Academic Year 2017-2018

Faculty Pharmaceutical, Biomedical and Veterinary Sciences

Biomedical Sciences

Exploring bacterial HPI networks using subgraph mining

By:

Lorenz Van de Veken

Project proposal in partial fulfillment of the requirements for the degree

Master in Biomedical Sciences

Promotor(en): Kris Laukens

Copromotor: Pieter Meysman

Coach: Pieter Moris

Campus Middelheim, Middelheimlaan 1, 2020 Antwerpen, Belgium

1. Abstract

Bacteria and humans are in a constant battle with each other. Bacteria try to take over niches in the human body to survive, manipulating host machinery, evading the immune system and potentially damaging the host tissue during the process, with adverse effects for the human host as a result. The key interface on which this battle takes place is the interactome between the host and the pathogen. Many human and pathogen proteins interact with one another, thus host-pathogen protein-protein interactions (PPIs) and the networks they form are an important tool to study bacterial infections. Due to recent advances in the biotechnology, high-throughput screening of these interactions can be done, resulting in a massive increase in interaction data. We intend to collect this data and analyse the host-pathogen protein-protein interaction (HPI) networks of human bacterial pathogens with several techniques. Specifically, we will use subgraph mining algorithms in order to find interesting patterns in these networks that may be linked with several characteristics of the bacteria, the course of infection and the disease progression.

2. Research question

Our hypothesis states that there exist enriched patterns within HPI networks of bacterial groups. This hypothesis covers a very broad and open. It is aimed at finding interesting features of bacterial HPI networks and describe them, rather than finding one particular type of interesting feature. Thus, this could be considered as a more hypothesis-generating study, that may lead to many, more specific, questions regarding the obtained interesting features.

We expect these features to be linked to several aspects of the bacterial life cycle and infection strategy such as immune evasion/suppression, host entry, dissemination, establishment of infection, rewiring of the host metabolism, etc.

Previous research has already indicated that bacteria have common infection strategies. They often target protein hubs and bottlenecks (Dyer et al., 2010). Furthermore, the targeted human proteins are often part of the immune system which leads to the hypothesis that infection strategy-linked patterns can be found in these HPI networks as immune evasion and/or suppression form a common survival strategy.

Our study has three major research goals:

First, we will use classical descriptive analytical methods to describe the general features of our dataset. Gene ontology (GO) enrichment and pathway enrichment analyses will also be done as they can provide a general contextual overview of our dataset.

Secondly, we will perform frequent subgraph mining on the entire dataset containing HPI networks of multiple bacterial groups in order to find general patterns across all species. The same technique can also be used on subsets of our dataset containing the HPI network of only one species to find patterns which occur frequently for that specific species.

Thirdly, mining for frequent subgraphs associated with selected interacting proteins of interest can be performed on HPI networks which also contain intraspecies and host-host interactions with the aim to discover extended patterns which potentially provide additional information such as the context (i.e. surrounding interactions) in which the found HPI interactions take place (Meysman et al., 2016).

3. Material and methods

Workflow

Data collection will be done by querying several databases for HPIs and exporting the data via either a MiTab file (Deutsch et al., 2017), or a Microsoft Excel file.

The databases that will be assessed are PHISTO, HPIDB 2.0 and IntAct.

PHISTO is a well-known web-based HPI database for human pathogens. Thus, the interactions in this database are only human-pathogen interactions. Given that this database extracts data from a

number of external databases, the amount of interactions can give an estimation on which pathogens are generally well characterised with regard to human-pathogen interactions and thus could be of interest to investigate. However, querying just one database could pose a problem: The architecture of the database could be a confounding factor for example as it determines which data is extracted from which external database and thus which data is presented to the querying user. To minimise this and other possible biases, two more databases will be queried for HPIs (Durmus Tekir et al., 2013).
HPIDB 2.0 is a curated database which is part of the IMEx consortium (International Molecular Exchange consortium) and thus provides expertly curated HPIs with a standardised output format. HPIDB also extracts data from a set of external databases broadening our data collection scope as some of these databases are not used by PHISTO (Durmus Tekir et al., 2013, Ammari et al., 2016).
IntAct is a more general PPI interaction database containing HPIs as well as non-HPIs. It has a sophisticated web-based curation tool which is used by many curation teams. These teams directly curate data into the IntAct database which in return disseminates their data to a many external databases. This leads to IntAct being a common source for a lot of PPI databases and intuitively it can be stated that IntAct 'summarises' all the other databases.(Orchard et al., 2014) Combined, these three databases will provide data which is appropriate for further analysis and investigation. More information is provided in Section A of the Appendix.
For every pathogen, the HPI data needs to be extracted from the databases and compiled into one dataset. For this, we will use the Python programming language to create a data parsing workflow (Oliphant, 2007).
Quality control will be performed to ensure that there is no missing or invalid data. Interaction entries are often accompanied by a miscore, which is a confidence value. This can be used to ensure only high confidence interactions are included in the further analysis.

Every database has different entry formatting standards concerning validity and completeness of the interaction data that is extracted or curated. Thus, we need to ensure that all the data is converted to one type of format. This could pose some problems if certain databases deviate from the format we intend to use.
Redundant entries need to be deleted: Due to the independent creation of and curation into the databases, some entries have been added to multiple databases. Thus, after merging the data, some entries will be present in duplicate or even triplicate. Aside from the different entry formatting standards of each database, these entries are identical to each other and need to be reduced to just one.
Other redundant entries include entries of one interaction confirmed by multiple experiments. Although, the multiple experimental validations add to the scientific accuracy and confidence of these interactions and might be added as an extra independent confidence value to the entry, only one entry is required to be present in the compiled data frame. In both cases, the entry with the highest confidence is chosen to be included.

