INVESTIGATING THE k-NEAREST NEIGHBORS RESOLUTION ALGORITHMS FOR PYROPRINTS AND CLUSTERING FOR BACTERIAL STRAINS

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Jeffrey D. McGovern

March 2016

COMMITTEE MEMBERSHIP

TITLE:	Investigating The k-Nearest Neighbors Resolution Algorithms for Pyroprints and Clustering for Bacterial Strains
AUTHOR:	Jeffrey D. McGovern
DATE SUBMITTED:	March 2016
COMMITTEE CHAIR:	Alexander Dekhtyar, Ph.D. Professor of Computer Science
COMMITTEE MEMBER:	Chris Kitts, Ph.D. Professor of Computer Science
COMMITTEE MEMBER:	Foaad Khosmood, Ph.D. Professor of Computer Science

ABSTRACT

Investigating The k-Nearest Neighbors Resolution Algorithms for Pyroprints and Clustering for Bacterial Strains

Jeffrey D. McGovern

Your abstract goes in here

ACKNOWLEDGMENTS

Thanks to:

- Andrew Guenther, for uploading this template

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Fecal contamination in public water sources is an issue that health officials and city and county governments must frequently combat. Pathogens present in fecal matter pose severe health risks to humans and pets and the decomposition of fecal coliform bacteria can upset the balance of aquatic ecosystems by depleting dissolved oxygen to low enough levels that it may kill other species in the water. Such severe threats to the health of humans, pets, and the local ecosystem motivates public health officials to take action in order to mitigate its consequences. Often times, no one observes the cause of the fecal contamination, but rising levels of fecal coliform bacteria indicate that fecal contamination has occurred. In these situations, usually the only course of action that natural resource managers have is to simply restrict public access until contamination levels reach an acceptable level, which does nothing to prevent further contamination. Identifying the source of fecal contamination in water supplies is an important initial step to prevent further contamination.

Microbial Source Tracking (MST) is the field of research that aims to discover the host-species that microbial lifeforms originate from and aids the process of sourcing fecal contamination. Microbes thrive inside the gut of animals, as well as in masses of plant matter, and routinely make their way into the environment via fecal matter deposition. Biologists conjecture that strains of microbes or bacteria present in fecal matter, called fecal indicator bacteria (FIB), remain relatively unique to the species of the host they originated from. A strain of a species of microbe is a subtype of that species where the microbes in that strain are closely related in some meaningful way. How researchers specifically define strains often differs, since each definition of a strain depends on the characterization of the microbe in question and the methods used to

derive such characterizations. Typically, the objective is to discover which microbes of a bacterial isolate came from the same parent microbe [34]. A strain, then, can be thought of as consisting of individuals descending from the same individual to generate a "group" or "family." Researchers put significant effort into choosing the relevant microbes and appropriately characterizing them in order to discover which strains tend to belong to which species.

A common method of MST known as library-based MST involves collecting fecal matter from a known host-species, culturing isolates of the relevant microbes in the fecal matter, and building a digital representation of the collected isolates for storage into a database and analysis. Storing an appropriate digital representation allows researchers to perform rigorous analysis and comparison between FIB isolates collected from different host-animals and host-species, as well as isolates collected from the same host-animal, but at different times. The data inserted into such a database may range from collection metadata about the microbiome, to a specific microbe characterization, or to any other useful set of metrics that can appropriately profile an entry [55].

In this way, researchers build a "library" of known-host-species isolates. Using this library, researchers can take an environmental sample with FIB from an unknown source, process the microbial isolates using the same procedure and the known-host-species isolates in the library, and compare the strain representation of the environmental sample to those in the library to find any close matches. Since the researchers know the host-species of the isolates in the library, they can make a reasonable determination of the source of the isolates in the environmental sample. The methods used to compare isolates and make assertions depends entirely upon the FIB, their method of collection, and their digital representation.

Library-based MST is usually only effective within the region in which the known-

host-species isolates came from, making it difficult to build a "one size fits all" library. Companies exist that can, for fees in the ballpark of $100, attempt to determine the host-species of a provided sample and while these companies exist nationwide, they are few in number and usually cannot build a representative sampling of every region for accurate host-species determination. Additionally, when investigating an incident of fecal contamination, investigators want to send out multiple samples to build reliable evidence for a determination of the source. As a result, the cost of outsourcing becomes too prohibitive and determinations too inaccurate for it to be an option. Thus, there exists a need for a cost-effective and accurate method of MST in order to properly tackle the problem of preventing fecal contamination in water supplies.

In 2009, the California Polytechnic State University San Luis Obispo (Cal Poly) Biology and Computer Science Departments built The Cal Poly Library of Pyroprints (CPLOP) [63], a database of *Escherichia coli* (*E. coli*) isolate fingerprints, called pyroprints. Students collect fecal samples from a variety of host-species from the San Luis Obispo area and build the pyroprints using a low cost DNA sequencing method called pyrosequencing on two intergenic regions of an *E. coli* isolate. Building pyroprints ends up costing roughly two orders of magnitude less than outsourcing samples, cutting the cost of building an effective MST library by as much as 60% [9]. It is through CPLOP that Cal Poly researchers hope to better understand bacterial strains, how to differentiate between them, and provide a cost-effective MST methodology.

