Chronology-Sensitive Hierarchical Clustering of Pyrosequenced DNA Samples of E. coli: A Case Study

Aldrin Montana
Alex Dekhtyar
Computer Science Department
California Polytechnic State University
San Luis Obispo, United States
{*amontana, dekhtyar*}*@calpoly.edu*

Emily Neal
Michael Black, Chris Kitts
Biology Department
California Polytechnic State University
San Luis Obispo, United States
{*erusch, ckitts, mblack*}*@calpoly.edu*

***Abstract*—Hierarchical clustering is used in computational biology as a method of comparing sequenced bacterial strain DNA and determining bacterial isolates that belong to the same strain. However, the results of the hierarchical clustering are, at times, difficult to read and interpret. This paper is a case study for the use of a modified hierarchical clustering algorithm, which takes into account the underlying structure of the bacterial DNA isolate collection to which it is applied.**

***Keywords*-bioinformatics; clustering; pyrosequence; pyrogram;**

I. Introduction

Escherichia coli (*E. coli*) are commensal inhabitants of the human gut[1][2] and also thrive in the intestinal tracts of most other mammals and some birds[1][3]. *E. coli* is frequently used as an indicator for fecal contamination in watersheds, lakes, beaches, and recreational water [4][5][6][7]. Because dangerous interspecific pathogens can be transferred through contaminated water, it is necessary for health and environmental protection agencies to be able to track a source of fecal contamination at the species level [5][6][7]. The general process linking microbes (in this case *E. coli*) to a host source is called Microbial Source Tracking (MST) [5][6][7].

Our research group is currently developing a cost-effective and efficient library-dependent MST method to create DNA fingerprints for different strains of *E. coli* using pyrosequencing; which we refer to as *pyroprinting*. In a pilot study, pyroprinting was used to investigate the variation in *E. coli*. Characterizing *E. coli* populations and their variation in humans is important not only to build an MST library but also to further understand the human interaction with this commensal organism.

In this paper, we compare the use of two clustering algorithms for *E.coli* strain detection in the context of a short-term single-host case study. The first algorithm, *Primer5*[8], is an implementation of the traditional hierachical clustering method, widely used in biology. To address the shortcomings of this method, we developed a hierarchical clustering algorithm that is sensitive to the internal organization of our collection of data. The rest of this paper is organized as follows. Section II provides a brief overview of the single-host study and the experimental data. Section III discusses related work and in section IV we introduce our chronology-sensitive hierarchical clustering algorithm. Finally, section V discusses the results of the case study.

II. Experimental Setup

Fecal samples from a single human subject were collected once a day for 14 days in September 2010. Samples were manually homogenized, and a sterile swab was inserted into the sample. An anal swab immediately after defecation and an anal swab a few hours later were also collected. All swabs were used to streak the samples onto MacConkey agar. Samples were not collected on day 7 due to the absence of defecation on that day. Figure 1 shows how many isolates were collected on each day.

All bacterial cultures were grown at 37° C overnight. Up to four pink isolates from each MacConkey plate were selected for biochemical *E. coli* confirmation. Isolates from the fecal sample were tagged *group F*, immediate swabs as *group I*, and late swabs as *group L*[1]. Each selected isolate was re-streaked onto MacConkey. Half of a pink isolated colony from the second MacConkey plate was patched on LB agar; the other half was patched on EMB agar. If a green, metallic sheen was observed on EMB, the isolate from LB was tested in tryptone broth for indole production and on citrate agar. *E. coli* were confirmed when isolates were positive for indole and negative for citric aid utilization.

Colony polymerase chain reaction[2] was performed on each confirmed *E. coli* isolate. Primers (listed below) designed to amplify the 23S rRNA - 5S rRNA Intergenic Transcribed Spacer (ITS) region were placed in consensus

[1] Isolate labelling further denotes the day the isolate was collected on, followed by an identifying number, i.e. *F1-1* corresponds to an isolate from a fecal sample on day 1, and is identified as the first of that group

[2] PCR parameters were: (1) 95°C, 2 minutes; (2) 95°C, 30 seconds; (3) 56°C, 30 seconds; (4) 68°C, 1 minute; (5) repeated steps (2)-(4) another 44 times; (6) 68°C, 5 minutes; (7) 4°C hold.

