Session:

Proceedings of the 4th conference of the
Hellenic Society for Computational Biology & Bioinformatics - HSCBB09
NHRF, Athens, 18-20/12/2009

XX/12/2009

Multi Level Clustering of Phylogenetic Profiles

Konstantinos M. Karagiannis [1], **Fotis E. Psomopoulos** [1,2,*] **and Pericles A. Mitkas** [1,2]

[1] Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki, GR541 24, Thessaloniki, Greece
[2] Intelligent Systems and Softare Engineering Lab, Informatics and Telematics Institute - CERTH, GR570 01, Thessaloniki, Greece

*Correspondence to: fpsom@issel.ee.auth.gr

ABSTRACT

Motivation: The prediction of gene function from genome sequences is one of the main issues in Bioinformatics. Most computational approaches are based on the similarity between sequences to infer gene function. However, the availability of several fully sequenced genomes has enabled alternative approaches, such as phylogenetic profiles (Pellegrini *et al.* (1999)). Phylogenetic profiles (*pp*) are vectors which indicate the presence or absence of a gene in other genomes. The main concept of pp's is that proteins participating in a common structural complex or metabolic pathway are likely to evolve in a correlated fashion. In this paper, a multi level clustering algorithm of pp's is presented, which aims to detect inter- and intra-genome gene clusters.

Methods: The clustering algorithm is an iterative method that accepts a number of pp's as input and returns a tree-like structure of clusters. The bottom level consists of all the pp instances as singleton clusters, whereas the top level (root of the tree) is a single cluster containing all pp's. Each iteration evaluates the centroid $C_i = \left[\frac{\sum_{j=1}^{n} p_{i,j}}{n}\right]$ of each cluster, where i is the cluster ID, n the number of pp's in the cluster, and $p_{i,j} = (p_{i,1}, p_{i,2}, \ldots, p_{i,n})$ the pp's in cluster i. Clusters with similar centroids are then merged. In the case of overlapping clusters, the common instances are assigned to a single cluster using the metric shown below.

$$U(p_i, k) = \frac{h(p_i, C_k)}{\widetilde{h}(k)}, \text{ where } \widetilde{h}(k) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} h(p_{k,i}, p_{k,j})}{n(n-1)},$$

$h(p_i, C_k)$ is the hamming distance of instance p_i from the centroid C_k of cluster k, and $\widetilde{h}(k)$ is the average distance of all members in cluster k.

Results: The algorithm was evaluated using a dataset of 3896 pp's from ProfUse (Goldovsky *et al.* (2005)) across five species (presented in Table 1). The result was an 8-level clustering tree (not including the top level), as shown in Fig 1. Each level of clustering presents unique characteristics. The common "signal" of the genes in the same species, which is prevalent in the lower levels, tends to be filtered out in higher levels and it is completely lost at the top level. The clusters in the same level present several cases with strong inter-genome or strong intra-genome cohesion (Fig 2).

Discussion: The algorithm addresses the problem of gene clustering in a less deterministic way and introduces the notion of levels in gene clustering. This method serves as a general computational tool for the annotation of large numbers of genes by highlighting evolutionary and functional patterns. The experiments showed that this method spots distinct patterns based on inter and intra-genomic signals. The outcome is a multilevel gene clustering, which attempts to capture at each level the different aspects of the affinity of a protein with another, in the same or in a different species.

REFERENCES

Goldovsky, L. *et al.* (2005). Cogent++: an extensive and extensible data environment for computational genomics. *Bioinformatics*, **21**(19), 3806 – 3810.

Pellegrini, M. *et al.* (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Nat. Acad. of Sc. USA*, **96**(8), 4285–4288.

Table 1. Input Genomes

Genome ID	# of genes	% of dataset	Rel. Pos. 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

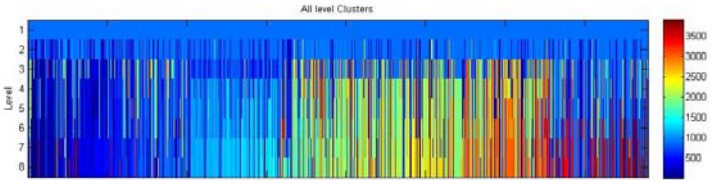

Fig. 1. Clusters of all different levels. Each cluster is designated a distinct color.

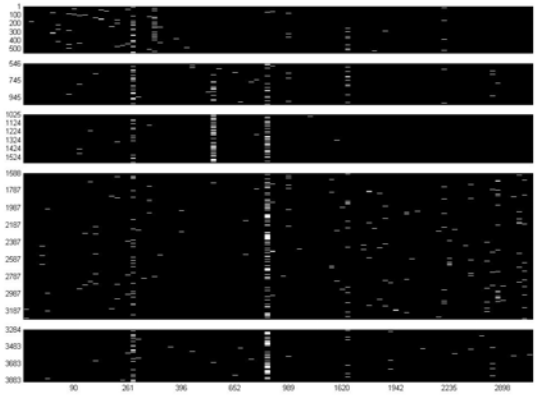

Fig. 2. Clusters created in the 3rd level. White spaces denote the presence of a gene (horizontal axis), whereas black color denotes the absence of the protein from the cluster (vertical axis)