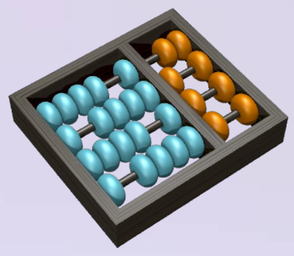
Comparison Between Complete Genomes of Vibrio Species



Patrícia P. Côgo[1], João Meidanis[1], Fabiano Thompson[2]
[1]University of Campinas, Institute of Computing, Campinas, Brazil
[2]Federal University of Rio de Janeiro, Institute of Biology, Departament 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 evolutive distance, which we believe represents an advantage over analyzes 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 obteined, we define our methology to compare complete genomes of vibrios, presented in part 2.

Part I

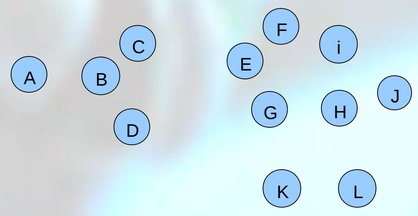
Preliminary Experiments

Six organisms, wich have their genomes completelly sequenced, were employed in our experiments, namely:

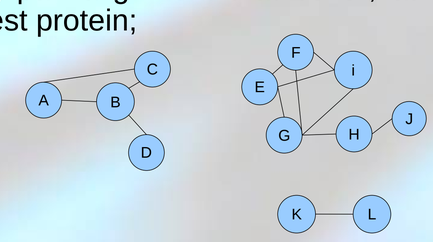
•*Photobacterium profundum;*
•*Vibrio choleare;*
•*Vibrio fisheri;*
•*Vibrio parahaemolyticus;*
•*Vibrio vulnificus* CMCP6 and YJ016**.**

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

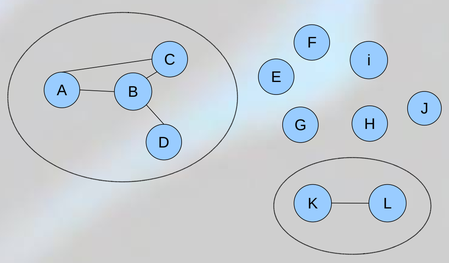
•Represent all proteins as graphs vertex;



•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;



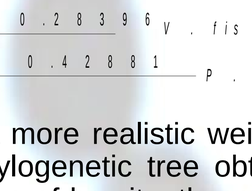
•Finally, employ a linkage method to group homolog genes: a group is manteined only if it has a member similar to all the other members; groups that do not satisfiy this criterion are dismantled in to singletons.



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

•Dependence on the organisms being comparing: families can be merged or depareted when an organismid add or excluded from the analysis

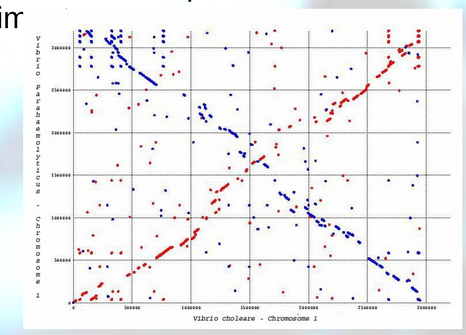
•Highly connected gene sets (exemplified by {A, B, C, D} in the following picture) could be dismantled , if the entire group does not have a member similar to all the others.



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 diferent scenarios, varying the weigth aplied 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 obteined agrees with the ones built by traditional analysis of gene markers.

Albeit, when a more realistic weithing scheme is aplied, the phylogenetic tree obteined is not so correct (for sake of brevity, these data will not be show here). It can indicate that the operations set is inadequate for the problem in hand. This conjecture is reinforced by the analysis of dotplots of these genomes, that show the large occurence of reversals in the evolution of these organisms, an operation not used in these experim



Part II

Proposal of a Comparison Methodology

In face of the problems founded in our initial analysis, we are proposing modifications in our scheme in order to obtein a more accurate comparison methodology.

Homology

The precision of rearrangement distances is extremely dependent on the homology detection accuracy. In order to define a consistent and accurate method, we delineated two criteria that it must observe, namely:

•*Transitivity*, that is, if A is homologous to B, and B is homologous to C, then A must be homologous to C.
•Independence of the gene set being analysed**.**

Complying these rules, we adopted a homology identification method based on ***Profile Analysis***, that is, weighted matrices that describe families of proteins,. as in the PROSITE and HAMAP projects [3,4] but with many more profiles, so that we can cover all vibrio proteins.

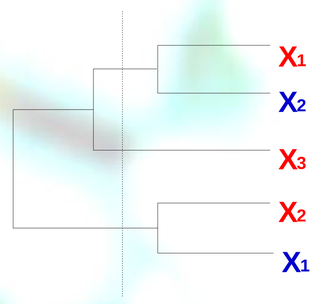
Model Description

We propose a set of rearrangement operations representing biological events relevant to vibrio evolution. The evolutionary distance between two genomes will follow the most parsimonious scenario approach, i. e., it will be defined as the total weight of a minimum-weight series of operations that transform a genome into the other. Operation weights will be selected taking into account both their frequency in evolution, and computation efficiency.

Our proposal is based on the occurrence of four operations:
•Fissions;
•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 find a gene family with more than one member in one or both genomes, a phylogenetic tree with all members 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 paralog genes. The following picture exemplifies this procedure: two organisms (represented by colors red an 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 ones.



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 subdivided them.
3. Eliminate genes that does 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 proposal 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

[1] J. Meidanis and Z. Dias. An alternative algebraic formalism for genome rearrangements. *Comparative Genomics Empirical and Analytical Approaches to Gene Order Dynamics Comparative Maps Multigenes Families*, 2000.
[2] A. Mclysaght, P. F. Baldi, and B. S. Galt. Extensive gene gain associated with adaptive evolution of poxviruses. *Proceedings of the National Academy of Sciences of the USA*, 100(26):15655–15660, 2003.
[3]C.J.A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinformatics, 3:265–274, 2002.
[4] A. Gattiker, K. Michoud, C. Rivoire, A. H. Auchincloss, E. Coudert, T. Lima, P. Kersey, M. Pagni, C. J. Sigrist, C. Lachaize, A. L. Veuthey, E. Gasteiger, and A. Bairoch. Automatic annotation of microbial proteomes in swiss-prot. Comput. Biol. Chem., 27:49–58, 2003.

This research is partially supported by grants by FAPESP (05/53279-6) and CNPq (470420/2004-9)