In order to be an effective MST fingerprinting method, pyroprinting must contain information that allows for the accurate discrimination between closely related strains of *E. coli* bacteria. Internal Transcribed Spacers (*ITS*) in bacteria are regions of DNA that do not contain instruction for building proteins and thus have high variability, since variability across generations of bacteria does not affect the survivability of

the microbe. Because of this high variability, researchers can use these regions to differentiate between strains of the same species of microbe. *E. coli* isolate pyroprints stored in CPLOP represent the polymerase chain reaction (PCR)-amplified regions of DNA between the *16S* and *23S* genes and *23S* and *5S* genes, referred to as *ITS-1* and *ITS-2* respectively. *ITS-1* and *ITS-2*, along with the entire *E. coli* genome, repeats seven times, giving us seven highly variable regions for each *ITS*. Any offspring inherit mostly accurate copies[1] of the *ITS* regions of the parent microbe, encoding the notion of a "group" or "family" and allowing researchers to use them to differentiate between strains [64]. By building pyroprints out of these regions, CPLOP researchers hope to gain a reproducible notion of an *E. coli* strain that they can use for MST.

A pyroprint is a vector comprised of the peak heights of pyrosequences of multiple copies of a repeated region of DNA. By dispensing a series of nucleotides at specific times and observing the resulting light emitted, CPLOP researchers can build a fingerprint of that DNA sequence. In traditional pyrosequencing, the DNA sequenced is an amplified version of a single sequence of DNA, allowing researchers to reconstruct the exact sequence of nucleotides that make up the DNA. Since CPLOP researchers pyroprint segments of DNA that repeat but are highly variable, researchers cannot reconstruct the exact sequences of the *ITS* sequences. Alternatively, CPLOP contains a pyroprint that represents the random variability in the entire genome of that particular *E. coli* isolate. Previous work in [61] optimized the pyroprinting process, including the dispensation sequence and peak height determination, for each *ITS* to best delineate between different strains of *E. coli* using the Pearson correlation coefficient to compare pyroprints.

The Pearson correlation coefficient ρ normalizes the covariance of two vectors by the standard deviation of each, providing a notion of relative co-variability between

[1]Some variation may occur, but it is assumed to be small for immediately related microbes and large for distantly related microbes.

the vectors that remains invariant of noise and scaling — a core reason why CPLOP researchers use it to compare pyroprints. In order to compare two *E. coli* isolates in CPLOP, researchers must separately compare the *ITS-1* pyroprints to each other and the *ITS-2* pyroprints to each other using Pearson correlation coefficient. It is meaningless to compare different *ITS* to each other since they represent entirely different sections of DNA that have been obtained through a different sequence of dispensations. This effectively gives us two comparison metrics between isolates: the Pearson correlation coefficient between two *ITS-1* and the Pearson correlation coefficient pyroprints between two *ITS-2* pyroprints — $\rho_{ITS\text{-}1}$ and $\rho_{ITS\text{-}2}$. Using these values, CPLOP researchers can rigorously define the notion of a strain.

CPLOP supports numerous research projects, ranging from longitudinal studies of a host-animal to large studies of one or more host-species, in order to understand the evolution and transmission of *E. coli* strains and verify that pyroprinting provides an accurate representation of *E. coli* strains. Previous work on CPLOP include formation and validation of the pyroprinting process, exploration of the evolution and transference of *E. coli* strains within and between host-animal and host-species, and new algorithms designed specifically for CPLOP to better understand its data.

Much of the work done so far using CPLOP has been exploring the composition, evolution, and transference of strains among host-animals and host-species. While part of this is to validate the MST methodology that leverages CPLOP data, researchers gain a large amount of insight into how *E. coli* strains get into and evolve in fecal matter by using pyroprints to rigorously study changes. Clustering methods become very useful in this case, owing their effectiveness to the notion of a strain being similar to a "group" or "family" of a closely related subtype of a species of microbe.

Two pieces of previous work, [39, 40] and [29], worked toward building clustering

algorithms that can provide meaningful insight into the *E. coli* isolates in CPLOP. The former, *OhClust!*, is an agglomerative clustering algorithm where a biologist-provided metadata-ontology guides the agglomeration. The latter, by Eric Johnson, is a density-based clustering algorithm — DBSCAN — optimized by a fast range query for nearby isolates.

While *OhClust!* takes advantage of all of the information available in CPLOP, DBSCAN encodes our notion of a strain the closest and allows for strain discovery without needing to guess what ontology provides the best insight. Moreover, the range query optimizations made in [29] allow for efficient, low-memory querying of isolates while still encoding the notion of Pearson correlation coefficient between isolates, an improvement over *OhClust!*'s need to precompute and store distances in order to mitigate the consequences of agglomerative clustering's need for a high number of distance computations. It may be that by preclustering isolates, we can speed up the computation of k-RAP.

In a nutshell, DBSCAN uses a distance metric, a minimum neighbors value `MinPts`, and an ε range to categorize data points as one of three types: core point, border point, or noise. A core point is a point that has at least `MinPts` data points within ε of it. A border point is a point that is within ε of a core point, but that does not have `MinPts` points within ε of it. Every other point is noise. The algorithm then defines a cluster as a group of neighboring core points with their associated border points.

