Author: Sulochana Bramhacharya

Affiliation: Hiram College, Hiram OH.

Address: P.O.B 1257
Hiram, OH 44234

Email: bramhacharyas1@my.hiram.edu

ACM number: 8983027

Category: Undergraduate research

Advisors: Prof. Louis T. Oliphant (OliphantLT@hiram.edu), Computer Science Department, Hiram College.

Prof. Brad W. Goodner (GoodnerBW@hiram.edu), Biology Department, Hiram College.

Metagenomic Data Analysis using Clustering

Problem and Motivation:

Recombinant DNA technology [2] and decrease in sequencing costs have revolutionized the area of genetics. Today, biologists can easily extract gene sequences of all organisms present in an environmental sample and create a fasta file [9]. One of the biggest challenges right now is to analyze these gene sequences, gene clusters and identify the differences in their cluster composition. There are two different approaches to measure diversity in gene clusters: a quantitative approach and a qualitative approach [8]. The quantitative approach measures the abundance of each taxon in a community and the qualitative approach measures only the presence or absence of the data in a community. The important parameters of the community are α-diversity that deals with the number of species found in an environment and ß-diversity which compares the structure of the community between two or more environmental samples, for instance healthy and diseased states [4].

Although numerous algorithms and approaches have been designed to cluster gene sequences, very less focus has been given to analysis of the composition of these clusters formed. The outliers in gene clusters are considered as noise and often ignored. Therefore, we tried to look into the differences in composition of two similar clusters from different samples i.e. find the outliers in these two similar clusters taken from two different samples. If any differences in the cluster composition are found from this comparative study, biologist can further study the samples to determine the factors causing the change in cluster composition. Comparing the micro bacterial samples from healthy and diseased people to understand the potential impact of a particular microbial community is an example of this approach. The biologists can also take the samples from different environmental conditions and compare their gene clusters to study how the organisms reacted to this change. Comparing the microbial community in sea water sample with and without excess nutrient to study how their community responds to perturbation is another example of this approach.

Background and Related Work:

Many algorithms have been proposed to study gene sequences using clustering method. K-means, hierarchical clustering, graph clustering theory are some popular clustering algorithms [6]. Best match algorithm, K-center method and interaction probability [5] are some cluster comparing algorithms. Jiang and Su proposed two-phased clustering process; modified k-means process (MKP) and Outlier-finding process (OFP) for outlier detection [6] in a cluster. However, for our metagenomic study, we used two different algorithms: Pairwise global sequence alignment algorithm [10] and Greedy incremental approach [7]. We used pairwise global sequence alignment to align the gene sequences and determined their degree of similarity and greedy incremental approach to cluster these sequences based on their similarity and the threshold value given by the user as well as to compare clusters to determine outliers. Cd-hit [7] is another gene clustering program that uses greedy incremental approach. Pairwise sequence alignment reveals the relationship between sequences and determines the correspondence between substrings in the sequences such that their similarity score is maximized. In greedy incremental approach, score of the sequence alignment is compared with the threshold value provided by the user to determine the best matching clusters.

Approach and Uniqueness:

Metagenomics is study of genetic material recovered directly from the environment. Biologists extract gene sequences from the given sample and create a 'fasta file' containing these sequences. Fasta file is the standard file format for gene sequences in genetics. It starts with '>' sign which is followed by a unique description for the gene sequence. This description usually contains the name of the sequence, length and other additional information if it is available. The genetic information follows after this unique description. An example of a bacterial gene sequence is given below.

```
>Axilla L_GPHQZAN04ILEVC rank=0031796 x=3407.0 y=1014.0 length=278

ACTCAAATGAATTGACGGGGACCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAA
CCTTACCAAATCTTGACATCCTCTGACCCCCTCTAGAGATAGTAGTTTTCCCCCGTTTCCGGGGGACGAG
AGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGC
CAACCCTTAAGCTTAGTTGCCACTCATTAAGTTGGGCACTCTAAGTTGACTGCCGGTGAC
```

