. 2012 Jan 12;8(1):e1002341.

67. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-protein interaction detection: methods and analysis. Int J Proteomics. 2014;2014:147648.

68. Davis FP, Barkan DT, Eswar N, McKerrow JH, Sali A. Host pathogen protein interactions predicted by comparative modeling. Protein Sci Publ Protein Soc. 2007 Dec;16(12):2585–96.

69. Chen X, Wang M, Zhang H. The use of classification trees for bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov. 2011;1(1):55–63.

70. Chen X-W, Liu M. Prediction of protein-protein interactions using random decision forest framework. Bioinforma Oxf Engl. 2005 Dec 15;21(24):4394–400.

71. Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell. 2014 Nov 20;159(5):1212–26.

72. Ahmed H, Howton TC, Sun Y, Weinberger N, Belkhadir Y, Mukhtar MS. Network biology discovers pathogen contact points in host protein-protein interactomes. Nat Commun. 2018 Jun 13;9(1):2312.

73. Pan A, Lahiri C, Rajendiran A, Shanmugham B. Computational analysis of protein interaction networks for infectious diseases. Brief Bioinform. 2016 May;17(3):517–26.

74. Van Opijnen T, Camilli A. Transposon insertion sequencing: A new tool for systems-level analysis of microorganisms. Nat Rev Microbiol. 2013;11(7):435–42.

75. Chao MC, Abel S, Davis BM, Waldor MK. The design and analysis of transposon insertion sequencing experiments. Nat Rev Microbiol. 2016;14(2):119–28.

76. Chaudhuri RR, Morgan E, Peters SE, Pleasance SJ, Hudson DL, Davies HM, et al. Comprehensive Assignment of Roles for Salmonella Typhimurium Genes in Intestinal Colonization of Food-Producing Animals. PLoS Genet. 2013 Apr;9(4).

77. Gawronski JD, Wong SMS, Giannoukos G, Ward D V., Akerley BJ. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. Proc Natl Acad Sci. 2009 Sep;106(38):16422–7.

78. Eckert SE, Dziva F, Chaudhuri RR, Langridge GC, Turner DJ, Pickard DJ, et al. Retrospective application of transposon-directed insertion site sequencing to a library of signature-tagged mini-Tn5Km2 mutants of Escherichia coli O157:H7 screened in cattle. J Bacteriol. 2011 Apr;193(7):1771–6.

79. Fu Y, Waldor MK, Mekalanos JJ. Tn-seq analysis of vibrio cholerae intestinal colonization reveals a role for T6SS-mediated antibacterial activity in the host. Cell Host Microbe. 2013 Dec;14(6):652–63.

80. Gao B, Vorwerk H, Huber C, Lara-Tejero M, Mohr J, Goodman AL, et al. Metabolic and fitness determinants for in vitro growth and intestinal colonization of the bacterial pathogen Campylobacter jejuni. PLoS Biol. 2017 May;15(5).

81. Bachman MA, Breen P, Deornellas V, Mu Q, Zhao L, Wu W, et al. Genome-wide identification of Klebsiella pneumoniae fitness genes during lung infection. mBio. 2015 Jun;6(3).

82. Wang N, Ozer EA, Mandel MJ, Hauser AR. Genome-wide identification of Acinetobacter baumannii genes necessary for persistence in the lung. mBio. 2014 Jun;5(3).

83. Grant AJ, Oshota O, Chaudhuri RR, Mayho M, Peters SE, Clare S, et al. Genes required for the fitness of Salmonella enterica serovar Typhimurium during infection of immunodeficient gp91-/- phox mice. Infect Immun. 2016 Apr;84(4):989–97.

84. Anderson MT, Mitchell LA, Zhao L, Mobley HLT, Welch R. Capsule Production and Glucose Metabolism Dictate Fitness during Serratia marcescens Bacteremia Downloaded from.

85. Hubbard TP, Chao MC, Abel S, Blondel CJ, Abel zur Wiesch P, Zhou X, et al. Genetic analysis of Vibrio parahaemolyticus intestinal colonization . Proc Natl Acad Sci. 2016 May;113(22):6283–8.

