Data Exploration PHP2610 Project

Breanna Richards

2022-10-11

Data Exploration

The primary data source that we're using for this study is the National Center for Biotechnology Information's (NCBI) Pathogen database. This database includes information about sequenced bacterial pathogens that originate from a food source. It records isolates with source and genetic information. Recall that our study aims to assess if we can predict future isolates' performance against antibiotics. More specifically, can we learn from past isolates to develop a system to recommend antibiotics to new isolates? What kind of patterns can we extract from our current data about behaviors of drug susceptibility and resistance?

The focus of the study started with the Salmonella enterica species. This was partially due to the fact that Salmonella is one of the most common germs that causes foodborne illness in the United States. Due to the nature of the NCBI data, the column containing information about drug resistance and susceptibility is more often then not missing. Since our research question is contingent upon information about how these isolates interact with antibiotics, we only consider cases with non-missing information for this particular variable. Thus, the records that we had available to us to assess our study question was sparse. This led to the decision to extend our analysis outside of Salmonella to also include isolates from the E.coli and Shigella species as well as the Campylobacter jejuni species not only to extend our records but to also possible assess the similarities and differences between the three species.

And so, filtering records from all three of the species, our final dataset contained 8,672 records for Salmonella, 3,880 records for E.coli and Shigella, and 3,497 records for Campylobacter jejuni for a total of 16,049 records. Let's start by examining our variables of interest.

Variables

There are several variables in the NCBI dataset that we are interested in considering in our study.

Organism group

This variable refers to the taxonomy group that the isolate belongs to. For the purpose of our study this is going to take on one of three values: Salmonella enterica, E.coli Shigella or Campylobacter jejuni.

Isolate

The isolate variable is unique for each record in our data and is how we identify individual organisms that have little genetic mixing.

Strain

The strain variable gives the microbial strain name. It's used to distinguish a genetically distinct lineage separated from another strain by one or two mutations. Different strings of strains indicate different genetic variants or subtypes of microorganisms.

Collection Date

Collection date gives the date that the sample was collected.

Location

The geographical origin of the sample.

Source type

Source type gives the general category that an isolate originates from. Some examples of values of source type are ‘Human’, ‘Animal feed’, ‘Environmental’ and ‘Food’.

Host and Host Disease

The ‘Host’ variable refers to the host species of the isolate. Examples include Homo sapiens, Bos taurus, pig, chicken, poultry, swine, and bovine. Host disease ties the isolate to a disease origin. Examples include gastroenteritis, salmonellosis, and diarrhea.

Serovar

The serovar gives the distinct variation within a species of bacteria. This column groups isolates based on their surface antigens, allowing the classification of isolates to the subspecies level.

Outbreak

The outbreak variable is a way to group isolates that originated due to the same breakout. It is a submitter-given name for the occurrence of more cases of a disease than expected among a specific group of people or within a specific area over a period of time.

Isolation Source and Isolation Type

Isolation source describes the physical, environmental and/or local geographical source of the biological sample from which the sampled was derived. Examples include boneless beef, turkey, river, creek water, and water filtration membrane. Isolation type generalizes the isolation source into either clinical or environmental.

AST Phenotypes and AMR Genotypes

The AST Phenotype column gives the antibiotic resistant phenotype of the isolate. This variable is key to our study and lists antibiotics that the isolate has demonstrated resistance, susceptibility, intermediate responsiveness, and/or dose-dependent susceptibility to. AMR genotypes gives the antimicrobial resistant genes found in the isolate.

Computed Types

This formula gives the antigen formula and serotype results from an in-silico experiment. The antigen formula gives the detected presence of specific viral antigen which indicates viral infection. The serotype groups isolates by their distinctive surface structures/antigens.

SNP Cluster

A single nucleotide polymorphism (SNP) is a genomic variant at a single base postion in DNA. It represents the difference in a single DNA building block. A SNP cluster is a group of isolates whose genome assemblies are closely related, depending on the clustering methodology used. This column gives each isolate’s pathogen SNP cluster accession.