In [36], we constructed the notion of a bacterial strain purely from the clusters produced by DBSCAN — i.e. we defined bacterial strains to be the clusters produced by DBSCAN. We studied the cluster purity — the proportion of isolates in a cluster that are of the same species — of the entire clustering at different `MinPts` values. In doing so, we observed the presence of so-called transient *E. coli* strains — strains of

E. coli that show up in many different host-species — that tend to confound MST. More importantly, it showed that CPLOP has relatively few of these transient strains and a large number of pure strains.

While the original purpose of CPLOP was to support MST and some manual MST studies have been conducted, little research has been done on building an automated MST method. Most studies performed with CPLOP focused on validating and exploring the various biological features captured by the pyroprinting process and the comparison metric used to compare pyroprints, the Pearson correlation coefficient. Building objective, repeatable classification metrics that use the data in CPLOP to assist MST can help biologists inform investigators of a possible source that caused, or is causing, fecal contamination.

As a first step towards building an effective classification technique, we chose to use the k-Nearest Neighbors (k-NN) classification algorithm on CPLOP to measure how accurately we can classify samples that we know the host-species of. k-NN classifies an unknown-class datum by querying a library of known data — each datum has a class, or classification — for "nearby" data; sorts the list by nearness, limiting it to k many "neighbor" data points; and classifies from this "k-nearest neighbors" list by picking the most plural classification present among the neighbors, breaking ties by average position.

A somewhat unique obstacle arises with the *E. coli* isolates in CPLOP: in order to compare isolates, we must use two different comparison metrics — $\rho_{ITS\text{-}1}$ and $\rho_{ITS\text{-}2}$. For k-NN on CPLOP *E. coli* isolates, this means that we produce two k-nearest neighbors lists that we must classify from. Resolving multiple k-NN lists can be useful for any data that has multiple meaningful-yet-exclusive ways to compare one datum to another. Biologists using k-NN will likely want to restrict the list further than k, since their definition of a strain relies heavily on bounding the Pearson correlation

coefficient between two isolates for both *ITS* — so we also add an α threshold to further limit the involved k-NN lists.

The four resolution algorithms, called the k-NN Resolution Algorithms for Pyroprints (k-RAP) and previously published in [35], are termed: Meanwise Resolution, Resolution by Winner, Resolution by Union, and Resolution by Intersection. Meanwise Resolution takes the average of the comparison value to form a single k-NN list. Resolution by Winner finds the most plural classification in each k-NN list and picks the classification with the most instances of that class in its list. Resolution by Union combines each k-NN list into a single set — performing effectively a union on all of the k-NN lists — and finds the most plural classification in the resulting set, breaking ties by average original position. Resolution by Intersection forms a new set that is exactly the isolates that appear in every k-NN list — effectively performing an intersection at the isolate level — expanding both lists and adding to the set until the set itself is of size k and choosing the most plural class of the new set.

Investigating k-RAP in [35] showed us that classification accuracy for the entire database stayed well above 50% with most of the resolution algorithms. Precision and recall for well-represented host-species also stayed safely above 0.30, which is far better than random and notably better than our outsourced baseline. Underrepresented species predictably performed poorly in classification. Furthermore, α thresholding noticeably improved performance on some resolution algorithms, causing one to perform better than the others with α, but worse without.

In this thesis, we investigate further the work done in [35] and [36] in depth and consider whether combining the two is useful for performing MST. Namely, the contributions of this paper are the following:

- k-NN Resolution Algorithms for Pyroprints: Modifications to the k-NN classification algorithm that can resolve multiple comparison metrics

- A modification to k-NN that adds α thresholding to further restrict the individual k-NN lists
- An empirical study measuring the accuracy of identifying the host-species for the *E. coli* isolates stored in CPLOP, investigating how values of k and α affect the accuracy with each resolution metric
- Revisions to work done in [35]
- An investigation of the efficient density-based clustering algorithm in [29] that is scalable and meaningfully encodes the comparison-metric used in CPLOP
- A set of validation measures for clustering bacterial isolates into strains
- An evaluation of our strain discovery procedure based on the defined set of measures

The rest of this thesis is organized as follows: Chapter 2 provides an overview of relevant work in the field of MST and an introduction to work done using CPLOP; Chapter 3 details CPLOP and the background necessary to understand the algorithms presented; Chapter 4 describes k-RAP and the use of DBSCAN as a clustering method for bacterial strains; Chapter 5 gives an overview of the structure of the code and how to use it; Chapter 6 defines the evaluation criteria that the algorithms are judged by and motivation for their use; Chapter 7 discusses the results of the investigation; and Chapter 8 concludes, offering suggestions for future work.

Chapter 2

RELATED WORK

Existing Microbial Source Tracking (MST) methodologies require a fecal indicator bacteria (FIB) fingerprinting method that allows for strain discrimination and a method of classification. Early work in MST [8] worked by measuring the ratio of fecal coliform bacteria to streptococci ratios, which fell out of use due to the "widely varying survival rates of the bacterial groups in the environment" [57]. In order to effectively use FIB for MST, researchers had to develop new methods to fingerprint and classify them with the appropriate host-species or find related strains.