We will initially exclude host-host or pathogen-pathogen interactions as it is our goal to look for patterns containing pathogen-host interactions and therefore only include interspecies interactions. However, we might extend our approach in the future by also incorporating the closest intra-interactome interaction neighbours. These interactions can provide even more contextual information relevant to the enriched pattern and the HPI contained in this pattern. It may also provide useful insights in the effects downstream of the HPI or in the upstream events leading up to the HPI. It should be noted though that HPI networks have no temporal characteristic, thus determining a sequence of interactions (e.g. a pathway or reaction cascades) solely based on these networks is not possible. Inclusion of homologs (such as paralogs and orthologs) can prove to be useful but once again will not be included initially. Inclusion of homologs can really expand the network and is very useful for finding conserved enriched patterns, which eventually could be correlated to essential steps in the infection process and disease progression.

Annotations have to be added to every entry as our subgraph mining algorithm relies on vertex labels, i.e. annotations for the interacting proteins. These annotations will contain metadata such as gene ontology or structural and homology data of the interacting proteins.
The gene ontology terms are extracted from a file per species, containing all of the GO terms related to the proteins of that species. The standard use of the GO terms is to annotate the protein with the most specific term (du Plessis et al., 2011). However, as it is our goal to find more general patterns in the networks, the proteins will be remapped to more general parent terms to homogenize the information content. This will achieve a more useful information density and will facilitate the subgraph mining algorithm. InterPro can be accessed as well for annotations which contain information on protein domains and families (Finn et al., 2017).
Taxonomy IDs are mapped to their respective species/strain to ensure the entry is representative for the queried pathogen. This mapping is based on the NCBI taxonomy database (Benson et al., 2009, Sayers et al., 2009).
To tackle all this a self-written Python script will be used. The script is based on the pandas module, which provides all the needed data frame handling tools needed for this task (McKinney, 2010). Most of the data manipulation is done using strings and regular expressions.
After compiling the dataset, the diversity data of the dataset will be evaluated, described and visualised using the visualisation and network analysis tools available in cytoscape (Shannon et al., 2003). GO term enrichment and pathway enrichment analyses will be done via available Python packages, e.g. GOA tools (available at: https://github.com/tanghaibao/goatools).
Next, some descriptive graph analyses will be done. Descriptive parameters such as degree, network betweenness centrality and network diameter will be calculated for the different bacterial groups in order to describe the obtained networks.
Afterwards, the subgraph analysis will be done using the subgraph discovery algorithm as presented in (Meysman et al., 2016). These analysis techniques are discussed in more detail in section 5.
Considering several of our analyses aim to determine the enrichment of objects (GO terms, pathway terms, subgraphs) with certain features in our dataset, we determine the statistical significance of our results by performing a hypergeometric test with a Bonferroni correction for multiple testing. These tests have been done in previous studies who performed similar analyses (Meysman et al., 2016, Durmus and Ulgen, 2017).
A general summary of our workflow has been provided in figure 1.

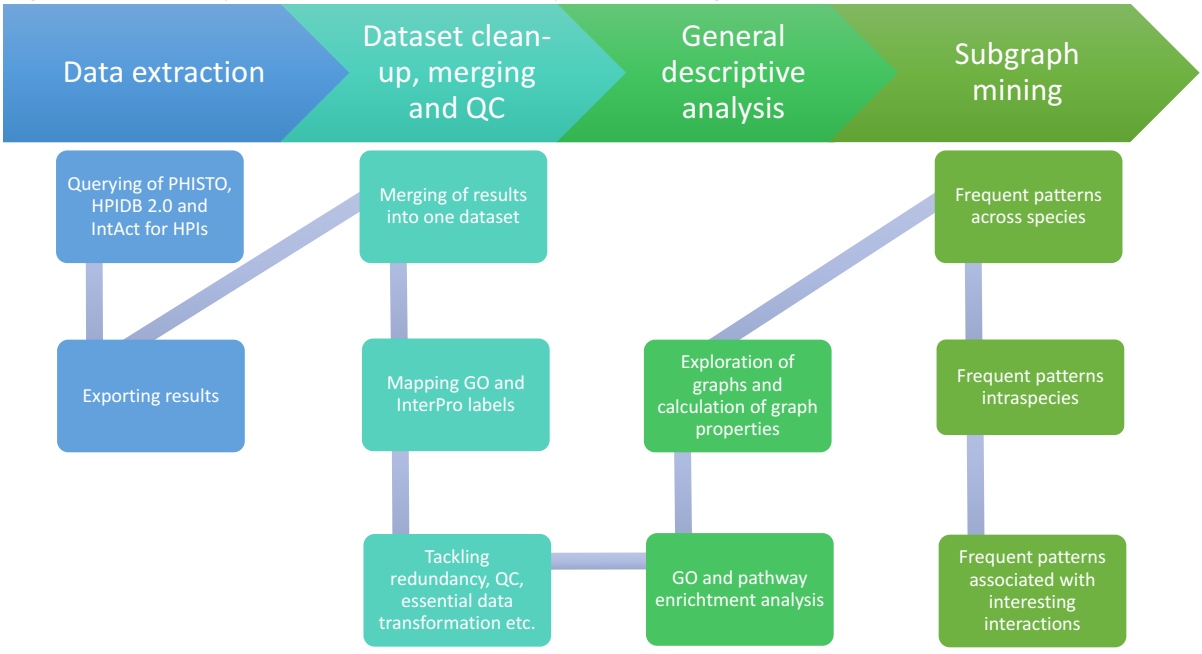

Figure 1: Visual representation of the workflow of the study
The four parts in the workflow are represented by the pointers in the header. A detailed step-by-step procedure is shown in the rectangles which have been ordered and color-coded to indicate during which part of the process they take place

