Academic Year 2017-2018

Faculty Pharmaceutical, Biomedical and Veterinary Sciences

Biomedical Sciences

Exploring bacterial HPI networks using subgraph mining

By:

Lorenz Van de Veken

Master Thesis 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

TABLE OF CONTENTS

Abbreviations

HPI: Host-pathogen protein-protein interaction

GO: Gene ontology

IPR: InterPro

PPI: Protein-protein interaction

Y2H: Yeas two hybrid assay

AP-MS : affinity purification – mass spectrometry

IMEx International Molecular Exchange

non-HPIs: non-host-pathogen protein-protein interactions

taxid: taxonomy identifier

GOC: gene ontology consortium

GAF: gene accession file

DAG: direct acyclic graph

MF: molecular function

BP: biological process

CC: cellular location

GOA: Gene ontology enrichment analysis

Summary

Bioinformatics has been a fast-growing field of study within the life sciences the last two decades. One of the reasons is the tremendous increase in publicly available data. Interaction data is no exception. Over the last couple of years, high throughput protein-protein interaction assays have been developed leading to a quick expansion of the public interaction databases. Interaction networks are becoming more and more part of the data landscape, especially within the field of infectious diseases. Host-pathogen protein-protein interaction networks (HPI networks) offer a unique perspective on the interplay between the biochemical machinery of both organisms as they form the key interface between the host and the pathogen. The multi-faceted aspect of these interaction networks also allows multi-faceted analysis approaches to be performed. Interaction data, especially when it is supplemented with biologically relevant metadata such as gene ontology (GO) and InterPro (IPR) terms, is a normal biological dataset on one hand and a biological network on the other hand. This allows for the use of descriptive statistical methods as well as enrichment analysis, but also has the potential to be analysed using descriptive network parameters or even pattern mining methods, such as subgraph mining or clustering methods. To investigate the potential of such a multi-faceted approach, we assessed three public databases for interaction data of *Francisella tularensis, Bacillus* anthracis and *Yersinia pestis*, 3 human pathogens of whom the infection strategy has readily been characterised. We constructed and visualised their HPI networks. We have further summarised the interaction datasets as well as evaluated the information content of biologically relevant metadata. We also performed a descriptive network analysis and executed multiple subgraph mining runs in search for biological patterns that occur in the networks. We have shown that these big datasets can be summarised using descriptive and enrichment analyses as well as the potential of subgraph mining approaches to find patterns in the biological networks. However, for this multi-faceted approach to reach its full potential, the analyses setups have to be further optimised.

Introduction

HPI networks have recently caught the attention of many researchers that specialise in the field of infectious diseases. They seem to form a new informative source, especially with regard to infection strategies as well as using them to find or infer new interactions.

Why study HPIs? And why use subgraph mining, or any data mining approach for that matter?

First of all, studying proteins as individual entities provides a lot of information about their structure, physicochemical properties, function, etc. However, to understand their role in a system such as a pathway or a biochemical process, it is almost mandatory to study proteins within the context of their interactions with other proteins. A big part of the functioning of this biological system is driven by proteins and their interactions. Therefore, this system can be represented using protein-protein interaction (PPI) networks.

Consider bacterial infections. When we get infected, most of the times, we will develop symptoms at one point or another. These symptoms are evidence that the normal functioning of our body is disrupted. Somehow the bacteria has interacted with our system and interfered with some of its processes. We already established that our body is a complex biological system in which PPIs play a big role. Therefore, you could assume that the bacteria interfere with these PPIs, probably using their own proteins. This could be done in many ways. The pathogen may produce a protein that disrupts one host protein via structural modification or competitive binding for example. As a result many, if not all, of the normal interactions with this host protein are disrupted.

And indeed, many studies have found that bacteria often use their proteins to interact with the host proteins to disrupt them, especially with regard to immune suppression or evasion.

Therefore, analysing HPI networks is a very valuable approach to study bacterial infections.

Many interaction studies are still done by performing a (high-throughput) interaction experiment and then constructing the network and analysing the data. However, due to the technical advancements with high-throughput assays, the amount of interaction data that is being generated has increased tremendously. Public interaction databases are growing every day. This increase in public data opens up new opportunities to tackle interaction networks from a new and promising angle, with the use of data mining techniques.

The goal of data mining is to study big (often uninterpretable) datasets and extract information from them. For this purpose many methods have been developed to tackle specific types of data and datasets. One such a method that has been developed to analyse networks is subgraph mining. The goal of subgraph mining is to find interesting or frequently occurring patterns (subgraphs) in large networks or even multiple networks.

Our hypothesis states that there exist frequently occurring patterns within HPI networks of bacterial groups.

This hypothesis covers a very broad and open, yet well-defined research goal. 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 as well as descriptive network parameters to describe the general features of our dataset.

Secondly, Gene ontology (GO) enrichment and pathway enrichment analyses will also be done as they can provide a general contextual overview of our dataset.

Thirdly, we will perform frequent subgraph mining on the bacterial HPI networks in order to find patterns that can be linked to the infection strategy using the subgraph mining tool described in (Meysman et al., 2016)

Material and methods

Data handling

Origin of data (PPI experiments) in the databases

Based on the previous section, 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. The study we performed does not contain any experimental technique whatsoever, but it is important to be conscious of the origin of the data that we will be working with. Therefore, two of the more commonly used methods will be described as well as the general features and problems of the resulting data.

The first technique is the yeast two-hybrid assay (Y2H assay). The Y2H is based on the physical separation of subunits of a transcriptional activation factor which inactivates the functionality of this factor. These subunits are fused to PPI candidate proteins. If the fused proteins interact, then the subunits will also get in close proximity to each other. This will restore the transcription activation functionality leading to the transcription of a reporter gene.

Another widely used method is affinity purification followed by mass spectrometry (AP-MS), which consists of two steps. First, a candidate protein is immobilised on a solid support which will be used as bait in a column to capture other candidate proteins from a soluble phase. After this 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.

There are many more techniques that are used, but as mentioned above, these are the two most commonly used. (Snider et al., 2015)

Databases

The obtained data from experimental assays gets curated into public 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. (Orchard et al., 2014, Ammari et al., 2016, Durmus Tekir et al., 2013)

The interaction data is exported in either a tab-delimited text or a comma-separated values format. All the data extraction and manipulation is done using the Python programming language. (Oliphant, 2007). The pandas module offers a wide range of data frame creation and manipulation methods making it possible to read these files and display the data in a format which is easy to assess (McKinney, 2010).

Every row in the data frame represents an interaction and relevant information such as detection method, pubmedID of the publication which found the interaction, the interaction type,… Depending on the source interaction database, the type and amount of relevant information that is provided can vary as well as the way this information is formatted. This complicates the data cleaning and quality control process, which we will discuss in the next paragraph.

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. This metadata will be extracted from files which contain all the GO and InterPro data that is currently available. (The Gene Ontology, 2017, Ashburner et al., 2000, Finn et al., 2017). Using a Python script, this file will be parsed and relevant data will be written out to a new file per species.

The goal of the data handling step is to obtain a uniform dataset per bacterial species which only contains HPIs. Therefore, we need to integrate all 3 sets of exported data, but, as said, the different formats complicate this integration. To solve this and several other problems that arose during this process, a workflow based on several self-written python scripts has been developed. These scripts also make use of 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 Uniprot also provides a RESTful API which allows for programmatic access of the UniprotKB database.

The exported datasets from IntAct and HPIDB 2.0 have a very similar format as they both follow the International Molecular Exchange consortium (IMEx consortium) which requires databases to follow certain rules pertaining to which information is provided in their database. (Orchard et al., 2012) PHISTO does not follow this consortium and as a result, the exported dataset is formatted differently. One of the big differences is that PHISTO only contains HPIs and only accepts Uniprot IDs as a protein identifier. IntAct and HPIDB 2.0 accept a variety of protein/gene IDs. Furthermore, these two databases also contain non-host-pathogen protein-protein interaction (non-HPIs). A non-HPI is an interaction between a pair of proteins which does not consist of one protein from the host and one from the (subject) pathogen. For example, an interaction between two human proteins is considered a non-HPI. Another example which was found in our datasets were interactions between a pathogen protein and a yeast protein.

Due to this difference between PHISTO and the other two databases, the exported results of IntAct and HPIDB 2.0 were merged in the beginning and underwent a different data clean-up procedure than the results PHISTO. Afterwards, the IntAct/HPIDB dataset and the PHISTO dataset were merged.

The next section will first describe the data clean-up process of the IntAct/HPIDB dataset, followed by the clean-up process of PHISTO and lastly, the merge of the two datasets.

Data extraction and clean-up

Interaction data

One of the major problems that occurred was due to the heterogeneity of the protein IDs. Most of the entries linked the interacting proteins to their respective Uniprot IDs, which is favourable because our source for GO and IPR annotations also linked these annotations to Uniprot IDs. However, some entries did not have these IDs. One entry for example had the following reference ID: intact:EBI-2810906. This ID could be used to trace back to a Uniprot entry. However, the web page told us that this entry was found to be obsolete and thus deleted from the UniprotKB. When we tried to trace back other entries we mainly got the same message. Some entries however were not found to be obsolete, but redundant and then the web page presented alternative Uniprot IDs. One example lead to three Uniprot IDs which referred to the 3 subunits of the HSP70. Thus, we could not simply ignore the entries without a proper Uniprot ID.

To overcome this, the first step was to search in the alternative ID column, which contained protein IDs that linked backed to other protein and gene databases such as ENSEMBL or ENTREZ. If this did not work, the gene name linked to that protein was extracted and queried against the Uniprot database using REST API which is a service that allows programmatic access to the Uniprot database, giving us the opportunity to integrate this solution in the data extraction script.

After a first look in the merged datasets, we noticed some entries were non-HPIs. An explanation as to why these non-HPIs are present in the dataset can be found in the database querying step that was done in the beginning of this workflow. All the databases were queried for interactions that contained at least one protein that was linked to the subject pathogen's taxonomy ID (taxid). Therefore, non-HPIs also fulfilled the query's request and were thus also presented as a match.

Therefore, the next step is to filter out these non-HPI, which requires 2 steps: the first one is to filter out all the interactions in the set that do not have the correct species and taxonomic information linking back to either the human host or the bacterial species. Then another filtering step is done to exclude all the host-host and pathogen-pathogen interactions.

However, there was a problem concerning the species and taxonomy information. First of all, the dataset contained entries whose provided taxid did not correspond to the provided species/strain. The discrepancy was not entirely 'wrong', but the taxid would refer to the general species (ex. *Francisella tularensis spp.*, taxid 263), but the species name referred to a very specific strain/subspecies (ex *F. tularensis subsp. tularensis SCHU S4*, taxid 177416) or the other way around.

Secondly, the pathogen proteins of some entries were linked to the general species, while others were linked to very specific strains.

To solve these issues, the NCBI taxonomy database was used. (Sayers et al., 2009, Benson et al., 2009). Using the ETE3 toolkit module, a copy of the NCBI taxonomy database was downloaded and used to perform the two filtering steps. (Huerta-Cepas et al., 2016) At first, all the entries which did not have a taxid corresponding to either the general species, one of its subspecies or the human host (i.e. 9606) were dropped. Thus, the remaining interactions are either host-host, pathogen-pathogen or host-pathogen interactions. The second filtering step removes the former two, resulting in a dataset containing only HPIs.

Due to their compliance to the IMEx consortium, IntAct and HPIDB 2.0 present many properties of the interactions using long standardised vocabulary and reference numbers. For our research, this standardised format is impractical to work with. Therefore, using string manipulation we clean this data and only retain the core information of every property.

After this clean up step, the IntAct/HPIDB dataset is ready to be merged.

The PHISTO dataset is relatively clean. It provides much less properties and only presents the core information of these properties. Due to the nature of the database, all the entries are HPIs thus a filtering step is not needed. The protein identifier used in this database is the primary Uniprot ID so with regard to adding metadata, this set is already primed perfectly.