Fingerprinting FIB usually falls into two categories: phenotyping and genotyping. Phenotypic methods of fingerprinting usually involve "morphology of colonies on various culture media, biochemical tests, serology, killer toxin susceptibility, pathogenicity, and antibiotic susceptibility," none of which allows researchers to reliably distinguish between closely related strains [34]. Genetic fingerprinting — genotyping — "has become widely used ... due to its high resolution" [34] and many methods exist that allow for effective discrimination [58, 57].

Classification methods use a variety of statistical measures to make determinations to either related strains or host-species, but most fall into library-dependent and library-independent. Library-independent MST searches for the presence of certain microbes in fecal matter or contaminated water. The presence of certain microbes can indicate what host-species may have deposited the fecal matter. Unfortunately, this method relies on prior knowledge of the types of microbes that may occur in the types of potential host-species[1], limiting the effectiveness of host-species determination [57].

[1]Often times, these methods can only detect whether the fecal content came from a human and maybe some domestic animal species [57].

Library dependent techniques work by building a database of FIB fingerprints that come from known host-species. These techniques usually differ in the fingerprinting process, which the classification technique is dependent upon. Using these libraries, researchers can handle common FIB from a variety of host-species, making it incredibly agile. Disadvantages of this technique come from: the need to build a large library size which can become cost-prohibitive; the transient nature of some *E. coli* strains, assuming *E. coli* is the FIB of choice; and the fact that the applicability of the database is limited to the region in which the database was built from [57]. C-PLOP is a library-based MST technique *E. coli* as FIB and pyroprints as cost-effective fingerprints and researchers use it to understand *E. coli* strains and determine the host-species of fecal matter.

2.1 Pyroprinting

A key component of strain-based Microbial Source Tracking is the representation of strains of fecal indicator bacteria (FIB). Numerous genotypic methods exist for differentiation between strains of *E. coli*. One can find a detailed discussion of why pyroprints perform better than these options in [31].

Cal Poly researchers introduce the concept and construction of pyroprints in [9, 31], describing the process through which they construct pyroprints from the multiple loci of isolated *E. coli* DNA. It discusses the work done in [61], which confirmed the reproducibility of pyroprinting and determined that a Pearson correlation coefficient correlation above 0.99 "could be a good threshold to minimize false separation of isolates from the same strain." These works explain in detail the advantages of using the pyroprinting methodology with respect to cost ("[p]yroprinting could reduce the cost of a library-based MST investigation by up to 60%" [9]), reproducibility (much of which can be found in [61]), and discrimination between (known) strains of *E. coli*,

compared to existing state of the art methods. It asserts that while *E. coli* were used, the pyroprinting process applies to a broad range of bacteria whose genome contain multiple loci.

In-silico simulations done in [10] delved into the sensitivity of using the Pearson correlation coefficient ρ to compare constructed pyroprints of known *E. coli* alleles gathered from the National Center for Biotechnology Information database. CUDA programming on a GPU sped up the ρ computation considerably, allowing the researchers to understand, given all possible combinations of seven known alleles to form a simulated isolate, how many isolates are "*hard to differentiate*," i.e. have a $\rho_{ITS\text{-}1}$ and $\rho_{ITS\text{-}2}$ above 0.99 [10]. The work in [10] supplements *in-vitro* work performed in [41].

Senior projects and master's theses [54], [63], and [68] discuss the development of many of the tools in CPLOP, from the backend database construction, to the frontend web view and usage for investigation. Cal Poly researchers have placed a large emphasis on validation of the methodologies included in building pyroprints from *E. coli* isolates. Biology students investigated how *E. coli* strains change in response to a variety of factors. Computer science students at Cal Poly have developed many tools to aid the biologists in both validating their methodologies and performing *E. coli* strain research on various host-animals and host-species

2.2 Empirical Strain Research

CPLOP has enabled numerous research projects in the field of biology. The following is a list of empirical strain research performed using CPLOP:

- Using Hadoop to Identify False Positives in Bacterial Strain Typing from DNA Fingerprints [2]

- Demographics and Transfer of *E. coli* Within Bos taurus Populations [17]
- *E. coli* Strain Demographics and Transmission in Cattle [18]
- Application of Pyroprinting for Source Tracking of E. coli in Pennington Creek [43]
- Demographics of *E. coli* Strains in the Human Gut Using Pyroprints: A Novel MST Method [44]
- *Escherichia coli* Strain Diversity in Humans: Effects of Sampling Effort and Methodology [45]
- Investigating the Dominant *Escherichia coli* Strain in Lambs and Ewes Using Pyroprinting: A Novel Method for Strain Identification [46]
- Source Tracking of Fecal Contamination Along San Luis Obispo (SLO) Creek [59]
- Short Communication: Typing and Tracking Bacillaceae in Raw Milk and Milk Powder Using Pyroprinting [66]

These studies provide significant insight into the evolution and transmission of *E. coli* strains and demonstrate the effectiveness of using pyroprints and CPLOP as a MST method. Many of the above studies provided a culminating experience for undergraduates and graduates in biology and computer science. What we aim to provide with k-RAP is a set of tools that students and researchers have at their disposal to make it easier to make reproducible discoveries and assertions about strains in CPLOP.

