AMR-TV: antimicrobial resistance transmission visualizer

Ivan S Gill
isgill93@student.ubc.ca

Introduction

Antimicrobial resistance, or AMR, describes microorganisms that have evolved resistance against antimicrobial medication (Levy, 1982). The evolution of AMR genes is largely driven by antimicrobial medication overuse, which occurs through channels such as self-medication (Rather *et al.*, 2017), clinical prescriptions (Luyt *et al.*, 2014), livestock management (Tang *et al.*), and the use of antimicrobial pesticides (Ramakrishnan *et al.*, 2019). After AMR evolves in a bacterial population, the associated genes can spread quickly to other populations via horizontal gene transfer, as many AMR genes occur on mobile genetic elements (Moser *et al.*, 2018). Due to both horizontal gene transfer and global antimicrobial misuse, AMR is currently evolving and spreading rapidly, with up to 10 million AMR-attributed deaths expected annually by 2050 (O'Neill, Jim, 2016). One method for mitigating the rate of AMR-induced illnesses and fatalities is through the improvement of AMR surveillance. By tracking the spread of AMR more effectively, clinicians and policymakers can make more informed decisions on the prescription and supply of appropriate antimicrobial medications in response to AMR outbreaks.

We will develop AMR-TV (antimicrobial transmission visualizer) to visualize potential routes of AMR transmission between groups of isolates sampled from various pathogenic organism groups. Using data from the NCBI Pathogen Detection Isolates Browser (National Center for Biotechnology Information, 2020), we will create an interactive adjacency matrix that summarizes levels of transmission within and across organism groups, and also allows users to narrow down the large NCBI dataset to a subset of interest. We will use node-link diagrams to explicitly illustrate potential routes of AMR transmission between groups of isolates from the user-specified subset, and we will also use table views to provide further information on the isolates within each group visualized in the node-link diagram.

My personal expertise with AMR is limited, as I have only begun investigating AMR in detail this term. The lab I started my MSc with has done research on AMR in the past, and proposed a project dedicated to visualizing AMR transmission for my thesis. This project I am proposing will be the first component of my MSc thesis. After completing this course project, I will expand AMR-TV in three ways. First, I will allow users to upload their own spreadsheets, rather than visualize a static rendition of the NCBI Pathogen Detection Isolates Browser. Second, I will integrate an ontology framework to automatically clean and standardize user-submitted data. Third, I will make changes to the user interface as requested when I present AMR-TV to the appropriate stakeholders, such as clinicians and microbiologists.

Related work

Most tools developed to visualize the movement of genetic material focus on vertical transmission. One example of this is T-REX (Boc *et al.*, 2012), a web server that uses genetic sequence data to construct a traditional phylogenetic tree, and then superimposes dashed edges between branches to indicate horizontal gene transmission. Given the degree to which AMR genes spread horizontally, such a representation of AMR pathogens could potentially produce a hairball of dashed edges over a relatively simple tree. To visualize large-scale horizontal gene transmission networks more clearly, we must move beyond attempts to confine them within traditional phylogenetic trees.

Within microbiology, visualizing transmission beyond vertical phylogenetic relationships has been attempted by tools designed to visualize the movement of microorganism populations across space. GenGIS is a software package that visualizes a network of nodes containing spatial information, by superimposing the nodes directly on a geographic map (Parks *et al.*, 2009). Although a system like this might accommodate more nodes than T-REX, scalability remains an issue. Assigning importance to the positional location of nodes in a node-link diagram prevents the usage of force-directed algorithms, which increases the possibility of occlusion with larger datasets. Microreact is an application that mitigates this problem by instead displaying single markers on a geographic map (Argimón *et al.*, 2016), which the users can select to see the force-directed node-link diagram underlying each marker in a separate view. This is a superior solution for two reasons. First, it allows the generation of much larger node-link diagrams, capable of elucidating non-rudimentary transmission routes. Second, it introduces the concept of summarizing data the user can filter down from, which is of critical importance when navigating AMR transmission routes from exceptionally large databases like the NCBI Isolates Browser.

