

# Analysis of Phylogenetic Profiles Using Bayesian Decomposition

Ghislain Bidaut<sup>1,2</sup>, Karsten Suhre<sup>2</sup>, Jean-Michel Claverie<sup>2</sup>, and Michael F. Ochs<sup>1</sup>

*1:Fox Chase Cancer Center, Bioinformatics, Department of Information Science and Technology, Philadelphia, PA, 19111, USA*

*2:Structural and Genetic Information Laboratory, (UMR 1889-CNRS-AVENTIS), Marseille, France*

G.Bidaut@fccc.edu, Karsten.Suhre@igs.cnrs-mrs.fr, jmc@igs.cnrs-mrs.fr, M.Ochs@fccc.edu

## Abstract

*Antibiotic resistance together with the side effects of broad spectrum antibacterials make development of targeted antibiotics of great interest. To meet the problem of identifying potential targets specific to some genuses, a dataset comprising a series of phylogenetic profiles was built for a series of pathogenic bacteria of interest. The profiles are the highest BLAST scores for genes compared to selected genes of *E. coli* and *M. tuberculosis*. The dataset reflects the past evolution of those genes due to adaptation to specific niches, marked by lateral gene transfer, duplication and mutation of existing genes, or merging of existing genes. Genes that function together will be constrained to evolve together, to maintain viability in the organism. However, a given gene may have a role in multiple functional groups through the evolutionary process. Analysis using Bayesian Decomposition helps to retrieve those relationships by retrieving fundamental patterns related to the evolutionary retained functions.*

## 1. Introduction

The routine use of antimicrobial agents in hospitals and in communities has led to an increasingly rapid emergence of antibiotic-resistant bacterial populations. Both Gram-positive and Gram-negative resistant strains have been reported recently [3], indicating an urgent need for approaches to discover new targets for potential antibiotics. Also, side effects of commonly used broad spectrum antibiotics caused by destruction of the native bacterial flora encourages the development of drugs with a narrowed spectrum of action. Although natural screening is still the primary method used for drug discovery, recent progress in the publication of full bacterial genomes has the potential to provide us with new insight on the bacterial biology, to help us identify unknown genes and to reveal pathways that could be targeted in specific genuses. The 106 fully com-

pleted genomes publicly available at The Institute for Genomic Research (Rockville, MD, USA, [www.tigr.org](http://www.tigr.org)) constitutes an impressive amount of data for which appropriate analysis and visualization tools need to be designed. The primary method of analyzing such data is still by sequence comparison using BLAST [1]. Databases regrouping orthologs have been built such as the Clusters of Orthologous Group of proteins database [6]. Phylogenetic profiles along with the use of Bayesian Decomposition appears to be a promising approach for comparison of whole-length bacterial genomes.

## 2. Methods

To identify potential genes of interest, a list of reference homologs from the *E. coli* and *M. tuberculosis* genomes have been scored against a set of 31 pathogenic bacterial genomes using BLAST. The reference set comprises all genes of length greater than 50 amino-acids having mutual maximal BLAST scores from the *E. coli* and *M. tuberculosis* genome sequences. This yields a list of 1073 homologs, roughly 50% of them with a known function. Each homolog has then been scored against the genome sequence of 31 pathogenic bacteria and normalized by their length, yielding a similarity matrix of 1073 phylogenetic vector profiles of size 31.

We used Bayesian Decomposition (BD) ([2], [4]) to separate the mixed fundamental signals that underlie the phylogenetic profiles and to retrieve functional units. BD is a bilinear decomposition algorithm that identifies for a data matrix  $\mathbf{D}$  a pattern matrix  $\mathbf{P}$ , which is a set of  $K$  basis vectors, and a distribution matrix  $\mathbf{A}$  that combine to form a mock matrix  $\mathbf{M}$  that reproduces the original data within the noise, i.e.

$$\mathbf{D} \sim \mathbf{M} = \mathbf{A} \cdot \mathbf{P}. \quad (1)$$

Since neither  $\mathbf{P}$  nor  $\mathbf{A}$  are known, BD uses a Bayesian Model with a Markov Chain Monte Carlo (MCMC) procedure to sample and find the properties of the solution space.