2.3 Clustering

Presented in [41, 42] are the comparison of two hierarchical clustering techniques. *Primer5* [12] and a chronology-sensitive hierarchical clustering algorithm. Using metadata about when researchers collected the samples used to build the isolates, the hierarchical clustering proceeds to first cluster isolates from samples collected on the same day and continues to cluster by increasing days away from the initial collection date. They found that the clusters built by the chronology-sensitive hierarchical clustering algorithm resembled the *Primer5* clusters, but were unsure of whether these clusters were appropriate.

The work in [41, 42] went on to become a part of *OhClust!* (**O**ntology-Based **h**ierarchical **Clust**ering**!**) [64, 40, 39], a metadata-aware hierarchical clustering algorithm that allows CPLOP researchers to provide a metadata ontology to guide the order of hierarchical clustering. Hierarchical clustering in general is a very calculation-intensive process, making it a problematic tool for servers with limited computational power. The computational crux comes with the number of comparisons needed between clusters — clusters of isolates in CPLOP's case.

Most hierarchical clustering algorithms compute the distances between clusters and agglomerate by picking clusters to combine into a cluster (made of clusters) for the next hierarchy. Cluster distances are merely the distance between some representative member — possibly an average of the actual members — of one cluster with a representative from the other cluster. The representatives used to compute distance may be the members in each cluster that are, for example, closest to each other, farthest from each other, or the centroid of each cluster.

Computationally intense distance metrics make implementing a performant hierarchical clustering algorithm problematic for programmers. The way *OhClust!* gets

around this difficulty is by precomputing the distances — Pearson correlation coefficients in CPLOP — beforehand and storing them in memory. This greatly speeds up the clustering, but requires at least 4GB of memory for the distance lookup table alone. Since the servers that host CPLOP only have 4GB of RAM in total, *OhClust!* cannot be directly incorporated into CPLOP.

In [29], Eric Johnson presents a density-based clustering algorithm for pyroprints in order to build an intuitive clustering method that uses density and nearness and an efficient range query algorithm to find nearby isolates. DBSCAN [20], short for Density-Based Spatial Clustering of Applications with Noise, can be efficient if the distance metric used satisfies the triangle inequality. Unfortunately, Pearson correlation coefficient does not satisfy the triangle inequality, but work in [29] adjusts the comparison metric to use the euclidean distance of z-score normalizations and optimizes further by organizing the pyroprints into a tree, making DBSCAN a viable method of clustering for the servers that host CPLOP.

An attempt to use [29] as a naïve MST method in [36] revealed that for the isolates that actually clustered (i.e. were not determined to be noise), the accuracy was fairly high. Essentially, [36] clusters an unknown-host-species isolate along with the rest of the known-host-species isolates CPLOP and classifies it as the most plural host-species in the resulting cluster. However, [29] clustered only about half of the isolates in CPLOP, while the rest remained unclustered and thus unclassified. The investigation in [36] was useful to confirm suspicions of so-called "transient" strains of *E. coli* bacteria. Subsection 3.3.1 discusses the details of [29] relevant to this thesis, Section 4.1 describes a potential methodology to use it as a MST method, and Section 7.1 expands upon the investigation in [36] and determines whether it can be useful to supplement a MST method like k-RAP.

The clustering methods presented in [41, 42, 64, 40, 39] and [29] are examples

of typical investigations into bacterial strain research. On their own, they do not constitute an actual MST methodology[2]. Ultimately, the goal of CPLOP is to be able to objectively classify the host-species of an *E. coli* isolate. Thus, merely clustering isolates is insufficient for MST. This thesis presents k-RAP, an MST technique that works with the pyroprints of isolates in CPLOP as a solution to MST.

2.4 k-NN Techniques

A plethora of k-Nearest Neighbors (k-NN) methods exist, but most are various attempts to optimize the search space — either with efficient range queries or by leveraging information about the space to improve search speed — or modifications to the neighbor list structure to improve classification. Surveys on k-NN techniques [7, 28] show that each variation builds data structures for efficient query, or abstracts the notion of the usually euclidean distance metric to build a more accurate classifier, while others may weight the neighbors or remove neighbors from consideration based off of some criteria An exception to the typically euclidean distance metric comes in the way of recommender systems [15, 47], which use a notion of similarity based off of scores. While efficient range query interests us, we have a solution for it in [29].

The method that most closely resembles what we are after comes from techniques that build multiple k-NN classifiers by generating feature subsets and polling the classifier to determine the class of the unknown datapoint [4, 5, 67]. Similar to bagging and bootstrapping techniques used train other classification algorithms, the feature-set of known datapoints is either reasonably partitioned into feature-subsets [4, 5], or clustered into subsets [67]. Some even perturb the data and group features to create multiple k-NN classifiers [30]. The resulting classification from the k-NN subset is then aggregated and the final classification is determined by majority voting. This

[2]The work in [36] attempts to use clustering as a MST technique.

approach will not apply to CPLOP, since we do not merely have a single comparison metric that we want to partition into multiple to improve classification. Isolates in CPLOP always have two entirely separate metrics that we must make a reasonable decision from.

