BACTOSOM-Viewer: a High-Throughput Bacterial Clustering and Mapping Tool Based on Landmark Proteins

Bilal Tamimi[1], Sami Salamin, Hashem Tamimi, Yaqoub Ashhab

Biotechnology Research Center, Palestine Polytechnic University, P.O-Box 198, Hebron, Palestine

Introduction

The availability of enormous amount of genomic data generated from various sequencing projects provides a great opportunity for researchers to study and infer the desired functional information. Bioinformatics-based approaches have proven powerful and cost-effective means to study and represent huge genomic and proteomic data from bacterial genomic projects.

Most system biology-based computational tools that aim at exploring and comparing biological sequences of prokaryotes focus on genomic data. Such tools usually perform similarity comparison that doest not take into account the differences between genomes. There is an increasing need to develop genomic and proteomic tools that shed light on the similarities and differences alike. Such methods would allow more accurate classification of bacteria that are based on whole proteome or genome instead of a single gene such as 16S-rRNA. In addition, a comprehensive analysis and comparison of various bacterial proteomes can help in assigning functions for many hypothetical proteins that are still annotated as proteins with unknown functions in the biological databases.

This paper presents a novel approach for developing an intelligent tool that can perform a high-throughput comparison and mapping to reveal similarities and differences among bacterial proteomes. The developed software is capable to visualizing the relationship between different types of bacteria based on Self Organizing Map method. The idea is to compare any given bacterial proteome with a set of essential proteins known here as a core proteome. The core proteome can be defined as the minimum set of proteins that are hypothetically essential to form a fully functional bacterium [1].

Material and methods

Land Mark Genes: The Mycoplasma genitalium 474 gene coding proteins were considered as land mark genes. These genes represent the minimal essential genes to constitute a core bacterial genome since M. genitalium is the smallest known bacteria. The protein sequences were obtained in Fasta format via GenBank [2].

Testing data: The full protein sequences of 6 bacteria that represent three different environmental niche were used to test the performance of our system. The protein sequences for each bacterium were downloaded from the GenBank as FASTA format with full annotations as shown in the following table:

Bacteria Name	Protein sequences
Brucella melitensis	2059
Brucella abortus	2029
Mycobacterium bovis	3944
Mycoplasma gallisepticum	763
Streptococcus agalactiae	1996
Streptococcus thermophilus	1709

Feature extraction: The feature table was extracted based on pairwise sequence alignment score between each protein (from the tested bacteria) and the land mark proteins using Smith-Waterman algorithm.

Self Organizing Map: The features' files represent the input for the Self Organizing Map [3]. Our SOM system was trained on the land mark proteins that are considered as a set of reference points to reveal similarities and differences among bacterial proteomes. The major output format of our system is shown as a 2-D graphical representation that allocate a position for each tested protein according to its degree of similarity to the land mark proteins.

Advance system options: An interactive querying system was developed to translate the graphical representation into a more useful output formats so as to facilitate advance data acquisition by biologists.

Results

The present developed system is a desktop application that uses MATLAB built-in tools and can be easily used by biologists. The BACTOSOM-Viewer is a Graphical User Interface, where user can select two species from a long list of available bacteria so as to plot their comparative proteomic as SOM graph. The to-ol was verified on the six selected bacteria.

Figure 1 shows the graphical result for our system when comparing Brucella abortus vs. Brucella melitensis proteomes. Depending on similarity degree, BACTOSOM-Viewer clusters each group of similar proteins in the same nodes. The clustering was in light with the gene ontology functional categorizations that are available in GenBank. The tool has the capability to identify all species-specific proteins as isolated small dots and their distance from the known core proteins.

To our knowledge, this is the first work that provides an SOM-based system that can perform large scale identification of proteome similarities and differences. It would allow a new model of bacterial taxonomy and it will further our understanding of prokaryotic genome evolution. In addition, this method can be used for the identification of species specific genes that can help in designing more accurate molecular diagnostic protocols.

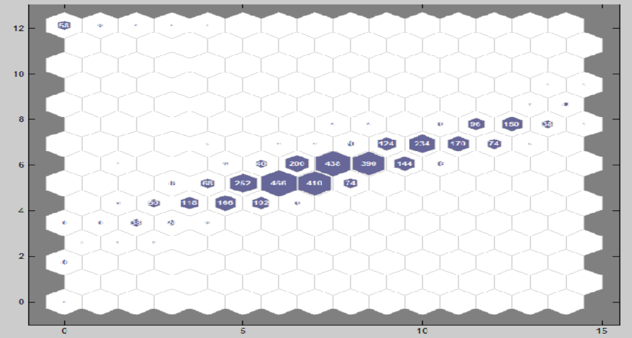

Fig.1: BACTOSM-Viewer results showing comparison between Brucella abortus vs. Brucella melitensis proteomes

References

John I. Glass, Nacyra Assad-Garcia, Nina Alperovich, Shibu Yooseph, Matthew R. Lewis, Mahir Maruf, Clyde A. Hutchison III, Hamilton O. Smith*, and J. Craig Venter. Essential genes of a minimal bacterium. Synthetic Biology Group, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850

The National Center for Biotechnology Information. Retrieved may 1,2010. http://www.ncbi.nlm.nih.gov

Wikipedia, the free encyclopedia. Retrieved may 31, 2010. http://en.wikipedia.org/wiki/Self-organizing_map