Interestingly, by attempting to mitigate these issues of scalability, AMR-TV will feature a solution that is most similar to several applications designed to visualize social networks. Similar to AMR transmission networks, the relationships between actors of a social network are most intuitively represented as node-link diagrams, but the density of these networks necessitates some mechanism for summarizing and filtering data. Many social network visualization applications accomplish this through an adjacency matrix coupled with a node-link diagram. However, this coupling is configured in various ways. VAIM is an application that uses its adjacency matrix to summarize the density of its coupled node-link diagram at every x and y position (Arleo *et al.*, 2020). Users can jump to positions of the node-link diagram with specific node densities by selecting cells from the matrix. This functionality would not make sense for AMR-TV, where the goal is to visualize clear paths between individual nodes of interest, rather than detect clusters. MatLink and MatrixExplorer both use their adjacency matrices to summarize connections between node groups (Henry and Fekete, 2007, 2006), which is similar to how AMR-TV's adjacency matrix will summarize connections between organism groups. However, MatLink uniquely replaces a separate node-link diagram view in favour of overlaying the links between nodes directly on the edges of the matrix. This is not an acceptable solution for AMR-TV, as it obscures the directionality of links, where paths between multiple nodes begin, and individual node attributes not encoded within the summarizing matrix. MatrixExplorer offers the solution most similar to AMR-TV's proposed solution, by allowing users to filter a subset of data via matrix cell selection prior to node-link diagram generation. However, MatrixExplorer does not group nodes until users select a node belonging to a specific cluster, while AMR-TV will always keep sampled isolates grouped by certain attributes, to reduce the large number of duplicate paths in the node-link diagram for some subsets of data.

Data abstractions

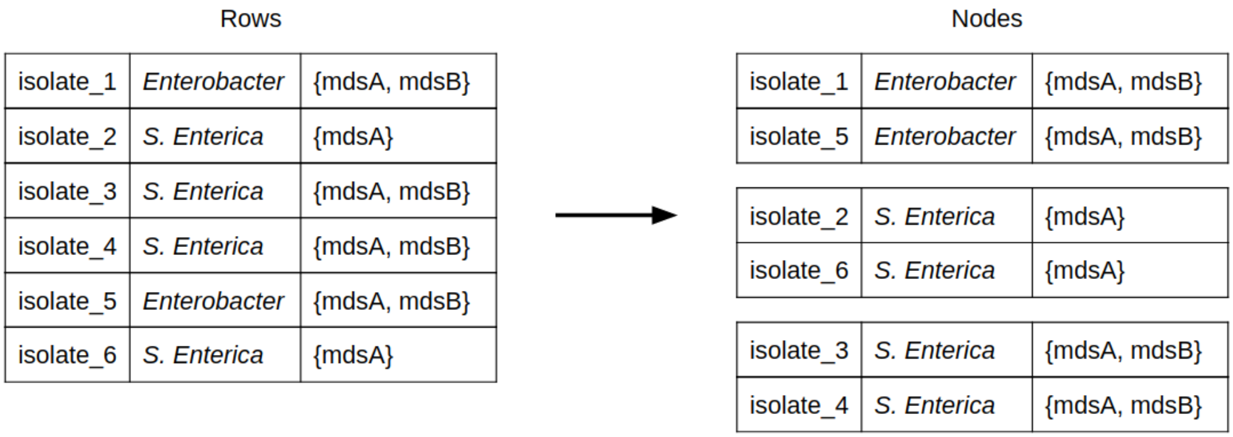
AMR-TV will use a static dataset, consisting of ~700k rows of data downloaded from the NCBI Isolates Browser on Oct 31, 2020. The NCBI Pathogen Detection Isolates Browser has seven columns relevant to AMR-TV's goal of visualizing potential AMR transmission events. These columns are “organism group”, “isolate”, “create date”, “location”, “isolation source”, “host”, and “AMR genotypes”. There are some missing and unstandardized values in this dataset, particularly under “isolation source” and “host”, but data cleaning is not within the scope of this project.

The primary dataset type modelled by AMR-TV will be a network. Nodes in this network will encode NCBI rows aggregated by “organism group” and “AMR genotypes”, as demonstrated by the example in Fig. 1. Each node will have three attributes: “organism group”, “AMR genotypes”, and “minimum create date”. “Organism group” is a categorical value, of which there are 33 distinct values in the NCBI Isolates Browser. “AMR genotypes” are collections of one or more categorical values, of which there are ~3k distinct unnested values in the NCBI Isolates Browser. “Minimum create date” is the minimum “create date” value of all NCBI rows aggregated within a node. It is an ordinal value, ranging from Dec 15, 2010 to Oct 31, 2020. Nodes will be directionally linked to nodes with both a higher “minimum create date” value, and an escapsulating collection of “AMR genotypes”. These links will represent potential routes of AMR gene transmission.