Min-Same and Min Diff

Min-same is the minimum SNP distance to another isolate of the same isolation type (clinical or environmental). Min-diff is the minimum SNP distance to another isolate of a different isolation type. For example, the minimum SNP difference from an environmental isolate to a clinical isolate.

We start by assessing the missingness in our data.

Table 1 gives the variables with missingness and orders them from most missing to least missing. As a rule of thumb, we may want to omit the variables with over 90% missingness as they may not be too helpful or

Table 1: Isolates Info from Most to Least Missing

	% Missing	Count Missing
outbreak	1.0000000	16049
host_disease	0.9768210	15677
host	0.9343261	14995
serovar	0.4806530	7714
computed_types	0.4596548	7377
min_diff	0.3587139	5757
min_same	0.2050595	3291
snp_cluster	0.1925354	3090
strain	0.0406256	652
amr_genotypes	0.0284130	456
collection_date	0.0216836	348
isolation_source	0.0195651	314
location	0.0070409	113

informative in our analysis. These variables are outbreak, host disease, and host. We recognize that our study will not be informed by information about an isolate's host disease, host species, and origin from an outbreak.

After removing the high-missingness variables, we note that 54.7% of the isolate records have at least one missing value for the 14 variables that we have left in our data.

We advance by assessing the distributions of our variables, starting with **strain**. There are 15,348 unique strains represented in our dataset, meaning that 95.6% of the isolates in our data are unique. We may want to consider dropping this variable as a result as it will not be too informative to our research problem.

Next, we looked at the **collection date** variable to examine the time periods that our data span. Since the raw values of this variable differ in terms of the specificity of the dates, i.e., some values only record year while others record the exact month and day as well, we'll reduce this variable to only the year for consistentcy purposes.

We observe the proportion of years represented within each organism group.

Interestingly, Salmonella is the only bacteria that retains isolate data for 2002-2008 (**Table 2**). Data starts to be retained for E.coli and Shigella in 2009 but shows the most records in 2017 and thereafter. Data collection for isolates for Campylobacter started in 2016 and retained records up to and including the year of 2020. As we advance in this project, we'll use this variable to understands trends in our data overtime, more specifically as it pertains to changes in drug resistance and susceptibility.

Next, let's look at the distribution of **serovar** within each bacteria of interest. Upon further inspection, we see that we only have non-missing serovar information for Salmonella. There are 150 distinct serovars in our Salmonella data. To reduce the number of groupings and simplify, values will only be distinguished as their own group if they represent over 2% of the isolates in the Salmonella data. Remaining serovar categories will be grouped into an 'Other' classification.

With this new subclassified serovar column, we are left with 13 distinct groupings. Aside from the 'Other' Classification group, Kentucky (~14%), Heidelberg (8%), and Typhimurium var. 5- (7%) are the most common serovar groupings.

The **AST phenotypes** variable, in its rawest form, come to us as one long string with all of the recorded antibiotics that each isolate is resistant, susceptible, intermediate in response, and susceptible-dose dependent to. We separated and classified this variable into four new variables (one for each of the drug responses mentioned) and listed the corresponding antibiotics for each category. This process was done for ease of use when we eventually go on to build our model and want to separate antibiotics by isolates' responses to them.

Table 2: Prop. of Years Represented Per Bacteria

Year	Campylobacter jejuni	E.coli and Shigella	Salmonella enterica
2002	0%	0%	1.76%
2003	0%	0%	2.32%
2004	0%	0%	3.6%
2005	0%	0%	4.05%
2006	0%	0%	3.97%
2007	0%	0%	3.6%
2008	0%	0%	5.62%
2009	0%	0.03%	4.94%
2010	0%	0%	3.44%
2011	0%	0.97%	4.32%
2012	0%	1.57%	4.29%
2013	0%	3.08%	4.2%
2014	0%	4.62%	8.07%
2015	0%	3.08%	1.29%
2016	37.88%	4.64%	12.22%
2017	22.54%	26.92%	6.87%
2018	19.78%	19.82%	8.34%
2019	19.78%	34.99%	17.08%
2020	0.03%	0.19%	0%
2021	0%	0.08%	0.01%
2022	0%	0.03%	0%

In Salmonella, tetracycline (23%), streptomycin (17%), sulfisoxazole (13%), ampicillin (12%), and ceftriaxone (6%) are the top antibiotics that our observed isolates are resistant to.

