Comparison Between Complete Genomes of Vibrio Species

Patrícia P. Cógó¹, João Meidanis¹, Fabiano Thompson²
¹University of Campinas, Institute of Computing, Campinas, Brazil
²Federal University of Rio de Janeiro, Institute of Biology, Department of Genetics, Rio de Janeiro, Brazil

Introduction

Genome sequencing has made possible the development of new comparative genomic methods, and hence, a revolution in microbial taxonomy. A new, modern microbial taxonomy is being established based on multiple loci or entire genomes.

With this purpose, we describe in this work an approach based on genomic rearrangements [1] to compare complete genomes, and apply it on the completely sequenced Vibrio Species. In our approach, we model evolutionary events like gene losses, lateral gene transfers, and chromosomal rearrangements to estimate evolutionary distance, which we believe represents an advantage over analyses based on gene markers or phenotypic characteristics.

Our methodology is strongly based on the same steps traditionally followed when measuring evolution in Bioinformatics, namely:

- Define the model;
- Identify homolog structures;
- Employ the model and these structures to find the most parsimonious sequence of events that can explain the differences between the genomes.

This presentation will be divided in two parts. The first of them report our preliminary experiments. From the analysis of the results obtained, we define our methodology to compare complete genomes of vibrios, presented in part 2.

Part I

Preliminary Experiments

Six organisms, which have their genomes completely sequenced, were employed in our experiments, namely:

- Photobacterium profundum;
- Vibrio cholerae;
- Vibrio fischeri;
- Vibrio parahaemolyticus;
- Vibrio vulnificus CMCP6 and Y1016.

In these experiments, we have applied the method described by McLysaght et al. [2] to identify and group homology structures. Briefly, it consists in:

- Represent all proteins as graphs vertex;
- Compare the complete set of proteins using the program BLASTP, linking two proteins as homologs if their comparison has e-value less or equal 10⁻⁵ and maximal scoring pair alignment that cover, at least, 40% of the longest protein;
- Finally, employ a linkage method to group homolog genes: a group is maintained only if it has a member similar to all the other members; groups that do not satisfy this criterion are dismantled in to singletons.

Analysing the families determined through this method, we have identified two deficiencies:

- Dependence on the organisms being comparing: families can be merged or depareted when an organism add or excluded from the analysis;
- Highly connected gene sets (exemplified by {A, B, C, D} in the following picture) could be dismantled, if the entire group does not have a member similar to all the others.

We have considered two operations in these experiments: gene loss and block interchange. Because we had not defined, at that moment, a proper method to treat duplications, two alternatives were tested: exclude all duplicated genes or randomly keep one of them.

We have construct distances matrices of different scenarios, varying the weight applied for each kind of rearrangement operation and the method employed to treat gene duplications. When gene loss and block interchange have the same weight and duplicated genes are discarted, the phylogenetic tree obtained agrees with the ones built by traditional analysis of gene markers.

Our proposal is based on the occurrence of four operations:

- Fusions;
- Reversals;
- Gene losses (modeled as special cases of fission)

Translocations, block-interchanges, and transpositions (which can be seen as special cases of block-interchanges), are not fundamental events, but rather a result of two successive operations, namely, a fission followed by a fusion.

A compromise is adopted to treat duplications: when we have a gene family with more than one gene member in one or both genomes, a phylogenetic tree with all genes in all genomes is built. The method consist in pruning the phylogenetic tree as close as possible to the root, in order to create subtrees without parallel genes. The following picture exemplifies this procedure: two organisms (represented by colors red and blue) are being comparing and five genes of a fictitious family X were found. The dashed line indicates how the family X will be subdivided in three new families.

Computing Rearrangement Distances

Once we have profiles that cover all vibrio proteins, we can identify homolog structures and apply our model to compute rearrangement distances. The steps necessary to calculate these distances are briefly described below.

1. Align all proteins that are being analyzed to the profiles, in order to identify homolog structures between them.
2. Identify families with multiple members in one or both genomes and subdivide them.
3. Eliminate genes that do not have a homolog in the other genomea and accounted the number of discarded block of consecutive genes for in the rearrangement distance.
4. Finally, apply rearrangement algorithms.

Conclusions

We proposed a new procedure to use when comparing whole genomes of vibrio species.

We intend to implement such procedure and gain from its use a deeper understanding of the evolution of vibrios. We also hope to help establish a sound, practical, and scalable methodology to compare complete genomes in other bacterial species.

References