P4: A New Approach to Cluster Pathogenic Bacteria Based on DNA Repeats and Segmental Genomic Duplications

Amjad khateeb[1], Khaldoun Halawnai, Hashem Tamimi, Yaqoub Ashhab

[1]Biotechnology Research Center, Palestine Polytechnic University, P.O.Box 198, Hebron, Palestine.

Introduction

Prokaryotic genomes are enormously diverse. The genome contents of prokaryotes are remarkably variable even among closely related species. Such genome plasticity is considered as a major player in the bacterial adaptation to various environmental and host constrains [1]. The driven forces of such variations are typically constituted of genomic repeats that engage in recombination events, either dependent on (homologous recombination) or independent of RecA (illegitimate recombination). Recombination among genomic repeats allows for antigenic and tissue tropism variation, but it may have negative consequences, such as replication pausing.

Recent reports have identified two major classes of genomic DNA repeats in bacteria; short low complexity repeats and long repeats. The first class constitute of different groups of relatively short-sequence repeats SSR such as microsatellite sequences. The second class includes transposable elements such as insertion sequences (ISs) [2]. Studies of DNA repeats in bacteria were done on small sets of samples that were studied separately, where the researchers were attempting to discover the relationship between these repeats and the biological function of a given bacterium.

This project aims to identify and analyze the DNA repeats and segmental genomic duplications in the genome of a wide range of pathogenic bacteria. The analysis will focus on discovering bacterial clusters based on DNA sequence repeats and explore its associations with biological features.

Material and Methods

The collected data in this project were two types. The first one was a set of whole genome sequence of 75 pathogenic and non-pathogenic bacteria. These DNA sequences were downloaded from NCBI (ncbi.nlm.nih.gov) genome database. The second type of data is a set of biological features of the selected bacteria such as gram stain, GC%, and cell division rate.

DNA repeats and segmental duplications from each genome sequence were extracted using a sequence alignment tool Mega-Blast by aligning each genomic sequence with itself. The resulting repeats and duplications were filtered according to their lengths and identities. The biological features of bacteria were represented into numerical values in order to be used in the processing stage.

The obtained sequence data were analyzed to investigate the relationship between DNA repeats and duplication features with the biological features of the relevant bacterium. We used the unsupervised clustering algorithms Fuzzy C-Mean (FCM), which used to divide the data into a set of groups based on the duplication features. The Subtractive Clustering was used to provide the FCM algorithm with the best number of clusters for the duplications features vector. The clustering process results were evaluated using a fitness function. This function correlates the biological features with the results of the clustering process, and then measure the correlation between these values and clusters.

In order to get the best result of clustering, the genetic algorithm was used to select a set of feature from the whole features vector to be used in the clustering process. Using the genetic algorithm we were able to obtain the best combination of features that significantly affect the clustering process. The previous clustering evaluation function was used as a fitness function in the genetic algorithm.

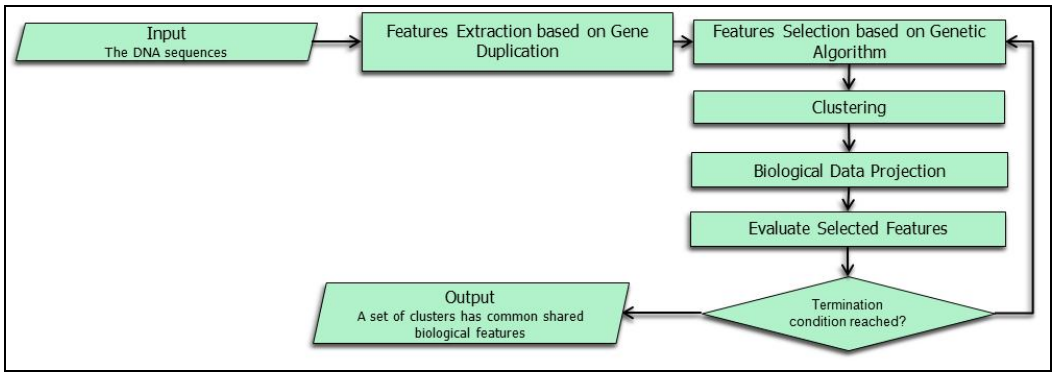

Figure 1: The Project Framework

Results

As results of this project, we introduced a new framework in clustering and studying the effects of DNA repeats on the bacterial adaptation to the environmental and host constrains. The present clustering approach was able to split the tested bacteria into different subgroups according to the correlation between the DNA repasts content and frequency and their biological features values. The following graphs shows the results of two experiments to cluster bacteria based on subtractive clustering radius 0.75 and number of clusters 5 and the correlation is measured with the Topology feature for the first figure. The second experiment was based on subtractive clustering radius 0.95 and number of clusters 2 and the correlation is measured with the Generation time feature

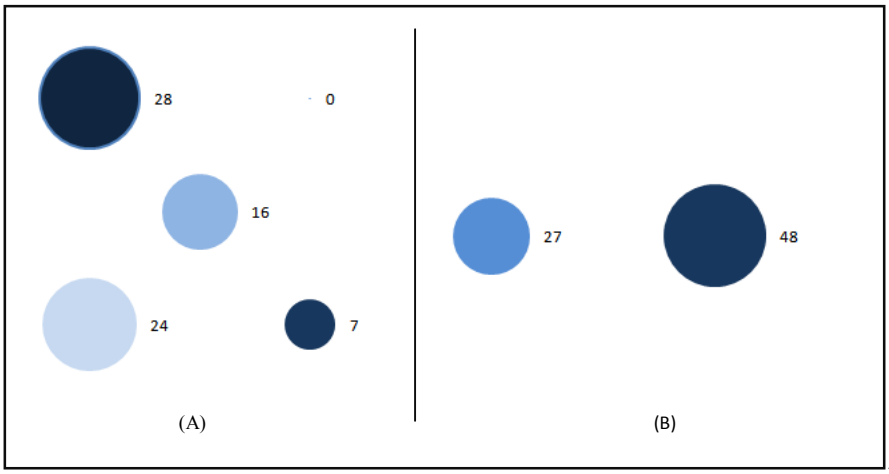

Figure 2: Clustering results of the 75 bacterial genomes. A. Clustering based on radius 0.75, number of clusters 5. B. Clustering based on 0.95, number of clusters 2.

Our proposed clustering methodology can help to understand biological relationship among bacteria based on their content of DNA repeats. In addition, this method will provide more biological insights to reveal the role of genomic DNA repeats in genome plasticity and bacterial evolution.

References

Achaz G, Rocha EP, Netter P, Coissac E. Origin and fate of repeats in bacteria. Nucleic Acids Res. 2002 Jul 1;30(13):2987-94. PubMed PMID: 12087185;

PubMed Central PMCID: PMC117046Mizuta, Satoshi, Koshino, Michimasa and Shimizu, Toshio. Seeking Genomic Duplication in Prokaryotic Genomes.

Vergara, Ismael , Allan K Mah, Jim C Huang, Maja Tarailo-Graovac, Robert C Johnsen and David L Baillie. "Polymorphic segmental duplication in the nematode Caenorhabditis elegans", BMC Genomics (2009) 10: 329.