In E.coli, the top drugs that these isolates are resistant to are tetracycline (15%), ampicilin (12%), streptomycin (8%), sulfisoxazole (8%), and ceftriaxone (7%).

In Campylobacter, tetracycline (48%), ciprofloxacin (21%), nalidixic acid (20%), clindamycin (3%), and erythromycin (2%) are the top drugs that isolates show resistance to.

In general, these bacteria share a lot of antibiotics that they've grown resistant to.

Regarding the discussion on drug susceptibility, Salmonella isolates saw the most susceptibility to tetracycline (23%), streptomycin (17%), sulfisoxazole (13%), and ampicillin (12%).

In E.coli, isolates and strains were found to be most susceptible to meropenem (8%), ciprofloxacin (7%), nalidixic acid (7%), azithromycin (7%), and cefoxitin (7%).

Finally, in Campylobacter, the drugs those isolates were most susceptible to were gentamicin (13%), erythromycin (13%), florfenicol (13%), azithromycin (12%), and clindamycin (12%).

Interestingly, we see some overlap in certain bacteria being both susceptible and resistant to the same antibiotics. It'll be interesting to explore this more and see if examine potential shifts from susceptible to resistant status for these illnesses.

Next, we look at the **locations** represented in our data. We'll reduce the location variable in our data to just be the countries. The vast majority of the records were collected in the United States with 99% of the Salmonella records hailing from the U.S., 85% of the E.coli and Shigella records from the U.S., and 95% of the Campylobacter jejuni records hailing from the U.S. The Netherlands and Germany are the next most popular countries represented. There are a total of 25 distinct countries represented in our dataset.

The **isolation source** variable originally hosted 180 unique values. Upon further inspection, we found that a lot of these distinctions were due to case sensitiveness as well as many variations of the same general

Table 3: Prop. of Sources Represented Per Bacteria

Organism	beef	chicken	other/unspecified	pork	stool	turkey	urine	water
Campylobacter jejuni	17.27%	67.44%	4.75%	4.15%	4.69%	1.7%	0%	0%
E.coli and Shigella	16.49%	16.73%	5.07%	12.55%	12.52%	26.76%	7.35%	2.52%
Salmonella enterica	5.44%	46.92%	0.75%	9.87%	1.18%	35.75%	0.09%	0%

Table 4: Prop. of Isolation Type Represented Per Bacteria

Organism	clinical	environmental/other
Campylobacter jejuni	4.83%	95.17%
E.coli and Shigella	24.85%	75.15%
Salmonella enterica	3.14%	96.86%

source. For example, 'pork', 'pork chops', 'pork chop', and 'Pork'. To simply, we grouped the isolates into broader categories based on the content we observed in the isolation source variable. We ultimately reduced this variable to 9 distinct catgeories: pork, chicken, turkey, beef, stool, urine, water, other/unspecified, and NA (for missingness).

The most common isolate source in our data is chicken for Campylobacter and Salmonella (**Table 3**). For E.coli and Shigella, it's turkey, with chicken being the second most common.

In terms of **isolation type**, there are 1,405 clinical isolates and 14,644 environmental/other isolates. This brings us to ~91% non-clinical (environmental/other) isolates. There's a difference in the distribution of isolation type in the E.coli bacterium group compared to the other two bacteria (Table 4).

There are 2,613 **SNP clusters** represented in our data, 811 clusters for Salmonella, 830 for E.coli and Shigella, and 974 for Campylobacter jejuni.

There are 3,085 unique sets of **AMR genotypes** represented in the data, which we may be interested in dissecting and using to supplement our AST phenotype information.

