Leveraging the k-Nearest Neighbors Classification Algorithm for Microbial Source Tracking Using a Bacterial DNA Fingerprint Library

Jeffrey D. McGovern
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
Email: jmcgover@calpoly.edu

Alexander Dekhtyar
Department of Computer Science
California Polytechnic State University
San Luis Obispo, CA 93407
Email: dekhtyar@calpoly.edu

Christopher Kitts
Department of Biological Sciences
California Polytechnic State University
San Luis Obispo, CA 93407
Email: ckitts@calpoly.edu

Michael Black
Department of Biological Sciences
California Polytechnic State University
San Luis Obispo, CA 93407
Email: mblack@calpoly.edu

Jennifer Vanderkelen
Department of Biological Sciences
California Polytechnic State University
San Luis Obispo, CA 93407
Email: jvanderk@calpoly.edu

Anya Goodman
Department of Biological Sciences
California Polytechnic State University
San Luis Obispo, CA 93407
Email: agoodman@calpoly.edu

***Abstract*—Fecal contamination in bodies of water is an issue that cities must combat regularly. Sourcing the species of the fecal matter helps curb the issue in the future, giving city governments the ability to mitigate the effects before they occur again. Microbial Source Tracking (MST) aims to determine source host species of strains of microbiological lifeforms and library-based MST is one method that can assist in sourcing fecal matter. The Biology Department in conjunction with the Computer Science Department at California Polytechnic State University San Luis Obispo (Cal Poly) teamed up to build a library called the Cal Poly Library of Pyroprints (CPLOP). Since 2009, students collect fecal samples, culture and pyrosequence the *E. coli* in the samples, and insert this data, called pyroprints, into CPLOP. Using two intergenic transcribed spacer regions of DNA, Cal Poly biologists perform studies on strain differentiation. We propose using k-Nearest Neighbors, a straightforward machine learning technique, to classify the host species of a given pyroprint, construct four algorithms to resolve the regions, and investigate classification accuracy.**

I. Introduction

Microbial Source Tracking (MST) aims to determine the source host species of strains — genetic subtypes — of microbiological lifeforms. Be it through the makeup of different microorganisms or the presiding strains of a particular microorganism in the source matter, researchers measure strain characteristics that distinguish source matter between different species. MST is a necessary step in a variety of applied studies: for example in order to properly address the issue of fecal contamination in a water source, identifying the root cause, i.e., the the species that causes of the contamination, is key.

Fecal provenance is one use for MST that attempts to delineate the strains of bacteria that reside within each species. In recent years biologists conjectured that the bacterial strains of fecal indicator bacteria, such as *E.coli* are usually specific to the species, outside of a certain percentage of so-called *transient strains*. Strain identification and sourcing allows researchers to characterize the species of provenance of the source matter that the microbiological lifeforms inhabit.

A common solution to MST is to collect relevant microbial cultures, obtain a strain-level digital representation of each collected culture, and store such representations in a database — henceforth referred to as a "library". The data inserted into these libraries can range from collected information about the microbiome, to a particular microbe characterization, or to any other useful set of metrics that can profile an entry appropriately.[4]

Following the creation of a microbial strain library, an MST method proceeds as follows. An environmental sample is taken and the microbial material in it is treated the same way as the library strains. The strain representation obtained in the result is compared to the library strain representations, and any close matches are shown to the biologists. As the provenance of the library strains is known, the biologists can make appropriate decisions regarding the provenance of the strains in the environmental sample.[8]

At the same time, library-based MST methods have significant limiting factors associated with them. These include cost to build and limited applicability outside of the environment the library was built from.[6] Companies do exist that will receive samples for a cost and classify the provenant species. Unfortunately, these tend to be costly and inaccurate — many times guaranteeing at best only 30% accuracy. Often, these commercial libraries are infeasibly costly and limited to particular regions of the world.

Over the past four years our research group developed a pyrosequencing-based method for bacterial strain identification. Based on this method, we have assembled a library of digital fingrerprints of *E.Coli* isolates collected from a variety of known host species. This library, called Cal Poly Library

of Pyroprints (CPLOP) [1][1], has served to support numerous research projects related to collection and analysis of *E.coli* as it occurs in host species and in the environment. However, while the original purpose of CPLOP was to support MST, up until this paper, no MST studies have been conducted.

This paper describes our first Microbial Source Tracking study that uses CPLOP. We use a slightly modified k-Nearest Neighbors classification method [2] to answer the following question: *for what percentage of CPLOP isolates can we properly identify the host species?* There is a variety of approaches one can take to answer this question. The k-Nearest Neighbors method is simple and straightforward: we elected to use it in our first study because it provides a nice baseline for all followup improvement attempts.

The main contributions of the paper are as follows:

- We modify the k-Nearest Neighbors method by adding one more parameter: the threshold α of similarity between the neighbors, beyond which no new neighbors are used. As such, the version of k-Nearest Neighbors used in this paper is the intersection of the regular nearest neighbor construct with a range query.
- The pyroprint representations of bacterial isolates give rise to multiple similarity scores between a pair of isolates. We describe a number of ways in which these multiple scores can be combined into a single decision procedure to select the "winner". In our study, we determine which of the ways to combine the similarity scores (we call the resolution methods) works best.
- We report on our empirical study to determine the accuracy of determining the host species for the bacterial isolates whose pyroprint representations are stored in CPLOP. In the study, we look at the best values of k (number of nearest neighbors to compare to), threshold α, as well as compare the results based on the four similarity score resolution methods.

