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

Computer Science ~ Software Engineering

Cal Poly, San Luis Obispo

Students: Aldrin Montana (CSC), Emily Neal (Bio)

Advisors: Alex Dekhtyar (CSC), Chris Kitts (Bio), Michael Black (Bio)

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

Background:

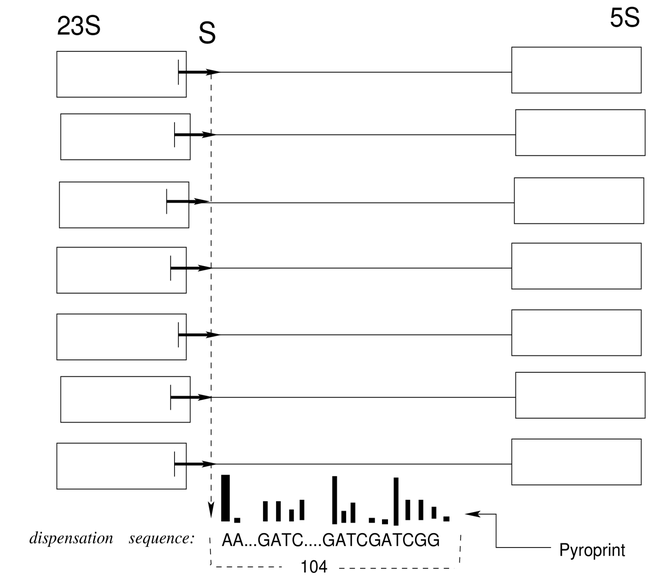
Our research group is developing an efficient, cost-effective, library-dependent Microbial Source Tracking method to create DNA fingerprints for different strains of ***E. coli*** using pyrosequencing; which we refer to as pyroprinting. In a pilot study, pyroprinting was used to investigate the variation 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.

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 collected via three particular methods:

- Direct sample (Fecal)
- Indirect sample immediately after defecation (Immediate)
- Indirect sample a few hours after defecation (Later)

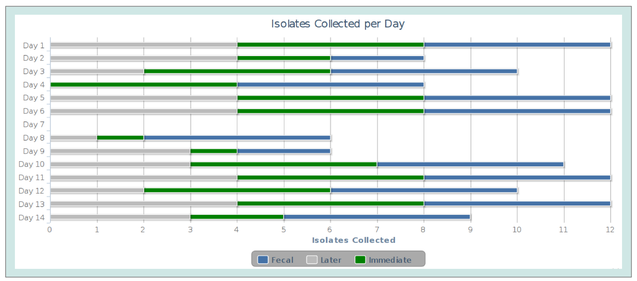
All samples were streaked onto MacConkey agar. Colony PCR was performed on each ***E. coli*** isolate. Primers designed to amplify the 23S rRNA – 5S rRNA Intergenic Transcribed Spacer region were placed in consensus regions of both rRNA genes. PCR products were used for pyrosequencing analysis.



Pyrosequencing 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**.

Results:

Threshold values of α = 0.997 and β = 0.955 were used for pyroprint similarity and a threshold value of **r** = .9977 was used for cluster integration. Our results show two large clusters and an additional small cluster. Primer5, a hierarchical clustering tool used by biologists, in comparison, portrays similar relationships in the data. We have reproduced results at least as sensitive as Primer5 and will test our method on more data to further investigate its value.



Primer 5

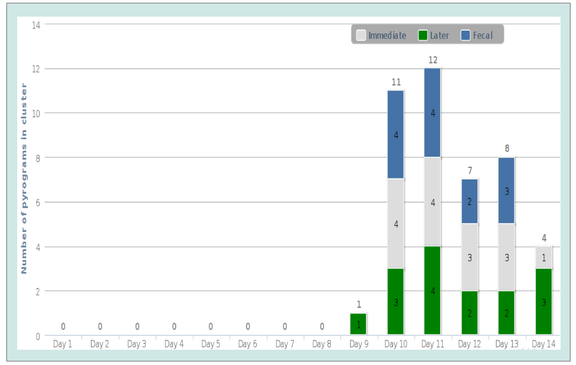
Chronology-Sensitive

	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

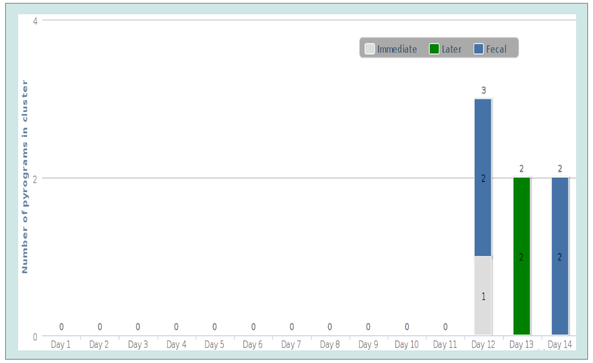
Table I
PYROPRINT CLUSTER CONFUSION MATRIX: CHRONOLOGY-SENSITIVE RESULTS (Y-AXIS) VS PRIMER5 RESULTS (X-AXIS)

Chronology-Sensitive Cluster 1

Chronology-Sensitive Cluster 2



Chronology-Sensitive Cluster 3



Algorithm:

Input: **Matrix M** of pairwise Pearson correlations between isolate pyroprints;
thresholds α ∈ [0, 1], β ∈ [0, 1];
distance relationship between collection groups.
Output: A dendrogram of isolates.

Step 1. Matrix transformation on similarity score M [i, j]:
Step 1.1. M [i, j] > α is replaced with **1**
Step 1.2. M [i, j] < β is replaced with **0**
Step 2. Clustering within days. For each day:
Step 2.1. Cluster within collection group.
Step 2.2. Cluster across collection group.
Note: 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 day 1, and chronologically adding one day at a time, combine clusters from different days. Continue until no two clusters have similarity of 1.
Step 4. Hierarchical clustering. Perform hierarchical clustering on isolate pyroprints with similarity greater than 0 and include the clusters constructed on Steps 2-3.