Exploring bacterial HPI networks using subgraph mining

Lorenz Van de Veken, Pieter Moris, Pieter Meysman, Kris Laukens

Introduction

Bacteria and **humans** are in a constant battle with each other. Bacteria try to take over niches in the human body to survive, manipulating host machinery, evading the immune system and potentially damaging host tissues during the process, with adverse effects for the human host as a result. The **key interface** on which this battle takes place is the **interactome** between the host and the pathogen. Many human and pathogen proteins interact with one another, thus **host-pathogen protein-protein interactions (HPIs) and the networks they form are an important tool to study the onset and progress of bacterial infections.**

Descriptive analysis

The **unit of analysis** consists of the **whole** of an **interaction** and the **annotations** of the interacting proteins. The interactions are used to build the network structure and the annotations provide biologically relevant information on the interacting proteins.

The interactions can be **visualised** using **Cytoscape**. However, due to the properties of interaction networks, a **messy, uninterpretable** graph is often the result.
Nevertheless, **descriptive graph analysis** can be performed to obtain useful information on the properties of this intangible network.

GO and pathway enrichment analysis shall be performed to determine in which **biological contexts** these HPIs most often take place. Many interactions in certain contexts might be indicative of an important role in the infection strategy.

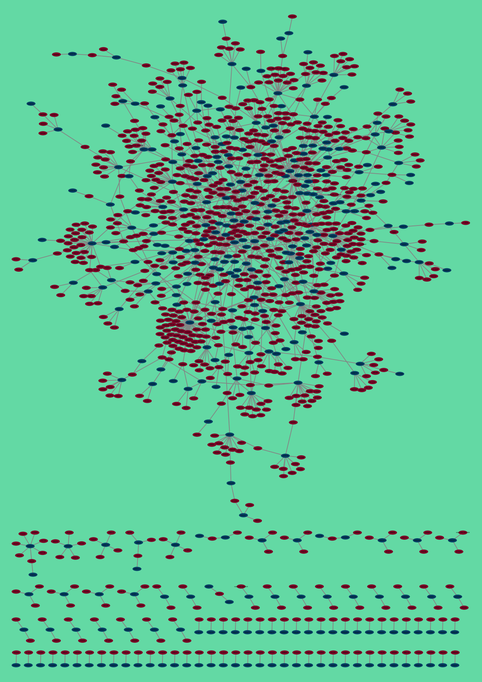

Figure 2: HPI network of Francisella tularensis. *Protein-protein interactions between F. tularensis and the human host. The data used to build the network was extracted from HPIDB 2.0. The red and blue vertices represent human and pathogen proteins, respectively.*

Expectations

We expect to find both general and species-specific interaction **patterns** linked to different aspects of the **infection strategy** and **life cycle** of the bacteria. Important biological processes that are likely to be targeted include the **immune system** and **metabolism** as the primary strategies of bacteria consist of evading or suppressing the immune system and rewiring the biochemical machinery of the host cell (in the case of bacteria with an intracellular phase) to fulfil the metabolic needs of the bacteria.

Data handling

Data is collected by querying **PHISTO, HPIDB 2.0** and **IntAct** for HPIs of *Yersinia pestis, Bacillus anthracis* and *Francisella tularensis* with their human host. The results are in the form of a MiTab, a tabular data format. Then, the datasets are tidied & homogenised, merged and quality control (QC) is performed using the Python programming language. The result is a dataset of interactions which form a biological interaction network. The nodes and edges of this network represent the proteins and the interactions between these proteins, respectively.
For each protein in this network, **GO and InterPro** terms are collected to use as node labels.

Key problems:

- Extraction of relevant features (tidying)
- Redundant entries (merging)
- Variable export file formatting (merging)
- Inherent false positivity of interaction data (QC)

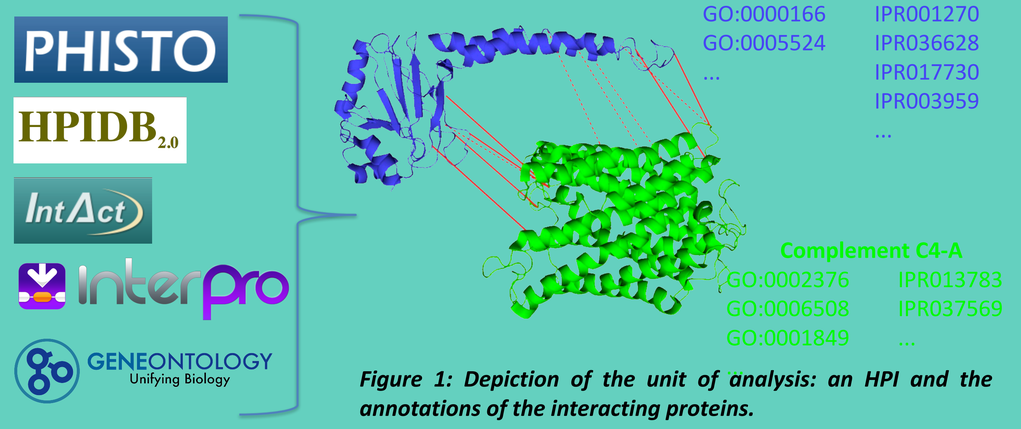

Figure 1: Depiction of the unit of analysis: an HPI and the annotations of the interacting proteins.

Subgraph mining

Subgraphs are in essence smaller graphs which are part of a given graph. Applied to our field of study, subgraphs represent **interaction patterns**. For example, a kinase of the human host which phosphorylates two bacterial proteins could be such a pattern. Subgraph **mining** refers to methods which **search** for **subgraphs with certain properties** in one or multiple given graphs. Most often **frequent subgraph mining** is performed which searches for frequently occurring subgraphs.

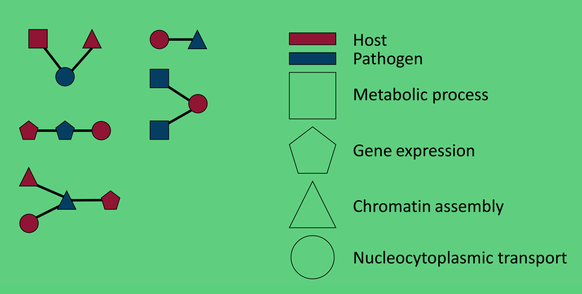

Figure 3: Subgraph examples
Some examples of subgraphs. The color indicates whether the vertex represents a pathogen or a host protein and the shape of the vertex represents the label the vertex was given.

However, it might be that the **interesting patterns** are not those that frequently occur in the given interaction network(s) but those that are **associated with certain proteins**, known to be relevant to disease susceptibility, pathogenesis, drug resistance, etc.
Therefore, **different subgraph analysis setups** will be used to explore the different aspects of the bacterial HPI networks.

Contact info: LorenzVandeVeken@student.uantwerpen.be

ADReM Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Belgium
Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital, Belgium

Universiteit Antwerpen