The rest of the paper is organized as follows. Section II introduces CPLOP and the pyroprinting process. Section III describes the variant of k-Nearest Neighbors with thresholding we used in the study and the similarity resolution techniques we used. Section IV shows the results of the study, and Section V concludes with the discussion of our results.

II. CPLOP: The Cal Poly Library of Pyroprints

In 2011, Cal Poly students and faculty built the Cal Poly Library of Pyroprints (CPLOP) [7] in order to facilitate MST. CPLOP stores information about multiple collected bacterial isolates of *E. coli*. The information stored in CPLOP is called *pyroprints*[1]: essentially, the peak heights of pyrosequences of specially constructed DNA products extracted from the *E. coli* DNA. In what follows we provide a brief description of the pyroprinting process and the CPLOP data.

Pyrosequencing is a DNA sequencing technique appropriate for sequencing short DNA fragments (up to around 150-200 base pairs). [5] While not as powerful as the Next-Generation Sequencing methods that can sequence full DNA of multiple samples in one run, pyrosequencing is significantly less expensive, with a single pyrosequencing run costing on the order of tens of dollars (not counting the cost of equipment). A pyrosequence of a DNA fragment can be represented as a vector of real values — one value per base pair, indicating the intensity of light emission that occurred during the sequencing reaction (light is emitted in response to a specific nucleotide reagent used in the sequencing process). *Pyroprinting* is a fingerprinting technique for bacteria that uses pyrosequencing to sequence a "mixed" DNA product, such that the resulting pyrosequence cannot be used to reproduce the DNA sequence, but instead becomes a "fingerprint" of the mixed product[1].

Pictured in Figure 1 is an abstracted segment of *E. coli* DNA used in CPLOP. *The shown pattern repeats around the ring of E. coli DNA seven times*. The internal transcribed spacers (ITS) between the *16S*, *23S*, and *5S* ribosomal DNA regions each contain non-coding DNA. These segments of DNA offer keen insight into strains of *E. coli*. Since they are non-coding, random variations occur that do not affect the survivability or reproducibility of the microbe. Since any offspring of a given *E. coli* strain inherits this DNA, the *16S-23S* and *23S-5S* regions offer the ability to differentiate strains.[7]

We pyroprint each ITS region separately. The DNA product that becomes sequenced *is a a PCR-amplified mix of the DNA from the seven loci of the ITS region in the E. coli DNA*. Each locus has a DNA sequence that may be different from the sequences in the other six loci, but all seven loci can be amplified jointly by selecting appropriate primers [1]. We refer to pyroprints from these regions as *16S-23S* and *23S-5S*.

Fig. 1. A diagram of a simplified segment of *E. coli* DNA, outlining the *16S-23S* and *23S-5S* ITS regions.

Isolates are the cultured samples of *E. coli* from a host's fecal sample. Once a fecal sample is collected, bacterial cultures are grown from it and preserved. Each isolate undergoes PCR procedures that "cut" and amplify the DNA in the two ITS regions, after which the pyrosequencing of each region produces the isolate pyroprints which are stored in CPLOP.

The data represented in CPLOP has a phylogeny, of sorts. Since the species, host, and isolate of each pyroprint is carefully and correctly tracked, there is a structure to the data, as shown in Figure 2.

species→host→isolate→pyroprint

Fig. 2. The structure of the data in CPLOP is such that a pyroprint comes from an isolate, which comes from a host, which is one instance of some species. In other words, species have hosts, which have isolates, which have pyroprints.

CPLOP's most intrinsic entry is the pyroprint. Each pyroprint comes from an isolate. Each isolate comes from a host. Since every host has been observed, it is of some species. Each host has any number of isolates designating the isolate. Each isolate has any number of pyroprints, tracked as pyroprints, each from either the *16S-23S* or *23S-5S* region of DNA.

[1] http://www.cplop.org

Each pyroprint has been observed as to be of some species, according to the host its isolate came from.

III. Classification Methodology

Over the past four years CPLOP has been used by faculty and students at Cal Poly as the *E. coli* pyroprint repository for a variety of studies. However, the studies conducted to date compared known-host *E. coli* to known-host *E. coli* in a variety of ways, and did not venture into Microbial Source Tracking — the original *raison d'être* for CPLOP. In order to understand how useful CPLOP is for MST studies, we need to determine whether the data currently stored in CPLOP can be used to properly identify host species of various isolates. A sufficiently straightforward approach to such an initial study is to use the part of CPLOP that stores isolates that came from known host species as both the training set and the test set, and determine if we can replicate the correct identification of the host species.[3][1]

For our study, we use a modified version of the k-Nearest Neighbors classification method, which uses host species names as category labels. The similarity between the isolates is computed by comparing isolate pyroprints from the same ITS region using Pearson Correlation Coefficient. These similarity scores are combined using four different resolution procedures to produce the winning category label. These three components of the process are described below.

A. Pearson Correlation

As mentioned in Section II, a pyroprint of a specific ITS region of a bacterial isolate is a vector $p = (x_1, \ldots, x_N)$ where, $i = 1, \ldots, N$ is called a *dispensation* - the position in the pyrosequencing process when the ith reagent was dispensed, and x_i is a numerical characteristic of the ith dispensation. CPLOP uses *peak heights*, the highest light emission values registered during the ith dispensation, as the values x_i of such characteristics.

For comparing two pyroprint vectors, we use the Pearson Correlation as defined in Definition III.1. Since it is the random variation in the DNA regions that create the light values in each vector, Pearson Correlation lends itself well toward our ends.

