Investigating Temporal Strain Diversity in Human *E. coli* Populations Using Pyroprinting: A Novel Strain Identification Method

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*

I. Introduction

Escherichia coli (*E. coli*) is a well-studied commensal bacterium which inhabits the intestinal tract of some birds and most mammals[1][2], including humans[1][3]. *E. coli* is also an indicator of fecal contamination when found in environments such as watersheds, lakes, beaches, and recreational water[4][5][6][7]. 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] because dangerous interspecific pathogens can be transferred through contaminated water. Thus, *E. coli* is often used for Microbial Source Tracking (MST), a field in which methods of strain differentiation are used to identify and associate strains with a host species.

Current strain differentiation methods include phenotypic and genotypic microbial analysis[8][4][9][5][6]. Genotypic methods are advantageous compared to phenotypic methods since they use differences in nucleic acids to distinguish between strains of *E. coli* with greater sensitivity[4][5][6]. However, current methods can be labor intensive or financially expensive, and many have issues in reproducibility when identifying known *E. coli* strains[4][5][6].

In order to address many of the common complaints regarding strain differentiation methods, our lab has developed a novel pyrosequencing based MST method. This new approach, *pyroprinting*, uses pyrosequencing to generate pyroprints (DNA fingerprints) based on the sequences from two highly variable regions called Intergenic Transcribed Spacers (ITS) located in each of seven copies of the ribosomal RNA operons in the *E. coli* genome[10][11][12]. These regions of DNA are non-coding, and thus are able to accumulate more nucleotide changes due to lack of selective pressures. Our pyroprinting method simultaneously sequences all seven copies of each ITS: the 16-23S rRNA region (ITS 1) and 23-5S rRNA region (ITS 2). This creates patterns of data which capitalize on the differences between the replicates of the ITS regions. The generated pyroprints serve as the basis of comparison between *E. coli* isolates. A “match” between *E. coli* isolates is determined when the pyroprints for both ITS 1 and 2 are identical between two isolates.

Longitudinal studies of microbial strains present in individuals are a key problem in the area of microbial source tracking[6]. These studies are concerned with the total number of strains of a single bacterium (e.g. *E. coli*) present in a host at a given moment in time, strain turnover over time, the presence of dominant strains, change of strain dominance over time, as well as corresponding cause-and-effect questions[8][13][14][6]. In longitudinal single-host studies, bacterial cultures are grown from samples collected in roughly even intervals (daily, weekly, etc.) from a single individual. We have established and piloted a novel framework for genotypic microbial analysis of *E. coli* in longitudinal studies[15]. This framework incorporates *pyroprinting* and an *in-silico* source strain differentiation method based on a novel clustering algorithm.

In the study described here, *E. coli* strain identification and diversity was investigated in three humans over a six-month time period. It was hypothesized that each individual would host one or two dominant strains over the entire six months but that minor strains would vary. To assess this hypothesis, an on-line clustering algorithm[1] was developed based on two main modifications to traditional hierarchical clustering: (a) pairwise similarity scores between collected isolates are updated using a set of threshold value pairs, and (b) the order in which the hierarchical clusters are built is controlled by the sampling scheme. Trends in the resulting clusters were observed for size and persistence over time. Each cluster was constructed using similar isolates which presumably were from the same strain. The size of a cluster is representative of the dominance of the corresponding strain.

II. Methods

Sampling and Confirmation. *E. coli* were sampled from three individuals (A, B, and C)[2] once a month for six months. Feces were collected, homogenized, and sampled with a sterile swab. Inoculated swabs were streaked onto MacConkey agar plates for single isolated colonies and were grown at 37° C overnight. From these initial plates, 15 single colonies from each sample were selected for a series of standard metabolic tests to confirm *E. coli*.

Molecular Analysis. Confirmed *E. coli* isolates were used in colony PCR to amplify both ITS 1 and ITS 2 in the ribosomal RNA operon of the *E. coli* genome[3]. Primers (Appendix A) were placed in consensus regions of both rRNA genes. PCR products were used for pyrosequencing analysis. Both ITS 1 and ITS 2 were pyroprinted using Qiagen Pyromark Q24 pyrosequencers. Pyrosequencing is a DNA sequencing process where a new strand of DNA is built incrementally by introducing nucleotides A, T, C, or G based on a dispensation sequence. For our method, a pyrogram acts as a DNA fingerprint (pyroprint) for each strain. The pyrosequencing process is described in more detail in [16].