The very last thing that is to be done in both datasets before the merge is the creation of a unique identifier for every entry. The idea is to create an identifier that can discriminate entries based on their uniprotIDs, the interaction detection method and the pubmedID of the publication where the interaction was found. After the merge this identifier then can be used to drop interactions which were present in both datasets as these will have the same unique identifier.

Metadata extraction

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 gene accession file (GAF), which can be downloaded from the Gene ontology consortium (GOC) website at http://www.geneontology.org/page/download-go-annotations. These files contain gene products and their annotations plus some additional information. The GOC provides smaller GAF files of certain species or subsets of species which are ready to use, instead of having to download one of the larger files containing much more species and extracting your desired data from that file. Our subject species, however, did not have such a file (except for the human host) and thus we had to download one of the larger files *(filename: goa_Uniprot_all.gaf.gz*) and extract the desired data using a self-written python script.

After the extraction from the larger GAF file, the GO terms are added to the interaction data frame using the Uniprot IDs as the reference variable to merge the metadata with the interaction data.)

This process was rendered complicated because many of the Uniprot IDs that were provided with the metadata were not the primary UniprotIDs whereas the ones that were used in the interaction dataset were. This is because UniprotKB works with accession numbers as IDs. Every entry has one unique primary accession number, which is used to refer to this protein and is referred to as the Uniprot ID. However, every entry also has one or multiple secondary accession numbers, which should not be used to refer to the protein. This convention is not always followed by everyone and thus it might be that these secondary accession numbers are used in papers and databases to refer to a protein in the UniprotKB. Therefore, the merge was not successful for most of the interactions. To solve this, we used the Uniprot REST API again. This service has a remap function which allows you to convert one type of protein ID to another type of protein ID. It also allows to convert secondary IDs to primary IDs within the same type of IDs which is the functionality needed for this particular situation.

After the conversion step, the annotations can be added to the interaction dataset using the primary IDs the reference variable.

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 specific GO terms are remapped to more general parent terms to homogenize the information content. This achieves 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). The data collection and annotation process is identical to the previously described procedure that was used for the GO terms: A large file containing IPR terms for proteins from multiple organisms is downloaded from the InterPro site. (https://www.ebi.ac.uk/interpro/download.html, filename: protein2ipr.dat.gz). This file is then parsed to extract the terms that are related to our subject species and finally these terms are added to the interaction dataset using the Uniprot ID as the reference variable. The IPR terms however will not be used for the final subgraph mining runs due to memory limitations of our computers. This will be explained in further detail in the Results section.

After compiling the dataset, the diversity data of the dataset is evaluated, described and visualised using the visualisation and network analysis tools available in Cytoscape. (Assenov et al., 2008, Shannon et al., 2003)

This descriptive analysis is the first arm of our entire analysis setup and will be used to describe what can be seen in the dataset: how many data is present, how is it distributed across the pathogens, how diverse is the metadata set that has been added and so on. The second arm is the subgraph mining analysis which is more aimed towards finding and describing patterns in the data.

Descriptive analyses

General descriptive analysis

For the descriptive analysis, basic descriptive and statistical parameters are calculated and visualised using a simple python script. We will look at variables such as number of unique interactions, number of unique proteins, number of unique labels, mean number of GO and IPR labels per protein,…

The information content of the GO labels will also be evaluated. As mentioned in my project proposal, the GO can be represented as a direct acyclic graph (DAG) which originates from 3 root terms. These 3 root terms are molecular function (MF), biological process (BP) and cellular compartment (CC). They also form 3 distinct categories within the GO and are often referred to as namespaces.

Every GO term, except for the 3 root terms, has at least one parent term in the DAG. Thus, every term can be traced back via parent terms to one of the root terms. Conceptually, terms that are farther away from the root are considered more specific than terms closer to the root. To capture this, we assigned every term a 'depth' characteristic which is calculated by finding the shortest path from the term to their respective root term. Thus, a term within the BP category with a depth of 4 can be traced back to the BP term via 3 parent terms which stepwise become more and more general in their information content as they are located closer to BP term.

After the depth assignment, information content of all the GO labels within a dataset can be investigated by calculating simple statistical parameters such as the minimum, maximum, average, standard deviation,… as well as by plotting the distribution of the depth values.

Descriptive network analysis

Protein interaction datasets inherently form a biological network which can be depicted using graphs. Therefore, general descriptive graph parameters also provide insight and information on our dataset. For consistency, the term 'network' will be used in the remainder of this dissertation, as it is more intuitive to link with biological concept of a protein-protein interaction network.

Different descriptive parameters such as degree, shortest path betweenness centrality and network diameter will be calculated for the different bacterial groups using the available tools in Cytoscape in order to describe these obtained networks. In the next paragraph a couple of these parameters will be explained how they are calculated and what they actually tell us.

A graph consists of vertices (aka vertices) and edges which connect these vertices. In our case, we are dealing with a protein-protein interaction network. The vertices of the network represent the proteins and the edges represent the interactions between the proteins.

The degree k_i of a vertex n_i is defined as the number of edges adjacent to this vertex. When considering a network without any self-loops or multiple links, k_i is equal to the number of neighbours. In a directed network, 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 networks: 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 networks as well as is directed networks. The diameter $d_m = max(d_{ij})$ of a network is defined as the biggest distance present in the network, i.e. the longest shortest path between two vertices.

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 centrality measures how often a vertex or edge is present in the set of all the shortest paths of that network. Eccentricity centrality and closeness centrality are also often used, albeit more in communication networks. The former measures how 'quick' you can get from one vertex to all the other vertices in a network, while the latter measures how 'quick' you can go from any vertex in the network to a certain vertex. The eccentricity centrality is often also reported as eccentricity on its own which is equal to the mathematical inverse of the eccentricity centrality value. For the exact formulas to calculate of these centralities, see Junker and Schreiber, 2008.

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. (Junker and Schreiber, 2008).

GOA

A set of GO terms can also be analysed using a GO enrichment analysis (GOA) test. This test counts how often a term appears within a certain study dataset by calculating its relative abundance and then compares this value to the relative abundance of the term in a given population set to see whether or not it is significantly more present (enriched) or less present (purified) in the study set. To perform this analysis, we used a downloadable tool which is written in Python and controlled using the bash command line. This tool is publicly available for download at:

https://github.com/tanghaibao/goatools/blob/master/README.md.

We will perform this analysis by taking all the GO terms of all the proteins from a network from one species and use the GO terms of the entire pathogen and human proteome as the population set. This enrichment test is based on Fischer's exact test and thus also requires us to choose a p-value threshold which we set at 0.01. This threshold is to be applied after a Bonferroni correction for multiple testing has been done.

Subgraph mining

Concept introduction

To perform the subgraph mining analyses, an in-house java command-line tool is used. This tool is publicly accessible via the following link: https://github.com/geraore/significantGraphMiner.

Conceptually, this tool searches for patterns that are frequently associated with the interesting vertices.

As described in the project proposal, a subgraph of a given graph has a certain support value within that graph which indicates how often this subgraph occurs in the graph. The algorithm will do this by counting how many vertices it can use as a source vertex to build the subgraph. Support values, however, can be calculated given any collection of vertices, not just all the vertices of the graph. An example is found in this algorithm. It requires that you indicate which vertices are of interest to you and will then separately count the support value of a subgraph within the whole graph as well as within the subset of interesting vertices. Then it essentially 'compares' these values to determine whether or not a subgraph is 'frequently associated' with the subset of interesting vertices. In practice, this is done by calculating an upper cumulative probability based on a hypergeometric distribution. This probability value is also known as the P-value for the statistical test of enrichment. When this p-value gets lower and lower for a given subgraph, it becomes less and less likely that we would find the calculated support value (or higher) if the subgraph was not associated with the selected interesting vertices, or in other words, the lower the p-value gets, the less likely it is that we would have found the calculated support value (or higher) by mere coincidence. Thus, if we then define a significance threshold for this p-value that is low enough (0.05 and 0.01 are most commonly used), we can state that when a subgraph has a p-value lower than this threshold, this subgraph is frequently associated with the interesting vertices. (Meysman et al., 2016)

We are mainly interested in finding patterns that occur frequently within the network without necessarily being frequently associated with a subpopulation of vertices. This can also be done with this tool by telling the algorithm that the entire graph is 'interesting'. This does require some different steps with regard to handling and interpreting the output, which will be discussed in the Results section.

Although, simple topological networks may be able to provide some information, considering our research goal, it is of a much bigger interest to find biologically relevant patterns. Therefore, the vertices will be labelled with GO terms and the tool will build subgraphs using these labels rather than the proteins themselves. This setup thus consists of two big components: a structural one which connects proteins through interactions and builds the entire topology of the network and an informative one which couples biologically relevant information to the proteins. The combination of these two components essentially forms a process which 'anchors' the biological information of a protein to the 'location' of the protein in the network and by doing that the algorithm can build subgraphs using this information.

Conceptual introduction of the algorithm

Understanding the building process of the subgraph patterns is an important to properly interpret the output of the subgraph mining tool. Subgraphs are built vertex by vertex. The initial subgraph is always built using two vertices connected with one edge. Then it will count the support of this pattern by counting how many of the vertices in the interesting vertices file can be used as a source vertex for this pattern. For a vertex to be a source vertex, the algorithm must be able to build the pattern by starting at that particular vertex. When this support value is high enough, i.e. it exceeds the support value that we provided in the setup, the algorithm will save this pattern in its memory as being frequent and add another vertex to the pattern connecting it with an edge. Then again, it checks all the vertices in the interesting vertices file to see how many vertices can be used as a source vertex for the bigger pattern. If this value exceeds the threshold, the algorithm will again save it in its memory and further expand the pattern until it either reaches the provided maximum motif size or until the support value no longer exceeds the threshold. Once it is has reached either of these criteria, the algorithm will go back to the last pattern it has saved and try to expand it in another way. If there are no other expansions of that pattern left to examine, it will trace back to last pattern it has saved before that one and try to expand it and so on until it finally gets back to the initial subgraph and thus has to create a new initial subgraph and repeat the whole procedure again.

When the tool has to analyse labelled graphs and thus search for label patterns (as is the case in our study), the procedure is somewhat the same, but the amount of subgraph patterns the algorithm has to check may be bigger, depending on the label setup. Since we allow a vertex to have more than one label, the algorithm cannot justt add a vertex to a pattern like we described above. It has to compute the support of the expansion of the pattern for every label of the vertex separately. Thus, if we already have a pattern that exceeds the support threshold and want to add a vertex X with labels a, b and c, the algorithm has to check the support for the pattern expanded with a, expanded with b and expanded with c. This will be shown with a practical example in the Results section Only those instances of the expanded pattern that exceed the support threshold will be saved in the memory of the algorithm and thus used for expanding the pattern even further.

Setting up the analysis

Offering the possibility to modify several parameters, the user can optimise the setup according to each type of analysis run he/she wants to perform. Modifiable parameters include variables such as maximum motif size, minimum support value, maximum P-value, which algorithm to use and so on. Distinct meticulous choices were made regarding the choice of certain parameter values and the presence or absence of certain analysis options. This was done by creating our own small test networks of which we knew the size, the patterns, the structure and so on and running the subgraph mining analysis with different setups. Nevertheless, this setup is not optimised and thus, it should be noted that other setups might give better results. In the paragraph below will be discussed how we chose to setup the analysis and why

We chose to provide an interesting vertices file that contained all the vertices present in the graph to find patterns that occur frequent within the HPI network.

The support value is a very important parameter to determine, considering that the algorithm uses this value as a cut-off. We chose to set this at 10%, meaning that pattern needs to be supported by at least 10% of the interesting vertices or in our particular case, by 10% of the entire graph.

The maximum motif size was chosen to be 4 vertices. This value is somewhat arbitrary as there is no evidence or rule on how big or small an interaction should be. During our testing runs, it became clear that motif size plays a big role in how much memory is needed for the algorithm to be able to run. Therefore, it is advisable to keep the motif size as small as possible.