Definition III.1. Suppose we have two vectors, $\vec{u}, \vec{p} \in \mathbb{R}^n$. The Pearson Correlation is a mapping, $S : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1] \subset \mathbb{R}$:

$$S(\vec{u}, \vec{p}) = \frac{\sum_{i=1}^{n}(u_i - E(\vec{u}))(p_i - E(\vec{p}))}{\sqrt{\sum_{i=1}^{n}(u_i - E(\vec{u}))^2}\sqrt{\sum_{i=1}^{n}(p_i - E(\vec{p}))^2}} = \frac{cov(\vec{u}, \vec{p})}{\sigma_{\vec{u}} \cdot \sigma_{\vec{p}}} \tag{1}$$

where $E(\vec{p})$ is the mean of $\vec{p} = (p_1, \ldots, p_n)$, $cov(\vec{u}, \vec{p})$ is the covariance of $\vec{u}$ and $\vec{p}$, and σ is the standard deviation.

In essence, the Pearson Correlation is the variance between two vectors normalized by the standard deviation of each. This provides the following properties:

For $\vec{u}, \vec{p}$ and S defined in Definition III.1:

- $S(\vec{p}, \vec{p}) = 1$
- $S(\vec{u}, \vec{p}) = S(\vec{p}, \vec{u})$
- Dissimilar $\vec{u}$ and $\vec{p}$ have $S(\vec{u}, \vec{p})$ close to 0

B. k-Nearest Neighbors (with α Threshold)

The k-nearest neighbors classification algorithm is a straightforward algorithm to classify an unclassified object using a library. We use the concept of a comparison metric to formulate an idea of "closeness." To outline the process:

Given an unclassified object u, a library of classified objects $\mathbb{L}$, and a comparison metric, C:

1) Compare u to each object in $\mathbb{L}$ using C
2) Add the classified object and the result to a list of neighbors, N
3) Sort N by most similar
4) Consider only the top k objects in N
5) Classify u as the most plural classification in the k-nearest N list

The idea is that we hope the unclassified object must be "close" to some of the classified objects in our database. By choosing the "most plural" classification — the classification that shows up in the k-nearest neighbors the highest number of times — we can, with some accuracy, classify our unknown object.

Our first modification to k-nearest neighbors is an additional condition at step 4:

4) Consider only the top k entries in N above threshold α

The α threshold allows biologists to filter out neighbors that are among the k closest, but too dissimilar to compare. For many other studies — not necessarily MST-focused — a Pearson Correlation of 0.99 or above defines a strain of *E. coli*. Filtering by some value near this may give more accurate results and provides an intuitive way to relate these lists to other studies.

C. Comparing Isolates

Of primary interest to the biologists using CPLOP is comparing isolates to each other. In CPLOP, each isolate is represented by a pair of pyroprints: one for each of the two ITS regions. As a result, the similarity between two isolates can be computed as the Pearson correlation between their 16S-23S pyroprints, as well as the Pearson correlation between their 5S-23S pyroprints. But a pair of similarity scores between two isolates is not a single similarity score. In order to accommodate the fact that two isolates need to be compared to each other in two different ways we need a resolution procedure. Rather than creating a new similarity measure out of a pair of similarity scores, we choose to update the k-Nearest Neighbors method with four different ways of selecting the winning category label based on how the pyroprints compare to each other. These four methods are described below.

In what follows, we generalize our problem. Let u and v are two library objects (isolates). Let $\mathbb{C}= (c_1, \ldots c_K)$ is a collection of comparison measures over library objects, with $K > 1$. All four resolution procedures described below work with such a generalized representation of isolates and comparison measures between them.

1) Meanwise Resolution: Given an unknown isolate u, a library of classified[2] isolates $\mathbb{L}$, and a set of comparison measures $\mathbb{C}$, we compare u to each object in $\mathbb{L}$ using each comparison metric in $\mathbb{C}$. For u and a $p \in \mathbb{L}$, we take the mean of the result of all of the comparison metrics and build a k-nearest neighbors list from it.

The mean can be any metric mapping $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and in the investigated implementation, we use the euclidean distance, also known as the L^2 norm. A single k-nearest neighbors list results from this algorithm that we filter by k and α and use to classify the unknown.

2) Resolution by Winner: Given an unknown object u, a library of classified objects $\mathbb{L}$, and a set of comparison metrics $\mathbb{C}$, we compare u to each object in $\mathbb{L}$ using each comparison metric in $\mathbb{C}$. For each comparison metric, we make a k-nearest neighbors list and filter by k and α accordingly.

Once we finish building each comparison metric's k-nearest neighbors list, we find the most plural classification from each list and track the number of times that classification shows up in that list. Then, we classify u based off the classification that has the highest number in its corresponding list.

3) Resolution by Union: Given an unknown object u, a library of classified objects $\mathbb{L}$, and a set of comparison metrics $\mathbb{C}$, we compare u to each object in $\mathbb{L}$ using each comparison metric in $\mathbb{C}$. For each comparison metric, we make a k-nearest neighbors list and filter by k and α accordingly.

After building each k-nearest neighbors list, we combine the lists into a set, keeping track of the original list position for tie-breaking. From this set, which we dub the union, we count the classifications and classify u as the most plural in the union of the lists, compared to the other lists.

4) Resolution by Intersection: Given an unknown object u, a library of classified object $\mathbb{L}$, and a set of comparison metric $\mathbb{C}$, we compare u to each object in $\mathbb{L}$ using each comparison metric in $\mathbb{C}$. For each comparison metric, we make a k-nearest neighbors list and filter by k and α accordingly, but ensure that we do not lose track of the entire sorted list of results.

After building each k-nearest neighbors list, we inspect each list for common isolates. We add isolates that appear in every list into a set that we call the intersection. If the size of the intersection is k, then we are done. Otherwise, we increase the length of our individual lists by δ and search for common isolate. This process repeats until the size of the intersection is k, or all of the isolates in the individual lists are below threshold α.