III. Chronology-Sensitive Clustering

Data Description. Given an isolate X, a specific pyroprinting site r_i in it, and a dispensation sequence $D = (d_1, \ldots, d_N)$, where $d_i \in \{A, T, C, G\}$, a pyroprint $\bar{X} = (x_1, \ldots x_N)$, where x_i are real numbers representing the light emitted during the pyrosequencing process when the reagent d_i was introduced on step i. Modern pyrosequencing machines sequence on the order of 100-150 nucleotides before the quality of sequencing starts deteriorating. In our experiments, we used two ITS regions, ITS 1 (r_1) and ITS 2 (r_2) for pyroprinting and used specially designed dispensation sequences for each site. The pyrosequencing machines are susceptible to some fluctuation in amplitude of light observed when sequencing. If pyroprints

[1] An algorithm that can run on input as it is given to the algorithm, without needing the entire input available

[2] Two females and one male, ranging in age from 20-25

[3] 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

of two different isolates differ substantially, then the two isolates belong to different strains. To compare pyroprints of two different isolates, we use Pearson correlation coefficient (see Equation 1). Pearson correlation coefficient is ideal for accommodating these fluctuations because it is concerned with comparing the variance between a pair of pyroprints $\bar{X}$, $\bar{Y}$ instead of a direct comparison between the values x_i, y_i. In order to appropriately interpret the correlation coefficient we utilize a pair of threshold values $\alpha = .950$, $\beta = .985$ for each ITS region r_i. If $sim(\bar{X}, \bar{Y})$ is sufficiently close to 1, we assume X and Y come from the same strain. However, if $sim(\bar{X}, \bar{Y})$ is sufficiently far from 1, then X and Y *definitely come from different strains*.

Data Transformation. Hierarchical clustering algorithms, including the algorithm we developed here, work by comparing, on each step, groups of data points (pyroprints, in our applications) and combining similar data points to each other. Our algorithm uses a *thresholded version* of the Pearson correlation coefficient to compare individual pyroprints to each other (see Equation 3. Anytime inter-cluster similarity measures are calculated, our algorithm applies a threshold function (Equation 3) which returns 1 if the two pyroprints are the same, and 0 if the respective pyroprints are definitely dissimilar. These cases are represented by the relationships $sim(\bar{X}, \bar{Y}) \geq \alpha$ and $sim(\bar{X}, \bar{Y}) < \beta$, respectively. Similarities between α and β are left unmodified. This transformation is the mechanism for ensuring strongly connected isolates at the core of each cluster. We represent already constructed clusters of pyroprints using a single pyroprint vector representing average values of the cluster known as the *average-link* hierarchical clustering method. Average-link inter-cluster similarity measures are calculated using Equation 2. In this function, $\alpha_r \in A$, $\beta_r \in B$ $\forall r \in R$ represents the threshold pair for each ITS Region r. C is the set of clusters such that $C_m \in C$ and $\exists i \in I$ such that $i \in C_m$.

Clustering Algorithm. Previous data sets used for analysis contained a large number of isolate pairs whose pyroprints had similarity between 0.995 and 1. In such situations, knowing that two pyroprints collected from the same sample have a high similarity is sufficient to put them into a single cluster *immediately*, even though a pyroprint from another month may have a higher similarity with one of them. In consideration of this, the algorithm we have developed constructs clusters in a two step process: (1) cluster all isolates that are *strongly connected*[4] in an order determined by the user; (2) cluster the remaining isolates with the clusters formed in step 1 using traditional hierarchical clustering. By integrating the remaining isolates into clusters in the last step, we minimize the negative effect that *weakly connected* isolates have on cluster construction.

The structure depicted in Appendix B dictates clustering order. Conceptually, this process occurs in the following manner: isolates located in the leaves of the structure are clustered using traditional hierarchical clustering. The resulting clusters are then propagated up the tree to the parent nodes. Optionally, the user may specify whether to incrementally grow clusters or to integrate all clusters together at once. This specific implementation of the approach incrementally grows clusters. This incremental growth is representative of the approximate chronology in which isolates were collected. This process repeatsfor each month in the study until clustering of all isolates in the structure is complete.

[4] A set of pyroprints P is considered *strongly connected* if each pyroprint $p \in P$ is sufficiently similar (i.e. *connected*) to every other pyroprint in P

IV. Results

Variation in *E. coli* populations was observed within the individual hosts. The total number of strains detected over six months varied for each of the three individuals. Pyroprinting detected a total of seven strains from individual A (Figure 2a), 15 from individual B (Figure 2b), and 14 from individual C (Figure 2c)[5]. Number of strains detected per month was different between individuals (Table I). However, most strains were detected only once during the experiment. For example, Strain 1 from individual A was detected only from the Month 1 sampling event. In the following sampling event, an entirely different strain (Strain 2) was detected with the pyroprinting method. Similar results were observed for the other individuals during the six month study. *E. coli* strains from different individuals were detected in different amounts, or number of occurrences, over the sixth month period. For example, individual A hosted one strain (Strain 5) which appeared in two consecutive months. Individual B hosted four strains detected in two separate months. Individual C hosted a strain detected in three separate non-consecutive months and two strains detected in two separate months. In all individuals, all other detected strains appeared only once.