The primary goal of the k-NN Resolution Algorithms for Pyroprints (k-RAP) is to resolve the two comparison metrics that CPLOP has for comparing isolates. That is, given an isolate, in order to find nearby isolates, one must separately compute the Pearson correlation coefficient ρ for each ITS, giving us two comparison metrics, $\rho_{ITS\text{-}1}$ and $\rho_{ITS\text{-}2}$. Typically, the vectors in k-NN techniques represent the entire set of features for a particular datapoint. Many k-NN algorithms assume that the distance metric used — usually euclidean distance — will encode a useful notion of distance.

k-RAP can apply to other datasets with separate comparison metrics, especially those that contain types of features that euclidean distance does not apply to. For example, in a demographic study for, say, a political study, subjects may have a multitude of features with different metrics of comparison. Location of residence may be one and favorite color another[3], with a goal of classifying a subject's political party. Euclidean distance may not be appropriate for the location metric, since great-circle distance on a globe may encode closeness more accurately. For color, while it may be straightforward to represent red, green, and blue values as a vector, euclidean distance may not be the best choice to gauge similarity in color, certainly not in the same way as the great-circle distance, especially for the reasons put forth in [37], which discusses the tremendous difficulties in building a uniform perceptive color space. Simply combining these two features into a single vector and performing euclidean distance may not produce the most appropriate results. Nevertheless, these metrics on their own are perfectly amenable to their own, accordant distance metric

[3]Certainly, many other psychological metrics can exist, but for simplicity's sake, let us consider only these two.

that cannot necessarily be used on other features, making it easy to create a k-NN for each feature separately. As such, when using k-NN on datasets with a complex set of features there is a need for the ability to resolve separate k-NN lists in order to usefully classify datapoints.

Chapter 3

BACKGROUND

3.1 Biological

3.1.1 Fecal Contamination

3.1.2 Microbial Source Tracking

3.1.3 Fecal Indicator Bacteria

3.1.4 Bacterial Strains

3.2 The Cal Poly Library of Pyroprints

3.2.1 *Escherichia coli* Isolates

3.2.2 Pyroprints

3.2.3 Pearson Correlation Coefficient

3.2.4 Database

3.3 Computational

3.3.1 Density-Based Clustering of Pyroprints

3.3.2 k-Nearest Neighbors

Chapter 4

METHODOLOGY

4.1 Clustering for Bacterial Strains

4.2 k-RAP

Chapter 5

IMPLEMENTATION

5.1 Graphing Cluster Metrics

5.2 Resolution Algorithms

Chapter 6

EVALUATION

6.1 Cluster Purity

6.2 Classification Metrics

Chapter 7

RESULTS

7.1 Clustering

7.2 Classifying

Chapter 8

CONCLUSION

BIBLIOGRAPHY

[1] *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2012, Philadelphia, USA, October 4-7, 2012*. IEEE, 2012.

[2] C. C. Adams. Using Hadoop to Identify False Positives in Bacterial Strain Typing from DNA Fingerprints. *California Polytechnic State University, San Luis Obispo*, 2016.

[3] J. M. Albert, J. Munakata-Marr, L. Tenorio, and R. L. Siegrist. Statistical evaluation of bacterial source tracking data obtained by rep-PCR DNA fingerprinting of Escherichia coli. *Environmental science & technology*, 37(20):4554–4560, 2003.

[4] S. D. Bay. Combining nearest neighbor classifiers through multiple feature subsets. In Shavlik [60], pages 37–45.

[5] S. D. Bay. Nearest neighbor classification from multiple feature subsets. *Intell. Data Anal.*, 3(3):191–209, 1999.

[6] L. Belanche-Muñoz and A. R. Blanch. Machine learning methods for microbial source tracking. *Environmental Modelling & Software*, 23(6):741–750, 2008.

[7] N. Bhatia and Vandana. Survey of nearest neighbor techniques. *CoRR*, abs/1007.0085, 2010.

[8] G. Bitton. Microbial indicators of fecal contamination. *Wastewater Microbiology, Third Edition*, pages 153–171, 2005.

[9] M. W. Black, J. VanderKelen, A. Montana, A. Dekhtyar, E. Neal, A. Goodman, and C. L. Kitts. Pyroprinting: A rapid and flexible genotypic

fingerprinting method for typing bacterial strains. *Journal of Microbiological Methods*, 105:121 – 129, 2014.

[10] D. Brandt, A. Montana, B. Somers, M. Black, A. Goodman, and C. Kitts. Pyroprinting sensitivity analysis on the GPU. In 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2012, Philadelphia, USA, October 4-7, 2012 [1], pages 951–953.

[11] Cal Poly. Cal Poly Github, 2016. `http://www.github.com/CalPoly`.

[12] K. R. CLARKE. Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1):117–143, 1993.

[13] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Information Theory*, 13(1):21–27, 1967.

[14] T. R. Desmarais, H. M. Solo-Gabriele, and C. J. Palmer. Influence of soil on fecal indicator organisms in a tidally influenced subtropical environment. *Applied and environmental microbiology*, 68(3):1165–1172, 2002.

[15] C. Desrosiers and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In Ricci et al. [53], pages 107–144.

