Clustering of discrete and fuzzy phylogenetic profiles

Fotis E. Psomopoulos[1,2,*], Pericles A. Mitkas[2] and Christos A. Ouzounis[1,3]
[1] Institute of Agrobiotechnology – CERTH, Thessaloniki GR-57001, Greece
[2] Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki GR-54124, Greece
[3] School of Natural & Mathematical Sciences, King's College London, London WC2R 2LS, UK
[*]Correspondence to: fpsom@issel.ee.auth.gr and christos.ouzounis@kcl.ac.uk or ouzounis@certh.gr

Motivation: Phylogenetic profiles have long been a focus of interest in computational genomics (Pellegrini et al. 1999). Encoding the subset of organisms that contain a homolog of a gene or protein, phylogenetic profiles are originally defined as binary vectors of n entries, where n corresponds to the number of target genomes. It is widely accepted that similar profiles especially those not connected by sequence similarity correspond to a correlated pattern of functional linkage. To this end, our study presents two methods of phylogenetic profile data analysis, aiming at detecting genes with peculiar, unique characteristics.

Methods: The Multi-level Clustering algorithm is an iterative method, reminiscent of k-means. It accepts a number of profiles as input and returns a tree-like structure of clusters (**Figure 1**). The bottom level consists of all profile instances as singleton clusters, whereas the top level (root of the tree) is a single cluster containing all profiles. Each iteration evaluates the centroid of each cluster, using an increasingly relaxed measure of similarity (distance measure).

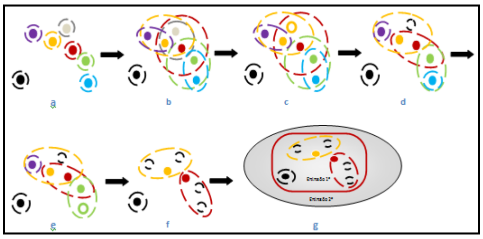

Figure 1:Diagram of six steps of Multi-level clustering.

The Fuzzy PhyloProf algorithm is a four-stage process; (a) construction of the fuzzy genome profile for each participating genome, (b) discretization of the fuzzy profiles, (c) de-noising of the initial profile data, and (d) calculation of the intra-/inter-genome distances for each gene. The produced distance diagram is decomposed into four areas: (i) top left quadrant, (ii) bottom right quadrant, (iii) bottom left corner, and (iv) the area along the main diagonal. It is argued that genes at the top left quadrant are highly genome-specific, whereas gene at the bottom right quadrant exhibit unexpected species distribution, possibly due to an exogenous nature.

Results: The methods were evaluated using a dataset of 3896 profiles from ProfUse (Goldovsky et al. 2005) across five species, as a benchmark dataset (**Table 1**). Applying the Multi-Level Clustering algorithm, the dataset produced five clusters at the 2nd level of the cluster tree. However, two of the five clusters (**Figure 2**) contained only 1.46% of the whole dataset, which is peculiar given the highly flexible distance measure at that level (similarity ≥ 31%).
The same dataset on the Fuzzy PhyloProf algorithm produced distance diagrams that clearly show interesting genes, based on the intra-/inter-genome distances in each case (**Figure 3**).

Genome ID	Number of genes	Percentage of dataset (%)	Relative Position in pp
BAPH-XSG-01	545	13.99	88
MGEN-G37-01	479	12.29	2
NEQU-N4M-01	563	14.45	148
SPYO-SF3-01	1696	43.53	50
UURE-SV3-01	613	15.74	39

Table 1: Input Genomes.

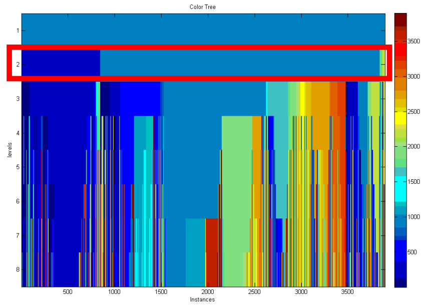

Figure 2: Clusters of all different levels. Each cluster is assigned to a distinct color. The five clusters of the second level are marked by a red horizontal bar.

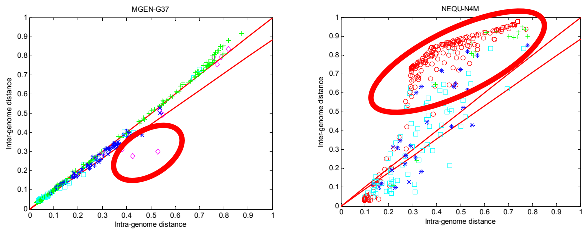

Figure 3: Fuzzy distance diagrams of genomes MGEN-G37-01 (left) and NEQU-N4M-01 (right). Marked are the highly genome-specific genes in NEQU-N4M-01, and the "peculiar" genes in MGEN-G37-01.

Discussion: Genes with similar phylogenetic profiles are likely to have similar structure or function, such as participating to a common structural complex or to a common pathway. Our two methods aim at detecting those outlier profiles of "interesting" genes, or groups of genes, with different characteristics from their parent genome.

References

Goldovsky, L. *et al.* (2005) *Cogent++: an extensive and extensible data environment for computational genomics.* Bioinformatics 21: 3806-3810.

Pellegrini M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.Proc. Natl. Acad. Sci. USA 96: 4285-4288.

5th Conference of the Hellenic Society for Computational Biology and Bioinformatics
17-19 October 2010, Thraki Palace Hotel, Alexandroupoli