D. Cross Validation with Holdout

To gauge accuracy of the results, we cross-validated against the library by separately holding out each isolate in CPLOP from CPLOP, classifying it against CPLOP, and verifying whether it is correct. Since each isolate in CPLOP has the correct species, we know whether a classification is correct or not.

[2]A "classified isolate" is an isolate for which the host species has been idenitifed in the database.

E. Library Makeup

CPLOP contains data from many different studies. Some investigate *E. coli* strain similarities and differences between species. Others are longitudinal, focusing on the change in *E. coli* strains within the same host. Some are even a mixture of the two, looking at the change in *E. coli* strains as a host of one species is exposed to the host of another species. Table I shows the breakdown of how many hosts, isolates, and pyroprints a given species has for the dataset we used to validate these algorithms.

We validated these algorithms using most of this data, filtering out pyroprint and species according to the following criteria:

- Pyroprints of environmental sources, such as lakes, rivers, and oceans.
  - Our focus of this study is to gauge the accuracy of animal species. Future studies will look into environmental sources and how animal species contribute to *E. coli* strains within.
- Pyroprints flagged as erroneous
- Isolates that no longer have any pyroprints in either the *16S-23S* or *23S-5S* ITS region
  - Our focus is on how well the resolution between two ITS regions work. Results focusing on the edge case of missing regions may be investigated in the future.
- Species that have fewer than 4 isolates
  - Our thought is that a species with fewer than isolates would have difficulty building a majority and we understand that there are problems with underrepresented speciess in our library.

Data removed according to the above criteria is not shown in Table I. One small note is that there may have been multiple species entries due either to typos, or nomenclature differences between studies. In Table I, we merged these different named species into the most familiar species name and counted accordingly.

IV. Results & Evaluation

There are a few areas of focus that we have when interpreting the results:

- What size k achieves the best results?
- What size α achieves the best results?
- Which metric resolution algorithm achieves the best results?

A. Evaluation Metrics

Indeed we can define "best" in many ways, but we choose to look at two metrics, recall and precision, and a combination of the two, the F-measure. The metrics look at the accuracy of the classification on the object and the object on the classification respectively, while F-measure hopes to represent a balance between the two.

TABLE I. A BREAKDOWN OF THE SPECIES REPRESENTED IN CPLOP, AND THE NUMBER OF HOST, ISOLATE, AND PYROPRINT ENTRIES FOR EACH.

species Name	Number of hosts	Number of isolates	Number of pyroprints
Barn Owl	3	5	13
Bat	1	37	74
Bear	1	6	12
Bobcat	1	4	8
Sea Lion	3	6	12
Cat	36	39	91
Chicken	15	40	82
Cliff Sparrow	14	28	59
Common Loon	2	4	8
Cow	427	1838	3772
Coyote	2	4	8
Deer	2	20	40
Dog	50	269	573
Elephant Seal	2	4	8
Great Horned Owl	2	4	11
Grey Fox	1	4	8
Ground Squirrel	50	196	401
Horse	49	51	102
Human	227	1590	4189
Mountain Lion	5	32	64
Opossum	5	12	24
Pelican	3	8	18
Pig	32	66	149
Pigeon	107	194	515
Rabbit	1	4	8
Racoon	2	4	9
Red Tailed Hawk	2	5	11
Red-shoulder Hawk	2	4	11
Sea Otter	3	10	20
Seagull	31	11	22
Sheep	50	94	200
Sparrow	8	15	30
Wild Turkey	36	72	177
Total species: 32	1236	4682	10732

1) Recall: In our study recall tracks how well we are able to discover all isolates from a given category, i.e., with a given host species. Given a category/host species name, the recall for that host species is the percentage of isolates taken from this host species that have been properly identified. For example, if our database had 100 cat isolates, and 74 of them were classified by our method as having come from a cat, the recall would be 74%. In this study, we compute both overall recall (what percentage of isolates were classified as their proper label) as well as species-level recall (what percentage of isolates that came from dogs/humans/sheep/etc. were classified as their proper label).

2) Precision: Precision tracks how well our method avoids misclassification errors. Given a category and a list of isolates our method classified as belonging to it, the precision of the method on the category is the percent of isolates from the list that has the correct label. For example, if our method returned 100 isolates labelled "Dog" of which 77 isolates really did come from dogs, the precision of the method is 77%. As with recall, we compute both overall precision, as well as the precision for each category/species label.

3) F-Measure: The F-measure, F_1, is the *harmonic mean* of the precision, P and the recall, R:

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2 \cdot \frac{P \cdot R}{P + R}$$

While we prefer maximizing this value, a value near 0.5 means we are doing well.

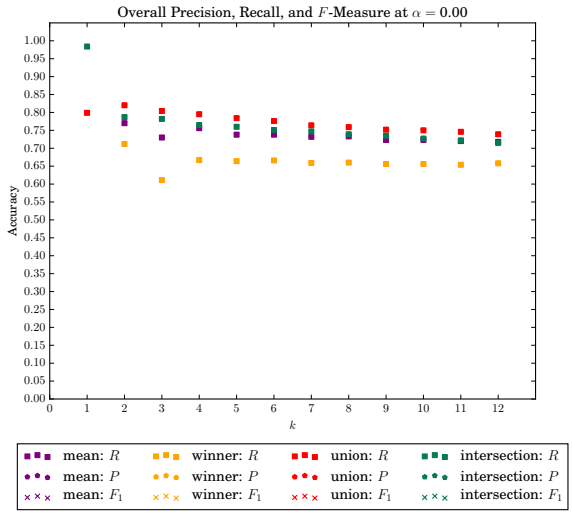

Fig. 3. The accuracy of all classifications performed with CPLOP across the four different algorithms with $\alpha = 0.00$ shows little improvement for $k > 5$. We look at only the percentage of correct classifications, since that value is equivalent to the precision and the recall.