[16] J. W. Dickerson Jr. *Evaluation, Development and Improvement of Genotypic, Phenotypic and Chemical Microbial Source Tracking Methods and Application to Fecal Pollution at Virginia's Public Beaches*. PhD thesis, Virginia Polytechnic Institute and State University, 2008.

[17] J. R. Dillard. Demographics and Transfer of Escherichia coli Within Bos taurus Populations. Master's thesis, California Polytechnic State University, San Luis Obispo, 2015.

[18] J. R. Dillard, J. J. VanderKelen, J. D. Kent, A. D. Frey, P. J. McCreesh, D. Britton, T. Branck, M. W. Black, and C. L. Kitts. E. coli Strain Demographics and Transmission in Cattle. *Strain*, 10(1):11, 2013.

[19] W. Ding, T. Washio, H. Xiong, G. Karypis, B. M. Thuraisingham, D. J. Cook, and X. Wu, editors. *13th IEEE International Conference on Data Mining Workshops, ICDM Workshops, TX, USA, December 7-10, 2013*. IEEE Computer Society, 2013.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data MiningI*, pages 226–231. AAAI Press, 1996.

[22] R. A. Finkel and J. L. Bentley. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9, 1974.

[23] E. Fix and J. L. Hodges Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, DTIC Document, 1951.

[24] E. Fix and J. L. Hodges Jr. Discriminatory analysis-nonparametric discrimination: Small sample performance. Technical report, DTIC Document, 1952.

[25] Q. Hua, A. Ji, and Q. He. Multiple real-valued K nearest neighbor classifiers system by feature grouping. In Proceedings of the IEEE International

Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10-13 October 2010 [49], pages 3922–3925.

[26] J. Huan, S. Miyano, A. Shehu, X. T. Hu, B. Ma, S. Rajasekaran, V. K. Gombar, M. Schapranow, I. Yoo, J. Zhou, B. Chen, V. Pai, and B. G. Pierce, editors. *2015 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2015, Washington, DC, USA, November 9-12, 2015*. IEEE Computer Society, 2015.

[27] *International Conference on Machine Learning and Cybernetics, ICMLC 2010, Qingdao, China, July 11-14, 2010, Proceedings*. IEEE, 2010.

[28] L. Jiang, Z. Cai, D. Wang, and S. Jiang. Survey of improving k-nearest-neighbor for classification. In Lei [33], pages 679–683.

[29] E. Johnson. Density-Based Clustering of High-Dimensional DNA Fingerprints for Library-Dependent Microbial Source Tracking. Master's thesis, California Polytechnic State University, San Luis Obispo, 2015.

[30] W. Juan. Multiple nearest neighbor classifiers system based on feature perturbation by mutual information. In International Conference on Machine Learning and Cybernetics, ICMLC 2010, Qingdao, China, July 11-14, 2010, Proceedings [27], pages 247–251.

[31] J. Kent, M. Alvarado, J. VanderKelen, A. Montana, J. Soliman, A. Dekhtyar, A. Goodman, C. Kitts, and M. Black. Pyroprinting: Novel Pyrosequencing-Based Method for Studying E. coli Diversity and Microbial Source Tracking (779.8). *The FASEB Journal*, 28(1 Supplement):779–8, 2014.

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

[33] J. Lei, editor. *Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007, 24-27 August 2007, Haikou, Hainan, China, Proceedings, Volume 1*. IEEE Computer Society, 2007.

[34] W. Li, D. Raoult, and P.-E. Fournier. Bacterial strain typing in the genomic era. *FEMS Microbiology Reviews*, 33(5):892–916, 2009.

[35] J. D. McGovern, A. Dekhtyar, C. Kitts, M. Black, J. Vanderkelen, and A. Goodman. Leveraging the k-nearest neighbors classification algorithm for microbial source tracking using a bacterial DNA fingerprint library. In Huan et al. [26], pages 1694–1701.

[36] J. D. McGovern, E. Johnson, A. Dekhtyar, M. Black, C. Kitts, and J. Vanderkelen. Library-based microbial source tracking via strain identification. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2016, Seattle, WA, USA, October 2-5, 2016 [48], pages 364–373.

[37] R. N. McLeod. A Proof of Concept for Crowdsourcing Color Perception Experiments. *California Polytechnic State University, San Luis Obispo*, 2014.

[38] D. J. Meagher. *Octree encoding: A new technique for the representation, manipulation and display of arbitrary 3-d objects by computer*. Electrical and Systems Engineering Department Rensseiaer Polytechnic Institute Image Processing Laboratory, 1980.

[39] A. Montana. Algorithms for Library-Based Microbial Source Tracking. Master's thesis, California Polytechnic State University San Luis Obispo, 2013.

[40] A. Montana, A. Dekhtyar, M. Black, C. Kitts, and A. Goodman. Ontological hierarchical clustering for library-based microbial source tracking. In Ding et al. [19], pages 568–576.

[41] A. Montana, A. Dekhtyar, E. Neal, M. Black, and C. Kitts. Chronology-sensitive hierarchical clustering of pyrosequenced DNA samples of e. coli: A case study. In Wu et al. [69], pages 155–159.