The algorithm also provides an option to choose with which mining algorithm the analysis should be performed. The three choices are base, FSG and gspan. Without going into further detail on the specifics of each algorithm, our test runs clarified that the base algorithm finds more patterns. This is not necessarily better because in that case the post-analysis filtering of the patterns requires the processing of more patterns, but it ensures that no possible interesting pattern is missed.

Then there are 3 other parameters which require no input, but rather their presence or absence in the command enables/disables certain functionalities. The 3 parameters are used to determine the label setup, the directedness of the network and the verbosity of the output respectively.

The first one is used to decide whether or not patterns need to have a label at every vertex. If this parameter is not present, every vertex will be assigned an extra ‘empty’ label to its set of GO labels. If the parameter is present, then this feature does not happen and the algorithm can only build patterns using the GO labels. This will be explained in more detail in the Results section when we talk about the analyses on the test networks.

The second presence/absence parameter determines how the algorithm will treat the network with regard to its directedness. If the parameter is present, the network is treated as undirected. If the parameter is absent, the network will be treated as directed. Since our specific research question deals with undirected networks, this parameter is present in our setup.

The presence of the third parameter is used to make the output ‘verbose’ which basically means that small explanatory phrases will be interspersed between the output to make it more readable and understandable for the user. A ‘verbose’ output might be more difficult to parse, but difficult to interpret. Even though our workflow requires an output parsing step, the utility of the verbosity of the output outweighs the small increase in parsing difficulty.

Results

Descriptive analysis

The results of the general descriptive analysis are presented in table 01, table 02 and table 03. We see an increase in number of interacting proteins and number of unique GO and IPR labels, going from *F. tularensis* to *B. anthracis* to *Y. pestis*. The distribution of the labels, however, is fairly equal across the three species.

Species	# pathogen proteins	# human proteins	# interactions
Francisella tularensis	346	982	1335
Bacillus anthracis	942	1713	3185
Yersinia pestis	1223	2149	4098

Table 01: General overview of the interaction datasets with regard to the proteins and PPIs
A significant difference between the species in number of interacting proteins as well as number of interactions can be seen.

Species	# unique labels	Range of # labels/pathogen protein	Average # labels/pathogen protein	Range of # labels/human protein	Average # labels/human protein
F. tularensis	5907	(1 - 22)	7,34	(1 - 234)	25,87
B. anthracis	7609	(1 - 22)	5,7	(1 - 260)	25,86
Y. pestis	8899	(1 -22)	6,18	(1 -260)	26,34

Table 02: General overview of the interaction datasets with regard to the GO labels
The number of labels between the three species varies, but the distribution in number of labels per protein stays fairly equal. The average number of labels per pathogen protein in the network of *F. tularensis* network is slightly higher than in the other two networks.

Species	# unique IPR labels	Range of # labels/pathogen protein	Average # labels/pathogen protein	Range of # labels/human protein	Average # labels/human protein
F. tularensis	2869	(1 -14)	6,1	(1 - 21)	4,84
B. anthracis	4562	(1 - 14)	4,26	(1 - 21)	5,01
Y. pestis	5299	(1 -22)	4,66	(1 - 20)	5,17

Table 03: General overview of the interaction datasets with regard to the IPR labels
The number of labels between the three species differs, but the distribution in number of labels per protein stays fairly equal. The average number of labels per pathogen protein in the network of *F. tularensis* network is slightly higher than in the other two networks while the range of number of labels per pathogen protein is slightly higher in the *Y. pestis* network.

The interaction networks were visualised using Cytoscape, as shown in figure 01 These visualisations often result in messy, dense 'hairball' structures, which are uninterpretable. One interesting feature that is already visible, however, is the disconnectedness of these interaction networks, i.e. you cannot start on any vertex of the network and travel to every single other vertex of the network, or in other words, for every possible pair of vertices, there does not necessarily exist a path between these vertices. All three HPI networks consist of a big, dense 'cloud' of interacting vertices accompanied by several smaller interaction 'networks'. Table 4 shows the diameter of the big 'cloud' in terms of edges, as this parameter is calculated by finding the longest shortest path in the entire network (as described in the Material and methods)

Species	diameter
F. tularensis	25
B. anthracis	15
Y. pestis	18

Table 04: Diameter of the bacterial HPI networks
The diameter of the three HPI networks, expressed in number of edges.

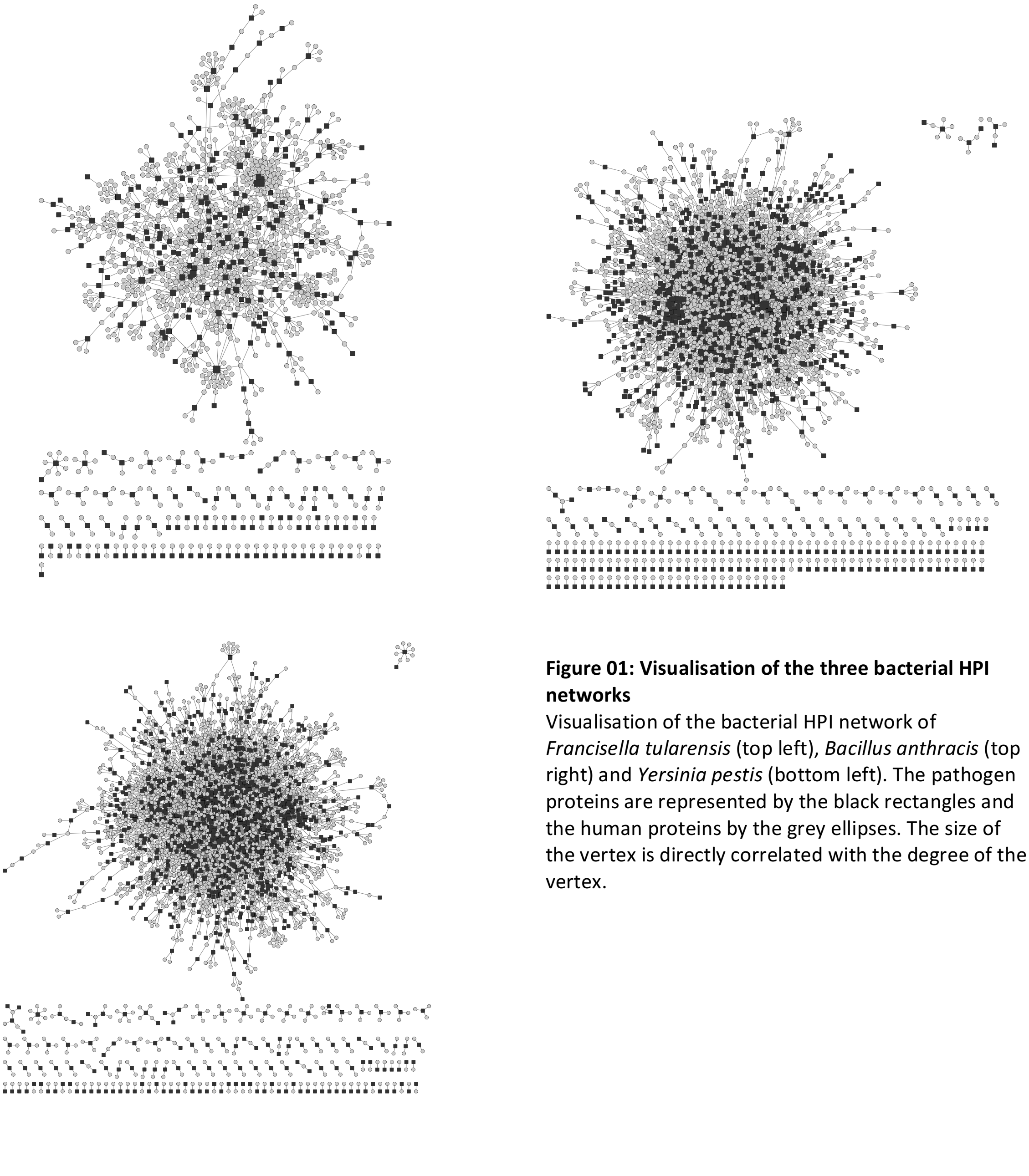


Figure 01: Visualisation of the three bacterial HPI networks
Visualisation of the bacterial HPI network of *Francisella tularensis* (top left), *Bacillus anthracis* (top right) and *Yersinia pestis* (bottom left). The pathogen proteins are represented by the black rectangles and the human proteins by the grey ellipses. The size of the vertex is directly correlated with the degree of the vertex.

We also evaluated these networks using several other descriptive network parameters. Unlike the diameter of a network, however, these parameters are often characteristics of vertices (e.g. degree, eccentricity, several centralities) and thus we would have to report these parameters for 1328 proteins, 2655 proteins and 3372 proteins for the *F. tularensis*, *B. anthracis* and *Y. pestis* network respectively.

Considering the goal of the descriptive analysis, namely making a summary of our dataset, we have not reported the full results, but rather described the features in the following paragraph and presented a table containing the 20 highest degree vertices in the *F. tularensis* network as an example (see table 05). The full overview of the network parameters can be found in the Appendix (Table A, B and C).

Protein	Avg. shortest path length	Betweenness centrality	Closeness centrality	Degree	Eccentricity
Q5NEC0	4,86915888	0,28221193	0,20537428	60	14
Q5NF74	6,55700935	0,06535184	0,15250855	29	17
Q5NID2	5,45981308	0,15351454	0,18315645	26	15
Q5NGV7	5,77757009	0,08665009	0,17308314	23	14
Q5NGF9	5,05607477	0,15673063	0,19778189	21	14
Q5NGW2	5,74392523	0,05768095	0,17409697	19	14
Q5NGF1	5,91962617	0,08250252	0,16892959	19	14
Q5NI89	5,06915888	0,17814529	0,19727139	16	15
Q5NFP9	7,0411215	0,04152194	0,14202283	15	16
Q5NF37	6,37570093	0,03773916	0,1568455	15	16
Q5NHX0	5,9588785	0,05568767	0,16781681	15	15
Q5NIP6	5,74579439	0,05766932	0,17404034	15	15
Q5NIJ3	6,54579439	0,0317167	0,15276985	13	16
Q5NFN4	6,20373832	0,03888289	0,16119313	13	14
Q5NEH1	5,97009346	0,03796434	0,16750157	13	15
Q5NIP5	6,43364486	0,01924528	0,15543289	12	16
Q5NEB6	6,55700935	0,03408718	0,15250855	12	16
Q5NGV3	7,48224299	0,03540494	0,13364976	12	18
Q5NFR9	5,88971963	0,03115721	0,16978737	12	14
Q5NF50	6,76261682	0,05849567	0,14787175	12	17

Table 05: Descriptive network parameters for the 20 highest degree vertices of the *F. tularensis network*
The results for the descriptive network parameters were obtained using the built-in NetworkAnalyzer tool from Cytoscape.

The first column contains the protein Uniprot ID. The second column and last column are expressed in number of edges. The third and fourth column are relative proportional values ranging from 0 to 1. The fifth column presents the degree which is expressed in number of vertices.

The results for the high degree nodes are very similar across the board. The average shortest path length is between 4 and 8 edges long. Overall, most have a very low betweenness centrality value, a low closeness centrality value and a rather high eccentricity value. We also tried to summarise the results by sorting the vertices using the values of different centralities as these also contain important information on the role of the vertices in the network. However, these centralities are more geared towards fully connected networks, as they are based on calculating the maximum shortest path in the network, starting from a given vertex or summing all the shortest paths in a network, again starting from a given vertex. As we have seen in figure 01, these bacterial HPI networks actually consist of one big connected network and several smaller 'networks'. The consequence of that specific structure leads to high centrality values for the vertices that are part of these smaller networks. Therefore, if we were to sort the results of the descriptive network analysis according to these centrality values, the vertices that will end up will all be part of these smaller networks. Of course, these vertices are part of the interaction network and have to be investigated as well, but these results would not representative for the entire network and are thus not included.

