Microbial Source Tracking by Molecular Fingerprinting

Jan Lorenz Soliman
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA
jsoliman@calpoly.edu

Aldrin Montana
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA
amontana@calpoly.edu

Kevin Webb
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA
kwebb01@calpoly.edu

Alex Dekhtyar
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA
dekhtyar@calpoly.edu

Michael Black
Biological Sciences Department
California Polytechnic State University
San Luis Obispo, CA USA
mblack@calpoly.edu

Chris Kitts
Biological Sciences Department
California Polytechnic State University
San Luis Obispo, CA USA
ckitts@calpoly.edu

Jennifer Vanderkellen
Biological Sciences Department
California Polytechnic State University
San Luis Obispo, CA USA
jvanderkelen@msn.com

Emily Neal
Biological Sciences Department
California Polytechnic State University
San Luis Obispo, CA USA
erusch@calpoly.edu

Anya Goodman
Department of Chemistry and Biochemistry
California Polytechnic State University
San Luis Obispo, CA USA
agoodman@calpoly.edu

ABSTRACT

To date, microbial source tracking (MST), i.e. determining the source of microbial contamination based on the specific strains observed in environment, is done using methods that are time-consuming, expensive and not always reliable. The biology department at Cal Poly, San Luis Obispo has developed a new method for MST called pyroprinting. Pyroprints are a result of pyrosequencing replicates of intergenic transcribed spacer (ITS) regions in a target bacterial genome. *E. coli* pyroprints can be used as DNA fngerprints of individual *E. coli* strains in identifying sources of fecal contamination and studying bacterial patterns in host animals. The MST method consists of two parts: the pyroprinting process and a database of sequenced pyroprints. The actual source tracking is achieved by comparing a newly obtained pyroprint to the pyroprints of known provenance from a database. In this paper, we describe the design and implementation of Cal Poly Library of Pyroprints (CPLOP). The CPLOP database provides storage and essential analysis of pyroprints for strain identification. Our current implementation contains pyroprints of bacterial isolates of *E. coli*, obtained by students and researchers from known hosts and from the environment. Users of CPLOP are able to organize pyroprints into groups, run analyses to find similarities between bacterial isolates, and cluster isolates into bacterial strains.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*scientific databases*;
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*clustering*;
J.3 [**Computer Applications**]: Life and Medical Sciences—*Biology and genetics*;

General Terms

Algorithms, Design

1. INTRODUCTION

Diseases caused by bacterial pathogens in food and water continue to be a major public health issue in California and around the world. Microbial Source Tracking (MST), a relatively new field of study related to epidemiology, is focused on identifying and tracking pathogen contamination in foods and resource waters. Because feces are a major source of pathogens, one common approach used for MST investigations is to collect a database of fingerprinted bacterial isolates from known fecal sources [1, 2]. The MST method consists of two parts: the bacterial isolation and fingerprinting process, and the construction of a database of fingerprints. By comparing fingerprints of isolated bacteria collected from food or water samples to the database of fecal isolates with known provenance, investigators can determine the contaminating source of animal feces in the food or water samples.

Many fingerprinting methods used for existing MST databases are time-consuming, expensive and difficult to reproduce across labs. Our research team has developed a new fingerprinting method for MST called pyroprinting. Pyroprints are produced by the simultaneous pyrosequencing of all copies

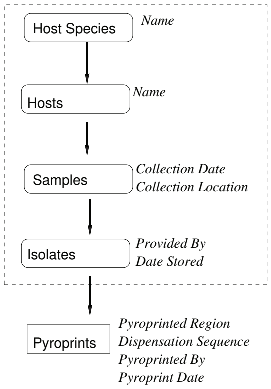

Figure 1: Provenance structure for isolate and pyroprint data in CPLOP.

of a specific intergenic transcribed spacer (ITS) region in a target bacterial genome. We are collecting pyroprints of *E. coli* made from the 23S-5S and 16S-23S ITS regions, both of which appear in *E. coli* in septuplicate.

2. CPLOP

CPLOP is a crucial part of the emerging molecular forensics program at Cal Poly. It provides cataloging and analytical support for various MST projects conducted by faculty and students in the biology program. CPLOP was initially designed and piloted by an undergraduate database application design course, and is currently under development by a small team of graduate and undergraduate students.

