Visual analytics on the spread of pathogens

Julien Herrmann
The Ohio State University
herrmann.136@osu.edu

Zachary L. Witter
UNC Charlotte
zwitter1@uncc.edu

Nakul Patel
UNC Charlotte
npate114@uncc.edu

Jonathan Kho
The Ohio State University
kho.19@osu.edu

Daniel A. Janies
UNC Charlotte
djanies@uncc.edu

Ümit V. Çatalyürek
The Ohio State University
umit@bmi.osu.edu

1. INTRODUCTION

Emerging infectious diseases represent critical concerns for force protection and public health. Thus far, most efforts have focused on single diseases and data types. In this project we focus on medical, genetic, host, temporal and geographic data for any type of infectious disease. We leverage the biomedical community's vast sequence based record of infectious disease variants as expressed in the National Center for Biotechnology Information's Genbank (NCBI) population genetic dataset collection (popset).

We have created a web-based application, the Pathogen Dynamic Graph (PDG), that allow users to build graphs joining these data types. The graph is pulled from the Genbank records and is efficiently stored in a local database. The nodes of the graph are entities of the popset sequence record as reported by subject matter experts. These nodes include organism, host, gene, protein, location and disease. The edges between nodes mean that for a popset record the nodes co-occur (e.g., an organism is sequenced for a particular gene). Taken as a whole, the linked data will allow a comprehensive understanding of what is known about an infectious diseases network. These graphs 1) expose new linkages among data types with visualization thus inspiring new hypotheses 2) allow users to reason on the data by graph mining.

2. AN EXAMPLE FROM MALARIA

We did a search of NCBIs Popset database with one search term "Malaria". From this search we produced a large graph with over 4,000 nodes and 11,000 edges. Once loaded in the PDG, the user can filter the data and search for terms of interest thus visualizing manageable subgraphs. In the case of zoonotic diseases the user may be interested in *Plasmodium knowlesi* - a potentially zoonotic malaria parasite. *Plasmodium knowlesi* typically infects macaques but *Plasmodium knowlesi* has begun to infect humans in Asia. We present a graph for *Plasmodium knowlesi* in figure 1. This graph shows the several host species (mosquito vectors, humans, and macaques), and many genes and proteins in the network for *Plasmodium knowlesi.* The genes and proteins may be functional genes (e.g., apical membrane antigen which is a key protein the parasite uses to invade host cells) or markers for diversity studies (e.g., cytochrome b).

The user can right click on any node to extend the reach of the graph beyond the current subgraph being visualized and thus navigate to other subgraphs of interest.

3. USAGE

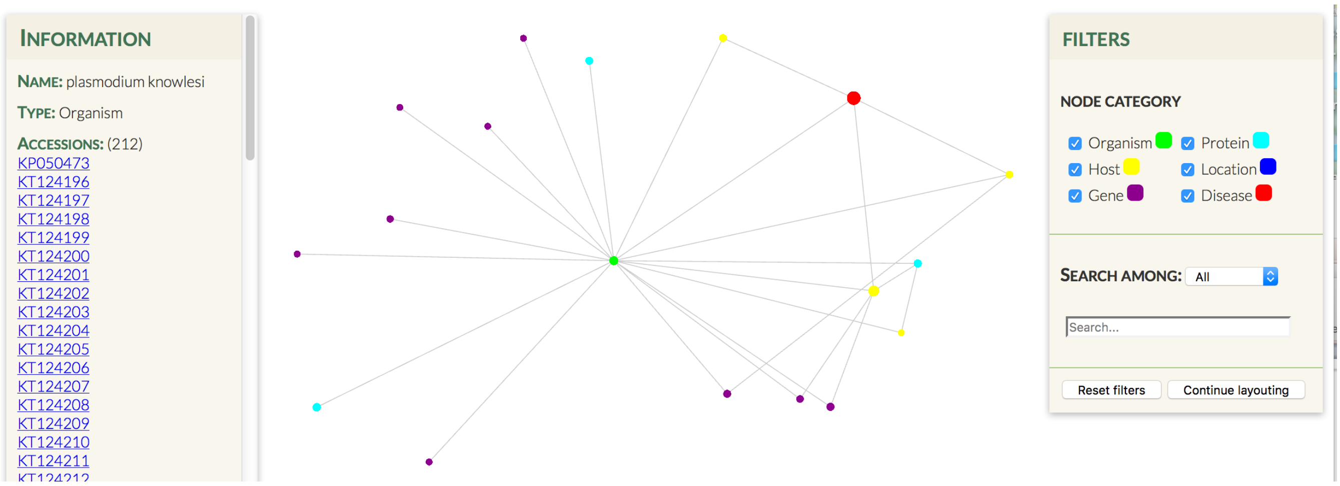
Depending on the size of the data, the entire PDG can be complex. The user of the PDG has several options to filter and query the PDG to ask questions about disease networks. First, there are several toggle buttons to turn on or off nodes of certain metadata type. Second, there is a text-based search field that can be used to find a subgraph with a node of interest at the center (Figure 1). Third, the user is presented with a superset of the population genetics data. The user can navigate back to the original data in Genbank by the links in the superset and or touch nodes of the graph to move to new subgraphs of interest. For example, *Plasmodium falciparum* a common malarial parasite for humans is shown in Figure 2.

4. CONCLUSIONS

As has been seen with many emergent pathogens, such small changes in the surface proteins of an animal virus can allow the virus to recognize human cells can effect an animal to human shift of the the virus. Much of the data required to assess these risks is in databases that do not lend themselves to query and response formats that can address these questions. The PDG allows the user to craft queries that provide insight on where and how pathogens emerge as zoonotic. Moreover the pathogen genes and human genes that mediate the interaction are presented. Thus we are developing new ways for the user to interact with a large complex data repository on pathogens like NCBI's Genbank.