The scale of this network is large, when the data is completely unfiltered prior to aggregation. There are ~77k nodes over the entire ~10 year period. The number of edges over this period is difficult to determine, as the underlying database takes a long time to perform this calculation. However, over the period of Oct 2020, there are ~5k nodes, with an average of ~17 edges per node. See Table 1 for some measurements of the network when filtered by other conditions.

The rationale behind aggregating the rows in the NCBI Isolates Browser to create individual nodes, rather than using the rows as nodes themselves, is to improve both readability and utility. On average, there are ~9 rows with the same values for “organism group” and “AMR genotypes”. Since transmission routes are determined solely by the values of “AMR genotypes”, this means that every unnested row would have an identical set of links, and a completely unnested node-link diagram would produce a large amount of identical paths. For the purposes of monitoring an AMR outbreak, we believe it more informative to visualize the unique routes of AMR transmission.

However, AMR-TV will allow users to gain further information on the rows nested inside each node by storing said rows in table dataset types. The attributes of these tables will be the remaining columns from the NCBI Isolates Browser mentioned earlier: “isolate”, “location”, “isolation source”, and “host”. “Isolate” is a categorical identifier, so there are ~700k distinct values. “Location”, “isolation source”, and “host” are all categorical, with ~4k, ~15k, and ~1k unique values respectively.



Rows

isolate_1	*Enterobacter*	{mdsA, mdsB}
isolate_2	*S. Enterica*	{mdsA}
isolate_3	*S. Enterica*	{mdsA, mdsB}
isolate_4	*S. Enterica*	{mdsA, mdsB}
isolate_5	*Enterobacter*	{mdsA, mdsB}
isolate_6	*S. Enterica*	{mdsA}

Nodes

isolate_1	*Enterobacter*	{mdsA, mdsB}
isolate_5	*Enterobacter*	{mdsA, mdsB}

isolate_2	*S. Enterica*	{mdsA}
isolate_6	*S. Enterica*	{mdsA}

isolate_3	*S. Enterica*	{mdsA, mdsB}
isolate_4	*S. Enterica*	{mdsA, mdsB}

Fig. 1. A visualization of how rows from the NCBI Isolates Browser are aggregated into the nodes used in AMR-TV’s network. In this example, 6 rows have been aggregated into 3 nodes. Aggregation occurs by organism group and AMR genotypes.

Filters	**# of nodes**	**Avg # of edges per node**
None	~77k	Unknown
Oct 1, 2020 to Oct 31, 2020	~5k	~17
Oct 1, 2020 to Oct 31, 2020 Enterobacter isolates only	190	~5
Oct 1, 2020 to Oct 31, 2020 *Enterobacter* and *Campylobacter jejuni* isolates only	384	~8
Oct 1, 2020 to Oct 31, 2020 *Salmonella Enterica* isolates only	439	~20

Table 1. The scale of networks derived by AMR-TV over various subsets of data. The number of edges with no filter is unknown, because it was too computationally expensive to determine.

Task abstractions

At a high-level, the intended actions of AMR-TV are to derive and discover. AMR-TV will derive nodes by aggregating rows from the NCBI Isolates Browser, and will derive links by examining the overlap of AMR genes between nodes. Users will discover potential AMR transmission routes between groups of sampled isolates by analyzing paths within this derived network. At a mid-level, AMR-TV will allow users to browse data, as users interested in examining the various transmission routes of a specific AMR outbreak will be able to filter the entire dataset by “create date” and “organism group”. At a low-level, AMR-TV will summarize data, by displaying all potential transmission routes between groups of similar isolates belonging to the user-specified subset of data.

The intended target of AMR-TV will be the paths present in the network data, which will indicate proposed routes of AMR transmission, but there are other features describing all data that will be targeted as well. Encoding “organism group” and “AMR genotypes” will inform clinicians of symptoms, outcomes, and which antimicrobial medications will not work during AMR outbreaks. Encoding “location”, “isolate source”, and “host” in the tables mapped to each node will allow users to consider AMR gene transmission routes in the context of any physical routes that may provide the proximity needed for horizontal gene transfer to occur, such as geographic routes and interspecies relationships.