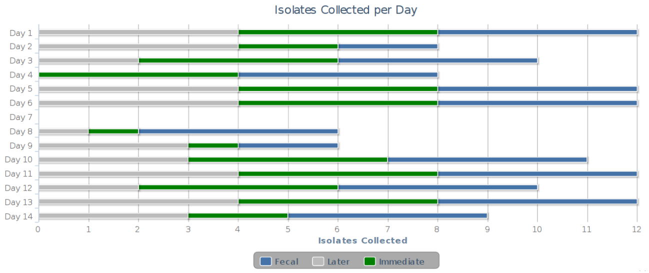

Figure 1. Isolates Per Day

regions of both rRNA genes. PCR products were used for pyrosequencing analysis (sequencing primer, S, is also listed below).

Primers

F:	5′- ATG AAC CGT GAG GCT TAA CCT T -3′
R:	5′-biotin- CTA CGG CGT TTC ACT TCT GAG T -3′
S:	5′- CGT GAG GCT TAA CCT T -3′

A. Pyroprinting method

Pyrosequencing is a DNA sequencing process where a new strand of DNA is built incrementally. Nucleotides A, T, C, or G are introduced to the DNA strand being sequenced in an order determined by a dispensation sequence. If the introduced nucleotide complements the next unbound nucleotide in the target DNA strand then the two nucleotides bind to each other, extending the new DNA strand and emitting light. This light is measured and used to construct a pyrogram of the DNA strand. [9] describes the pyrosequencing process in more detail.

In our case study we pyrosequenced the 23S rRNA - 5S rRNA ITS region of the *E. coli* genome. Because this region is non-coding, it accumulates more mutations than regions of DNA that code for important proteins or RNAs. This ITS region is present in *seven* locations in the *E. coli* genome. In this study, all seven copies of this ITS region were sequenced together to create a *pyroprint*[3]. Under the assumption that ITS sequences vary for different strains of *E. coli*, a pyroprint acts as a DNA fingerprint for each strain. This process is depicted in Figure 2.

B. Data Description

For each isolate, a mix of the 23S rRNA - 5S rRNA was extracted and amplified using PCR. The intragenic region was then pyrosequenced and the obtained pyroprints were compared to each other.

Given an isolate X, its pyroprint is a sequence $\bar{X} = (x_1, \dots x_N)$ of real numbers[4] reported by the pyrosequencing equipment for each of the dispensations. In our experiments, pyroprints of length $N = 104$ were generated following the same dispensation sequence of nucleotides for each bacterial isolate. Pyroprint similarity was calculated using pearson correlation coefficient:

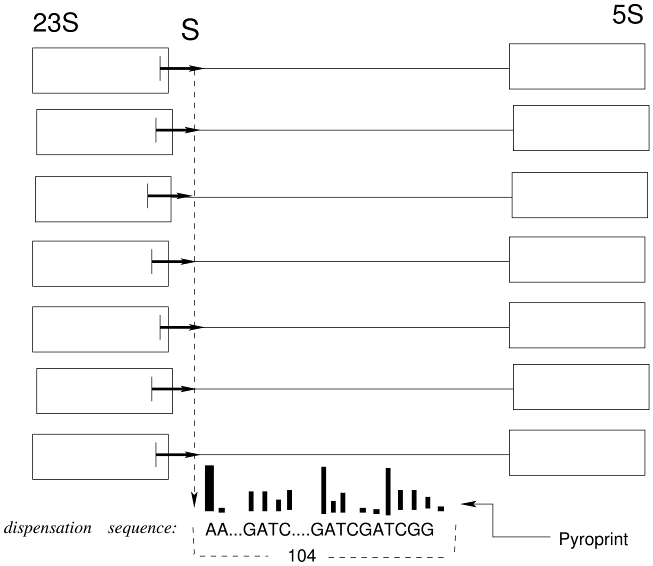

Figure 2. Pyroprinting process: light intensities (black bars at the bottom) are reported for each nucleotide in the dispensation sequence. Open boxes represent conserved DNA sequences in the 23S and 5S rRNA genes and S indicates the point at which the sequencing primer binds to the DNA, beginning the sequencing process.