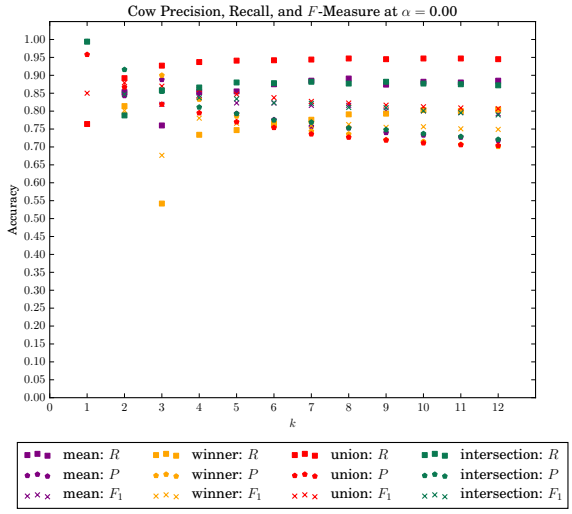

Fig. 4. There are 1838 Cow isolates in CPLOP. For most resolution algorithms, we observe little improvement when $k > 5$.

B. Adjusting k

Adjusting k is an important first step. We investigate k values ranging from 1 to 17, but focus primarily on $k \leq 12$. At this point, we do not filter the results in order to focus primarily on the affect of the size of the k-nearest neighbors list. Thus, α is 0, allowing for the full k list to factor into classification.

Overall, for $k \geq 5$, the accuracy does not improve, but instead levels off. Depending on the resolution algorithm, this value is between 65% and 75% accuracy, as shown in Figure 3. By "overall," we mean that for every classification, we validated if it was correct and calculated what proportion to all classifications made that represents to determine accuracy. When looking at all classifications, precision and recall are identical values, as is F-measure.

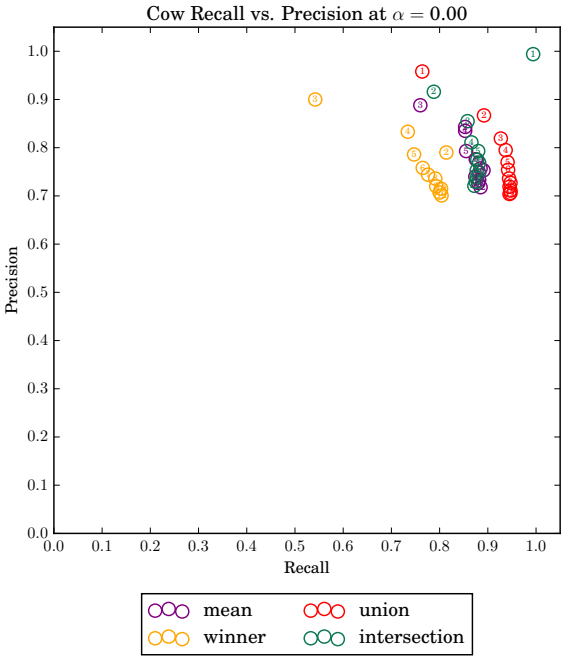

Fig. 5. There are 1838 Cow isolates in CPLOP.There are 1838 Cow isolates in CPLOP. Looking at the Recall as it compares to the Precision for $\alpha = 0.99$ allows us to visualize the tradeoffs we make when picking a k value. Labeled within each datapoint is the k value at that point

One good example is the Cow. As Figure 4 shows, Cow follows a trend similar to the overall accuracy, staying roughly between 70% and 95% accurate. Certain algorithms get worse for $k > 5$, while other improve.

Figure 5 examines the relationship between R and P. This can help us understand the trade offs of choosing one k over another. We will later build a meaningful strategy for how confident we are at recalling a species versus our confidence in a classification of a species.

C. Adjusting α

By adding a threshold value, we investigated whether this further limitation improves the accuracy by restricting outliers from populating a k-nearest neighbors list. We investigate $\alpha = \{0.00, 0.98, 0.99\}$. Outside of this study, $\alpha = 0.99$ defines the boundary between strains. One reason we investigate 0.98 is to see whether loosening our definition of strain differentiation gives us a better accuracy.

Overall, we observe that the accuracy slightly improves as we increase the α threshold. Figure 6 shows that overall, the accuracy increases as we increase α.

Adding the α made minimal changes to the accuracy of Cow classifications, so only the recall versus precision is shown in Figure 7. More details into how α affect the classification accuracy can be seen in Tables II, III, and IV.

D. Adjusting the Algorithm

Choosing which algorithm too resolve the two different regions of each isolate is an important step. We investigate the differences between the aforementioned four algorithms

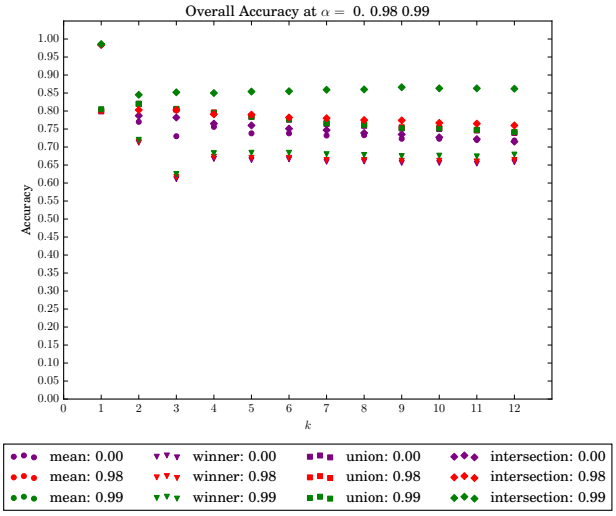

Fig. 6. Shown is the accuracy of all classifications performed with CPLOP across the four different algorithms. We find that the accuracy of certain resolution algorithms perform better with higher α values.