86. Le Breton Y, Belew AT, Freiberg JA, Sundar GS, Islam E, Lieberman J, et al. Genome-wide discovery of novel M1T1 group A streptococcal determinants important for fitness and virulence during soft-tissue infection. PLoS Pathog. 2017 Aug;13(8).

87. Miller DP, Hutcherson JA, Wang Y, Nowakowska ZM, Potempa J, Yoder-Himes DR, et al. Genes Contributing to Porphyromonas gingivalis Fitness in Abscess and Epithelial Cell Colonization Environments. Front Cell Infect Microbiol. 2017 Aug;7.

88. Subashchandrabose S, Smith SN, Spurbeck RR, Kole MM, Mobley HLT. Genome-Wide Detection of Fitness Genes in Uropathogenic Escherichia coli during Systemic Infection. PLoS Pathog. 2013;9(12):1–15.

89. Wang J, Pritchard JR, Kreitmann L, Montpetit A, Behr MA. Disruption of Mycobacterium avium subsp. paratuberculosis-specific genes impairs in vivo fitness. BMC Genomics. 2014 May;15(1).

90. Olson MA, Siebach TW, Griffitts JS, Wilson E, Erickson DL. Genome-wide identification of fitness factors in mastitisassociated Escherichia coli. Appl Environ Microbiol. 2018 Jan;84(2).

91. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.

92. Jehl P, Manguy J, Shields DC, Higgins DG, Davey NE. ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. Nucleic Acids Res. 2016;44(W1):W11–5.

93. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–32.

94. Riley M, Abe T, Arnaud MB, Berlyn MKB, Blattner FR, Chaudhuri RR, et al. Escherichia coli K-12: A cooperatively developed annotation snapshot - 2005. Nucleic Acids Res. 2006;34(1):1–9.

95. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

96. Collignon P, Beggs JJ, Walsh TR, Gandra S, Laxminarayan R. Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis. Lancet Planet Health. 2018;2(9):e398–405.

97. Dual RNA-seq of pathogen and host | Nature Reviews Microbiology [Internet]. [cited 2021 Aug 15]. Available from: https://www.nature.com/articles/nrmicro2852

98. Aprianto R, Slager J, Holsappel S, Veening J-W. Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection. Genome Biol. 2016 Sep 27;17(1):198.

99. Nuss AM, Beckstette M, Pimenova M, Schmühl C, Opitz W, Pisano F, et al. Tissue dual RNA-seq allows fast discovery of infection-specific functions and riboregulators shaping host–pathogen transcriptomes. Proc Natl Acad Sci. 2017 Jan 31;114(5):E791–800.

100. Westermann AJ, Förstner KU, Amman F, Barquist L, Chao Y, Schulte LN, et al. Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. Nature. 2016 Jan 28;529(7587):496–501.

101. Westermann AJ, Barquist L, Vogel J. Resolving host-pathogen interactions by dual RNA-seq. PLoS Pathog. 2017 Feb;13(2):e1006033.

102. Saliba A-E, C Santos S, Vogel J. New RNA-seq approaches for the study of bacterial pathogens. Curr Opin Microbiol. 2017 Feb;35:78–87.

103. Dual RNA-seq of Nontypeable Haemophilus influenzae and Host Cell Transcriptomes Reveals Novel Insights into Host-Pathogen Cross Talk | mBio [Internet]. [cited 2021 Aug 15]. Available from: https://journals.asm.org/doi/full/10.1128/mBio.01765-15

104. Choi Y-J, Aliota MT, Mayhew GF, Erickson SM, Christensen BM. Dual RNA-seq of parasite and host reveals gene expression dynamics during filarial worm-mosquito interactions. PLoS Negl Trop Dis. 2014 May;8(5):e2905.

105. Kachroo P, Eraso JM, Olsen RJ, Zhu L, Kubiak SL, Pruitt L, et al. New Pathogenesis Mechanisms and Translational Leads Identified by Multidimensional Analysis of Necrotizing Myositis in Primates. mBio. 2020 Feb 18;11(1):e03363-19.