Next, we looked at the **min-same** and **min-diff** variables. We observed that the distributions of minimum SNP distance to another isolate whether that isolate be of the same isolation type or different isolation type, are generally right-skewed (**Figure 1**). This may speak to the presence of outbreaks, isolates closely related to one another that may even have occurred near the same time. Our E.coli and Shigella records seem to be slightly distinct from the other two bacteria in its distributions as if presents a bit more variance in its min-same and min-diff values.

We can create a new variable to indicate the number of isolates that are "close" to a given isolate. We'll define a closely related isolate as an isolate within 7 SNPs of another. If an isolate is within 7 SNPs of another clinical isolate and another environmental isolate, that counts as 2. Thus, only going off of the data that is available to us, this variable will only take on discrete values from 0-2.

Briefly looking at the distribution of this new count variable, we observed that ~66% of isolates from E.coli and Shigella are not "close" to any other isolates while this proportion is around 25-26% for Campylobacter and Salmonella. Our Campylobacter jejuni isolates are the most close with ~44% having a min-diff and min-same value less than 7. This folowed by Salmonella with 35% of records having a min-diff and min-same value less than 7, and then E.coli with only 9% of records having a min-diff and min-same value less than 7.

The last variable that we'll look at broadly is computed types. We start by separating computed types into the two pieces of information that the variable provides: antigen formula and serotype. Perhaps the most important piece of information that we want to extract from this variable is the antigen formula because we already have serotype information from the serovar varialbe. We checked to see if this information serotype and serovar from the two different variables, was comparable or not and found that it was. So, from the computed types variable, we'll only reserve the antigen formula information.

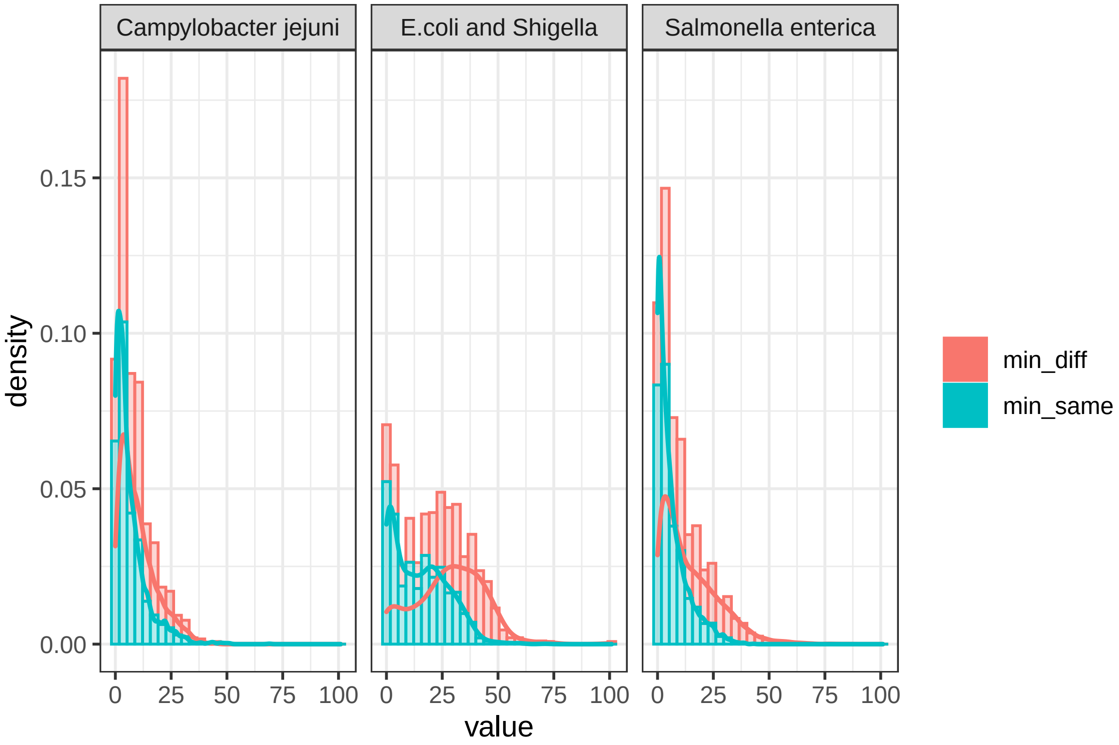

Figure 1: Distributions of min-same and min-diff of isolates belonging to each bacteria

This leaves us with 126 unique antigen formulas. We can reduce this further by looking at the distribution of antigen formulas represented. Upon further investigation, we found that the only isolates that don't have missingness for antigen formula are isolates that belong to Salmonella. Categories representing at least 2% of the Salmonella isolates will remain their own distinct columns while categories representing less than 2% will be grouped into an 'Other' category.