The results for the *B anthracis* and *Y. pestis* networks are very similar to the results for the *F. tularensis* network.

GOA

In the previous section, we already presented some basic descriptive statistics involving the number of GO terms as well as their depth. Besides using them as labels for the subgraph mining (of which the results will be presented in the next section), we also performed an enrichment analysis as described in the Materials and methods.

Species	MF	CC	BP	Total
Francisella tularensis	71	114	188	373
Bacillus anthracis	110	161	334	605
Yersinia pestis	110	157	355	622

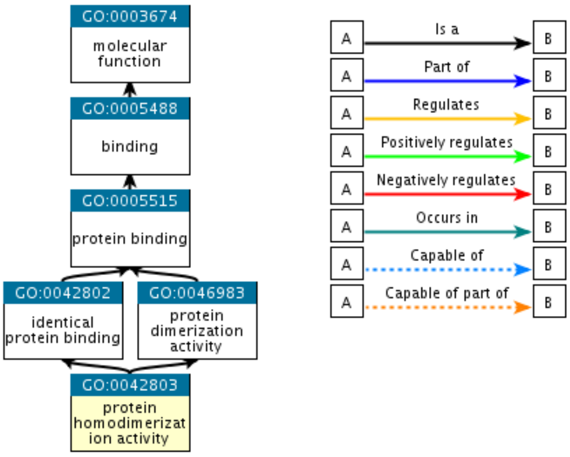
Table 06: Counts of the number of significantly enriched and purified GO terms
A count was performed on the results of the GO enrichment analysis. We counted the number of significantly enriched and purified terms per namespace (MF, CC and BP) per species.

Table 06 summarizes the results of the analysis. The BP namespace accounts for a significantly larger fraction of the enriched and purified terms. Furthermore, the total number of enriched and purified terms in the *F.* tularensis network is lower, especially with regard to the BP namespace. To maintain the overview and readability of this section, we decided to not present the full overview of the results, but rather only the overview of the top 20 enriched and purified terms of the MF namespace for *F. tularensis* as an example. We have described the features of the results within every namespace for all three species in the next paragraphs and refer to the appendix for the full overview of all the results (Appendix, table D, E, F, G, H, I, J K and L)

GO ID	enrichment	name	depth	p_bonferroni
GO:0005515	e	protein binding	2	0
GO:0042802	e	identical protein binding	3	1,71632E-37
GO:0003723	e	RNA binding	4	8,99688E-33
GO:0019901	e	protein kinase binding	5	1,45295E-23
GO:0019899	e	enzyme binding	3	8,4914E-22
GO:0045296	e	cadherin binding	4	6,13107E-21
GO:0042803	e	protein homodimerization activity	4	6,28179E-20
GO:0008022	e	protein C-terminus binding	3	5,93996E-19
GO:0000166	e	nucleotide binding	4	4,87553E-18
GO:0008137	p	NADH dehydrogenase (ubiquinone) activity	6	1,31469E-17
GO:0003677	e	DNA binding	4	1,3159E-16
GO:0046982	e	protein heterodimerization activity	4	2,71626E-16
GO:0046872	e	metal ion binding	4	1,81742E-15
GO:0047485	e	protein N-terminus binding	3	3,93572E-15
GO:0008134	e	transcription factor binding	3	4,42985E-12
GO:0005102	e	receptor binding	3	7,45753E-12
GO:0019903	e	protein phosphatase binding	5	1,51837E-11
GO:0016787	e	hydrolase activity	2	5,0409E-09
GO:0000978	e	RNA polymerase II proximal promoter sequence-specific DNA binding	9	2,36562E-08
GO:0003682	e	chromatin binding	2	3,3703E-08

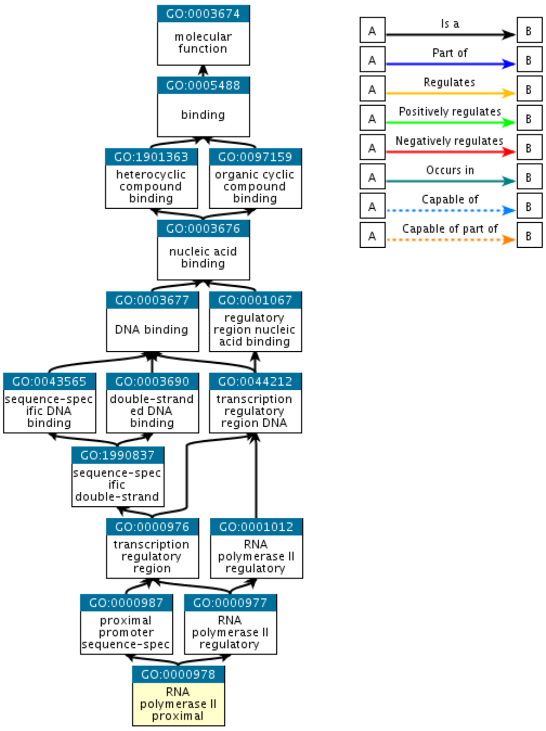
Table 07: Most significantly enriched and purified GO terms in the MF namespace, linked to the *F. tularensis* network
The 20 most significant enriched and purified GO terms within the MF namespace were extracted from the results of the GOA analysis of the *F. tularensis* dataset. In the first column we find the GO ID, in the second column the type of enrichment is indicated. an 'e' stands for enrichment, indicating that the term is significantly more represented in the study set, while a 'p' indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

In the MF namespace, most of the enriched terms are of a low specificity depth and a lot of them actually fall within each other’s ancestor chart. An ancestor chart visualizes the remapping that can be done with each term (see figure 02).



QuickGO - https://www.ebi.ac.uk/QuickGO

Figure 02: GO term ancestor charts
The ancestor charts of GO:0042803 (left) and GO:0000978 (right). The black arrows indicate child to parent relations (also referred to as ‘is_a’ relations). The child term is the term from which the arrow originates.


GO:0003674
molecular function
GO:0005488
binding
GO:1901363
heterocyclic compound binding
GO:0097159
organic cyclic compound binding
GO:0003676
nucleic acid binding
GO:0003677
DNA binding
GO:0001067
regulatory region nucleic acid binding
GO:0043565
sequence-specific DNA binding
GO:0003690
double-stranded DNA binding
GO:0044212
transcription regulatory region DNA
GO:1990837
sequence-specific double-strand
GO:0000976
transcription regulatory region
GO:0001012
RNA polymerase II regulatory
GO:0000987
proximal promoter sequence-spec
GO:0000977
RNA polymerase II regulatory
GO:0000978
RNA polymerase II proximal
A
Is a
B
A
Part of
B
A
Regulates
B
A
Positively regulates
B
A
Negatively regulates
B
A
Occurs in
B
A
Capable of
B
A
Capable of part of
B

QuickGO - https://www.ebi.ac.uk/QuickGO

For example, protein homodimerization activity (at depth 4) can be remapped to the term identical protein binding (at depth 3) as well as to the term protein binding (at depth 2). Another example is RNA polymerase II proximal promotor sequence-specific DNA binding (at depth 9). This term can be remapped to DNA binding (at depth 2). Many terms refer back to protein binding(-like) features which is not unexpected considering that all of the proteins are part of a protein-protein interaction network.

For the CC namespace, the average specificity depths are higher, but still fairly general and there are less terms that are involved with each other. 2 interesting terms are the purified ‘MHC class II protein complex’ term and enriched ‘endosome’ term, as these strengthen the current state-of-the-art regarding *Francisella tularensis* which comprises that the pathogen has an intracellular phase that plays a key role in its infection strategy.

The BP namespace contains the most specific terms of the three namespaces. Many of them are not directly present in each other's ancestry chart, but their charts do converge at some point or, in other words, they describe different aspects of one general biological process. In this particular case most terms can be linked to either immune response or gene transcription, which is also logical as immune evasion/suppression and DNA transcription are very important processes for survival.

Interestingly enough, we find similar results for *B. anthracis* and *Y. pestis*, whose infection strategy also revolves around intracellular survival and immune evasion/suppression.

Subgraph mining

Test networks

As mentioned in the material and methods section, the subgraph mining tool provides the use of several parameters to modify the setup of the analysis such as maximum motif size, the support threshold, which algorithm to use and so on. To ensure that we understood properly how this tool worked in practice, we made several simple test networks as well as basic variants of these networks to test out how each parameter influenced the results of the subgraph mining. The next paragraphs will deal briefly with the general working principles of this tool.

Consider a small network consisting of 166 vertices. As can be seen in figure 03, 30 of these vertices are arranged in '3-vertices-2-edges' interaction patterns, while the rest of network are simple '2 vertices-1-edge' interaction patterns. We have visually rearranged the 3-vertices patterns to V-shaped patterns (and from now on, they will be referred to as such) to make the understanding of the algorithm more intuitive. It should be noted that the visual structure of the subgraphs is purely arbitrary and used to ease visualisation, but is not a reference with regard to the way the algorithm builds and treats subgraphs (as we will see later on)

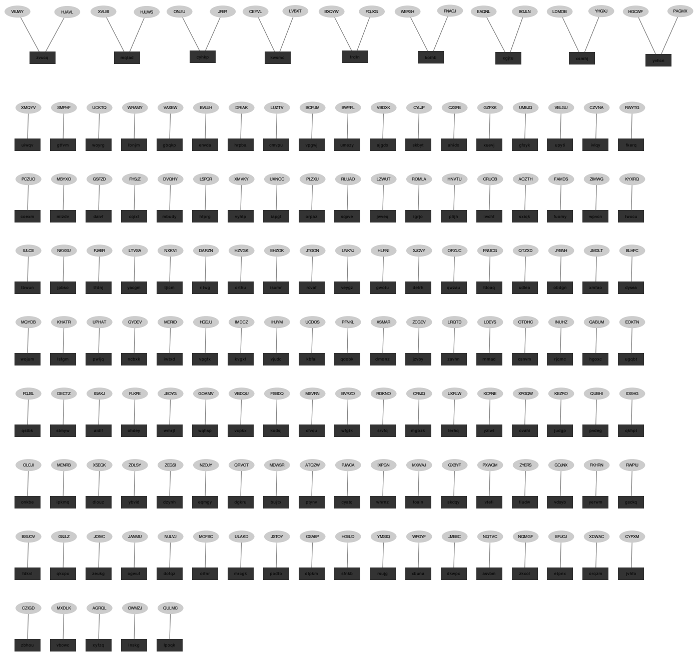


Figure 03: Visualisation of test network The test network is arbitrarily constructed to contain 10 'V-shaped' patterns. The black rectangles represent theoretical pathogen proteins while the grey ellipses represent theoretical human proteins. We have also labelled the proteins in such a way that all the human and pathogen proteins that are part of the 'V-shaped' pattern share the same labels respectively.

To stay within the subject of the thesis, we will treat these as HPI networks and thus the grey vertices may be interpreted as being human proteins and the black vertices as pathogen proteins. Since the subject HPI networks of our thesis work with labels and patterns made up of these labels, we've also added our own labels to this test network. This is done arbitrarily in such a way that all the vertices that are part of a V-shaped pattern have the same labels. More specifically, the pathogen protein has a P label and the 2 human proteins have H1 and H2 respectively. We also provided an interesting vertices file in which we've listed all the pathogen proteins that are part of a V-shaped pattern. We run the analysis using the following command:

```
java -jar ./build/jar/subgraphmining.jar -g ../input/Graph-file.txt -p ../input/interesting-vertices.txt -l
../input/Label-file-pattern-revealed.txt -o ../output/output-standard -s 10 -m 4 -i -u -d -a base
```

The highlighted part of the command is where we define our parameters as described in the Materials and methods: We choose a support threshold of 10, which in this case is equal to 100% of the interesting vertices. We will use a maximum motif size of 4 and also make sure that only fully labelled patterns will be build. Furthermore, we provide the directedness of our network to the tool, which in this case is undirected. We ask for a verbose output for easier output interpretability and last but least tell the tool to use the base algorithm.