We used 16S gene sequences for both clustering and cluster comparison purposes. 16S gene sequences are highly conserved gene sequences. These gene sequences are also involved in production of protein. As the protein synthesis mechanism is similar in all the organisms, we can say that the RNA sequences that assist in protein synthesis do not vary either. These sequences also mark the evolutionary distance and relatedness of organisms. Hence these 16s gene sequences act as finger print of an organism. We can compare these 16S gene sequences with a database of known organisms to identify the organism and other closely related organisms. If these sequences do not match with any of the known sequences then we can conclude that something is new in this gene sequence and possibly an evolution of new organism [3].

We divided our program into two phases for this metagenomic study. First phase includes separate clustering of gene sequences from two different samples and second phase includes comparison of the clusters formed in each sample to find the best matching pair.

We used greedy incremental approach [7] to cluster the gene sequences. The sequences from a fasta file are first sorted in descending order of the sequence length. The longest sequence becomes the representative sequence of the first cluster. The remaining sequences in the file are compared against this representative sequence using pairwise global sequence alignment algorithm [10] to calculate their percent match. According to this algorithm, we set up the scoring scheme by providing match, mismatch value and gap penalty as user input. We also construct scoring matrix [10] corresponding to the length of the two sequences being compared. The first row and column of this matrix are initialized with the multiples of gap penalty as shown in the figure below. If the bases from two sequences match at a particular location, we add the match value to the score, mis-match value if they do not match as well as calculate the score when paired with the gap by adding the gap-penalty. We also keep track of the parent row and column of each cell that determined the score so that we can back track from the bottom right cell to the top left cell to find the sequence alignment. We obtain this sequence alignment from the scoring matrix using high road alignment. Thus, we calculate the score of the alignment and out of all the possible outcomes, we select the best alignment with highest score and least number of gaps.

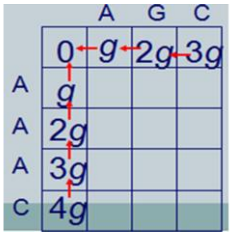

Fig: Matrix initialization

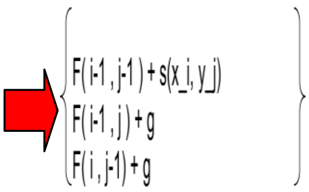

Fig: calculating possible scores

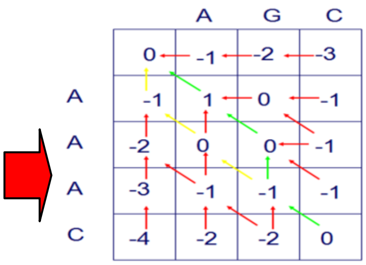

Fig: complete score matrix

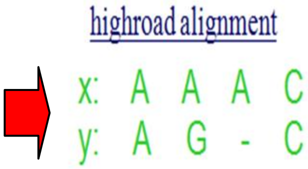

Fig: Best sequence Alignment

Next step is to calculate the percent match. We take this best sequence alignment to see if there are any bases aligned with gaps at both edges. If there are, then we cut off the edges until we find the first base-pair in the alignment and ignore this segment. As shown in the figure below once we chop off the ends, we count the total number of matched base pairs and divide it by the total base pairs in this sequence length that is being considered. This will be the percent match of the alignment and it is compared with the threshold value provided by the user to determine how closely related they are.