$$sim(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^{N}(x_i - E(\bar{X}))(y_i - E(\bar{Y}))}{\sqrt{\sum_{i=1}^{N}(x_i - E(\bar{X})}\sqrt{\sum_{i=1}^{N}(y_i - E(\bar{Y})}},$$

where $E(\bar{X})$ and $E(\bar{Y})$ are means of the respective sequences. The goal of the study was to determine which bacterial isolates belong to the same strain. If $sim(\bar{X}, \bar{Y})$ is sufficiently close to 1, we assume X and Y come from the same strain. If this assumption is too strong — it is possible X and Y belong to different strains, yet appear to match. However, if $\bar{X}$ and $\bar{Y}$ are *sufficiently different*, then X and Y *definitely come from different strains.*

A set of pyroprints is considered *strongly connected* if each pyroprint in the set is sufficiently similar (i.e. *connected*) to every other pyroprint in the set. We discuss the notions of *sufficient similarity* and *sufficient dissimilarity* more formally in Section IV.

III. Related Work

Traditional clustering focuses on direct relationships between data. Distance functions are defined and the process of clustering is as simple as grouping data with the highest similarity measures, then applying thresholds to maximal dissimilarity measures in a cluster.

Temporal clustering, as described by Kamath and Caverlee [10], is a variation of clustering applied to a communica-

[3] A vector of light emission values based on combinations of all seven ITS region sequences for the pyrosequenced *E. coli* isolate

[4] representing light intensities

tion network. Nodes represent members in the communication network while edge weights represent communication between said nodes. The edge weights in the network are based on when the messages are exchanged.

PoClustering (partially ordered clustering) clusters data into *PoSets* (partially ordered sets) by finding all clique clusters for all possible diameters W(D) where D is the maximal dissimilarity in a dissimilarity matrix[11]. PoClustering is a generalization of both heierarchical and pyramidal clustering that allows overlaps between clusters such that a PoCluster P is defined as P = {cliqueset $_{\delta}$ (d) | ∀ d ∈ W(D)}[12]. PoClusters can be represented as a directed acyclic graph, with each node representing a clique cluster and its diameter, and each edge representing subset relationships between nodes. PoClusters are able to successfully preserve the majority of relationships present in the data whereas hierarchical clustering is unable to do so.

Temporal clustering modifies similarity measures in context of the temporal locality of particular events of interest. This is slightly different from the method we propose in this paper, as we do not modify similarity measures of *E. coli* isolates based on chronological distance. Instead, we simply enforce a particular ordering on cluster candidates. Similarly, PoClustering enforces a particular ordering on the clustering process without modifying similarity measures. Although PoClustering is not time-sensitive, it is important related work for our method regarding connectivity constraints between *E. coli* isolates.

Primer5 [8] is a hierarchical clustering tool commonly used by biologists. Hierarchical clustering works by iteratively combining clusters until there is one cluster remaining[13]. We use Primer5 as a benchmark to compare our algorithm against. Section V-A contains our analysis.

IV. Chronology-Sensitive Clustering

During pyrosequencing, individual light intensities are subject to fluctuation which arises from the difference in the experimental conditions under which individual isolates are sequenced. Our data contains a large number of pairs of isolates, whose pyrosequences have similarity between 0.995 and 1. In such situations, knowing that two pyroprints from the same day have a high similarity is sufficient to put them into a single cluster *right away*, even though a pyroprint from another day may have a higher similarity with one of them. Our approach leverages chronology-related information from the dataset to change the order in which hierarchical clustering combines clusters. In our algorithm, clusters are first formed out of isolates collected on the same day (further subdivided by collection method), then grown in chronological order across days.