[42] A. Montana, A. Dekhtyar, E. Neal, M. Black, and C. Kitts. Investigating temporal strain diversity in human e. coli populations using pyroprinting: A novel strain identification method. Technical report, Technical report, California Polytechnic State University, San Luis Obispo, CA, 2012.

[43] C. Moritz, D. Shapiro, and C. Pann. Application of Pyroprinting for Source Tracking of E. coli in Pennington Creek. *California Polytechnic State University, San Luis Obispo*, 2015.

[44] E. Neal, C. Sabatini, W. Tang, M. Black, and C. Kitts. Demographics of E. coli Strains in the Human Gut Using Pyroprints: A Novel MST Method. In *CSUPERB, Poster*. Jan, 2012.

[45] E. R. Neal. Escherichia coli Strain Diversity in Humans: Effects of Sampling Effort and Methodology. Master's thesis, California Polytechnic State University, San Luis Obispo, 2013.

[46] J. Nguyen, J. Vanderkelen, M. Black, and C. Kitts. Investigating the Dominant Escherichia coli Strain in Lambs and Ewes Using Pyroprinting: A Novel Method for Strain Identification. *California Polytechnic State University, San Luis Obispo*, 2015.

[47] X. Ning, C. Desrosiers, and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In Ricci et al. [52], pages 37–76.

[48] *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2016, Seattle, WA, USA, October 2-5, 2016*. ACM, 2016.

[49] *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10-13 October 2010*. IEEE, 2010.

[50] V. Ramachandran, editor. *Proceedings of the Fourth Annual ACM/SIGACT-SIAM Symposium on Discrete Algorithms, 25-27 January 1993, Austin, Texas*. ACM/SIAM, 1993.

[51] S. Ranka, iTamer Kahveci, and M. Singh, editors. *ACM International Conference on Bioinformatics, Computational Biology and Biomedicine, BCB' 12, Orlando, FL, USA - October 08 - 10, 2012*. ACM, 2012.

[52] F. Ricci, L. Rokach, and B. Shapira, editors. *Recommender Systems Handbook*. Springer, 2015.

[53] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors. *Recommender Systems Handbook*. Springer, 2011.

[54] C. Ricketts. Cal Poly Library of Pyroprints: Quality Control Analysis and Web Development. Master's thesis, California Polytechnic State University, San Luis Obispo, 2014.

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

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

[57] D. Sargeant, W. R. Kammin, and S. Collyard. *Review and critique of current microbial source tracking (mst) techniques*. Environmental Assessment Program, Washington State Department of Ecology, 2011.

[58] T. M. Scott, J. B. Rose, T. M. Jenkins, S. R. Farrah, and J. Lukasik. Microbial source tracking: Current methodology and future directions. *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, pages 5796–5803, 2002.

[59] D. Shapiro, J. Kent, M. Zuleta, C. Kitts, M. Black, and J. VanderKelen. Source Tracking of Fecal Contamination Along San Luis Obispo (SLO) Creek. *The FASEB Journal*, 29(1 Supplement):575–12, 2015.

[60] J. W. Shavlik, editor. *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconsin, USA, July 24-27, 1998*. Morgan Kaufmann, 1998.

[61] D. Shealy. Exploration of PyroPrinting for Environmental Forensics. Technical report, California Polytechnic State University, San Luis Obispo, California, June 2012.

[62] J. M. Simpson, J. W. Santo Domingo, and D. J. Reasoner. Microbial source tracking: state of the science. *Environmental science & technology*, 36(24):5279–5288, 2002.

[63] J. L. Soliman. CPLOP: The Cal Poly Library of Pyroprints. Master's thesis, California Polytechnic State University San Luis Obispo, 2013.

[64] J. L. Soliman, A. Dekhtyar, J. Vanderkellen, A. Montana, M. Black, E. Neal, K. Webb, C. Kitts, and A. Goodman. Microbial source tracking by molecular fingerprinting. In Ranka et al. [51], pages 617–619.

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

[66] J. J. VanderKelen, R. D. Mitchell, A. Laubscher, M. W. Black, A. L. Goodman, A. K. Montana, A. M. Dekhtyar, R. Jimenez-Flores, and C. L. Kitts. Short Communication: Typing and Tracking Bacillaceae in Raw Milk and Milk Powder Using Pyroprinting. *Journal of Dairy Science*, 99(1):146–151, 2016.

[67] L. Wang, Q. Hua, X. Wang, and Q. Chen. Combination of multiple nearest neighbor classifiers based on feature subset clustering method. In Yeung et al. [70], pages 538–547.

[68] K. Webb. Cplop-cal poly's library of pyroprints. *California Polytechnic State University, San Luis Obispo*, 2011.

[69] F. Wu, M. J. Zaki, S. Morishita, Y. Pan, S. Wong, A. Christianson, and X. Hu, editors. *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2011, Atlanta, GA, USA, November 12-15, , 2011*. IEEE Computer Society, 2011.

[70] D. S. Yeung, Z. Liu, X. Wang, and H. Yan, editors. *Advances in Machine Learning and Cybernetics, 4th International Conference, ICMLC 2005, Guangzhou, China, August 18-21, 2005, Revised Selected Papers*, volume 3930 of *Lecture Notes in Computer Science*. Springer, 2006.

[71] P. N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In Ramachandran [50], pages 311–321.