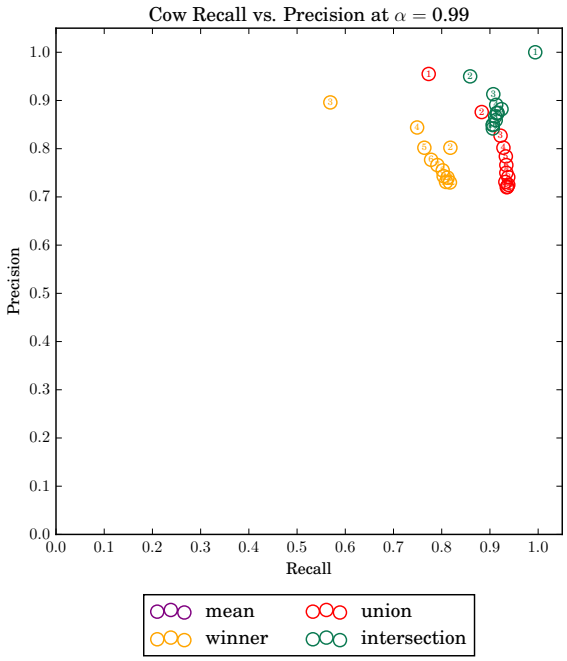

Fig. 7. There are 1838 Cow isolates in CPLOP. Increasing the α for a species with this mane isolates made minimal improvements to the accuracy on all but the resolution by intersection algorithm, which noticeably improved.

as they relate to k and α values and how each differ among species of different representation. With library-based-MST, it is important to realize the representation of a species in the library may heavily skew the accuracy of the library.

While interpreting the data, we state that there may be some "%" increase or decrease which we intend to mean the increase in the raw value of the percentage. Additionally, values in the tables represent the proportion of the three metrics, but are easily interpreted as percentages. Section IV-A explains the meaning of each precision (P), recall (R), and F-measure (F_1).

TABLE II. PRECISION (P), RECALL (R), AND F-MEASURE (F_1) OVERALL AND FOR PARTICULAR SPECIES AT k=7, $\alpha = 0.00$.

		Meanwise			Winner		
$species$	$\|isolates\|$	P	R	F_1	P	R	F_1
Overall	4682	0.732	0.732	0.732	0.659	0.659	0.659
Human	1471	0.857	0.922	0.888	0.771	0.861	0.814
Cow	1718	0.757	0.885	0.816	0.744	0.776	0.760
Pigeon	194	0.420	0.242	0.307	0.280	0.253	0.266
Dog	149	0.596	0.436	0.504	0.449	0.356	0.397
Wild Turkey	72	0.383	0.250	0.303	0.277	0.181	0.219
Chicken	40	0.182	0.050	0.078	0.143	0.075	0.098
Cat	39	0.571	0.308	0.400	0.438	0.359	0.395
Bat	37	0.857	0.973	0.911	0.692	0.973	0.809
Seagull	11	0.000	0.000	0.000	0.000	0.000	0.000
Grey Fox	4	0.143	0.250	0.182	0.400	0.500	0.444
Racoon	4	nan	0.000	nan	nan	0.000	nan

		Union			Intersection		
$species$	$\|isolates\|$	P	R	F_1	P	R	F_1
Overall	4682	0.764	0.764	0.764	0.747	0.747	0.747
Human	1471	0.843	0.930	0.884	0.839	0.925	0.880
Cow	1718	0.736	0.944	0.827	0.769	0.882	0.822
Pigeon	194	0.569	0.170	0.262	0.470	0.284	0.354
Dog	149	0.761	0.450	0.566	0.649	0.497	0.563
Wild Turkey	72	0.688	0.306	0.424	0.510	0.361	0.423
Chicken	40	0.000	0.000	0.000	0.250	0.100	0.143
Cat	39	0.889	0.410	0.561	0.571	0.308	0.400
Bat	37	0.857	0.973	0.911	0.857	0.973	0.911
Seagull	11	nan	0.000	nan	nan	0.000	nan
Grey Fox	4	1.000	0.500	0.667	0.800	1.000	0.889
Racoon	4	nan	0.000	nan	nan	0.000	nan

TABLE III. PRECISION (P), RECALL (R), AND F-MEASURE (F_1) OVERALL AND FOR PARTICULAR SPECIES AT k=7, $\alpha = 0.98$.