The second modification in the algorithm is the transformation of the similarity scores. Our algorithm takes as input two parameters $\alpha > \beta$ representing similarity thresholds. A Pearson correlation score above α is replaced with the score of 1, indicating the two pyroprints are the same. Similarity scores below β indicate the respective pyroprints are definitely dissimilar and are replaced with 0. Scores between α and β are left intact. This transformation is performed any time when intercluster distances are computed.

A. Algorithm

For each pyroprint we use two additional attributes: the day it was collected on and the group it was collected in. Different collection groups may be considered closer or further away from each other: e.g., in our study, isolates from groups *F* and *I* are closer to each other than to the isolates from group *L*. We provide a distance relationship between the collection groups as one of the inputs to the algorithm. The algorithm proceeds as follows.

- Input: Matrix M of pairwise Pearson correlations between isolate pyroprints; thresholds $\alpha \in [0, 1]$, $\beta \in [0, 1]$; distance relationship between collection groups.
- Output: a dendrogram of isolates.
- Step 1. Matrix transformation. Each similarity score $M[i, j] > \alpha$ is replaced with 1; each similarity score $M[i, j] < \beta$ is replaced with 0;
- Step 2. Clustering within days. For each day:
  - Step 2.1. Cluster within collection group. For each collection group cluster all the isolates with similarity scores of 1.
  - Step 2.2. For each pair of collection groups in order of the distance relationship: combine clusters from individual groups with similarity scores of 1.
    Each time two clusters are combined, recompute the similarity matrix; apply the Step 1 procedure to it.

  At the end of this step, all isolates within each day will be partitioned into *strongly connected* clusters.
- Step 3. Chronological clustering. Starting from days 1 and 2, adding one day at a time combine clusters from different days for clusters with similarity scores of 1.
  Each time two clusters are combined, recompute the similarity matrix; apply the Step 1 procedure to it.
  Continue until no two clusters have similarity of 1.
- Step 4. Hierarchical clustering. Perform hierarchical clustering on the clusters constructed on Steps 2-3.

Our algorithm can be used with different inter-cluster distance measures. We implemented average link, complete link, single link, and Ward's method. For this study we used average-link distance. Average link cluster distance was calculated by averaging across all possible pairwise similarities between the isolates from the two clusters whose similarity is being calculated.

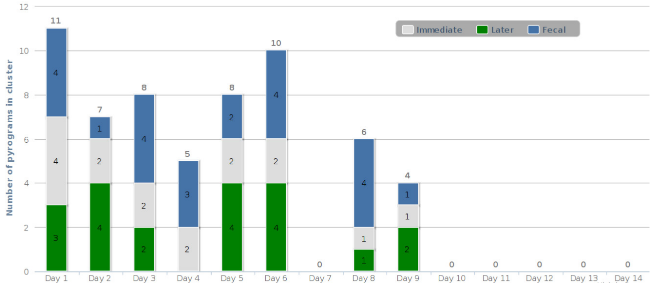

Figure 3. Cluster 1

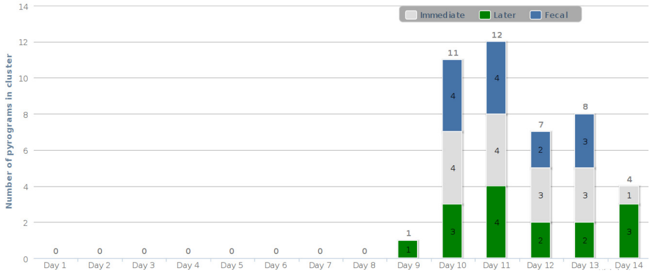

Figure 4. Cluster 2

	Cluster A	Cluster B	Cluster C	No Cluster
Cluster 1	58	1	0	0
Cluster 2	0	42	0	1
Cluster 3	0	3	4	0
Cluster 4	1	0	0	1
No Cluster	0	1	0	16