After running this analysis, we get the following output

Motif	FreqS	FreqT	Pvalue
1P-2H2	10	10	1,01E-02
1P-2H2,1P-3H1	10	10	1,01E-02
1P-2H1,1P-3H2	10	10	1,01E-02
1P-2H1	10	10	1,01E-02

Table 08: Frequent fully labelled motifs found in the test network
The motifs that were found to be frequent after a subgraph mining run. The building patterns are presented in the first column. The second column indicates the subgroup frequency of the pattern. This is calculated based on how many vertices of the interesting vertices subpopulation could be used as a source vertex for the subgraph. The third column indicates the total frequency of the pattern and is calculated in the same manner as the subgroup frequency but uses the vertices of the entire graph. The last column provides the p-value that was calculated to determine whether or not the pattern is frequently associated with the interesting vertices.

The motifs are represented in a very specific way as they also reflect the building pattern of the subgraph. Thus, in this particular instance, it first found the motif 1P-2H2 to be frequent because it has a support value of ten, which is equal to our support threshold. The notation of the motif also tells us that the algorithm used a vertex with the P label as the source vertex, hence the ‘1’and then connected a second vertex (hence the ’2’) with the H2 label to the source vertex. This already indicates that the algorithm might have problems taking symmetry into account, which becomes even more clear with the second and third result in the table. The second motif should be read as follows: the algorithm started the motif with a vertex with label P as the source vertex and connected it to a vertex with the label H2. Then it added a third vertex to the motif with the label H1 and connected it to the vertex with the P label. This motif was also found ten times. Next, the third motif has again a vertex with label P as the source vertex, but this time it is connected to a vertex with the H1 label. Then, the algorithm added a third vertex with the label H2 and connected it to the vertex with the P label. It goes without saying that in an undirected network this third motif is equal to the second motif. However, the algorithm does not compute it as such due to the building pattern which further demonstrates that visualisation and visual interpretation of subgraphs should be done carefully. This is important to keep in mind when interpreting the results of the subgraph mining analysis on the bacterial networks.

If we do the same setup as the previous analysis run, but allow not fully labelled patterns to also be build, we find the same patterns as before plus every possibility in which one or more labels are left out. For example, the first motif in table 02 is exactly the same as the second motif in table 01, but during the building process, instead of adding the third vertex with the H1 label, the algorithm added it with no label. As such, allowing not fully labelled patterns intuitively is the same as adding an 'empty' label that's common to all the vertices in the graph which the algorithm can use during the building process.

Motif	FreqS	FreqT	Pvalue
1P-2H2,1P-3	10	10	1,01E-02
1P-2H2,1P-3H1	10	10	1,01E-02
1-2H1,1-3	10	10	1,01E-02
1P-2H1,1P-3H2	10	10	1,01E-02
1-2H1	10	10	1,01E-02
1-2H2	10	10	1,01E-02
1P-2	10	10	1,01E-02
1P-2H1,1P-3	10	10	1,01E-02
1P-2H2	10	10	1,01E-02
1P-2H1	10	10	1,01E-02
1-2H2,1-3H1	10	10	1,01E-02
1-2H1,1-3H2	10	10	1,01E-02
1P-2,1P-3	10	10	1,01E-02
1-2H2,1-3	10	10	1,01E-02
1-2,1-3	10	10	1,01E-02

Table 09: Frequent motifs found in the test network after performing a modified subgraph mining run
The motifs that were found to be frequent after the modified subgraph mining run. The original 4 results from the first run can be found in this table, as well as every possible variation in which one or more labels are left out.

To verify this, we did a run which once again only allowed fully labelled patterns, but we provided a label file in which every vertex that's part of the V-shaped pattern has a second label, namely, O. Thus, we have added a common label which the algorithm can use to build the patterns. As can be seen in table 03, the resulting motifs are exactly the same as in table 02 with the only difference that in table 03 the 'empty' label is not empty, but rather 'O', the label we assigned to all the vertices.

Motif	FreqS	FreqT	Pvalue
1P-2H2,1P-3H1	10	10	1,01E-02
1P-2H1,1P-3H2	10	10	1,01E-02
1P-2O	10	10	1,01E-02
1O-2H1,1O-3O	10	10	1,01E-02
1O-2H1,1O-3H2	10	10	1,01E-02
1O-2H2,1O-3H1	10	10	1,01E-02
1P-2H2	10	10	1,01E-02
1O-2H1	10	10	1,01E-02
1O-2H2	10	10	1,01E-02
1P-2H1	10	10	1,01E-02
1P-2O,1P-3O	10	10	1,01E-02
1O-2H2,1O-3O	10	10	1,01E-02
1O-2O,1O-3O	10	10	1,01E-02
1P-2H1,1P-3O	10	10	1,01E-02
1P-2H2,1P-3O	10	10	1,01E-02

Table 10: Frequent motifs found in the test network after adding a common label
The results are practically identical to the results presented in table 09 except for the ‘empty’ label being replaced the letter ‘O’.

Bacterial HPI networks

After having performed the necessary test runs on self-made networks, multiple analysis runs were done on the bacterial HPI networks which were constructed from the interaction data during the data handling step. We performed 4 runs per species with different label file setups to vary the information content. For every species, we made label files in which all the GO terms were remapped to their parent terms at a specificity depth of 2, 3 and 4 as well as a label file where no remapping was done (i.e. containing all the terms at their respective specificity depth as they can be found in the database.) In three of the four files, we also modified the GO terms to indicate if the term is linked to a pathogen protein or a human protein. This will help with interpreting the patterns.

Run	setup	# frequent motifs
1	F_tul; GO_2	660386
2	F_tul; GO_4; H/P	39058
3	F_tul; GO; H/P	3298
4	B_ant; GO_2	48394
5	B_ant; GO_4; H/P	120
6	B_ant; GO; H/P	4
7	Y_pes; GO_2	51366
8	Y_pes; GO_4; H/P	0
9	Y_pes; GO; H/P	0
10	F_tul; GO_3; H/P	294788
11	B_ant; GO_3; H/P	/
12	Y_pes; GO_3 ;H/P	4076

Table 11: Overview of the analysis setups and their number of results
An overview of all the analysis runs that were performed. The first second column contains information on the setup. First, the species is mentioned (F_tul = *Francisella tularensis;* B_ant = *Bacillus anthracis*; Y_pes = *Yersinia pestis*), then the depth of the labels (GO_x= terms remapped to depth x; GO = terms not remapped/kept at their original depth) and lastly, it is mentioned if the labels were modified to indicate whether or not the protein linked with the term is a pathogen (_P) or host (_H) protein. *Small note about the odd numbering of the runs: We kept the numbering of the runs in chronological order of execution. We added the runs with GO terms that have been remapped to depth 3 later on during the research to gain more insight in the effect of GO term depth on the frequent motifs*.

It is clear that the number of motifs that are found to be frequent varies a lot between different analysis setups as well as between the three species. Take note that run 6, 8 and 9 resulted in little to no frequent subgraphs. It was not possible to perform run 11 with the maximum computing power we had at our disposal. The failure was due to insufficient memory capacity.

To tackle the sheer amount of motifs found by the subgraph mining tool for some runs, a summarising approach was taken to further analyse the output. To achieve this, the motifs were grouped with respect to the first edge of their building pattern and then a count was done to evaluate how many subgraphs in the results are contained within these groups. We also did a second grouping step based on the second edge of the subgraph building patterns, which subdivided the groups and we performed a count on these subgroups as well. Since the amount of results is too high to be able to present them all in this thesis, we have made a top 10 selection of motifs based on the highest counts per grouping based per run. These represent the core patterns that are most frequently occurring within the network. This was done based on the second grouping for all runs except for runs 6,8, 9 and 11. The latter three due to the absence of proper results, as abovementioned. With regard to run 6, we presented the raw results because only two motifs were found, each consisting of 1 edge. Therefore, no group subdivision could be done. Once again, we have described the features of the results of all the runs per species, but only provided the table overview of the first analysis run that was performed on the *F. tularensis* network and refer to the appendix for the full overview of all the runs. (Appendix, table M, N, O, P, Q, R, S, T and U)

Francisella tularensis (run 1, run 2, run 3 and run 10):

Edge 1	Edge 2	Count
1intracellular part -2cellular metabolic process	2cellular metabolic process -3immune system process	113
1intracellular part -2cellular metabolic process	2cellular metabolic process -3leukocyte migration	102
1intracellular part -2cellular metabolic process	2cellular metabolic process -3side of membrane	108
1intracellular part -2heterocyclic compound binding	2heterocyclic compound binding -3leukocyte migration	108
1intracellular part -2organic substance metabolic process	2organic substance metabolic process -3immune system process	110
1membrane-bounded organelle -2cellular metabolic process	2cellular metabolic process -3leukocyte migration	100
1protein binding -2cellular metabolic process	2cellular metabolic process -3immune system process	111
1protein binding -2heterocyclic compound binding	2heterocyclic compound binding -3leukocyte migration	101
1protein binding -2organic cyclic compound binding	2organic cyclic compound binding -3leukocyte migration	101
1protein binding -2organic substance metabolic process	2organic substance metabolic process -3leukocyte migration	105

Table 12: Top 10 most frequent motifs found after performing analysis run 1
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

As can be seen in table 12, the variety in labels present in the highest count motifs of run 1 is rather limited. It is worth noting that run 1 is performed with GO labels that have been remapped to depth

2, thus if many of the remapped terms have one or multiple shared parent terms that have a depth of 2, all the different labels at their respective depth will have been remapped to that (those) shared parent term(s). Furthermore, considering the low depth, these results should always be viewed in the context of higher specificity runs. In run 10, we worked with GO labels at depth 3, which also have been modified to indicate if it is linked to a human or pathogen protein. the overall counts are lower, but the variety in the edge 2 column is a bit higher. And this trend continues, as can be seen in the results of run 2, which is performed with GO terms remapped to depth 4. The counts went down again, but the variety in the edge 2 column has also risen even more. Then lastly, run 3, which is performed without remapping the GO terms, gives very small counts compared to the other 3 runs and a fairly different set of labels.

Bacillus anthracis (run 4, run 5, run 6 and run 11)

The counts for the motifs found after run 4 are significantly lower when compared to run 1 which had the same setup as run 4, but was performed on the network of *F. tularensis*. The set of labels in the edge 2 column also differs, but again, considering the depth of the labels, these results should not be interpreted on their own. Unfortunately, run 5 and 6 did not provide many results. Run 5 only gave us 5 motifs with counts of 2 after the grouping procedure, and run 6 only gave us 2 motifs after the subgraph mining run. This may be due to a wide variety in the labels within the network or an unstructured distribution of the labels throughout the network as both scenarios will not lead to many patterns being formed within the network. Running run 11 with GO terms at depth 3 could've provided useable results but our computer was not able to perform this run due to memory issues.

Yersinia pestis (run 7, run 8, run 9 and run 12)

Run 7, which is the equivalent in setup to run 4 and run 1, also gave a different label set when compared with the results of run 1 and run 4 as well as a different range of counts. Once again, this setup is done with depth 2 labels and should thus not be viewed and interpreted on their own due to the low specificity. However, run 8 and run 9 gave no results at all after the subgraph mining. This is therefore again a scenario in which we would suspect that the variety of the labels in the network is too high or that the distribution of the labels is too unstructured for patterns to arise. Run 12, in contrast to run 11, did produce results. However, the namespace distribution across the labels of the motifs changed with regard to run 7. The labels present in the motifs after run 7 are mainly part of the BP namespace, whereas after run 12 a lot of labels are part of the CC namespace. This makes it more difficult to interpret the results from both runs.

Discussion

Performing a multi-faceted analysis of these bacterial HPI networks, gave us a lot of data covering a wide range of aspects. Now, we can use this information to find links, explain peculiarities within each aspect by using information from another as well as discuss what could've been done differently or what we could do in the future to gain even more insight in these networks.