The number of isolates in each strain in a given month can be used as a measure of abundance for each strain detected. Therefore, the strain comprised of the most isolates in a sampling period may represent the dominant strain. The data indicate that dominant strains can and do vary over time. Person A hosted a new dominant strain every sampling event except the last. Person B and C also hosted new dominant strains frequently, although some dominant strains (e.g. Strain 1 from C) re-appeared as dominant strains in subsequent sampling events. Other dominant strains (e.g. Strain 9 from B) in persons B and C were detected as minor strains in either previous or later months.

V. Related Work

Clustering Algorithms. Primer5 [17] is a software package commonly used by biologists that performs hierarchical clustering. Hierarchical clustering works by iteratively combining clusters until there is one cluster remaining[18].

Some extensions to hierarchical clustering are described by Liu et al.[19] and Kamath and Caverlee[20]. Kamath and Caverlee apply a variation of clustering to a communication network where edge weights in the network are adjusted based on when the messages are exchanged. Whereas Liu et al. use *PoClustering* (partially ordered clustering), which clusters data into *partially ordered sets* by finding all clique clusters for all possible diameters $W(D)$ where D is the maximal dissimilarity in a dissimilarity matrix[21]. PoClustering is a generalization of both hierarchical and pyramidal clustering that allows overlaps between clusters such that a PoCluster P is defined as $P = \{\text{cliqueset}_{\delta}\ (\text{d}) \mid \forall\ \text{d} \in \text{W(D)}\}$[19].

In contrast to Kavath and Caverlee, we 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.

[5] For the results reported in this paper, strain identification is only within a single host, e.g. Strain 1 for person A is not necessarily the same as Strain 1 for person B

VI. Conclusion

The results partially support our hypothesis. Although minor strains varied as predicted, individuals did not host dominant strains for the entire sampling period. According to the new pyroprinting method, the *E. coli* populations in all individuals were neither stable nor static. Dominant strains can change on a monthly basis such that the *E. coli* populations present in detectable amounts appear entirely different six months later. Additionally, *E. coli* populations fluctuate differently in different individuals. These findings compliment studies like Anderson et al.[8] that also showed *E. coli* populations are diverse over time and depending on the individual host. However, the data in that study were collected using a genotypic method that is considered labor intensive[4]. Furthermore, the results presented here show a greater amount of diversity over time than some studies using phenotypic differentiation methods[13][14], suggesting pyroprints are more sensitive than those particular methods.

VII. Future Work

Biological. Further analysis to identify and compare shared strains across all three individuals is currently in progress. Additonally, data have been collected from sample sets with greater numbers of isolates to estimate the sampling effort, or number of isolates, required to accurately represent *E. coli* populations. Lastly, a pyroprint database is currently under construction to house *E. coli* pyroprints from human hosts in addition to a multitude of other host species. In all, the findings presented here suggest that since human *E. coli* populations can change within a month, the database may need continual updates with *E. coli* pyroprints from human populations of interest. Conclusions regarding the source of fecal contamination may need to carefully consider temporal variation for the host species to interest.

Computational. The presented algorithm does not yet directly address the problem of weakly connected pyroprints[6]. PoClustering[19][21] will be important when addressing this problem as cliquesets[7] in the *E. coli* data may obviate relationships between weakly connected pyroprints. Further analysis of the accuracy and performance of the clustering method will be conducted using a large quantity of real data that is being made available by the biologists, and software for simulating our lab's pyroprinting method (developed in a class by computer science students). Lastly, although the approach used by our method is applicable to any data collection with internal structure, our implementation is specific to the described study. Work is underway to formalize the algorithm to be applicable to any type of implicit data structure.

VIII. Acknowledgements

This work was supported in part by an undergraduate education grant from the W.M. Keck foundation, CSUPERB, and College Based Fees from the College of Science and Mathematics and Cal Poly, San Luis Obispo. Thank you to the individuals from our laboratories in the EBI and under Dr. Michael Black in the Department of Biological Sciences, especially C. Sabatini, W. Tang, and J. VanderKelen.