4. Alternative research strategies

Our hypothesis contains a major exploratory aspect. This study aims mainly at observing the HPI networks of bacterial groups and describing their features and patterns, as either a descriptive analysis of currently available public interaction data or potentially as a stepping stone for further research by considering patterns of interest more in depth with regards to the underlying biology.
Analysing interactions as individual entities or in small interconnected clusters is often done and can provide detailed information on these interactions.
Network formation is an intrinsic property of interaction datasets. Therefore, intuitively, network analysis seems the only proper strategy to describe these datasets as a whole. Of course, there are other methods besides descriptive analytics and subgraph mining to tackle this.
One such a method is the use network clustering models. For example, one study analysed the interaction network of the TNF-α/ NFκB pathway and decomposed this into functional modules and protein complexes using minimum clique partitioning, an exclusive type I clustering algorithm. (More information on network clustering models is provided in Section 5.7) This method can easily be applied to HPI networks (Junker and Schreiber, 2008).

5. State-of-the-Art

5.1 Introduction

The research question entails two big fields of study, a biological and a computational one Considering this, a multidisciplinary state-of-the-art will be provided, touching upon biological aspects of infectious pathogens their strategies and PPIs with the human host as well as data sources, data annotation and data analysis strategies.

5.2 Infection strategies

As mentioned in section 3, the databases were queried to determine which pathogens are well-characterised with regard to HPIs. Three of them have substantially more HPI data available:
Yersinia pestis, Bacillus anthracis and *Francisella tularensis*. Therefore, the infection strategies of these three pathogens will be described in the following paragraphs, as these will become our primary subjects of analysis.
The inclusion of other pathogens in our study might be beneficial but further investigation would be needed to ensure that the data that is available can provide sufficient information for analysis.

Yersinia pestis

Yersinia pestis, the etiological agent of pneumonic, bubonic and septicaemic plague, is a facultative anaerobe, Gram-negative, non-motile, non-spore-forming coccobacillus. The bacillus is transmitted through infected flea bites, contact with contaminated fluid or tissues and through contact with infectious droplets. The latter one is the only way human-to-human transmission can occur and is rarely seen nowadays.
After initial contact, *Y. pestis* enters local macrophages which fail to kill the bacteria and migrate to local lymph nodes. The internal bacteria escape from the macrophage and proliferate extracellularly in these lymph nodes causing swelling and inflammation thus causing the nodes to swell, and form the so-called buboes, typical for the bubonic plague.
Some of the bacteria will also disseminate to the bloodstream and cause systemic disease. This phenotype is called the septic plague. If the bacterial infection initially takes place in the lungs, a necrotizing pneumonia can be induced, also known as pneumonic plague.
Once *Y. pestis* is localised to the regional lymph nodes, it evades the host's immune cells using the well-known type III secretion system (T3SS) to inject Yersinia effector proteins (Yops) into the cytosol of the host cell. These Yops can subvert cell signalling pathways, disrupt the metabolism and induce apoptosis, as well as block phagocytosis and inhibit cytokine production. By doing this, *Y.pestis* will

not only survive the encounter with the immune cells, but also undermines the immune response (Atkinson and Williams, 2016, Perry and Fetherston, 1997).

Bacillus anthracis

Bacillus anthracis is a Gram positive, aerobic, spore bearing bacillus and the causative agent of anthrax. Anthrax infection occurs after inhalation, ingestion or cutaneous contact with the endospores of the agent. These endospores do not go unnoticed by the innate immune system. Macrophages will phagocytose them and migrate to the local lymph nodes. Yet, the endospores are not killed in the phagosome. On the contrary, they germinate and become vegetative. Subsequently, they are released from the cells and enter the lymphatic system and then into the bloodstream to induce a systemic infection.

Despite the innate immune system responding to the initial contact with *B. anthracis*, most infections go unnoticed until the incubation period is over which is quickly followed by septic shock and sudden death. Thus *B. anthracis* must have some strategies to evade the host immune response.

Anthrax lethal toxin (LT) plays a key role in this strategy. This toxin is known to interact with macrophages and dendritic cells. LT interferes with the activation of macrophages and induce apoptosis, causing release of the phagocytosed bacteria as well as preventing propagation of anti-bacterial signals which normally are induced by macrophages. Dendritic cells on the other hand act as a very important mediator between the innate and adaptive immune system. LT causes abnormal maturation of dendritic cells. LT-exposed dendritic cells induce an anergy-like phenotype in CD4+ T cells impairing the development of proper humoral and cell-mediated immune responses to the anthrax infection (Fukao, 2004, Spencer, 2003).