First and foremost, when performing analyses using public databases, one should always take caution that whatever insights you get from this data, these insights only hold up provided that the data you have is of a proper confidence as well as representative with regard to the population you are getting insights about. To apply this to our study, we need to assume that the data that we downloaded from the public databases is correct and that the collective of interactions that we found for each species is representative for the actual interactome. Of course, there are measures one can take to assess this, we performed quality control and filtering steps for example. Several interactions were found by multiple studies or different techniques were used to discover the same interaction.

We also want to assess some biases that inevitably are present in our study. Well-studied proteins will have more information present in the databases. With regard to interaction studies, this has a rather important impact as well-studied proteins have a tendency to have more interactions compared to lesser known proteins. It is then difficult to determine whether this well-known protein indeed engages in more interactions or whether the lesser known protein's interaction partners just haven notbeen tested yet. This bias can also be seen at the species level. We found that the HPI network of *Y. pestis* has more interactions than that of *B. anthracis* and that the HPI network of *B. anthracis* has more interactions *F. tularensis*. (Table 01) Is this because *Y. pestis* engages in more HPIs or is this because it has been studied more extensively than the other two pathogens?

The size of the networks is also a good example of the importance to investigate multiple parameters that convey information on the same concept. As mentioned above, the size of an interaction can be interpreted by counting the number of interactions. However, when you calculate the network diameters, the *F tularensis* network is the largest of the three. (Table 04).

There was a difference in number of unique GO labels between the three species as well a small difference in the average number of GO labels with *F. tularensis* having less unique labels, but more labels per pathogen protein which already indicates that, intuitively, the intrinsic characteristics network of *F. tularensis* are more favourable with regard to finding patterns. And indeed, in the *F. tularensis* network, patterns with informative labels were found across all the runs, while the runs with labels that had a higher depth specificity performed on the other networks produced little to no results.

When visualising and describing the networks from a mathematical network point of view, it becomes clear that the disconnectedness of the networks pose some difficulties with regard to calculating several descriptive network parameters, especially centralities. These difficulties may be overcome by analysing the big dense connected part of the network separately from the accompanying smaller 'networks'.

With regard to the results of the GOA, as mentioned, we find across all three species similar results which is interesting as it shows that this enrichment analysis is able to find the general biological aspects that contribute to the infection strategy of all three pathogens. Someone with a critical view might point out that it was not able to pick up on the differences of the three pathogens, which is correct. However, this might be due to the choice of population set against which the enrichment is tested. For every species, we picked all the GO terms of the entire interactome, which means that the GO terms of the entire human proteome are also included in the population set. The terms that would give indications to the pathogen specific strategy are probably only linked to pathogen proteins of the interaction network, which are small minority compared to the human proteins, while the terms referring to the general infection strategy can also be linked to the human proteins. Therefore, the general terms may have a much higher likelihood of being enriched.

The tool we used also provided the proteins of the study set (in other words, the proteins that were part of the interaction network) that were labelled with the significantly enriched or purified GO term. Sometimes no proteins were reported. It is worth looking into this, as this may provide extra information on the enrichment/underrepresentation. For example, if a term referring to the glycolysis is purified and has no human proteins linked to it, then it could be hypothesised that the glycolysis is not a target of the bacterial attack. On the other hand, if multiple terms linked to our metabolism have many human proteins linked to them and little to no pathogen proteins, then this might be a representation of the hijacking of our metabolism by the bacteria.

Using the GOA can also prove to be useful with regard to the subgraph mining, because it gives insight in the information content of different specificity levels that is present in the network. This insight might then be used to construct label files that will provide more informative patterns s this now proved to be a critical point in the subgraph mining analysis. As already mentioned, the information distribution was better suited for pattern mining in the *F. tularensis* network compared to the other two networks.

Conclusion

We have performed 3 different analysis approaches in order to describe bacterial HPI networks, evaluate the biological information content in these networks and find biologically relevant patterns.

The general descriptive analysis as well as the GOA provided very useful tools to summarise our datasets. The descriptive network analysis, however, should not be performed unless the connectedness of the network is evaluated as this influences many network parameters.

The subgraph mining approach has shown potential in finding biologically relevant patterns, but to unlock its full potential an optimisation study has to be done.

Proper insights in the information content and distribution of biological labels are needed to ensure a good outcome when searching for labelled patterns.

This explorative study showed that multifaceted approaches have the potential to find biologically relevant information with regard to bacterial infection strategies. However, the different setups and parameters of every analysis that is included in such a multifaceted approach need sufficient exploration, evaluation and testing to make sure every aspect of the approach can perform an optimal analysis.

Appendix

In this appendix, all the mentioned tables can be found that accompany the descriptions in the Results section. For the full result overviews, we refer you to our github, which can be found at https://github.com/Lorenz-VdV/Master-thesis

Francisella tularensis

Protein	Avg. shortest path length	Betweenness centrality	Closeness centrality	Degree	Eccentricity
Q5NEC0	4,86915888	0,28221193	0,20537428	60	14
Q5NF74	6,55700935	0,06535184	0,15250855	29	17
Q5NID2	5,45981308	0,15351454	0,18315645	26	15
Q5NGV7	5,77757009	0,08665009	0,17308314	23	14
Q5NGF9	5,05607477	0,15673063	0,19778189	21	14
Q5NGW2	5,74392523	0,05768095	0,17409697	19	14
Q5NGF1	5,91962617	0,08250252	0,16892959	19	14
Q5NI89	5,06915888	0,17814529	0,19727139	16	15
Q5NFP9	7,0411215	0,04152194	0,14202283	15	16
Q5NF37	6,37570093	0,03773916	0,1568455	15	16
Q5NHX0	5,9588785	0,05568767	0,16781681	15	15
Q5NIP6	5,74579439	0,05766932	0,17404034	15	15
Q5NIJ3	6,54579439	0,0317167	0,15276985	13	16
Q5NFN4	6,20373832	0,03888289	0,16119313	13	14
Q5NEH1	5,97009346	0,03796434	0,16750157	13	15
Q5NIP5	6,43364486	0,01924528	0,15543289	12	16
Q5NEB6	6,55700935	0,03408718	0,15250855	12	16
Q5NGV3	7,48224299	0,03540494	0,13364976	12	18
Q5NFR9	5,88971963	0,03115721	0,16978737	12	14
Q5NF50	6,76261682	0,05849567	0,14787175	12	17

Table A: Descriptive network parameters for the 20 highest degree vertices of the *F. tularensis* network The results for the descriptive network parameters were obtained using the built-in NetworkAnalyzer tool from Cytoscape. The first column contains the protein Uniprot ID. The second column and last column are expressed in number of edges. The third and fourth column are relative proportional values ranging from 0 to 1. The fifth column presents the degree which is expressed in number of vertices.

Bacillus anthracis

Protein	Avg. shortest path length	Betweenness centrality	Closeness centrality	Degree	Eccentricity
P19838	3,92876344	0,17883084	0,25453301	60	9
Q81SN0	3,89516129	0,19237841	0,25672878	60	9
Q81KT8	4,46953405	0,08744633	0,22373697	54	9
Q81VT8	4,30107527	0,12634153	0,2325	48	9
A0A1A9IJH2	4,07616487	0,07946646	0,24532864	48	8
A0A1V4BGP0	4,07616487	0,07946646	0,24532864	48	8
Q81YE8	4,38351254	0,06486397	0,22812756	32	9
Q81XC3	5,11648746	0,03756399	0,19544658	29	10
Q9NY15	4,30062724	0,09199764	0,23252422	27	9
Q81TT4	5,05734767	0,03668982	0,1977321	26	9
Q81LD0	4,89516129	0,02599257	0,20428336	24	9
P17931	4,28270609	0,06867541	0,23349723	23	8
Q81S14	4,41487455	0,0451707	0,226507	23	9
Q81U80	4,99910394	0,02147409	0,20003585	22	10
Q6KNA9	4,3297491	0,05319756	0,23096026	22	9
Q81S59	4,36111111	0,05043155	0,22929936	21	9
Q81JL0	4,67831541	0,03562179	0,21375215	21	9
P61769	4,44578853	0,05564915	0,22493198	21	10
Q81Y62	4,53942652	0,04881154	0,22029214	21	9
Q81QG7	4,98566308	0,0245974	0,20057513	20	10

Table B: Descriptive network parameters for the 20 highest degree vertices of the *B. anthracis network* The results for the descriptive network parameters were obtained using the built-in NetworkAnalyzer tool from Cytoscape. The first column contains the protein Uniprot ID. The second column and last column are expressed in number of edges. The third and fourth column are relative proportional values ranging from 0 to 1. The fifth column presents the degree which is expressed in number of vertices.

Yersinia pestis

Protein	Avg. shortest path length	Betweenness centrality	Closeness centrality	Degree	Eccentricity
P19838	4,00810144	0,21623141	0,24949468	104	11
Q7ARD3	3,86509334	0,2738403	0,25872596	100	12
P17778	4,72525537	0,06579489	0,21162877	39	12
Q9NY15	4,23494188	0,08663786	0,23613075	38	11
Q9GZM7	4,26734766	0,07751409	0,2343376	30	11
P04233	4,39697076	0,07101687	0,2274293	27	11
Q7CGS9	5,12116943	0,02035118	0,1952679	26	11
P04275	4,52518492	0,03632693	0,22098544	26	13
P61769	4,74286721	0,06045248	0,21084293	25	11
E7E587	4,95632265	0,01722477	0,20176249	22	12
O00165	4,54068334	0,04698368	0,22023117	22	13
P46379	5,05988024	0,02342459	0,19763314	22	11
P68587	4,95632265	0,01722477	0,20176249	22	12
Q7CHZ0	5,44170483	0,01418979	0,18376594	20	12
P48995	4,77668193	0,04653486	0,20935034	20	11
Q10587	5,1880944	0,02128401	0,192749	19	12
O00560	5,04085946	0,01866156	0,19837887	19	13
Q8NDV7	4,84219796	0,03185016	0,20651779	19	11
Q12882	5,0211342	0,02584444	0,19915819	18	13
Q9HCC8	5,19513913	0,02910116	0,19248763	18	11

Table C: Descriptive network parameters for the 20 highest degree vertices of the *Y. pestis* network
The results for the descriptive network parameters were obtained using the built-in NetworkAnalyzer tool from Cytoscape. The first column contains the protein Uniprot ID. The second column and last column are expressed in number of edges. The third and fourth column are relative proportional values ranging from 0 to 1. The fifth column presents the degree which is expressed in number of vertices.