Solutions

AMR-TV will consist of three views: an adjacency matrix, node-link diagram, and table.

The adjacency matrix view, as seen in Fig. 2, will have rows and columns encoding organism groups such as *Salmonella enterica*, *E.coli and Shigella*, and *Neisseria*. Cells will use the luminance channel to encode the number of potential AMR transmission routes within the “organism group” values represented by each cell. Darker cells will represent a larger number of shared AMR transmission routes. Users will be able to interact with the matrix in two ways. First, users can specify a date range to narrow down the amount of data visualized by the adjacency matrix. Second, users can select cells from the adjacency matrix to produce a coupled node-link diagram. To utilize colour-encoding, users will not be able to select cells from more than 12 organism groups. It is unlikely that users concerned with AMR outbreaks will need to examine more than 12 organism groups, but we may consider removing this limitation and producing the low-level view in black and white.

The node-link diagram view will resemble the visualization seen in Fig. 3A, although usually with many more nodes and edges. Nodes will encode groups of sampled isolates. Directed links will encode encapsulated AMR genotypes, or potential AMR transmission events. Dotted links will indicate genes shared by isolates belonging to the same organism group. The luminance channel will encode the minimum “create date” of each node, with lighter nodes representing older nodes. Colour hue will encode the “organism group” of nodes. Selecting a node will display the “AMR genotypes” of that node, as well as a table containing further information on the original NCBI rows aggregated within that node, as seen in Fig. 3B.

While designing AMR-TV, we considered replacing the node-link diagram with a second adjacency matrix. This was primarily driven by the fear that some subsets of the NCBI data have more than an average of four edges per node, which can reduce readability. However, we decided against this for three reasons. First, the ultimate goal of AMR-TV is to trace potential paths of AMR transmission, and a node-link diagram is by far the most intuitive way to do this. Second, we also believe that eventually standardizing fields like “location” and “host” in the future, and then allowing users to filter by these fields, will significantly reduce the number of edges in many problematic subsets of data. Third, many current subsets of data have a reasonable number of edges, with most problems stemming from select organism groups that are sampled much more often than others, such as *S. Enterica*.