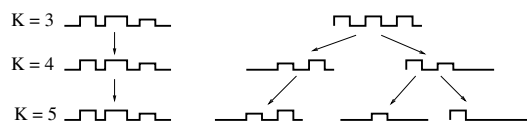
We decomposed the phylogenetic profiles using BD for a number of solutions  $K$  positing between 3 and 30 basis vectors. Each decomposition yielded a distribution matrix  $A$  and a pattern matrix  $P$  of basis vectors. We classified those into a hierarchical tree according to their *Pearson* correlation. For any  $K$ , basis vectors can be related to the corresponding bacteria using the distribution matrix  $A$ , to show which bacteria contain each functional unit. The tree was built iteratively by connecting the basis vectors for a decomposition in  $N$  basis vectors, to the one for a decomposition in  $N + 1$  basis vectors (see figure).

### 3. Results

The obtained tree has been created and explored using BDTree, a cluster visualization tool. The tree configuration that appears is extremely informative about the nature of the data. 16 branches of stable patterns have appeared. *Mycobacterium leprae* and *Salmonella typhi* have been isolated each one in their own branch. Strains from the *Chlamydiae* genus have been grouped in the same branch. Longer and more complex genomes such as *Agrobacterium tumefaciens*, *Brucella suis* and *Brucella melitensis* are present in many patterns suggesting that their genome is explained by many functional units shared with other species. For instance, the *Rickettsia* have been isolated but their functional units are partially shared with *Agrobacterium tumefaciens*, the *Brucella*, *Coxiella burnetii* and *Neisseria meningitidis*. The *Staphylococcus* and *Mycoplasma* are tight together in the same pattern, although the *Staphylococcus* share many other functional units with other strains with long genomes. One of the stable branches includes all the bacteria, and is comprised of genes that provides a backbone of functions necessary to the survival of all organisms. The tree can also help to infer how many basis vectors are needed to describe the data. Estimating the dimensionality of a given dataset is still an open problem in complex data mining problems such as in the Rosetta microarray data [2]. In our case, the consistency of the bacteria groups suggests that 24 solutions are appropriate for describing the data since at this level bacteria characterized by cell wall functions are isolated from *Mycobacteria*. BDTree permits the validation of known biological results such as the fact that *M. leprae* seems to have strongly reduced certain metabolic pathways such as oxidative and anaerobic respiration pathways. In the functional groups found by BD that are linked to *M. leprae*, the genes involving oxidative and anaerobic respiration are virtually absent.

### 4. Conclusion

In this work, phylogenetic profiles have been created from 1073 homologs across 31 bacterial species and an-



**Figure 1. Organization of the basis vectors into a tree to infer stable basis vectors and estimate the optimal number of solutions.**

alyzed with BD to identify potential antibacterial targets. Bayesian Decomposition grouped genes from functional units that have evolved together. This provides insight on how genes are related and how they have evolved among species. This has the potential to identify protein targets that are critical to functions specific to individual bacterial species. At this step, we are still comparing results with known biological knowledge and assessing the viability of the method. BDTree is a first step toward reducing the overwhelming amount of genomic data by classification within a hierarchical tree and integration of biological knowledge from various databases to aid interpretation. These early results are encouraging so far, since genes group together consistently, based uniquely on the separation of the functional units necessary for their survival. Proper filtering methods with finer annotations should improve the ability of this approach to identify unique targets. As an extension to this work, we intend to include additional genomes to further define and specify the functional groups.

### References

- [1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- [2] G. Bidaut, T. Moloshock, J. Grant, F. Manion, and M. F. Ochs. Application of bayesian decomposition to gene expression analysis of deletion mutant data. In *Methods of Microarray Data Analysis II*, Dec. 2001.
- [3] M. H. Kollef. Selective digestive decontamination should not be routinely employed. *Chest*, 123(5 suppl):464S–468S, May 2003.
- [4] T. Moloshok, R. Klevecz, J. D. Grant, F. Manion, W. F. Speier, and M. F. Ochs. Application of bayesian decomposition for analysing microarray data. *Bioinformatics*, 4(18):566–75, Apr. 2002.
- [5] S. Sisibi and J. Skilling. Prior distribution on measure space. *J. R. Statist. Society. B*, 59(1):217–235, 1997.
- [6] R. Tatusov, E. Koonon, and D. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct. 1997.