GOA

Francisella tularensis

GO	enrichment	name	depth	p_bonferroni
GO:0005515	e	protein binding	2	0
GO:0042802	e	identical protein binding	3	1,71632E-37
GO:0003723	e	RNA binding	4	8,99688E-33
GO:0019901	e	protein kinase binding	5	1,45295E-23
GO:0019899	e	enzyme binding	3	8,4914E-22
GO:0045296	e	cadherin binding	4	6,13107E-21
GO:0042803	e	protein homodimerization activity	4	6,28179E-20
GO:0008022	e	protein C-terminus binding	3	5,93996E-19
GO:0000166	e	nucleotide binding	4	4,87553E-18
GO:0008137	p	NADH dehydrogenase (ubiquinone) activity	6	1,31469E-17
GO:0003677	e	DNA binding	4	1,3159E-16
GO:0046982	e	protein heterodimerization activity	4	2,71626E-16
GO:0046872	e	metal ion binding	4	1,81742E-15
GO:0047485	e	protein N-terminus binding	3	3,93572E-15
GO:0008134	e	transcription factor binding	3	4,42985E-12
GO:0005102	e	receptor binding	3	7,45753E-12
GO:0019903	e	protein phosphatase binding	5	1,51837E-11
GO:0016787	e	hydrolase activity	2	5,0409E-09
GO:0000978	e	RNA polymerase II proximal promoter sequence-specific DNA binding	9	2,36562E-08
GO:0003682	e	chromatin binding	2	3,3703E-08

Table D: Most significantly enriched and purified GO terms in the MF namespace, linked to the *F. tularensis* network The 20 most significant enriched and purified GO terms within the MF namespace were extracted from the results of the GOA analysis of the *F. tularensis* dataset. In the first column, we find the GO ID, in the second column the type of enrichment is indicated. an 'e' stands for enrichment, indicating that the term is significantly more represented in the study set, while a 'p' indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

GO	enrichment	name	depth	p_bonferroni
GO:0005737	e	cytoplasm	3	1,3728E-203
GO:0005829	e	cytosol	4	2,9108E-127
GO:0070062	e	extracellular exosome	5	6,1732E-110
GO:0005634	e	nucleus	5	2,3711E-100
GO:0005654	e	nucleoplasm	5	3,08141E-82
GO:0005856	e	cytoskeleton	5	4,31329E-47
GO:0016021	p	integral component of membrane	3	6,05212E-43
GO:0005886	e	plasma membrane	2	1,39697E-42
GO:0005925	e	focal adhesion	5	6,67907E-41
GO:0043234	e	protein complex	2	4,84252E-32
GO:0048471	e	perinuclear region of cytoplasm	4	5,72607E-31
GO:0042613	p	MHC class II protein complex	6	9,93527E-31
GO:0005615	e	extracellular space	2	5,61555E-24
GO:0005768	e	endosome	6	1,2244E-23
GO:0005783	e	endoplasmic reticulum	5	4,58244E-22
GO:0005813	e	centrosome	6	1,67484E-21
GO:0070469	p	respiratory chain	2	3,32519E-21
GO:0005576	e	extracellular region	1	2,71195E-20
GO:0005794	e	Golgi apparatus	5	4,4813E-20
GO:0042995	e	cell projection	2	7,82697E-20

Table E: Most significantly enriched and purified GO terms in the CC namespace, linked to the *F. tularensis* network The 20 most significant enriched and purified GO terms within the CC namespace were extracted from the results of the GOA analysis of the *F. tularensis* dataset. In the first column we find the GO ID, in the second column the type of enrichment is indicated. an 'e' stands for enrichment, indic,ating that the term is significantly more represented in the study set, while a 'p' indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

GO	enrichment	name	depth	p_bonferroni
GO:0043312	e	neutrophil degranulation	9	1,76545E-42
GO:0006351	e	transcription, DNA-templated	9	7,37876E-38
GO:0016032	e	viral process	4	4,76744E-37
GO:0019882	p	antigen processing and presentation	2	7,38361E-34
GO:0006397	e	mRNA processing	8	2,98576E-28
GO:0045944	e	positive regulation of transcription by RNA polymerase II	11	1,64725E-27
GO:0006366	e	transcription by RNA polymerase II	10	1,9191E-25
GO:0008380	e	RNA splicing	8	2,41145E-23
GO:0006955	p	immune response	2	2,15711E-22
GO:0000122	e	negative regulation of transcription by RNA polymerase II	11	7,01046E-22
GO:0045893	e	positive regulation of transcription, DNA-templated	10	6,91365E-19
GO:0006915	e	apoptotic process	4	1,1753E-17
GO:0000398	e	mRNA splicing, via spliceosome	11	4,80815E-17
GO:0043066	e	negative regulation of apoptotic process	7	1,26849E-16
GO:0006357	e	regulation of transcription by RNA polymerase II	10	5,48013E-15
GO:0043161	e	proteasome-mediated ubiquitin-dependent protein catabolic process	9	1,0535E-14
GO:0050852	e	T cell receptor signaling pathway	10	1,12046E-14
GO:0015031	e	protein transport	7	1,88327E-14
GO:0007049	e	cell cycle	2	2,18346E-14
GO:0006355	e	regulation of transcription, DNA-templated	9	3,53274E-13

Table F: Most significantly enriched and purified GO terms in the BP namespace, linked to the *F. tularensis* network
The 20 most significant enriched and purified GO terms within the BP namespace were extracted from the results of the GOA analysis of the *F. tularensis* dataset. In the first column, we find the GO ID, in the second column the type of

enrichment is indicated. an ’e’ stands for enrichment, indicating that the term is significantly more represented in the study set, while a ‘p’ indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

Bacillus anthracis

GO	enrichment	name	depth	p_bonferroni
GO:0005515	e	protein binding	2	0
GO:0003674	e	molecular_function	0	1,3663E-121
GO:0042802	e	identical protein binding	3	5,2545E-105
GO:0003723	e	RNA binding	4	6,82182E-66
GO:0045296	e	cadherin binding	4	1,38299E-46
GO:0046872	e	metal ion binding	4	8,24479E-43
GO:0019899	e	enzyme binding	3	2,13526E-42
GO:0042803	e	protein homodimerization activity	4	2,45896E-40
GO:0019901	e	protein kinase binding	5	2,01503E-38
GO:0000166	e	nucleotide binding	4	1,97726E-29
GO:0008137	p	NADH dehydrogenase (ubiquinone) activity	6	2,29014E-29
GO:0032403	e	protein complex binding	3	4,29702E-27
GO:0003682	e	chromatin binding	2	4,13928E-24
GO:0019904	e	protein domain specific binding	3	1,86524E-22
GO:0003713	e	transcription coactivator activity	5	9,23928E-22
GO:0031625	e	ubiquitin protein ligase binding	5	2,61199E-21
GO:0008134	e	transcription factor binding	3	2,48041E-20
GO:0047485	e	protein N-terminus binding	3	4,17441E-17
GO:0046982	e	protein heterodimerization activity	4	2,26422E-16
GO:0003714	e	transcription corepressor activity	5	2,42417E-16

Table G: Most significantly enriched and purified GO terms in the MF namespace, linked to the *B. anthracis* network The 20 most significant enriched and purified GO terms within the MF namespace were extracted from the results of the GOA analysis of the *B. anthracis* dataset. In the first column, we find the GO ID, in the second column the type of enrichment is indicated. an ’e’ stands for enrichment, indicating that the term is significantly more represented in the study set, while a ‘p’ indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

GO	enrichment	name	depth	p_bonferroni
GO:0005737	e	cytoplasm	3	1,0506E-301
GO:0005829	e	cytosol	4	1,7351E-252
GO:0070062	e	extracellular exosome	5	1,0928E-218
GO:0005634	e	nucleus	5	2,9095E-191
GO:0005654	e	nucleoplasm	5	1,7916E-168
GO:0005886	e	plasma membrane	2	9,6916E-103
GO:0005856	e	cytoskeleton	5	4,62561E-78
GO:0043234	e	protein complex	2	5,70466E-67
GO:0042613	p	MHC class II protein complex	6	2,18688E-60
GO:0005925	e	focal adhesion	5	1,19037E-52
GO:0042995	e	cell projection	2	3,14812E-48
GO:0048471	e	perinuclear region of cytoplasm	4	3,65548E-48
GO:0005794	e	Golgi apparatus	5	2,67612E-46
GO:0016021	p	integral component of membrane	3	2,58096E-45
GO:0005575	e	cellular_component	0	3,05883E-42
GO:0005783	e	endoplasmic reticulum	5	6,30819E-41
GO:0005730	e	nucleolus	5	8,14866E-38
GO:0031410	e	cytoplasmic vesicle	5	2,41395E-37
GO:0005768	e	endosome	6	7,1867E-36
GO:0005615	e	extracellular space	2	1,13888E-34

Table H: Most significantly enriched and purified GO terms in the CC namespace, linked to the *B. anthracis* network The 20 most significant enriched and purified GO terms within the CC namespace were extracted from the results of the GOA analysis of the *B. anthracis* dataset. In the first column, we find the GO ID, in the second column the type of enrichment is indicated. an ’e’ stands for enrichment, indicating that the term is significantly more represented in the study set, while a ‘p’ indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

GO	enrichment	name	depth	p_bonferroni
GO:0008150	e	biological_process	0	5,9942E-125
GO:0045944	e	positive regulation of transcription by RNA polymerase II	11	3,31802E-73
GO:0043312	e	neutrophil degranulation	9	1,22153E-72
GO:0019882	p	antigen processing and presentation	2	1,41439E-64
GO:0006351	e	transcription, DNA-templated	9	1,2292E-63
GO:0016032	e	viral process	4	6,71023E-59
GO:0000122	e	negative regulation of transcription by RNA polymerase II	11	2,23628E-47
GO:0006915	e	apoptotic process	4	2,35053E-44
GO:0045893	e	positive regulation of transcription, DNA-templated	10	5,11871E-40
GO:0006955	p	immune response	2	9,01996E-40
GO:0007049	e	cell cycle	2	1,95896E-36
GO:0006366	e	transcription by RNA polymerase II	10	6,68846E-36
GO:0045892	e	negative regulation of transcription, DNA-templated	10	1,62284E-34
GO:0006397	e	mRNA processing	8	3,88185E-33
GO:0015031	e	protein transport	7	2,286E-31
GO:0008380	e	RNA splicing	8	9,36287E-30
GO:0008283	e	cell proliferation	1	1,63563E-28
GO:0016569	e	covalent chromatin modification	5	3,33481E-27
GO:0006355	e	regulation of transcription, DNA-templated	9	1,32089E-24
GO:0006357	e	regulation of transcription by RNA polymerase II	10	3,2154E-24

Table I: Most significantly enriched and purified GO terms in the BP namespace, linked to the *B. anthracis* network The 20 most significant enriched and purified GO terms within the BP namespace were extracted from the results of the GOA analysis of the *B. anthracis* dataset. In the first column, we find the GO ID, in the second column the type of enrichment is indicated. an 'e' stands for enrichment, indicating that the term is significantly more represented in the study set, while a 'p' indicates a significant underrepresentation of the term. The third column contains the name of the GO

term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

Yersinia pestis

GO	enrichment	name	depth	p_bonferroni
GO:0005515	e	protein binding	2	0
GO:0042802	e	identical protein binding	3	2,7495E-103
GO:0003723	e	RNA binding	4	7,93997E-70
GO:0045296	e	cadherin binding	4	1,46623E-53
GO:0042803	e	protein homodimerization activity	4	2,41849E-48
GO:0008137	p	NADH dehydrogenase (ubiquinone) activity	6	7,43638E-47
GO:0019899	e	enzyme binding	3	2,78767E-46
GO:0031625	e	ubiquitin protein ligase binding	5	8,94352E-36
GO:0000166	e	nucleotide binding	4	1,18795E-35
GO:0003682	e	chromatin binding	2	1,7559E-32
GO:0019901	e	protein kinase binding	5	3,5979E-32
GO:0046872	e	metal ion binding	4	4,81668E-28
GO:0016787	e	hydrolase activity	2	2,88376E-26
GO:0008134	e	transcription factor binding	3	3,8734E-26
GO:0000978	e	RNA polymerase II proximal promoter sequence-specific DNA binding	9	1,31501E-25
GO:0008022	e	protein C-terminus binding	3	4,1072E-22
GO:0032403	e	protein complex binding	3	1,52848E-17
GO:0004129	p	cytochrome-c oxidase activity	8	3,82292E-17
GO:0004527	e	exonuclease activity	5	9,95E-17
GO:0019904	e	protein domain specific binding	3	1,25045E-15

Table J: Most significantly enriched and purified GO terms in the MF namespace, linked to the *Y. pestis* network The 20 most significant enriched and purified GO terms within the MF namespace were extracted from the results of the GOA analysis of the *Y. pestis* dataset. In the first column, we find the GO ID, in the second column the type of enrichment is indicated. an 'e' stands for enrichment, indicating that the term is significantly more represented in the study set, while a 'p' indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