106. Schulte LN, Schweinlin M, Westermann AJ, Janga H, Santos SC, Appenzeller S, et al. An Advanced Human Intestinal Coculture Model Reveals Compartmentalized Host and Pathogen Strategies during Salmonella Infection. mBio. 2020 Feb 18;11(1):e03348-19.

107. Damron FH, Oglesby-Sherrouse AG, Wilks A, Barbier M. Dual-seq transcriptomics reveals the battle for iron during Pseudomonas aeruginosa acute murine pneumonia. Sci Rep. 2016 Dec 16;6:39172.

108. Griesenauer B, Tran TM, Fortney KR, Janowicz DM, Johnson P, Gao H, et al. Determination of an Interaction Network between an Extracellular Bacterial Pathogen and the Human Host. mBio. 2019 Jun 18;10(3):e01193-19.

109. Rienksma RA, Suarez-Diez M, Mollenkopf H-J, Dolganov GM, Dorhoi A, Schoolnik GK, et al. Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. BMC Genomics. 2015 Feb 5;16:34.

110. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-745.

111. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

112. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019 Aug;37(8):907–15.

113. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014 Aug 1;30(15):2114–20.

114. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019 May 7;47(8):e47.

115. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

116. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016 Jan 4;44(D1):D7-19.

117. de Groot NS, Torrent Burgas M. A Coordinated Response at The Transcriptome and Interactome Level is Required to Ensure Uropathogenic Escherichia coli Survival during Bacteremia. Microorganisms. 2019 Sep;7(9):292.

118. Rendón JM, Lang B, Tartaglia GG, Burgas MT. BacFITBase: a database to assess the relevance of bacterial genes during host infection. Nucleic Acids Res. 2020 Jan 8;48(D1):D511–6.

119. Wolf T, Kämmer P, Brunke S, Linde J. Two's company: studying interspecies relationships with dual RNA-seq. Curr Opin Microbiol. 2018 Apr;42:7–12.

120. OECD. Bacteria: Pathogenicity factors. 2016 Apr 5;27–79.

121. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. The front line of host defense. Immunobiol Immune Syst Health Dis 5th Ed [Internet]. 2001 [cited 2021 Aug 15]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK27105/

122. Louche A, Salcedo SP, Bigot S. Protein–Protein Interactions: Pull-Down Assays. In: Journet L, Cascales E, editors. Bacterial Protein Secretion Systems: Methods and Protocols [Internet]. New York, NY: Springer; 2017 [cited 2021 Aug 22]. p. 247–55. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-7033-9_20

123. Tiessen A, Pérez-Rodríguez P, Delaye-Arredondo LJ. Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes. BMC Res Notes. 2012 Feb 1;5(1):85.

124. Nakai K, Kidera A, Kanehisa M. Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng. 1988 Jul;2(2):93–100.

125. Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 1996 Jan;9(1):27–36.

126. Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. Nucleic Acids Res. 1999 Jan 1;27(1):368–9.

127. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 2000 Jan 1;28(1):374.

128. Macho Rendón J, Lang B, Ramos Llorens M, Gaetano Tartaglia G, Torrent Burgas M. DualSeqDB: the host–pathogen dual RNA sequencing database for infection processes. Nucleic Acids Res. 2021 Jan 8;49(D1):D687–93.

129. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May;411(6833):41–2.

130. Joy MP, Brock A, Ingber DE, Huang S. High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol. 2005 Jun 30;2005(2):96–103.

131. Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, Hennig H, Wolkenhauer O, Mirzaie M, et al. A systematic survey of centrality measures for protein-protein interaction networks. BMC Syst Biol. 2018 Jul 31;12(1):80.

132. Csardi G, Nepusz T. The Igraph Software Package for Complex Network Research. InterJournal. 2005 Nov 30;Complex Systems:1695.

133. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504.

134. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.

135. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan;37(1):1–13.

136. Garcia-Garcia J, Schleker S, Klein-Seetharaman J, Oliva B. BIPS: BIANA Interolog Prediction Server. A tool for protein–protein interaction inference. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W147–51.

137. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Jul 15;1–11.