

Poster I-11

Retrieving Bacterial Functional Units by Bayesian Decomposition Analysis of Phylogenetic Profiles

Bidaut Ghislain^{*1,2}, Suhre, Karsten², Claverie, Jean-Michel², Ochs Michael¹

¹*Bioinformatics, Information Science and Technology, Fox Chase Cancer Center, Philadelphia, PA, USA;*

²*Structural and Genetic Information Laboratory, Marseille, France*

Antibiotic resistance together with the side effects of broad spectrum antibacterials make development of targeted antibiotics of great interest. The availability of the sequences from several bacterial genomes can potentially reveal genus-specific targets. We present a new method for finding potential targets from genomic data based on the creation of a similarity dataset and its analysis with Bayesian Decomposition. This scheme provides a classification of genes related to their presence or absence in functional units that exist in certain bacterial species. We constructed a dataset of phylogenetic profiles (vectors that encode the similarity, measured by BLAST scores, of a gene across many species) for a series of 31 pathogenic bacteria of interest with 1073 genes taken from the reference organisms *E.coli* and *M.tuberculosis*. Since genes that function together have most likely evolved together to maintain viability in the organism, they should be linked into what we define as a *functional unit*. However, a gene may have acquired a role in multiple functional units through the evolutionary process. These multiple functions remain difficult to identify with classic clustering methods that force a gene into a single group. We therefore used Bayesian Decomposition (BD) to analyze the data. BD acts as a matrix factorization algorithm, finding fundamental vectors that describe the data and leading to an unsupervised classification of genes into functional units and bacteria into related species, i.e. bacteria that share similar functional units. Since the estimation of the dimensionality of the data is a remaining problem in this case, we organized results from successive BD analyzes with increasing numbers of patterns into a tree to identify those fundamental vectors that are most robust across multiple proposed numbers of patterns. BD has grouped the bacteria phylogenetically consistently on the basis of the proteins necessary for their survival. Complex genomes that are closer to the reference *E.coli* have been described by many fundamental vectors, suggesting that groups of genes that form functional units have been separated. Strains that are more distant from the reference in terms of evolution, such as *M.leprae* or *S.typhi* (characterized by particular metabolisms) have been isolated in their own fundamental vector. Also, bacteria having cell wall functions have been isolated. The analysis has also revealed that genes involved in anaerobic respiration are related to the bacteria that use this particular metabolism, and a similar conclusion has also been drawn for aerobic respiration. A set of genes constituting the backbone of genes necessary to survival has also been isolated. We are presently extending this work with a larger dataset including more bacteria.

This work has been supported in part by grant CA06927 to R. Young.