Aside from the "other" category, '8:i:z6' (15%), '4:i:1,2' (13%), and '4:r:1,2' (8%), are the top three most antigen formulas represented in the Salmonella isolates that we have records on.

Initial Subquestions:

The first question that we want to consider is: 1. Are isolates within the same SNP cluster susceptible/resistant to the same antibiotics? How about isolates within the same isolation source?

In order to assess this, we will compute the similarity scores between the drugs that isolates within the same clusters and isolation source have. To do this, we'll use the `stringsim` function in the `stringdist` package in R (van der Loo M, 2014). The `stringsim` function computes pairwise string similarities between elements of a vector. It returns values between 0 and 1 (inclusive), corresponding to perfect dissimilarity at a value of 0, and perfect similarity at a value of 1. We used the default method for distance calculation, optimal string alignment (restricted Damerau-Levenshtein distance). We'll assess the average similarity values for isolates within the same clusters. These values are broken down by drug resistance and drug susceptibility.

Table 5 shows that on average, the drug behaviors of isolates within the same cluster tend to be fairly similar with mean and median values hovering between 0.75-0.86. Drug behaviors like susceptible dose dependent response and intermediate response have a lot more missingness compared the the data that we have on

Table 5: Drug Behavior Similarity between Clusters

	Susceptible Sim.	Resistant Sim.	Intermediate Sim.	Suscept. Dose Dependent Sim.
Min.	0.0903446606874123	0.0573248407643312	0.0967741935483871	0.157894736842105
1st Qu.	0.73724030113124	0.571300422139967	0.484884615384615	1
Median	0.851528384279476	0.823170731707317	1	1
Mean	0.8264932049237	0.757366412928215	0.780827591539239	0.893162633608281
3rd Qu.	0.939886652175522	1	1	1
Max.	1	1	1	1
NA's	1326	3392	4463	4459

Table 6: Average drug responses for close and distant clusters with closeness being defined as an average min-diff or min-same SNP distance of < 10 SNPS

Cluster Type	Average Susceptibility	Average Resistance	Average Intermediate	Average SDD	Num Clusters
close	0.9276393	0.8976005	0.8247229	0.8795887	1502
distant	0.8590357	0.7819367	0.7333333	NaN	314

susceptibility and resistance and so with smaller sample sizes, if the data that we do have are similar and not very variable, these specific behaviors will have high similarity scores. We may consider removing the intermediate and/or susceptible dose dependent behaviors due to their high volumes of missingness. Despite high averages, there are some clusters that don't seem to have low similarity, based on our data.

We decided to group clusters into two categories, 'close' and 'distant'. Close clusters are defined as clusters with average min-diff or average min-same less than or equal to 10 SNPS. All other clusters are categorized as distant. We looked at the average similarity measures for close and distant clusters and found that in general, clusters that we considered 'close' had high similarity when it came to their isolates' drug responses (**Table 6**).

The second question that we want to consider is: 2. Does drug resistance and susceptibility change over time?