Francisella tularensis

Francisella tularensis causes acute, often lethal, pneumonic disease. It consists of four subspecies, three of which are classified as type B and the remaining one as type A. The latter one is the most lethal for humans. Inhaling as few as 10 organisms is enough to cause acute, lethal pneumonic disease, making the organism a potentially very lethal biological weapon. Our focus will consequently be aimed at investigating this lethal type A strain. However, studies investigating *F .tularensis* have been using an attenuated type B strain, known as live vaccine strain (LVS) most of the time. Thus, it is possible that there is not enough data available to solely focus on the lethal type A strain. Nonetheless, the infection survival strategy of both types essentially entails suppression of the host's immune system. The difference between type A and B is that type A induces a broader and more intense suppression of multiple immune response pathways. One of the effects is the failure of dendritic cells and macrophages to undergo phenotypic activation and consequently the failure of mobilisation of effector cells. Another effect is the failure of the immune system to produce inflammatory cytokines. Type B on the other hand does activate dendritic cells and macrophages phenotypically, but still suppresses the production of inflammatory cytokines, which are also associated with maturation of dendritic cells and macrophages. Because type B induces weaker immune suppression, the immune system will recover itself more quickly, giving the bacteria less time to disseminate and proliferate freely, explaining why this type is less virulent than type A (Bosio et al., 2007, D'Elia et al., 2011, Celli and Zahrt, 2013).

5.3 Protein-protein interactions

Based on the previous paragraph, it is obvious that one of the key elements in the infection strategies of the pathogens is based on interaction between the biochemical machineries of the pathogen and the human host. The key players in these systems are proteins and thus protein-protein interactions are of utmost importance.

To study these interactions several experimental methods have been developed. In the following paragraph, two of the more commonly used techniques will be explained:

Y2H

The yeast two hybrid (Y2H) assay involves two physically separated units of a transcription factor. These units are a DNA-binding domain (BD) and a transcriptional activation domain (AD), both fused with one of the 2 candidate interacting proteins. If the candidate proteins come in close proximity to each other or physically interact, the AD and BD are able to function together and a reporter gene can be expressed. (Figure 2A)

The Y2H assay is one of the easiest and most straight-forward approaches one can take to study PPIs. It is carried out *in vivo* which helps to avoid complications and artefacts associated with cell lysis.

This assay, however, is prone to a high false-positive rate. A couple of reasons are:

- This assay is done using a yeast host, which means that the PPIs with other organisms may not be detectable due to poor expression or lack of necessary posttranslational modifications, cofactors, or other binding partners.
- The candidate proteins are coupled to two parts of a transcriptional factor which has to interact with the DNA to drive the transcription of a reporter gene. This means that the candidate proteins must have access to and interact in the nucleus to induce expression of the reporter gene. Compartment-confined proteins, therefore, cannot be studied in full-length form.

The readout is also indirect, which prevents spatial or temporal analysis of PPIs (Snider et al., 2015).

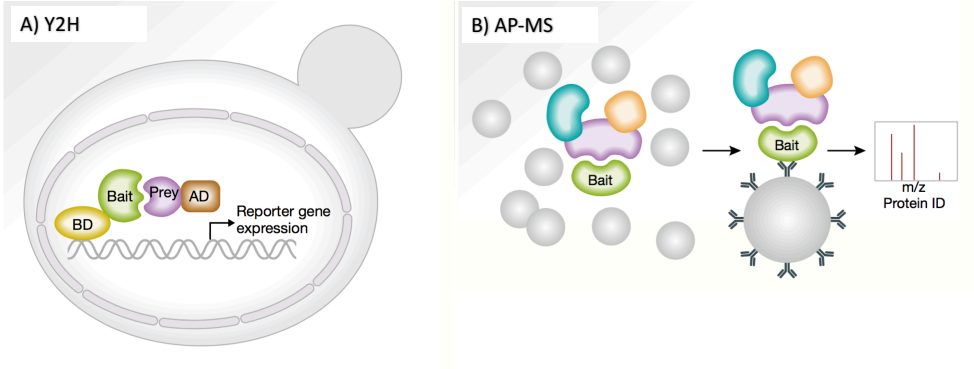

Figure 2: Y2H screen and AP-MS assay

Panel A: Yeast two-hybrid screening assay for protein interaction detection. A library of prey proteins fused with the activation domain (AD) of a yeast transcription factor is screened against a bait protein, which is fused with the binding domain (BD) of the same transcription factor. If the bait and prey proteins interact, the BD and AD can work together to induce the transcription of a reporter gene.

Panel B: Affinity purification – mass spectrometry assay. A bait protein is immobilised on a bead and subsequently used to screen a library of candidate proteins. After this purification step the formed complexes and thus the protein interactions can be identified using HPLC-MS.

AP-MS

Another commonly used detection method is affinity purification-mass spectrometry (AP-MS). A candidate protein is chosen as 'bait' and is immobilised on a solid support (agarose or magnetic beads are most often used). Then this 'bait' is used to capture other candidate target proteins from a soluble phase. After this affinity purification step, the captured proteins are usually digested with proteases. The generated peptides are then sub-fractionated using high-pressure liquid chromatography (HPLC). After the HPLC, an ionisation step is done and the ionised peptides are detected using a mass spectrometer. (Figure 2B)

This procedure can be done using endogenous, native proteins as bait. These native proteins can be immobilised using specific antibodies. Another method is tagging the bait proteins with standardised epitope tags (such as His-, protein A-, etc.) and then use antibodies against the epitope tag to immobilise the target.

The AP-MS assay using endogenous native baits has one big advantage: Proteins that are purified retain in their natural form. This eliminates problems associated with protein tagging. It also allows the screening of multiple isoforms.