Acknowledgments

This work was partially supported by the Defense Threat Reduction Agency grant HDTRA1-16-C-0010.



INFORMATION

NAME: plasmodium knowlesi

TYPE: Organism

ACCESSIONS: (212)

KP050473
KT124196
KT124197
KT124198
KT124199
KT124200
KT124201
KT124202
KT124203
KT124204
KT124205
KT124206
KT124207
KT124208
KT124209
KT124210
KT124211
KT124212

FILTERS

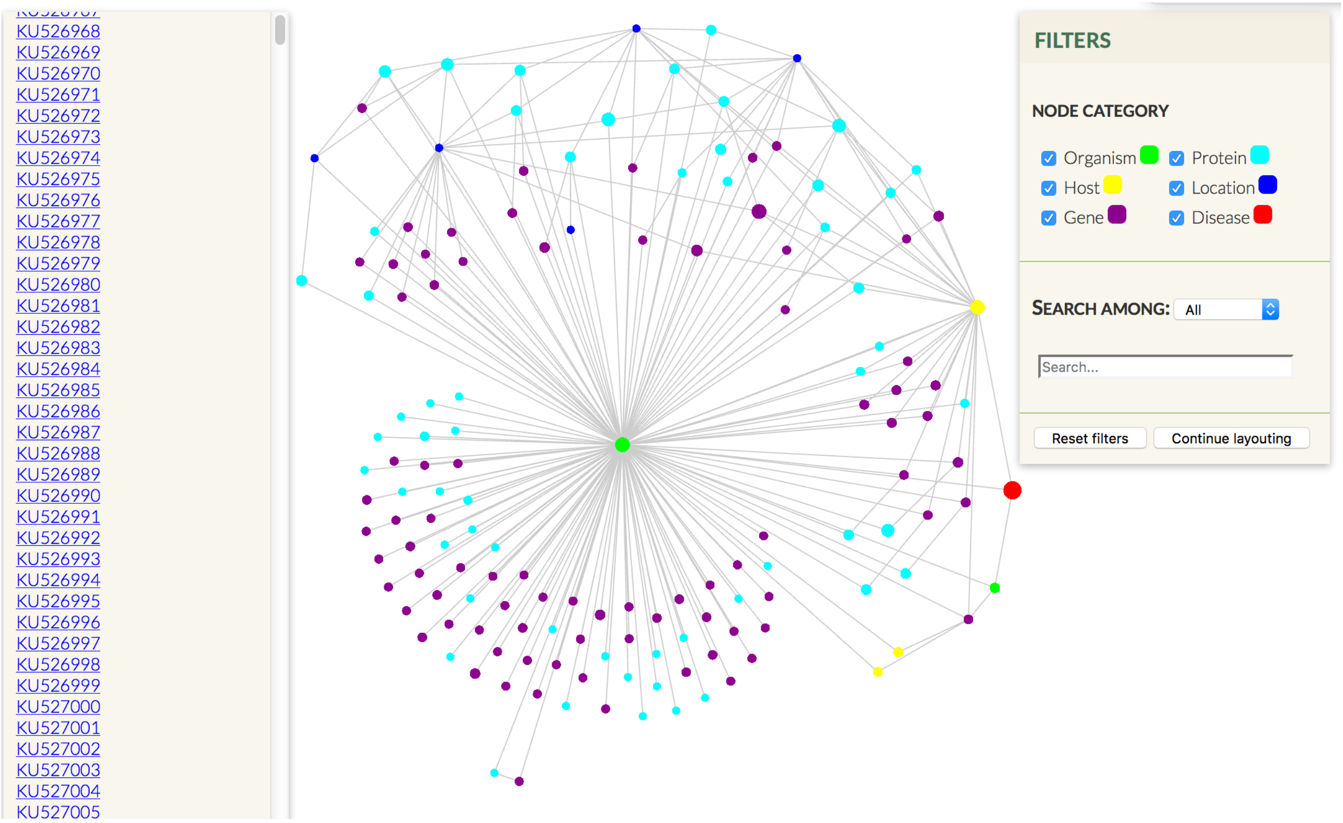
NODE CATEGORY

☑ Organism 🟩 ☑ Protein 🟦
☑ Host 🟨 ☑ Location 🟦
☑ Gene 🟪 ☑ Disease 🟥

SEARCH AMONG: All

Search...

Figure 1: Screenshot of the PDG application in which a user has identified a network in which *Plasmodium knowlesi* interacts with *Homo sapiens*. On the left side of the screen are relevant links to popset data. On the right side of the screen is the legend for the graph. When the user clicks on any node in the graph, the subgraph of that node and its neighbors is displayed.



KU526967
KU526968
KU526969
KU526970
KU526971
KU526972
KU526973
KU526974
KU526975
KU526976
KU526977
KU526978
KU526979
KU526980
KU526981
KU526982
KU526983
KU526984
KU526985
KU526986
KU526987
KU526988
KU526989
KU526990
KU526991
KU526992
KU526993
KU526994
KU526995
KU526996
KU526997
KU526998
KU526999
KU527000
KU527001
KU527002
KU527003
KU527004
KU527005

FILTERS

NODE CATEGORY

☑ Organism 🟩 ☑ Protein 🟦
☑ Host 🟨 ☑ Location 🟦
☑ Gene 🟪 ☑ Disease 🟥

SEARCH AMONG: All

Search...

Figure 2: Screenshot of the PDG application showing a network for *Plasmodium falciparum*, a malaria parasite, several human and animal hosts, and the genes and proteins that mediate their interactions.