Table I
PYROPRINT CLUSTER CONFUSION MATRIX: CHRONOLOGY-SENSITIVE RESULTS (Y-AXIS) VS PRIMER5 RESULTS (X-AXIS)

V. RESULTS

Our results are shown in Figures 3, 4 and 5. Threshold values of $\alpha = 0.997$ and $\beta = 0.95$[5] were used for pyroprint similarity and a threshold value of $r = .9977$ was used for cluster integration. Our results show two large clusters, each spanning multiple days, and an additional small cluster. Additionally, a significant number of individual isolates on various days appeared to represent completely separate *E. coli* strains. The largest cluster, Cluster 1, pictured in Figure 3 exists on days 1 - 9. On day 10, another cluster, Cluster 2, pictured in Figure 4, became prevalent, lasting the remainder of the experiment. A possible third strain appeared on days 12, 13, and 14 in Cluster 3 (Figure 5).

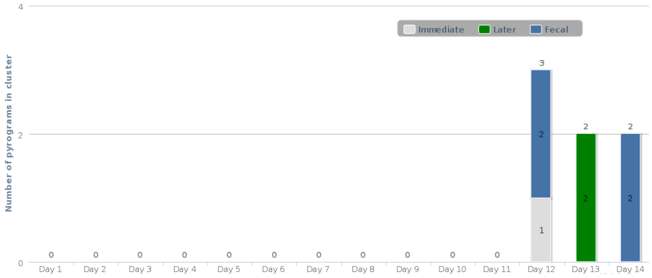

Figure 5. Cluster 3

A. *Primer5*

On our dataset, Primer5 produced the output shown in Figure 6. Hierarchical clustering discounts the chronology-related information from the dataset. Consequently, the output of Primer5 in such situations is hard to read and organize. Using $r \approx .997$ as a threshold in Figure 6 yields two clusters, A and B, each containing $\approx 40\%$ of the pyroprints. The remaining 20% of pyroprints contain a tiny cluster of four pyroprints (cluster C) and 18 unclustered pyroprints. The first cluster, A, appears to reside in the human host from day 1 until day 9, then dies out on day 10. At the same time when cluster A disappears the second cluster, B, starts to appear. Cluster B becomes established on day 10 and persists until the end of the study.

B. *Comparison of results*

Both methods capture two large clusters and one smaller cluster, presumably representing different *E. coli* strains residing in the human host. Our results also show a cluster of two pyroprints. Table I compares the results of the two methods. Each cell of the table shows how many pyroprints belong to it: e.g., there were 58 pyroprints which our method put in Cluster 1 and Primer5 put in Cluster A. Both methods almost agree on one cluster (Cluster 1/Cluster A), with only two mismatches; they also have a good agreement on the second large cluster, although our algorithm puts three more pyroprints from Cluster B into Cluster 3.

To better understand the cause for discrepancies, consider the pyroprint *F10-1*. According to Figure 6, *F10-1* is first clustered together with pyroprint *L9-3*. Our algorithm first clusters *F10-1* with *F10-2*, *F10-3*, and *F10-4*. *F10-1* is most similar to *L9-3*, however the similarity between *F10-1* and *F10-2*, *F10-3*, and *F10-4* is above our threshold α. Thus, the eventual position of *F10-1* in the cluster hierarchy will differ in the two algorithms.

During this pilot study the human subject was experiencing uneasiness in the stomach area, peaking around days 9 – 11. This event serves as some validation for initial analysis of the pyroprint clusters. The timing of the human subject's ailment coincides with the second dominant strain present in each clustering method's results.

[5] These threshold values were supplied by the Biology co-authors. The values were established empirically following an experiment in which biological material drawn from the same isolate was pyrosequenced separately multiple times.