		Winner		
$species$	$\|isolates\|$	P	R	F_1
Overall	4682	0.664	0.664	0.664
Human	1471	0.773	0.865	0.816
Cow	1718	0.749	0.777	0.763
Pigeon	194	0.287	0.254	0.269
Dog	149	0.448	0.349	0.392
Wild Turkey	72	0.308	0.222	0.258
Chicken	40	0.150	0.075	0.100
Cat	39	0.467	0.359	0.406
Bat	37	0.692	0.973	0.809
Seagull	11	0.000	0.000	0.000
Grey Fox	4	0.400	0.500	0.444
Racoon	4	nan	0.000	nan

		Union			Intersection		
$species$	$\|isolates\|$	P	R	F_1	P	R	F_1
Overall	4682	0.767	0.767	0.767	0.780	0.780	0.780
Human	1471	0.845	0.930	0.885	0.876	0.950	0.912
Cow	1718	0.742	0.943	0.831	0.799	0.894	0.844
Pigeon	194	0.538	0.181	0.271	0.521	0.333	0.406
Dog	149	0.756	0.456	0.569	0.698	0.536	0.606
Wild Turkey	72	0.697	0.319	0.438	0.571	0.387	0.461
Chicken	40	0.000	0.000	0.000	0.308	0.121	0.174
Cat	39	0.889	0.410	0.561	0.632	0.353	0.453
Bat	37	0.857	0.973	0.911	0.878	0.973	0.923
Seagull	11	nan	0.000	nan	0.000	0.000	0.000
Grey Fox	4	1.000	0.500	0.667	0.800	1.000	0.889
Racoon	4	nan	0.000	nan	nan	0.000	nan

Overall, with $\alpha = 0.00$, Figure 3 illustrates that the resolution by union algorithm consistently performs better. For $k = 7$ and $\alpha = 0.00$, Table II shows that using the resolution by unions algorithm performs with 76.4% accuracy with meanwise and resolution by winner and intersection respectively achieving 73.2%, 65.9%, and 74.7 accuracy%.

Poorly represented species, like the Cat, Chicken, Seagull, and Racoon did not benefit from the resolution by union algorithm, each achieving no classifications, correct or otherwise.

Once we restrict with a somewhat loose threshold of 0.98, overall we see that the intersection method provides the best accuracy, improving on non-thresholded values. For $k = 7$ and $\alpha = 0.98$, the intersection algorithm achieves 78.0% accuracy%, while resolution by winner and union respectively achieve 66.4% and 76.7% accuracy.

TABLE IV. PRECISION (P), RECALL (R), AND F-MEASURE (F_1) OVERALL AND FOR PARTICULAR SPECIES AT k=7, $\alpha = 0.99$.

		Winner		
$species$	$\|isolates\|$	P	R	F_1
Overall	4682	0.680	0.680	0.680
Human	1471	0.780	0.872	0.823
Cow	1718	0.766	0.791	0.778
Pigeon	194	0.314	0.263	0.286
Dog	149	0.527	0.401	0.455
Wild Turkey	72	0.320	0.222	0.262
Chicken	40	0.167	0.100	0.125
Cat	39	0.433	0.333	0.376
Bat	37	0.720	0.973	0.828
Seagull	11	0.429	0.273	0.334
Grey Fox	4	0.500	0.500	0.500
Racoon	4	nan	0.000	nan

		Union			Intersection		
$species$	$\|isolates\|$	P	R	F_1	P	R	F_1
Overall	4682	0.766	0.766	0.766	0.859	0.859	0.859
Human	1471	0.843	0.925	0.882	0.926	0.979	0.952
Cow	1718	0.750	0.934	0.832	0.874	0.914	0.894
Pigeon	194	0.476	0.205	0.287	0.611	0.468	0.530
Dog	149	0.739	0.456	0.564	0.838	0.738	0.785
Wild Turkey	72	0.719	0.319	0.442	0.667	0.455	0.541
Chicken	40	0.000	0.000	0.000	0.000	0.000	0.000
Cat	39	0.938	0.385	0.546	0.909	0.588	0.714
Bat	37	0.837	0.973	0.900	0.973	1.000	0.986
Seagull	11	0.000	0.000	0.000	nan	0.000	nan
Grey Fox	4	1.000	0.500	0.667	0.667	1.000	0.800
Racoon	4	nan	0.000	nan	nan	nan	nan

Table III shows that a handful of poorly represented species achieved slightly better results when $\alpha = 0.98$. Notably, the intersection algorithm F-measure increased slightly for Wild Turkey, Cat, and Chicken on the order of 3%.

Unfortunately, the meanwise algorithm fails to classify when we use a large enough α and thus we have ommited the results in Tables III and IV. In certain cells of the tables, including Table II, "nan" values in either P or F_1 mean no classifications were made of that species.

Restricting with $\alpha = 0.99$, our definition of strain differentiation, overall accuracy improves more with resolution by intersection and less so with resolutions by winner and union, garnering 85.9%, 68.0%, and 76.6% accuracy respectively. Again, meanwise resolution fails to produce any classifications.

For poorly represented species, we see some similar improvements for P, R, and F_1, but also some exceptions. Wild Turkey for example, improves by about 2%-3% for resolutions by winner and union and 11% for resolution by intersection, while Cat decreases by 3% for resolution by winner, but improves by 6% and 47% for resolution by union and intersection.

E. Poorly Represented Species

Some species had worse accuracy than the overall accuracy. In particular, species such as "Chicken" with only 40 isolates representing it showed similar leveling of accuracy for $k > 5$, but had far poorer accuracy, as shown in Figure 8. For $k > 5$, the accuracy of classifying chicken ranges from as low as 10% to a peak of 26%. The classification accuracy for many species in CPLOP heavily relies on its representation in CPLOP.

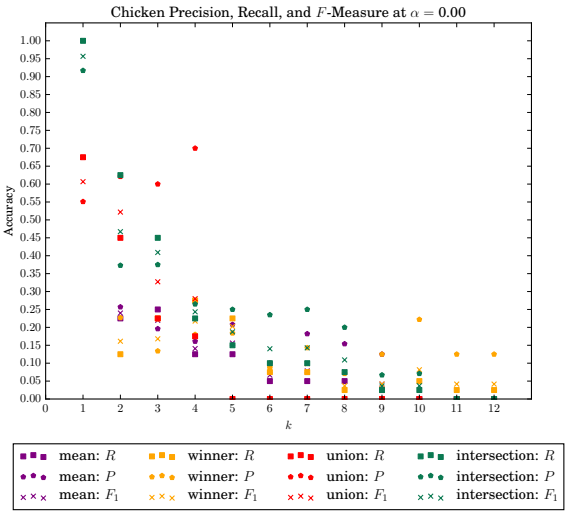

Fig. 8. There are 40 Chicken isolates in CPLOP. Unfortunately, due to their low representation in CPLOP, classification accuracy is low.