[6] A weakly connected pyroprint is a pyroprint p_i that is not connected to every other pyroprint $p_j \in P$

[7] For example, a set of pyroprints P such that each pyroprint $p_i \in P$ is similar

Appendix A.

Primers

Primers designed to amplify the 23S rRNA - 5S rRNA Intergenic Transcribed Spacer (ITS) region:

Forward: $5'$- ATG AAC CGT GAG GCT TAA CCT T -$3'$

Reverse: $5'$-biotin- CTA CGG CGT TTC ACT TCT GAG T -$3'$

Sequencing: $5'$- CGT GAG GCT TAA CCT T -$3'$

ITS 1 Dispensation Order: CCTCTACTAGAGCG 20(TCGA) TT

ITS 2 Dispensation Order: AACACGCGA 23(GATC) GAA

Appendix B.

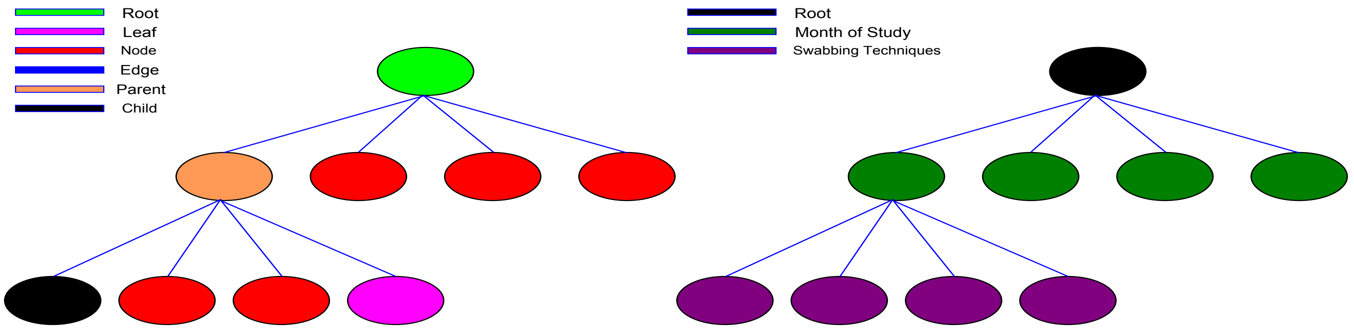
Structural Organization

In general, the hierarchical tree is represented as a general n-ary tree as seen in Figure 1a. The parts of the tree are described as follows: (node) Any item in the tree, depicted as circles in Figure 1a, (root) First node in the tree, located at the top, (leaf) A bottom-most node in the tree, (edge) A connection between any two nodes, (parent) In a pair of nodes with a direct connection, the parent node is the top node, or is closer to the root, (child) In a pair of nodes with a direct connection, the child node is the bottom node, or is further from the root, (depth) The number of edges to get from the root to a particular node, (level) Set of all nodes with the same depth.

Appendix C.

Definitions

Let I be the set of all isolates. Let R be the set of ITS regions associated with i $\forall i \in I$ R. Let A be the set of all upper thresholds, and B be the set of all lower thresholds. $\forall r \in R$, $\forall \alpha \in A$, $\forall \beta \in B$ let α_r, β_r be the upper and lower threshold pair corresponding to the region r such that $\alpha_r > \beta_r$. Let P be the set of all pyroprints. $\forall i \in I$, $\forall r \in R$ let $P_{ir} \in P$ be the pyroprint corresponding to isolate i at ITS region r. Let M be the set of all similaritiy matrices. $\forall r \in R$ let $M_r \in M$ be the matrix containing all pairwise Pearson correlations such that $\forall i, j \in I \times I$ $M_r[i, j] = sim(P_{ir}, P_{jr})$



(a) General

(b) Experimental

Figure 1: Isolate Tree - Structural Organization: The general structure illustrates the parts of the organizational structure, whereas the experimental structure represents the month that the isolates were collected in the first level and the swabbing technique used to collect the isolate in the second level.

Appendix D.