2.1 Data Collection

Initial collection of pyroprint data has been conducted in the classroom, student project, and research settings. In the classroom setting, students taking Cal Poly's General Microbiology course (MICRO 224) isolate, identify and store *E. coli* from fecal samples from a variety of hosts (e.g. canine, feline and bovine) or their own anal swab samples[1]. Individual senior projects involved Biology and Microbiology majors collecting environmental samples from a number of locations throughout San Luis Obispo County and isolating *E. coli* from them, as well as collecting fecal samples from a variety of farm animals hosted on Cal Poly's farm units. Faculty and student research projects yielded collections of bacterial isolates in longitudinal studies of *E. coli* diversity in humans and other host species.

2.2 Pyroprint Database

CPLOP stores information about individual pyroprints, as well as their provenance information, broken into the following components: isolate of origin, sample, host and host species. Each of the components is outlined below. The overall provenance structure is shown in Figure 1.

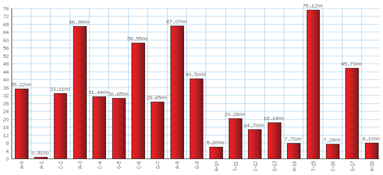

Figure 2: A pyroprint as a vector of peak heights.

Pyroprints. A pyroprint is a vector of values describing light intensity at each reagent dispensation during the pyrosequencing process. Individual dispensations are represented as tuples consisting of the peak height intensity value and the corresponding dispensed nucleotide reagent. An example peak height graph of a pyroprint is shown in Figure 2.

With each pyroprint, we associate additional meta-data describing its nature: the ITS region that was pyrosequenced, the dispensation sequence used, the date and time of the pyroprinting and the name of the person who conducted the process.

Isolates. Pyroprints are grouped into the isolates that they were sequenced from. From each sample, individual isolates of *E. coli* are frozen and maintained by the Biology department. The number of isolates from a single sample can vary widely. Some samples may provide only one or two isolates, while others may have hundreds.

Samples. Samples are gathered from the environment. The data from each sample, which is normally a fecal sample is stored as Isolates and metadata associated with the sample itself. Each sample is produced by a specific host. Like the number of isolates in a sample, the number of samples from each host can vary widely.

Hosts. A host is a unique individual belonging to one of the host species. CPLOP contains information about multiple hosts from which *E.coli* Isolates were drawn. For example, there may be three different horse hosts: horse1, horse2 and horse3. With each host, we associate the following information: a name of the specific host (e.g., 'horse1' or 'Mr.Ed') and the host species.

Host species. Hosts belong to Host Species, which is the species of the host that the sample was gathered from. In CPLOP, the Host Species has two fundamental features used to distinguish between them: a common name, such as Cat, Dog or Seagull and a Latin name, such as *Felis catus*, *Canis lupus familiaris*, *Larus californicus*.

CPLOP currently contains over 4,500 pyroprints from the DNA of *E. coli* isolates that have been isolated and pyrosequenced by students and researchers at the Cal Poly Biology department. These 4,500 pyroprints make up 2,900

The Pyroprint Database provides a relational database for storing pyroprints and their associated information (isolate, sample, host, host species). In addition, the Pyroprint Database provides a framework for linking host species to pyroprints. Once strains are developed, they can be associated to host

[1] Approximately 120 students across campus take this course, offered twice a year. Thus we obtain around 200-240 new isolates each year for sequence analysis from this course alone.

```
function CLUSTERDATASTRUCTURE(N)
    $C \leftarrow 0$
    if $N = null$ then
        return $C$
    end if
    if $|children(N)|! = 0$ then
        for $n \in children(N)$ do
            $C \leftarrow C \cup clusterDataStructure(n_i)$
        end for
        $C \leftarrow C \cup performHierarchical(C)$
    else
        $C \leftarrow performHierarchical(C_n)$
    end if
    return $C$
end function
```

Figure 3: Traversing the data structure for Clustering

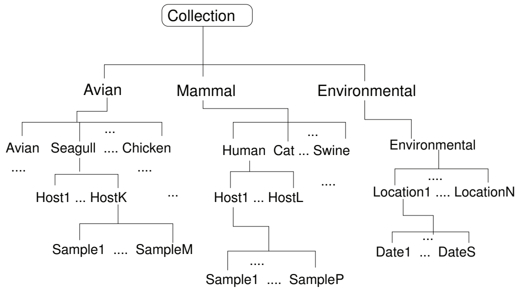

Figure 4: CPLOP Ontology

species.