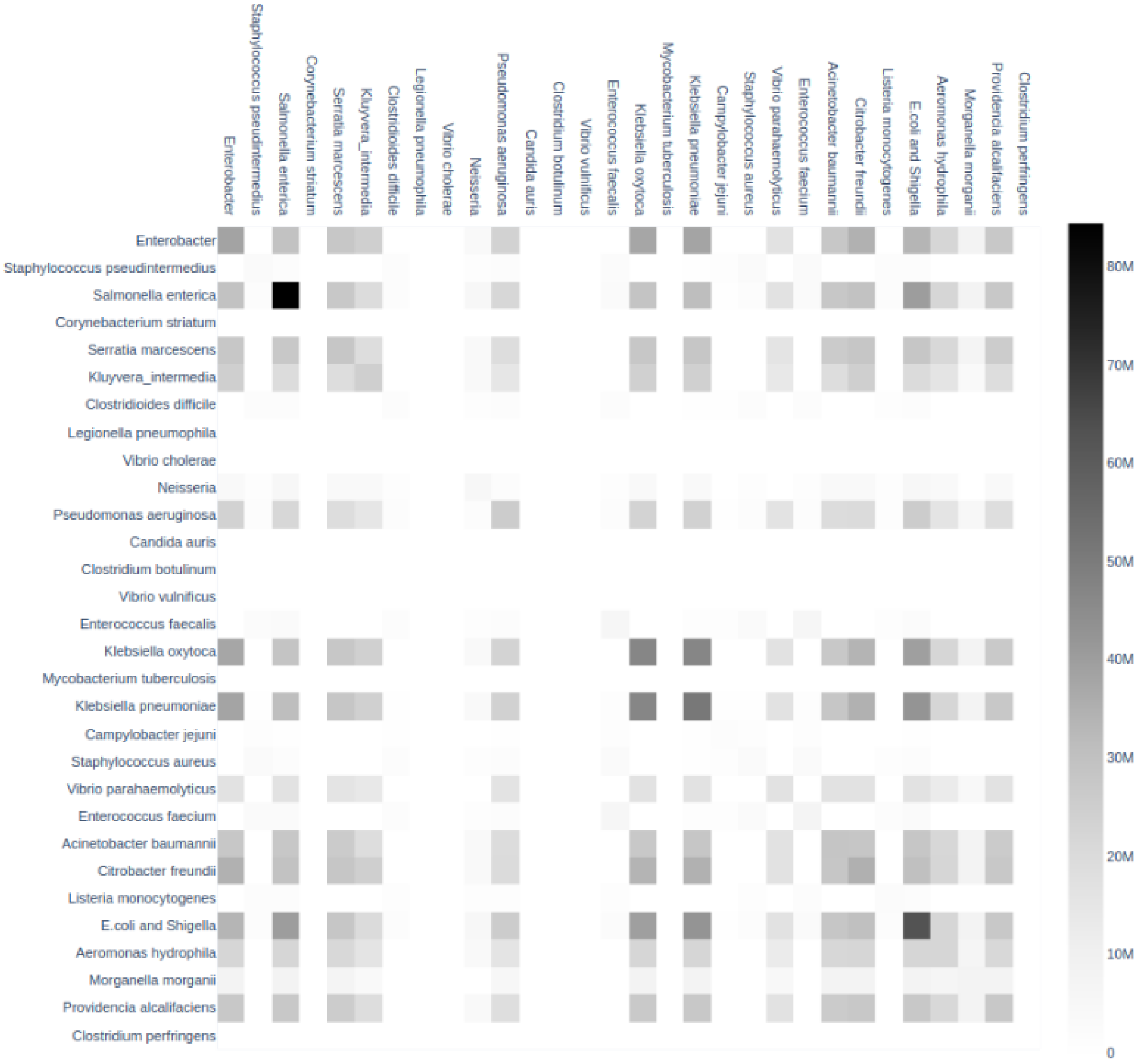
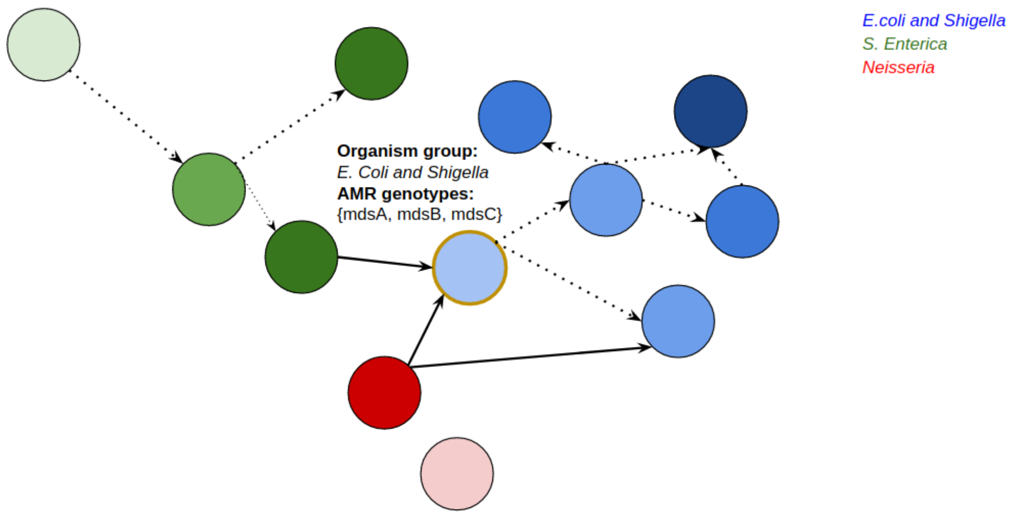
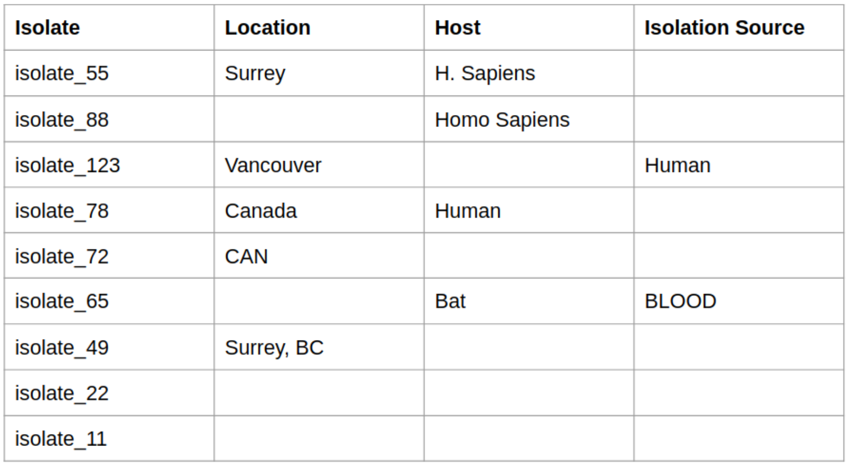


