Session:

Proceedings of the 6th conference of the
Hellenic Society for Computational Biology & Bioinformatics - HSCBB11
University of Patras Conference Center, 7-9/10/2011

XX/10/2011

Detecting species evolution through metabolic pathways.

Dimitrios Vitsios[1], **Fotis E. Psomopoulos**[1,2,*], **Pericles A. Mitkas**[1] **and Christos A. Ouzounis** [2]

[1] Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki, GR541 24, Thessaloniki, Greece, and
[2] Institute of Agrobiotechnology (INA),Center for Research and Technology Hellas (CERTH), GR570 01, Thessaloniki, Greece

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

ABSTRACT

Motivation: The emergence and evolution of metabolic pathways represented a crucial step in molecular and cellular evolution. With the current advances in genomics and proteomics, it has become imperative to explore the impact of gene evolution as reflected in the metabolic signature of each genome (Zhang *et al.* (2006)). To this end a methodology is presented, which applies a clustering algorithm to genes from different species participating in the same pathway.

Methods: The algorithm accepts as input a KEGG (Kanehisa and Goto (2000)) pathway map identifier $mapID$, and list of n genome identifiers that consist the target data set. By retrieving the k_i genes ($i = 1 \ldots n$) from each genome that participate in $mapID$, n blastable databases are constructed. In the next step, $n \times \sum_{i=1}^{n} k_i$ blast searches are performed, leading eventually to a matrix $P[\sum_{i=1}^{n} k_i][n]$, where each row is an n-bit vector that denotes the presence or absence of a homologue in each genome for the corresponding gene (reminiscent of phylogenetic profiles). Finally, the data in this homology matrix P are clustered using the MCL (Enright *et al.* (2002)) or EM algorithm and a custom similarity metric $sim_{jac} = \frac{m_{11}+m_{00}}{m_{01}+m_{10}+m_{11}}$ based on the jaccard metric.

Results: The proposed method was applied as a test case on the Glycolysis / Gluconeogenesis metabolic pathway (*KEGG identifier map00010*), which is well known and extensively documented. The three genomes participating in the test case, namely *Escherichia Coli K-12 MG1655*, *Arabidopsis Thaliana* and *Homo Sapiens*, were selected for sufficient phylogenetic diversity. The total number of genes in the dataset is 209 (eco:39, ath:105 and hsa:65).The selected clustering algorithm is MCL, and the results are presented in Figures 1 and 2. The produced clusters were evaluated using both the modified similarity metric and the gene homology, and the intra-cluster cohesiveness (≈ 1.813) was significantly higher than the inter-cluster similarity (≈ 0.479).

Discussion: Although these are only preliminary results, some interesting observations can be made. The presented test case was one among several different experimental setups. In all cases however, the first cluster always contained EC identifiers along the main reaction chain of the pathway, leading to the tentative conclusion that it may correspond to the highly conserved genes. Moreover, by superimposing the highlighted pathway diagrams along the implied phylogenetic distance of the genomes, one may infer sub-chains of the pathway that have been transformed or evolved across the species. A thorough investigation of this problem, together with rigorous experimentation on several different sets of pathways / genomes may provide more information in these areas.

REFERENCES

Enright, A. J., Dongen, S. V., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, pages 1575–1584.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, pages 27–30.

Zhang, Y., Li1, S., Skogerb, G., Zhang, Z., Zhu, X., Zhang, Z., Sun, S., Lu, H., Shi, B., and Chen, R. (2006). Phylophenetic properties of metabolic pathway topologies as revealed by global analysis. *BMC Bioinformatics*, pages 1–13. doi:10.1186/1471-2105-7-252.

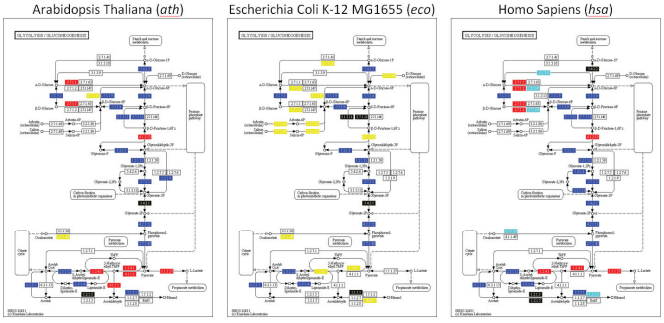

Fig. 1. The Glycolysis/Gluconeogenesis pathway for the three genomes in the case study. In each pathway the EC identifiers of the corresponding genome are highlighted according to the cluster their genes belong to. A special case are the EC identifiers highlighted in black: they contain genes from more than one clusters. However, it must be noted that each gene is assigned to a single cluster, whereas an EC identifier, corresponding to several genes, may in turn belong to different clusters.

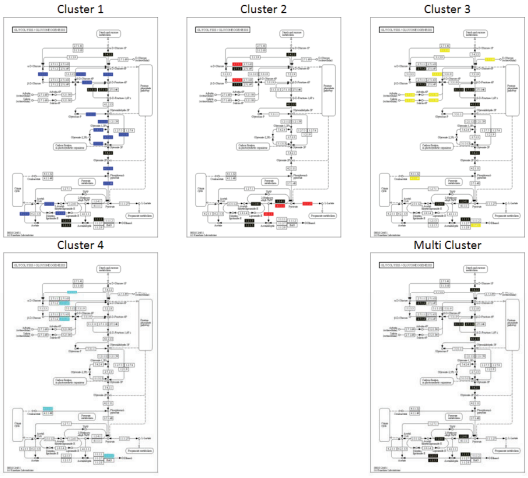

Fig. 2. The Glycolysis/Gluconeogenesis pathway for each of the four produced clusters, and the case of EC identifiers with genes from multiple clusters (black highlights). It is interesting to note that the first cluster contains genes that constitute the main process of the pathway, whereas the EC identifiers of the fourth cluster contain genes only from the human genome.