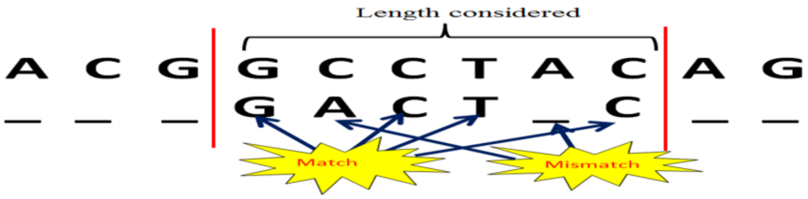

Fig: Actual alignment segment for calculating the percent match

The threshold value is always between 0 and 1 inclusive and can vary depending on how user wants the clusters to be formed. Threshold value of 0.9 means clustering the gene sequences at family level i.e. all the sequences belonging to a family in the taxonomical hierarchy are clustered together, 0.95 means clustering at genus level and 0.99 means clustering at species level. If the percent match is greater or equal to the threshold value provided by the user, we conclude that the sequence being compared is similar to the representative sequence of the cluster and thus becomes the part of the first cluster. Otherwise a new cluster is formed and this sequence becomes the representative sequence of this new cluster. We continue to compare all the remaining sequences from the file with the representative sequence of all the clusters formed to find the matching cluster or start a new cluster if no matching cluster can be found.

Once all the sequences are clustered, we enter the second phase of this comparative study. These clusters are compared with the similar clusters from another sample. To find the similar cluster pair from two samples, we take the representative sequence of each cluster from first sample and compare it with representative sequence of each cluster from second sample. Using pairwise global alignment, we calculate the percent match for each of this alignment and compare it with the threshold value. If the percent match is greater or equal to the threshold value, we consider the cluster pair to be similar. Once we find the similar cluster pair, we compare all the other remaining sequence in cluster from second sample with the representative sequence of cluster from first sample. Any sequence with the percent match less than the threshold value is considered as outlier of the cluster pair. Thus we obtain two different sets of data from each corresponding pair of clusters, i.e. a set of similar sequences in both clusters and a set of dissimilar sequences that are considered outliers. We repeat this process by comparing all the other remaining sequences in the cluster from first sample with the representative sequence of the cluster from second sample. Pictorial representation of this comparative study is given below.