One notable exception is the Bat. In everyone application of our k-Nearest Neighbors algorithms, Bat has above 95% accuracy. It is possible that due to their small size and relative dietary segregation from the surrounding environment that the strains of *E. coli* stay particularly unique. It may also be a quirk of the fact that each isolate comes from a single host, making it difficult to draw conclusions from such results.

V. Discussion & Conclusion

Generally, when using k-Nearest Neighbors, it is preferred to use single digit k values. Through our investigation of these various k-Nearest Neighbors classification algorithms, we find that that general advice holds true. For our dataset, using $k \geq 5$ does not produce much different results. Choosing $k < 5$ is a dangerous notion, since it is likely that an outlier may make its way into the k-nearest neighbors list, confounding the results. Staying with $5 \leq k \leq 9$ appears to be a safe and reasonable option, providing a good balance between accuracy and filtering of isolates.

Outside of this study, we choose to differentiate between strains of *E. coli* using $\alpha = 0.99$. It appears that using these values is advantageous. There were, however, some exceptions to those results, motivating us to look at non-thresholded k-nearest neighbors lists during classification as well.

The four resolution algorithms — meanwise, winner, union, and intersection — each have their own quirks and behaviors as we alter k and α.

Meanwise, which currently uses the Euclidean norm to resolve different metrics, did not respond to the α threshold and completely stopped classifying anything for α near 1. This is very likely due to Euclidean norm mapping $([0,1],\ldots,[0,1]) \to [0,\sqrt{1+\cdots+1}]$. To get around this, we multiplied the resulting norm by a factor of $\sqrt{2}$, which may have unexpected results. We may investigate this further, or choose a more natural norming method, like arithmetic or geometric mean. With no α filtering, it performed third best with an overall 73.2% classification accuracy.

Winner performs worst, classifying accurately between 65% and 68% of the time. Some alterations to this algorithm may make it more reliable, such as only counting the species that appear in all lists.

Unionwise performs very well. Without filtering the k-nearest neighbors lists by α, we find that the unionwise method classifies best, with an overall accuracy of 76.4%. However, once we add in α filtering, the unionwise does not improve, staying relatively close to 76%.

Intersection performs best when we use α. This is likely due to the "list" actually being a set of common isolates. Overall, without filtering, the accuracy was 74.7%, 78%, and 85.9% for α =0.00, 0.98, and 0.99 respectively.

Overall, we find that the intersection algorithm performs the best and recommend moving forward with it. While unionwise did perform well, it did not respond well to thresholding and still did not perform as well as the intersection algorithm overall. Meanwise and winner may be more useful with previously mentioned modifications and we may investigate these in the future.

Poorly distributed representation of species and environmental incomparabilities are issues endemic to library-based MST. CPLOP has an overabundance of Cow and Human isolates, and an underrepresentation of many of the species in the database. This dilutes the k-nearest neighbors list considerably for species like the Racoon, Seagull, and Chicken.

Library population issues aside, environmental limitations are another concern for accuracy. Nearly every sample in the library comes from a 30 mile radius around Cal Poly, making much of the data incomparable to data collected from a different region.

References

[1] Michael W. Black, Jennifer VanderKelen, Aldrin Montana, Alexander Dekhtyar, Emily Neal, Anya Goodman, and Christopher L. Kitts. Pyroprinting: A rapid and flexible genotypic fingerprinting method for typing bacterial strains. *Journal of Microbiological Methods*, 105:121 – 129, 2014.

[2] Daniel T Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2005.

[3] Emily Neal, Collin Sabatini, Winnie Tang, Michael Black, and Chirs Kitts. Demographics of e. coli strains in the human gut using pyroprints: A novel mst method, 2012.

[4] K Ritter, Ethan Carruthers, C Carson, R Ellender, V Harwood, Kyle Kingsley, Cindy Nakatsu, Michael Sadowsky, Brian Shear, Brian West, et al. Assessment of statistical methods used in library-based approaches to microbial source tracking. *J Water Health*, 1:209–223, 2003.

[5] Mostafa Ronaghi, Mathias Uhlén, and Pål Nyrén. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365, 1998.

[6] Debby Sargeant, William R Kammin, and Scott Collyard. *Review and critique of current microbial source tracking (mst) techniques*. Environmental Assessment Program, Washington State Department of Ecology, 2011.

[7] Jan Lorenz Soliman, Alex Dekhtyar, Jennifer Vanderkellen, Aldrin Montana, Michael Black, Emily Neal, Kevin Webb, Chris Kitts, and Anya Goodman. Microbial source tracking by molecular fingerprinting. In Sanjay Ranka, Tamer Kahveci, and Mona Singh, editors, *ACM International Conference on Bioinformatics, Computational Biology and Biomedicine, BCB' 12, Orlando, FL, USA - October 08 - 10, 2012*, pages 617–619. ACM, 2012.

[8] J Stewart, R Ellender, J Gooch, Sunny Jiang, S Myoda, and S Weisberg. Recommendations for microbial source tracking: lessons from a methods comparison study. *J Water Health*, 1:225–231, 2003.