CPLOP is deployed at the url http://cplop.cosam.calpoly.edu. At present, it hosts information about all pyroprints built by our research group over the past 15 months, and provides the functionality for comparing individual pyroprints from the database and determining which pyroprints match, i.e., are likely to have originated from the same bacterial strain.

3. STRAIN MAINTENANCE

To identify and maintain strains we use a variation of hierarchical clustering. We developed a new algorithm, *OH Clust* (Ontology-based Hierarchical Clustering), which is a modification of the traditional agglomerative clustering algorithm adapted to take advantage of known bacterial isolate metadata (isolate identity, host species, date/time and location of collection). We organize our isolate clustering by point of origin. The OH Clust algorithm starts by clustering the isolates belonging to the leaf nodes in the ontology. Once all of the leaf nodes are clustered, the clusters are propagated to the parent. At this point the inter-cluster distances are recomputed. Hierarchical clustering is then performed again on the recomputed clusters. This is then done recursively for each parent node until it gets to the top of the ontology. The psuedo-code for the OH Clust algorithm is provided in Figures 3 and 5. Figure 3 illustrates how the ontology is traversed. Figure 5 describes how the clustering is implemented and Figure 5 shows how the inter-cluster distance is recomputed at each level.

```
function PERFORMHIERARCHICAL(C)
    $C' \leftarrow C$
    $(c_a, c_b) \leftarrow 0$
    $S[j,k] \leftarrow recomputeDistances(C')$
    for $c_j, c_k \in C'$ do
        $S[j,k] \leftarrow sim(c_j, c_k)$
    end for
    while $|C'| > 1$ and $sim(c_a, c_b) \geq \alpha$ do
        $(c_a, c_b) \leftarrow argmin(S[j,k])$
        if $sim(c_a, c_b) \geq \alpha$ then
            $C' \leftarrow C' \cup combineClusters(c_a, c_b)$
            $S[j,k] \leftarrow recomputeDistances(C')$
        end if
    end while
    for $c_i \in C$ do
        if $|c_i = 1$ then
            $W \leftarrow W \cup c_i$
        end if
    end for
    return C
end function
function RecomputeDistances(C)
    for $c_j, c_k \in C$ do
        if $sim(c_i, c_k) \geq \alpha$ then
            $S[j,k] \leftarrow 1$
        else if $sim(c_i, c_k) < \beta$ then
            $S[j,k] \leftarrow 0$
        end if
    end for
    return $S$
end function
```

Figure 5: Hierarchical Clustering

OH Clust then uses the ontological structure to partition a given dataset and to cluster these partitions in a directed order. Using this ontology provides flexibility in user-specified ontologies, scalability in incremental updates, and access to more information in being able to query sub-relationships in the analyzed data.

4. CONCLUSION

CPLOP currently is able to store pyroprints and all relevant information, store pyroprinting protocols and parameters, store isolate metadata, maintain user-specified groups of pyroprints, and match pyroprints against other pyroprints. CPLOP will have support for clustering to allow for strain identification and strain maintenance by integrating the OH Clust algorithm into the application.

5. ACKNOWLEDGMENT

This work was supported in part by an undergraduate education grant from the W.M. Keck foundation.

6. REFERENCES

[1] T. M. Scott, J. B. Rose, T. M. Jenkins, S. R. Farrah, and J. Lukasik. Microbial source tracking: Current methodology and future directions. *Appl. Environ. Microbiol.*, 68:5796 – 5803, dec 2002.

[2] D. M. Stoeckel and V. J. Harwood. Performance, design, and analysis in microbial source tracking studies. *Applied and Environmental Microbiology*, 73:2405 – 2415, apr 2007.