To answer this question, we'll just look at drug responses categorized as either susceptible or resistant instead of also considering intermediate and susceptible-dose dependent.

```
##        organism_group      strain collection_date   serovar        isolate
## 1 Salmonella enterica  CVM N50422            2013 Saintpaul PDT000078107.1
## 2 Salmonella enterica FSIS1605829            2016 Saintpaul PDT000114571.2
## 3 Salmonella enterica FSIS1606072            2016 Saintpaul PDT000123179.2
## 4 Salmonella enterica   NY-N19886            2008  Adelaide PDT000101356.2
## 5 Salmonella enterica   NY-N19887            2008  Adelaide PDT000101357.2
## 6 Salmonella enterica   NY-N19888            2008  Adelaide PDT000101358.2
##   location          isolation_source      isolation_type     snp_cluster
## 1  USA: GA                pork chops environmental/other PDS000032576.34
## 2   USA:PA          animal-swine-sow environmental/other PDS000032576.34
## 3   USA:GA animal-swine-market swine environmental/other PDS000032576.34
## 4   USA:OR                 pork chop environmental/other PDS000003480.41
## 5   USA:OR                 pork chop environmental/other PDS000003480.41
## 6   USA:OR                 pork chop environmental/other PDS000003480.41
##   min_same min_diff
## 1        4        5
## 2        8       10
## 3        6        8
## 4        2       23
## 5        2       23
## 6        5       26
##
## 1
## 2
## 3
## 4                            aac(3)-IV=COMPLETE,aadA7=COMPLETE,aph(4)-Ia=COMPLETE,blaTEM-1=COMPLETE,
## 5 aac(3)-IV=COMPLETE,aadA7=COMPLETE,aph(4)-Ia=COMPLETE,blaTEM-1=COMPLETE,bleO=COMPLETE,bleO=PARTIAL_
## 6                                          aac(3)-IV=COMPLETE,aadA7=COMPLETE,aph(4)-Ia=COMPLETE,blaT
##                                  computed_types serovar_new
## 1 antigen_formula=4:e,h:1,2,serotype=Saintpaul   Saintpaul
## 2       antigen_formula=-:-:-,serotype=I -:-:-   Saintpaul
## 3 antigen_formula=4:e,h:1,2,serotype=Saintpaul   Saintpaul
## 4   antigen_formula=-:f,g:-,serotype=I -:f,g:-       Other
## 5   antigen_formula=35:f,g:-,serotype=Adelaide       Other
## 6   antigen_formula=-:f,g:-,serotype=I -:f,g:-       Other
##                                                               resist
## 1                                                               <NA>
## 2                                                               <NA>
## 3                                                               <NA>
## 4 ampicillin, gentamicin, streptomycin, sulfisoxazole, tetracycline
## 5               ampicillin, gentamicin, sulfisoxazole, tetracycline
## 6 ampicillin, gentamicin, streptomycin, sulfisoxazole, tetracycline
##
## 1 amoxicillin-clavulanic acid, ampicillin, azithromycin, cefoxitin, ceftiofur, ceftriaxone, chloramp
## 2            amoxicillin-clavulanic acid, ampicillin, azithromycin, cefoxitin, ceftriaxone, chloramp
## 3            amoxicillin-clavulanic acid, ampicillin, azithromycin, cefoxitin, ceftriaxone, chloramp
## 4                                                                        amikacin, amoxicillin-clavu
## 5                                                          amikacin, amoxicillin-clavulanic acid, ce
## 6                                                                        amikacin, amoxicillin-clavu
##   inter suscept_dose_dep country isolation_source_new suscept_sim resist_sim
## 1  <NA>             <NA>     USA                 pork   0.9141631         NA
## 2  <NA>             <NA>     USA                 pork   0.9570815         NA
## 3  <NA>             <NA>     USA                 pork   0.9570815         NA
## 4  <NA>             <NA>     USA                 pork   0.7585825  0.6832875
## 5  <NA>             <NA>     USA                 pork   0.7315341  0.5997657
## 6  <NA>             <NA>     USA                 pork   0.7585825  0.6832875
##   inter_sim sdd_sim
## 1        NA      NA
## 2        NA      NA
## 3        NA      NA
## 4        NA      NA
## 5        NA      NA
## 6        NA      NA
```

Since we've established that clusters tend to be very similar in terms of their drug responses

- dataset just with isolate, year, and column for each drug in drug resistance - 1 if resistant 0 otherwise
- do the same for susceptible

References

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2871933/

van der Loo M (2014). “The stringdist package for approximate string matching.” *The R Journal*, *6*, 111-122. https://CRAN.R-project.org/package=stringdist.