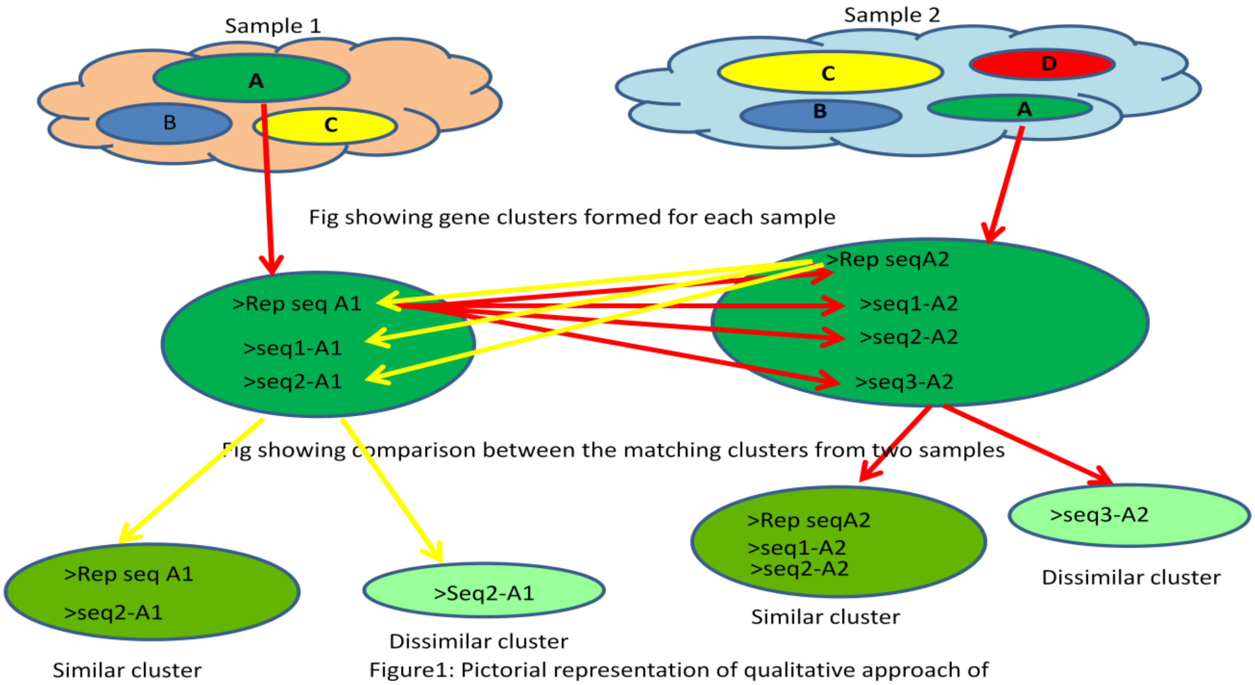

Fig showing gene clusters formed for each sample

Fig showing comparison between the matching clusters from two samples

Figure1: Pictorial representation of qualitative approach of cluster analysis

Results and Contributions:

Our project is an approach to design a simple gene clustering and comparing tool that can be used in classroom by the students and researchers. The users can cluster two fasta files from two samples at a time and view the results of cluster comparison. Our program first outputs all the clusters formed for each fasta file and then outputs sets of similar and dissimilar clusters for every pair of similar clusters from two samples. These sets of data are very useful to start exploring factors contributing to the presence and absence of the gene sequence in the particular sample. This cluster comparison tool is coded in Java.

We ran test run with two small sets of bacteria sequences to obtain the sets of similar and dissimilar sequences in each matching cluster pairs. When our results were compared with the BLAST [1] results, the resulting percent match of the sequence alignment from both programs were somewhat similar. However we did not expect our results, percent match, to be exactly similar to the BLAST results as the algorithm used in both tools were completely different.

In future, we plan to optimize our program so that it allows users to compare fasta files from multiple samples at a time. We also plan to work on optimizing our cluster comparison algorithm to improve its efficiency. We also plan to improve the graphical interface and format the results so that the output is easier to read. We plan to test our program with larger metagenomic data sets as well.

References:

1. Basic Local Alignment Search Tool." *BLAST:*. National Library of Medicine, n.d. Web. 14 Apr. 2013. http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web
2. Chedrese, Pedro J. "Recombinant DNA Technology." Reproductive Endocrinology. N.p.: Springer US, 2009. 75-82. Print.
3. Clarridge, Jill E. "Impact of 16S RRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases." *PubMed Central* 17.4 (2004): 840-62. Print.
4. Fabrice A, Dider R (2009) Exploring Microbial Diversity Using 16S rRNA High-Throughput Methods. J Compt Sci Syst Biol 2:074=092. doi10.4172/jcsb.1000019
5. Goldberg, Mark K., Mykola Hayvanovych, and Malik M. Ismail. Measuring Similarity between Sets of Overlapping Clusters. Rensselaer Polytechnic Institute, 2010. Web. 27 Sept. 2012.
6. Jiang, M. F., and C. M. Su. Two-phase Clustering Process for Outliers Detection. National Chiao Tung University, 25 Apr. 2000. Web. 27 Sept. 2012.
7. Li, Weizhong, and Adam Godzik. "Cd-hit: A Fast Program for Clustering and Comparing Larger Sets of Protein and Nucleotide Sequences." Oxford Journals 22.13 (2006): 1658-659. Print.
8. Lozupone, C., et al.(2007) Quantitative and Qualitative ß Diversity Measures "Lead to Different Insights into Factors that Structure Microbial Communities, 78, 1578- 1585.
9. *NCBI*. U.S. National Library of Medicine, n.d. Web. 12 Apr. 2013. <http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml>.
10. Pair Wise Global Alignment of Sequences. Web. 25 Sept. 2012. <http://media.wiley.com/product_data/excerpt/91/04708483/0470848391.pdf>.