Fig. 2. Preliminary adjacency matrix using the luminance channel to illustrate the number of potential AMR transmission events between isolates of varying organism groups, over the period of Oct 1, 2020 to Oct 31, 2020. The numbers on the colour bar are not currently accurate. This matrix was constructed without aggregating isolates, so the numbers are magnitudes higher than they should be. In the future, users will be able to select the cells of this matrix to produce the node-link diagrams seen in Fig. 3A. There will also be widgets to specify the period of time.

A)



B)



Isolate	Location	Host	Isolation Source
isolate_55	Surrey	H. Sapiens	
isolate_88		Homo Sapiens	
isolate_123	Vancouver		Human
isolate_78	Canada	Human	
isolate_72	CAN		
isolate_65		Bat	BLOOD
isolate_49	Surrey, BC		
isolate_22			
isolate_11			

Fig. 3.
A) Node-link diagram generated after selecting cells from the adjacency matrix. Three different organism groups are differentiated by colour hue. Darker colours represent nodes with a higher “minimum create date” value. Outgoing links join nodes with AMR genotypes overlapped by the AMR genotypes of darker nodes. The user has selected an *E. coli and Shigella* node, revealing the AMR genotypes of that particular node, and generating the table in Fig. 3B.
B) Table detailing the aggregated isolates of the node selected in Fig. 3A. With the exception of “isolate” values, users can expect the values inside this table to be unclean and unstandardized. However, it is possible that some tables could provide enough information to consider AMR transmission routes in the context of geospatial and interspecies relationships.

Implementation

We will implement AMR-TV using Docker, with a Django container to display the interactive visualizations, and a Postgres container to store the NCBI Pathogen Detection Isolates data. We will use the Python library NetworkX to generate networks. NetworkX contains many different algorithms for generating graphs, but we may try other algorithms in the literature if time permits. We will use the Python library Plotly to draw the adjacency matrix and table views, and by using the data supplied by NetworkX, we will use Plotly to draw the node-link diagram view too. Plotly is interactive and reactive, so we should be able to implement the cell and node selection functionalities described in Fig. 2 and Fig. 3.

Results

Users that want to monitor potential transmission events during an AMR outbreak of interest will begin by inputting start- and end-date values under the adjacency matrix view. Users can then select cells from organism groups they are interested in, or click cells that are dark enough to warrant further exploration. AMR-TV will then generate a node-link diagram view of the user's selected cells. It should be noted that when users consider the darkness of cells, they would expect darker cells on the main diagonal of the adjacency matrix, due to pathogens from the same organism group possessing an increased chance of sharing AMR genes due to common ancestry.

Using the node-link diagram, users then begin to infer potential AMR transmission events between groups of isolates. By examining the colour hue of the nodes, users can infer transmission events between different organism groups. Users can infer the order of transmission, by observing the luminance of nodes, and the directionality of link arrowheads. To infer potential routes of transmission facilitated by geography, or interspecies relationships, users can select nodes for a table providing further details on the isolates nested within each node.