On the other hand, AP-MS with standardised tags allow us to study proteins for which native antibodies are not available and also enables the analysis of different proteins using one single defined process with a specific antibody as many different proteins can be tagged with one single epitope.
Of course, there are some downsides to this method. One big limitation that the necessity of the cell lysis and affinity purification steps do not allow the detection of spatial and temporary PPIs.
Abundant proteins have the tendency to co-purify during the AP step. Artefacts can also occur as a result from the exposure of proteins to each other in a non-physiological environment (i.e. cell lysate). Epitope tags and ectopic expression may lead to improper folding. Ectopic expression also may have a disruptive effect on the interaction due to mislocalisation. These problems, however, can be overcome. Appropriate use of negative controls, further enrichment using tandem affinity purification, quantification approaches etc. can all help filtering out the true positives.
Data analysis of AP-MS data requires expertise with MS and specific bioinformatics tools for analysis as well as the need to address the limitations listed above (Snider et al., 2015).

False positives

All interaction detection methods are known to have one common problem, especially when used in a high throughput setting, as is often the case in interaction studies: False positives.
This is inherent to the assays that are being used, as explained specifically for the Y2H assay and AP-MS assay in the previous paragraphs. Several experimental and computational methods have been proposed and developed to assess this. The computational methods encompass strategies such as validating interacting protein characteristics (i.e. same subcellular location, similar functional and process annotations, shared GO terms, etc.) or the use of prediction methods to score the detected protein pairs as well as discovering undetected pairs (Snider et al., 2015).
Nevertheless, when doing interaction network analysis, this false positive nature of the data should be taken into account when analysing and discussing the results and drawing conclusions.

5.4 Databases

The obtained data from experimental assays gets curated into so-called interaction databases, such as PHISTO, HPIDB 2.0 and IntAct. Each of them has their own architectural structure, manual curation activity and external databases from which they extract their data.
The three queried databases have an easy-to-use interface with different functionalities, allowing advanced searches and sometimes even offer visualisation tools and basic statistic calculators.
Combined these databases provide data which is relevant, highly qualitative and representative for most, if not all, the PPI data that is available at the moment. More information on these databases is provided in Section A of the Appendix.
To achieve a higher information content, the core interaction data that is extracted from the interaction databases will be supplemented with metadata, GO and InterPro terms more specifically, to describe the interacting proteins.

5.5 Gene ontology

GO is a structured and controlled annotation vocabulary which can be used to describe different aspects of a gene or gene product functionality and the relation between different annotations. The annotations can be divided into three non-overlapping categories: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). The relations consist of ‘is_a’, ‘part_of’, ‘has_part’ and ‘regulates’ relationships. Some of these relationships are subdivided. All the relationships form edges of a Directed Acyclic Graph (DAG) while the annotation terms form the nodes.
Some examples of GO terms related to YPO2245, a protein from *Y. pestis,* are:
MF: electron transfer activity (GO:0009055)
BP: electron transport chain (GO:0022900)
CC: plasma membrane (GO:0005886)

Annotations originate from different sources. Like interaction data, annotations can be directly inferred from experimental evidence. These annotations are also considered to be the most reliable type of annotations. Other sources are computational methods, author statements, curator statement, automatically-assigned, etc. (du Plessis et al., 2011).

5.6 InterPro

InterPro contains information on several protein characteristics such as protein domains, homologous proteins, protein families. This information can be used to improve the information content of the vertex labels (Finn et al., 2017).
Some examples of InterPro terms related to the abovementioned *Y. pestis* protein, YPO2245, are:
Family: Electron transport complex, RnfB/RnxB (IPR010207)
Domain: 4Fe-4S domain (IPR0072024)

5.7 Biological network analysis

Interaction data can be integrated with each to form a biological interaction network. Studying proteins as individual entities provides a lot of information about their structure, physicochemical properties, function, etc. But to understand its role in a system such as a pathway or a biochemical process, it is almost mandatory to study the protein within the context of its interactions with other proteins. A big part of the functioning of a biological system is driven by proteins and their interactions. Therefore, this functional framework can be represented using protein-protein interaction networks. These networks enable the use of several analytical approaches to interpret the data they contain.

Protein function prediction using GO

As mentioned before, interaction data is often complemented by gene ontology data. This data allows us to describe the function of a protein within the context of the biological system it is involved in. Protein function prediction is a very important tool to expand our knowledge as manually determining protein functions is expensive and time-consuming. Therefore, computational strategies need to be developed to produce high-quality annotations. Protein interaction networks are a very solid basis as proteins often achieve their function by co-operating with other proteins. Strategies to determine protein functions based on PPI networks.
can roughly be classified into two categories: global network topology and local neighbourhood based approaches. Both types have their advantages and disadvantages and are often subject to debates within the bioinformatics community.
When using gene ontology in protein function predictions, it is very important to incorporate measures that represent the functional similarity of GO terms based on the structure of the GO DAG. One such a measure is semantic similarity. This measure can be used to propagate annotations from characterised proteins to uncharacterised proteins.

PHI networks

PPI networks can be used to study the interplay between a pathogen and its host during an infection. This infection is usually the result of interaction between the two, mostly via specific PPIs. Computational prediction of these interactions has become a big source of data. These prediction methods can be categorised into three categories: interologs based, domain & structure based and machine-learning based (Mulder et al., 2014).

Graph theory

These networks can also be represented as graphs. A graph is a network representation of relations (called edges) between elements (called nodes or vertices) of a certain group.
In our case, the interacting proteins and their interactions can be represented by vertices and edges, respectively.

Representing the data as graphs allows the use of several descriptive and analytical tools. In figure 3 an HPI network of *F. tularensis* is depicted.