Equations

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

$$sim(C_m, C_n) = i, j \in C_m \times C_n, \frac{\sum_{\forall r \in R} M_r[i,j]}{|C_m \times C_n|} \tag{2}$$

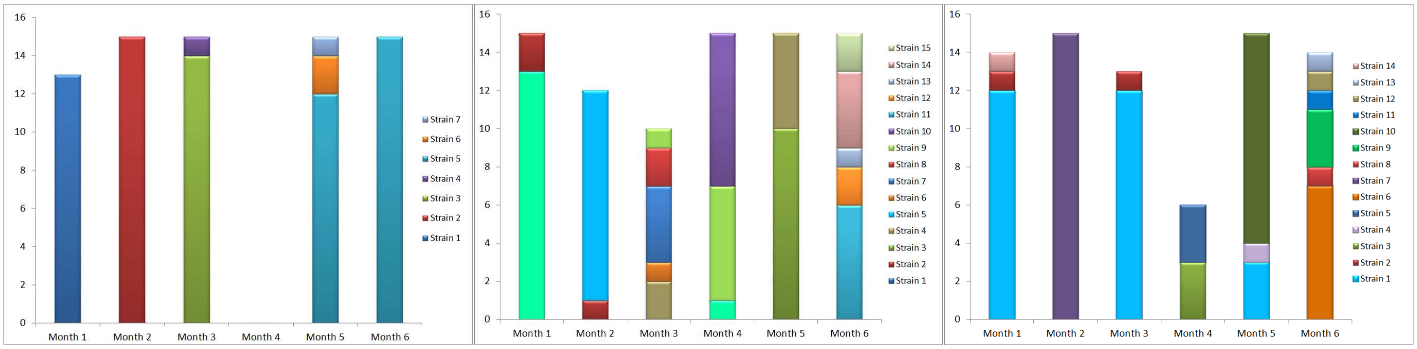
$$thr(sim(C_m, C_n)) = \begin{cases} 0 & if sim(C_m, C_n) < \beta_r \\ 1 & if sim(C_m, C_n) > \alpha_r \\ sim(C_m, C_n) & otherwise \end{cases} \tag{3}$$

Appendix E.

Results

Table I: Number of Strains Detected Per Month for Each Individual

Individual	Average Number of Strains Detected per Month	Minimum Number of Strains Detected per Month	Maximum Number of Strains Detected per Month
A	1.33	1	3
B	3.17	2	5
C	2.83	1	6



(a) person A

(b) person B

(c) person C

Figure 2: Strains over time from all three individuals. Only fecal isolates were used for this temporal analysis. Strain values are not evaluated between individuals, e.g. Strain 1 in Person A is not necessarily the same as Strain 1 in Person B.

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. 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.

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

[4] D. M. Gordon, “Strain typing and the ecological structure of escherichia coli,” *Journal of AOAC International*, vol. 93, no. 3, 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, december 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, march 2002.

[8] M. Anderson, J. Whitlock, and V. Harwood, “Diversity and distribution of escherichia coli genotypes and antibiotic resistance phenotypes in feces of humans, cattle, and horses,” *Applied and Environmental Microbiology*, vol. 72, pp. 6914 – 6922, november 2006.

[9] H. Ochman, T. Whittam, D. Caugant, and R. Selander, “Enzyme polymorphism and genetic population structure in escherichia coli and shigella,” *Microbiology*, vol. 129, no. 9, p. 2715, 1983.

[10] S. Boyer, V. Flechtner, and J. Johansen, “Is the 16s-23s rrna internal transcribed spacer region a good tool for use in molecular systematics and population genetics? a case study in cyanobacteria,” *Molecular Biology and Evolution*, vol. 18, no. 6, pp. 1057 – 1069, june 2001.

[11] A. Roth, M. Fischer, M. Hamid, S. Michalke, W. Ludwig, and H. Mauch, “Differentiation of phylogenetically related slowly growing mycobacteria based on 16s-23s rrna gene internal transcribed spacer sequences,” *Journal of Clinical Microbiology*, vol. 36, no. 1, pp. 139 – 147, january 1998.

[12] S. Tyler, C. Strathee, K. Rozee, and W. Johnson, “Oligonucleotide primers designed to differentiate pathogenic pseudomonads on the basis of sequencing of genes coding for 16s-23s rrna internal transcribed spacers,” *Clinical and Diagnostic Laboratory Immunology*, vol. 2, no. 4, pp. 448–453, july 1995.

[13] D. Caugant, B. Levin, and R. Selander, “Genetic diversity and temporal variation in the e. coli population of a human host,” *Genetics*, vol. 98, no. 3, p. 467, 1981.

[14] H. Sears and I. Brownlee, “Further observations on the persistence of individual strains of escherichia coli in the intestinal tract of man,” *Journal of Bacteriology*, vol. 63, no. 1, pp. 47–57, january 1952.

[15] A. Montana, E. Neal, A. Dekhtyar, M. Black, and C. Kitts, “Chronology-sensitive hierarchical clustering of pyrosquenced dna samples of e. coli: A case study,” in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, ser. BIBM ’11. IEEE, 2011, pp. 155 – 159.

[16] 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

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

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

[19] 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.

[20] 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

[21] 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