Discussion and Future Work

There are two limitations to this proposed implementation of AMR-TV. The first limitation is the lack of standardization on attributes other than “create_date”, “organism_group”, and “amr_genotypes”. This prevents any reliable data filtering based on geospatial and host information, which are all useful metrics to consider when monitoring a specific AMR outbreak, or for refining any dense node-link diagram to improve readability. The second limitation is the estimation of horizontal gene transmission without sophisticated bioinformatics methods. Ordinarily, horizontal gene transmission would be inferred through the sequence composition and evolutionary history of AMR genes (Ravenhall *et al.*, 2015). AMR-TV's method of inferring horizontal gene transmission by simply comparing AMR genotypes will produce transmission routes of varying likelihood. Ideally, AMR-TV would be the final component of a pipeline, with earlier components using sequence analysis to assign a likelihood to each node-link relationship within a dataset, which could then be used to filter the visualized network for more likely transmission routes. However, the primary purpose of this project within the scope of this course is to conduct a technique driven exploration of encoding and interaction ideas when visualizing AMR transmission, which we believe we can still accomplish by considering transmission routes of varying likelihood.

Milestones

Setup Django+Postgres framework in Docker container, with NetworkX and Plotly installed

- Hours: 4
- Completion: Oct 26

Create Django fixture containing NCBI data, and load it into Postgres database

- Hours: 8 hours
- Completion: Nov 3

Create API calls for querying database that will be needed for user-specified visualizations

- Hours: 16
- Completed: Nov 13

Using plotly, create adjacency matrix with shaded cells and colour bar, using data from some stub, static period of time

- Hours: 16
- Completion: Nov 16

Refine stub matrix to use aggregated groups of isolates, instead of individual isolates

- Estimated hours: 1
- Estimated completion: Nov 20

Using plotly and networkx, generate directionless and colourless node-link diagram for stub data used in adjacency matrix

- Estimated hours: 8
- Estimate completion: Nov 24

Add colour, luminance, and directionality to node-link diagram

- Estimated hours: 9
- Estimated completion: Nov 26

Set-up ability to to select nodes in stub node-link diagram, and see tables

- Estimated hours: 8
- Estimated completion: Nov 28

Set-up time widgets to allow users to select time period of interest, and dynamically update adjacency matrix and node-link diagram

- Estimated hours: 8 hours
- Estimated completion: Nov 30

Bibliography

Argimón,S. *et al.* (2016) Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics*, **2**.

Arleo,A. *et al.* (2020) VAIM: Visual Analytics for Influence Maximization. *ArXiv200808821 Cs*.

Boc,A. *et al.* (2012) T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.*, **40**, W573–W579.

Henry,N. and Fekete,J. (2006) MatrixExplorer: a Dual-Representation System to Explore Social Networks. *IEEE Trans. Vis. Comput. Graph.*, **12**, 677–684.

Henry,N. and Fekete,J.-D. (2007) MatLink: Enhanced Matrix Visualization for Analyzing Social Networks. In, Baranauskas,C. *et al.* (eds), *Human-Computer Interaction – INTERACT 2007*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 288–302.

Levy,S.B. (1982) Microbial resistance to antibiotics. An evolving and persistent problem. *Lancet*, **2**, 83–88.

Luyt,C.-E. *et al.* (2014) Antibiotic stewardship in the intensive care unit. *Crit. Care*, **18**, 480.

Moser,K.A. *et al.* (2018) The Role of Mobile Genetic Elements in the Spread of Antimicrobial-Resistant Escherichia coli From Chickens to Humans in Small-Scale Production Poultry Operations in Rural Ecuador. *Am. J. Epidemiol.*, **187**, 558–567.

National Center for Biotechnology Information (2020) Isolates Browser. Bethesda (MD): National Library of Medicine (US).

O'Neill, Jim (2016) Tackling Drug-resistant Infections Globally: Final Report and Recommendations. HM Government and the Wellcome Trust, London.

Parks,D.H. *et al.* (2009) GenGIS: A geospatial information system for genomic data. *Genome Res.*, **19**, 1896–1904.

Ramakrishnan,B. *et al.* (2019) Local applications but global implications: Can pesticides drive microorganisms to develop antimicrobial resistance? *Sci. Total Environ.*, **654**, 177–189.

Rather,I.A. *et al.* (2017) Self-medication and antibiotic resistance: Crisis, current challenges, and prevention. *Saudi J. Biol. Sci.*, **24**, 808.

Ravenhall,M. *et al.* (2015) Inferring Horizontal Gene Transfer. *PLOS Comput. Biol.*, **11**, e1004095.

Tang,K.L. *et al.* Restricting the use of antibiotics in food-producing animals and its associations with antibiotic resistance in food-producing animals and human beings: a systematic review and meta-analysis. *Lancet Planet. Health*, **1**, e316.