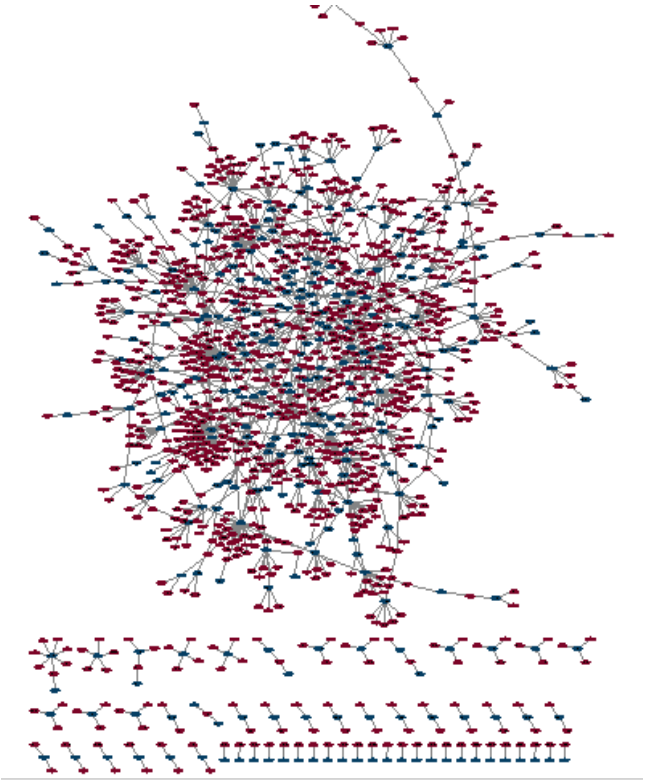

Figure 3: HPI network of *F. tularensis*
HPI network of *F.* tularensis with the human host, visualised using cytoscape. The data used to build the network was extracted from HPIDB 2.0. The red and blue vertices represent human and pathogen proteins, respectively.

A graph G is defined as a pair G(V,E) where V is a set of nodes and E is a set of edges, and $E \subseteq V \times V$. There are many types of graphs. Nodes and edges can be labelled for example. In that case, the graph is called a labelled graph. Edges in a graph can represent an ordered pair of nodes. If this is the case, the graph is said to be directed. If there is no edge orientation, then the graph is deemed undirected. Connected graphs are graphs where there is a path along the edges that links each pair of nodes. And lastly, the edges of a graph can be assigned numeric values. Then we call it is considered a weighted graph (Mrzic et al., In preparation).

Every graph is characterised by different global properties which can be described and computed using mathematical equations. Some commonly used properties are degree, distance, diameter, centrality and the clustering coefficient.

The degree k_i of a vertex n_i is defined as the number of edges adjacent to this vertex. When considering a graph without any self-loops or multiple links, k_i is equal to the number of neighbours. In a directed graph, we distinguish the incoming edges and outgoing edges using the input degree k_i^{in} and the output degree k_i^{out}, respectively.

The distance d_{ij} between any two vertices n_i and n_j defined as the shortest path between those vertices or in other words, the minimal number of edges that has to be traversed to go from one vertex to the other. When assessing distances in directed graphs: the distance from n_i to n_j may be different from the distance from n_j to n_i. The distance between two vertices can be infinite when there exists no traversable path to go from one vertex to the other. This can occur the in unconnected graphs as well as is directed graphs.

The diameter $d_m = max(d_{ij})$ of a graph is defined as the biggest distance present in the graph, i.e. the longest shortest path between two nodes. Centrality is a very important and commonly used measure which aims to describe a vertex or edge based on its position within the network. There are different types of centrality. The most well-known centrality is the betweenness centrality. This one measures how often a vertex or edge is present in the set of all the shortest paths of that network.

Lastly, the clustering coefficient is related to the local cohesiveness and measures the probability that two vertices with a common neighbour are connected. The clustering coefficient of a vertex is calculated as the ratio of the actual number of edges present between the neighbours of a given vertex and the maximal number of edges possible between the neighbours of that vertex. Note that this differs from the network clustering models mentioned in Section 4. These models will be discussed in more detail in the next paragraph (Junker and Schreiber, 2008).

Network clustering models

Network clustering models try to group the elements of a dataset in smaller sets of elements which are similar in some way.

Clustering models are classified based on two characteristics:

- Firstly, the relation of the clusters it produces is a very important property. The clusters can be disjoint or overlapping. This is also referred to as exclusive clustering and overlapping clustering, respectively.
- Secondly, the goal of the clustering is also used: Either the number of clusters is minimised or the cohesiveness or cluster similarity of each cluster is maximised. These characteristics are referred to as type I clustering and type II clustering, respectively.

The clustering model described in Section 4, called the minimum clique partioning algorithm, is an exclusive type I clustering method. This means that it will form disjoint clusters (=exclusive), minimising the number of clusters while ensuring that the formed clusters have a cohesiveness value above a certain threshold (=type I) (Junker and Schreiber, 2008).

Frequent subgraph mining

A graph G_s is considered a subgraph of graph G if the set of all nodes and the set of all edges of graph G_s are subsets of the set of all nodes and the set of all edges of graph G, respectively.