GO	enrichment	name	depth	p_bonferroni
GO:0005737	e	cytoplasm	3	0
GO:0005829	e	cytosol	4	0
GO:0070062	e	extracellular exosome	5	8,1929E-242
GO:0005634	e	nucleus	5	7,0402E-184
GO:0005654	e	nucleoplasm	5	5,7423E-174
GO:0005886	e	plasma membrane	2	2,44384E-85
GO:0042613	p	MHC class II protein complex	6	1,08583E-84
GO:0016021	p	integral component of membrane	3	1,68343E-72
GO:0005856	e	cytoskeleton	5	2,74021E-64
GO:0005925	e	focal adhesion	5	1,17159E-57
GO:0005768	e	endosome	6	2,78565E-57
GO:0070469	p	respiratory chain	2	5,15091E-52
GO:0048471	e	perinuclear region of cytoplasm	4	5,42442E-51
GO:0043234	e	protein complex	2	2,16406E-50
GO:0005794	e	Golgi apparatus	5	1,39069E-48
GO:0005743	p	mitochondrial inner membrane	6	5,08726E-46
GO:0031410	e	cytoplasmic vesicle	5	7,18873E-45
GO:0005730	e	nucleolus	5	2,58722E-39
GO:0005783	e	endoplasmic reticulum	5	8,18427E-37
GO:0016607	e	nuclear speck	7	1,45296E-36

Table K: Most significantly enriched and purified GO terms in the CC namespace, linked to the *Y. pestis* network The 20 most significant enriched and purified GO terms within the CC namespace were extracted from the results of the GOA analysis of the *Y. pestis* dataset. In the first column, we find the GO ID, in the second column the type of enrichment is indicated. an 'e' stands for enrichment, indicating that the term is significantly more represented in the study set, while a 'p' indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

GO	enrichment	name	depth	p_bonferroni
GO:0019882	p	antigen processing and presentation	2	5,25576E-92
GO:0006351	e	transcription, DNA-templated	9	2,8099E-73
GO:0045944	e	positive regulation of transcription by RNA polymerase II	11	2,99548E-66
GO:0006955	p	immune response	2	7,14676E-65
GO:0016032	e	viral process	4	8,94374E-65
GO:0043312	e	neutrophil degranulation	9	2,1715E-56
GO:0008380	e	RNA splicing	8	2,41144E-47
GO:0000122	e	negative regulation of transcription by RNA polymerase II	11	2,46742E-43
GO:0006397	e	mRNA processing	8	1,76179E-42
GO:0015031	e	protein transport	7	9,22776E-42
GO:0006915	e	apoptotic process	4	1,94242E-40
GO:0006366	e	transcription by RNA polymerase II	10	4,31133E-40
GO:0043066	e	negative regulation of apoptotic process	7	2,22545E-32
GO:0000398	e	mRNA splicing, via spliceosome	11	2,53605E-32
GO:0045893	e	positive regulation of transcription, DNA-templated	10	1,78594E-30
GO:0016569	e	covalent chromatin modification	5	2,50889E-29
GO:0007049	e	cell cycle	2	4,24775E-29
GO:0045087	e	innate immune response	4	1,05294E-27
GO:0008285	e	negative regulation of cell proliferation	5	1,70661E-26
GO:0006355	e	regulation of transcription, DNA-templated	9	7,36498E-26

Table L: Most significantly enriched and purified GO terms in the BP namespace, linked to the *Y. pestis* network The 20 most significant enriched and purified GO terms within the BP namespace were extracted from the results of the GOA analysis of the *Y. pestis* dataset. In the first column, we find the GO ID, in the second column the type of enrichment is indicated. an 'e' stands for enrichment, indicating that the term is significantly more represented in the study set, while a 'p' indicates a significant underrepresentation of the term. The third column contains the name of the GO term. The fourth column presents the depth of the GO term within the DAG and in the last column the Bonferroni-corrected p-value is provided.

Subgraph mining

Francisella tularensis

Run 1:

Edge 1	Edge 2	Count
1intracellular part -2cellular metabolic process	cellular metabolic process -3immune system process	113
1intracellular part -2cellular metabolic process	cellular metabolic process -3leukocyte migration	102
1intracellular part -2cellular metabolic process	cellular metabolic process -3side of membrane	108
1intracellular part -2heterocyclic compound binding	heterocyclic compound binding -3leukocyte migration	108
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3immune system process	110
1membrane-bounded organelle -2cellular metabolic process	cellular metabolic process -3leukocyte migration	100
1protein binding -2cellular metabolic process	cellular metabolic process -3immune system process	111
1protein binding -2heterocyclic compound binding	heterocyclic compound binding -3leukocyte migration	101
1protein binding -2organic cyclic compound binding	organic cyclic compound binding -3leukocyte migration	101
1protein binding -2organic substance metabolic process	organic substance metabolic process -3leukocyte migration	105

Table M: Top 10 most frequent motifs found after performing analysis run 1
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

Run 2

Edge 1	Edge 2	Count
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3immune system process _H	69
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3negative regulation of immune system process _H	81
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3negative regulation of multicellular organismal process _H	86
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3regulation of cell activation _H	77
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3regulation of cell motility _H	67
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3regulation of cellular component movement _H	68
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3regulation of immune response _H	63
1protein binding _H-2cellular nitrogen compound biosynthetic process _P	cellular nitrogen compound biosynthetic process _P-3regulation of leukocyte activation _H	90

Table N: Top 10 most frequent motifs found after performing analysis run 2
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

Run 3

Edge 1	Edge 2	Count
1protein binding _H-2ATP binding _P	ATP binding _P-3DNA binding _H	18
1protein binding _H-2ATP binding _P	ATP binding _P-3cytoskeleton _H	18
1protein binding _H-2ATP binding _P	ATP binding _P-3hydrolase activity _H	18
1protein binding _H-2ATP binding _P	ATP binding _P-3multicellular organism development _H	18
1protein binding _H-2ATP binding _P	ATP binding _P-3positive regulation of transcription	18
1protein binding _H-2ATP binding _P	ATP binding _P-3transcription	17
1protein binding _H-2ATP binding _P	ATP binding _P-3microtubule organizing center _H	16
1protein binding _H-2ATP binding _P	ATP binding _P-3extracellular exosome _H	16
1protein binding _H-2nucleotide binding _P	nucleotide binding _P-3perinuclear region of cytoplasm _H	18
1protein binding _H-2nucleotide binding _P	nucleotide binding _P-3cytoskeleton _H	16
1protein binding _H-2metal ion binding _P	metal ion binding _P-3cytoskeleton _H	16

Table O: Top 10 most frequent motifs found after performing analysis run 3
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

Run 10

Edge 1	Edge 2	Count
1protein binding _H-2cellular biosynthetic process _P	cellular biosynthetic process _P-3adaptive immune response _H	82
1protein binding _H-2cellular biosynthetic process _P	cellular biosynthetic process _P-3endoplasmic reticulum part _H	85
1protein binding _H-2cellular biosynthetic process _P	cellular biosynthetic process _P-3external side of plasma membrane _H	81
1protein binding _H-2cellular biosynthetic process _P	cellular biosynthetic process _P-3homeostatic process _H	97
1protein binding _H-2cellular biosynthetic process _P	cellular biosynthetic process _P-3regulation of homeostatic process _H	83
1protein binding _H-2cellular nitrogen compound metabolic process _P	cellular nitrogen compound metabolic process _P-3immune system process _H	81
1protein binding _H-2organic substance biosynthetic process _P	organic substance biosynthetic process _P-3adaptive immune response _H	82
1protein binding _H-2organic substance biosynthetic process _P	organic substance biosynthetic process _P-3endoplasmic reticulum part _H	85
1protein binding _H-2organic substance biosynthetic process _P	organic substance biosynthetic process _P-3external side of plasma membrane _H	81
1protein binding _H-2organic substance biosynthetic process _P	organic substance biosynthetic process _P-3homeostatic process _H	97

Table P: Top 10 most frequent motifs found after performing analysis run 10
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

Bacillus anthracis

Run 4

Edge 1	Edge 2	Count
1intracellular part -2cellular metabolic process	cellular metabolic process -3response to chemical	41
1intracellular part -2cellular metabolic process	cellular metabolic process -3response to stress	40
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3cellular response to stimulus	40
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3extracellular organelle	43
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3membrane-bounded organelle	40
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3plasma membrane	47
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3protein binding	39
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3regulation of molecular function	41
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3response to chemical	41
1intracellular part -2cellular metabolic process	cellular metabolic process -3cellular response to stimulus	38

Table Q: Top 10 most frequent motifs found after performing analysis run 4
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

Run 5

Edge 1	Edge 2	Count
1protein binding _H-2biological_process _P	biological_process _P-3membrane _H	2
1protein binding _H-2biological_process _P	biological_process _P-3regulation of cellular metabolic process _H	2
1protein binding _H-2biological_process _P	biological_process _P-3regulation of macromolecule metabolic process _H	2
1protein binding _H-2molecular_function _P	molecular_function _P-3membrane _H	2
1protein binding _H-2molecular_function _P	molecular_function _P-3regulation of macromolecule metabolic process _H	2

Table R: All the frequent motifs found after performing analysis run 5
All the frequent motifs after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

Run 6

Edge 1
1cytoplasm _H-2molecular_function _P
1cytoplasm _H-2biological_process _P

Table S: Results of analysis run 6
Run 6 resulted in 2 frequent motifs consisting of 1 edge each.

Yersinia pestis

Run 7

Edge 1	Edge 2	Count
1intracellular part -2cellular metabolic process	cellular metabolic process -3regulation of biological quality	51
1intracellular part -2cellular metabolic process	cellular metabolic process -3response to chemical	49
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3biosynthetic process	51
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3cellular metabolic process	50
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3extracellular organelle	51
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3regulation of biological quality	53
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3regulation of biological process	48
1intracellular part -2organic substance metabolic process	organic substance metabolic process -3response to chemical	50
1intracellular part -2primary metabolic process	primary metabolic process -3cellular response to stimulus	50
1protein binding -2organic substance metabolic process	organic substance metabolic process -3response to chemical	49

Table T: Top 10 most frequent motifs found after performing analysis run 7
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

Run 12

Edge 1	Edge 2	Count
1intracellular membrane-bounded organelle _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3cytoplasmic part _H	31
1intracellular membrane-bounded organelle _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3regulation of cellular process _H	32
1protein binding _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3cytoplasmic part _H	28
1protein binding _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3intracellular membrane-bounded organelle _H	25
1protein binding _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3positive regulation of biological process _H	26
1protein binding _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3regulation of cellular process _H	28
1regulation of cellular process _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3cytoplasmic part _H	26
1regulation of cellular process _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3intracellular membrane-bounded organelle _H	22
1regulation of cellular process _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3positive regulation of biological process _H	34
1regulation of cellular process _H-2macromolecule metabolic process _P	macromolecule metabolic process _P-3protein binding _H	27

Table U: Top 10 most frequent motifs found after performing analysis run 12
The 10 most frequent motifs based on counts done after having performed a grouping step based on the first and second edge of the building patterns of the motifs. The first and second column contains the first and second edge of the core motifs that represent the subgroup and the third column indicates how many subgraphs can be found in the results that are built from this core motif.

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.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ASSENOV, Y., RAMIREZ, F., SCHELHORN, S. E., LENGAUER, T. & ALBRECHT, M. 2008. Computing topological parameters of biological networks. *Bioinformatics,* 24**,** 282-4.

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

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.

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.

HUERTA-CEPAS, J., SERRA, F. & BORK, P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol,* 33**,** 1635-8.

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*.

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.

ORCHARD, S., KERRIEN, S., ABBANI, S., ARANDA, B., BHATE, J., BIDWELL, S., BRIDGE, A., BRIGANTI, L., BRINKMAN, F. S., CESARENI, G., CHATR-ARYAMONTRI, A., CHAUTARD, E., CHEN, C., DUMOUSSEAU, M., GOLL, J., HANCOCK, R. E., HANNICK, L. I., JURISICA, I., KHADAKE, J., LYNN, D. J., MAHADEVAN, U., PERFETTO, L., RAGHUNATH, A., RICARD-BLUM, S., ROECHERT, B., SALWINSKI, L., STUMPFLEN, V., TYERS, M., UETZ, P., XENARIOS, I. & HERMJAKOB, H. 2012. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods,* 9**,** 345-50.

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.

THE GENE ONTOLOGY, C. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res,* 45**,** D331-D338.