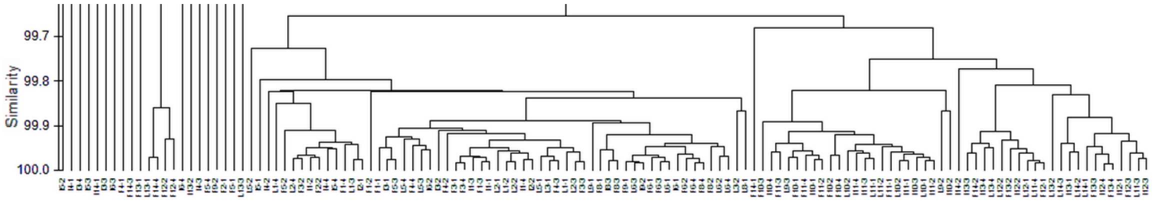

Figure 6. Primer5 Clustering Results

VI. Conclusion

This paper describes our pilot case study for a clustering algorithm applied to pyrosequenced ITS regions of *E. coli*. This algorithm produces a hierarchical cluster organization for biological data collections which possess a distinct internal structure. Using this algorithm we have successfully produced clusters that portray similar relationships in the data as Primer5's implementation of hierarchical clustering. However, without more data we cannot determine if the structure imposed by our algorithm yields better clusters.

VII. Future Work

Large quantities of real data is being made available for us to analyze which will obviate strengths and flaws in our algorithm not yet identified. Software for simulating our lab's pyroprinting method is being developed by the computer scientists and statisticians in our group. When this is completed, it will be possible to use real data to generate simulated data for further analysis.

Our algorithm does not directly address the problem of pyroprints that are *weakly connected*. PoClustering[12][11] will be important when addressing this problem. Cliquesets in the *E. coli* data may obviate obscure relationships between *weakly connected* pyroprints.

This clustering algorithm, as conceived, is applicable to any data collection with internal structure, but our implementation is specific to the described pilot study. Work is underway to formalize the algorithm to be applicable to any type of implicit data structure.

Acknowledgment

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

References

[1] P. E. Paramo and et al., "Large-scale population structure of human commensal escherichia coli isolates," *Applied and Environmental Microbiology*, vol. 70, pp. 5698 – 5700, sep 2004.

[2] D. L. Hartl and D. E. Dykhuizen, "The population genetics of escherichia coli," *Annual Review of Genetics*, vol. 18, pp. 31 – 68, 1984.

[3] D. M. Gordon and A. Cowling, "The distribution and genetic structure of escherichia coli in australian vertebrates: host and geographic effects," *Microbiology*, vol. 149, pp. 35 – 75, dec 2003.

[4] D. M. Gordon, "Strain typing and the ecological structure of escherichia coli," *Journal of AOAC International*, vol. 93, pp. 974 – 984, may 2010.

[5] 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.*, vol. 68, pp. 5796 – 5803, dec 2002.

[6] J. M. Simpson, J. W. S. Domingo, and D. J. Reasoner, "Microbial source tracking: State of the science," *Environmental Science and Technology*, vol. 36, pp. 5279 – 5288, dec 2002.

[7] 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*, vol. 68, pp. 1165 – 1172, mar 2002.

[8] K. Clarke, "Non-parametric multivariate analyses of changes in community structure," *Australian Journal of Ecology*, vol. 18, pp. 117 – 143, 1993.

[9] M. Ronaghi. (2001, january) Pyrosequencing sheds light on dna sequencing. Genome Technology Center, Stanford University. Palo Alto, California. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11156611

[10] K. Y. Kamath and J. Caverlee, "Transient crowd discovery on the real-time social web," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 585–594. [Online]. Available: http://doi.acm.org/10.1145/1935826.1935909

[11] J. Liu, Q. Zhang, W. Wang, L. McMillan, and J. Prins, "Clustering pair-wise dissimilarity data into partially ordered sets," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 637–642. [Online]. Available: http://doi.acm.org/10.1145/1150402.1150480

[12] J. Liu, Q. Zhang, W. Wang, L. Mcmillan, and J. Prins, "Poclustering: Lossless clustering of dissimilarity data," in *SIAM International Conference on Data Mining*, 2007.

[13] B. Liu, *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.