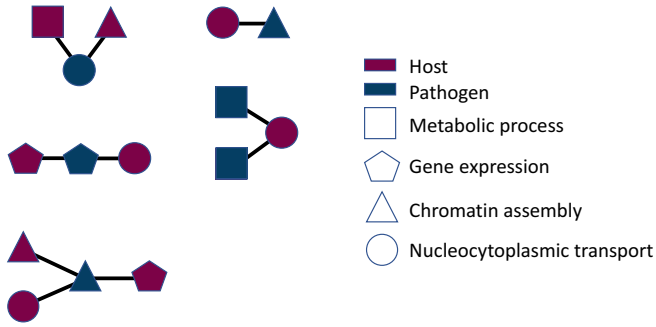

Figure 4: Subgraph examples
Some examples of potential outcomes of the subgraph mining algorithm. The color indicates whether the vertex represents a pathogen or a host protein and the shape of the vertex represents the label the vertex was given.

Frequent subgraph mining is used to find 'interesting' subgraphs. The most straightforward approach is to find subgraphs which occur more often than a given threshold. This number of occurrences is often called the 'support'. Depending on whether you are searching for frequent subgraphs through multiple graphs or in one single graph, the way the support is computed differs. Other approaches to determine whether a subgraph is interesting exist but the concept of counting the number of occurrences is the most commonly used one and often utilised as the first step in many subgraph mining procedures (Mrzic et al., In preparation).

Figure 4 depicts some examples of subgraphs which could be found in our HPI networks as a result of the subgraph mining algorithm.

Most subgraph mining algorithms search for a subgraph pattern which occurs more frequently in a graph dataset than a given support threshold. However, often the interesting graphs are not the ones who are simply frequently occurring but are most often associated with a specific set of vertices. In a biological context, the most interesting subgraphs would be those that could be associated with vertices linked to pathogenesis, for example. It requires a different, more specific algorithm to discover these frequent subgraphs associated with selected vertices. Meysman et al. have developed such a "subgroup discovery algorithm to find subgraphs in a graph that are associated with a given set of vertices." (Meysman et al., 2016).

Appendix

A. Databases

PHISTO purely contains HPIs of human pathogens and extracts data from several valuable external databases (APID, IntAct, DIP, MINT, iRefIndex, STRING, MPIDB, BIND, and Reactome). Even though it doesn't use manual curation to expand their data, it does have a text mining module which is used extract and assign experimental methods to HPIs, who are lacking this data
An easy-to-use user interface assists in executing various types of searches, browse the database and performing certain data analyses.
As of 12/06/2017, it contains 8894 HPIs between 3689 human proteins and 2675 pathogen proteins covering 59 pathogenic strains (Durmus Tekir et al., 2013).

HPIDB 2.0 is a curated database which is part of the IMEx consortium and conforms to the IMEx consortium standards for biocuration. The most abundant host species in the database is the human host, thus a sizeable number of qualitatively annotated HPI data of human pathogens should be available for extraction. HPIDB also extracts data from a set of external databases.
As of 14/03/2017, it contains 55,505 unique protein interactions between 55 host and 523 pathogen species. However, 97% of the PPI in the database are human-pathogen interactions and 16% of all the PPI are bacteria-host interactions (Ammari et al., 2016).

IntAct is a more general PPI interaction database containing HPIs as well as non-HPIs. It is also part of the IMEx consortium. Even more so, IntAct has a sophisticated web-based curation tool which is used by many curation teams, database-related as well as independent teams. These teams directly curate data into the IntAct database which in return disseminates their data to a lot of external databases. This leads to IntAct being a common source for a lot of PPI databases (Orchard et al., 2014).

In summary, these three databases combined provide data which is relevant, highly qualitative and representative for most, if not all, the PPI data that is available at the moment.

References

AMMARI, M. G., GRESHAM, C. R., MCCARTHY, F. M. & NANDURI, B. 2016. HPIDB 2.0: a curated database for host-pathogen interactions. *Database (Oxford),* 2016.

ATKINSON, S. & WILLIAMS, P. 2016. Yersinia virulence factors - a sophisticated arsenal for combating host defences. *F1000Res,* 5.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2009. GenBank. *Nucleic Acids Res,* 37**,** D26-31.

BOSIO, C. M., BIELEFELDT-OHMANN, H. & BELISLE, J. T. 2007. Active suppression of the pulmonary immune response by Francisella tularensis Schu4. *J Immunol,* 178**,** 4538-47.

CELLI, J. & ZAHRT, T. C. 2013. Mechanisms of Francisella tularensis intracellular pathogenesis. *Cold Spring Harb Perspect Med,* 3**,** a010314.

D'ELIA, R., JENNER, D. C., LAWS, T. R., STOKES, M. G., JACKSON, M. C., ESSEX-LOPRESTI, A. E. & ATKINS, H. S. 2011. Inhibition of Francisella tularensis LVS infection of macrophages results in a reduced inflammatory response: evaluation of a therapeutic strategy for intracellular bacteria. *FEMS Immunol Med Microbiol,* 62**,** 348-61.

DEUTSCH, E. W., ORCHARD, S., BINZ, P. A., BITTREMIEUX, W., EISENACHER, M., HERMJAKOB, H., KAWANO, S., LAM, H., MAYER, G., MENSCHAERT, G., PEREZ-RIVEROL, Y., SALEK, R. M., TABB, D. L., TENZER, S., VIZCAINO, J. A., WALZER, M. & JONES, A. R. 2017. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J Proteome Res,* 16**,** 4288-4298.

DU PLESSIS, L., SKUNCA, N. & DESSIMOZ, C. 2011. The what, where, how and why of gene ontology--a primer for bioinformaticians. *Brief Bioinform,* 12**,** 723-35.

DURMUS, S. & ULGEN, K. O. 2017. Comparative interactomics for virus-human protein-protein interactions: DNA viruses versus RNA viruses. *FEBS Open Bio,* 7**,** 96-107.

DURMUS TEKIR, S., CAKIR, T., ARDIC, E., SAYILIRBAS, A. S., KONUK, G., KONUK, M., SARIYER, H., UGURLU, A., KARADENIZ, I., OZGUR, A., SEVILGEN, F. E. & ULGEN, K. O. 2013. PHISTO: pathogen-host interaction search tool. *Bioinformatics,* 29**,** 1357-8.

DYER, M. D., NEFF, C., DUFFORD, M., RIVERA, C. G., SHATTUCK, D., BASSAGANYA-RIERA, J., MURALI, T. M. & SOBRAL, B. W. 2010. The human-bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis. *PLoS One,* 5**,** e12089.

FINN, R. D., ATTWOOD, T. K., BABBITT, P. C., BATEMAN, A., BORK, P., BRIDGE, A. J., CHANG, H. Y., DOSZTANYI, Z., EL-GEBALI, S., FRASER, M., GOUGH, J., HAFT, D., HOLLIDAY, G. L., HUANG, H., HUANG, X., LETUNIC, I., LOPEZ, R., LU, S., MARCHLER-BAUER, A., MI, H., MISTRY, J., NATALE, D. A., NECCI, M., NUKA, G., ORENGO, C. A., PARK, Y., PESSEAT, S., PIOVESAN, D., POTTER, S. C., RAWLINGS, N. D., REDASCHI, N., RICHARDSON, L., RIVOIRE, C., SANGRADOR-VEGAS, A., SIGRIST, C., SILLITOE, I., SMITHERS, B., SQUIZZATO, S., SUTTON, G., THANKI, N., THOMAS, P. D., TOSATTO, S. C., WU, C. H., XENARIOS, I., YEH, L. S., YOUNG, S. Y. & MITCHELL, A. L. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res,* 45**,** D190-D199.

FUKAO, T. 2004. Immune system paralysis by anthrax lethal toxin: the roles of innate and adaptive immunity. *Lancet Infect Dis,* 4**,** 166-70.

JUNKER, B. R. H. & SCHREIBER, F. 2008. *Analysis of biological networks,* Hoboken, N.J., Wiley-Interscience.

MCKINNEY, W. 2010. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference,* **,** 51-56

MEYSMAN, P., SAEYS, Y., SABAGHIAN, E., BITTREMIEUX, W., VAN DE PEER, Y., GOETHALS, B. & LAUKENS, K. 2016. Mining the Enriched Subgraphs for Specific Vertices in a Biological Graph. *IEEE/ACM Trans Comput Biol Bioinform*.

MRZIC, A., MEYSMAN, P., BITTREMIEUX, W., MORIS, P., CULE, B., GOETHALS, B. & LAUKENS, K. In preparation. Grasping frequent subgraph mining for bioinformatics applications.

MULDER, N. J., AKINOLA, R. O., MAZANDU, G. K. & RAPANOEL, H. 2014. Using biological networks to improve our understanding of infectious diseases. *Comput Struct Biotechnol J,* 11**,** 1-10.

OLIPHANT, T. E. 2007. Python for Scientific Computing. *Computing in Science & Engineering,* 9**,** 10-20.

ORCHARD, S., AMMARI, M., ARANDA, B., BREUZA, L., BRIGANTI, L., BROACKES-CARTER, F., CAMPBELL, N. H., CHAVALI, G., CHEN, C., DEL-TORO, N., DUESBURY, M., DUMOUSSEAU, M., GALEOTA, E., HINZ, U., IANNUCCELLI, M., JAGANNATHAN, S., JIMENEZ, R., KHADAKE, J., LAGREID, A., LICATA, L., LOVERING, R. C., MELDAL, B., MELIDONI, A. N., MILAGROS, M., PELUSO, D., PERFETTO, L., PORRAS, P., RAGHUNATH, A., RICARD-BLUM, S., ROECHERT, B., STUTZ, A., TOGNOLLI, M., VAN ROEY, K., CESARENI, G. & HERMJAKOB, H. 2014. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res,* 42**,** D358-63.

PERRY, R. D. & FETHERSTON, J. D. 1997. Yersinia pestis--etiologic agent of plague. *Clin Microbiol Rev,* 10**,** 35-66.

SAYERS, E. W., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., EDGAR, R., FEDERHEN, S., FEOLO, M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., LANDSMAN, D., LIPMAN, D. J., MADDEN, T. L., MAGLOTT, D. R., MILLER, V., MIZRACHI, I., OSTELL, J., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T., SHUMWAY, M., SIROTKIN, K., SOUVOROV, A., STARCHENKO, G., TATUSOVA, T. A., WAGNER, L., YASCHENKO, E. & YE, J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res,* 37**,** D5-15.

SHANNON, P., MARKIEL, A., OZIER, O., BALIGA, N. S., WANG, J. T., RAMAGE, D., AMIN, N., SCHWIKOWSKI, B. & IDEKER, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res,* 13**,** 2498-504.

SNIDER, J., KOTLYAR, M., SARAON, P., YAO, Z., JURISICA, I. & STAGLJAR, I. 2015. Fundamentals of protein interaction network mapping. *Mol Syst Biol,* 11**,** 848.

SPENCER, R. C. 2003. Bacillus anthracis. *J Clin Pathol,* 56**,** 182-7.