Minimum metabolic information for the reconstruction of the evolution of metabolisms

Irene García Mosquera[1], *Bessem Chouaia*[2], *Mercè Llabrés*[3], *Marta Simeoni*[4]

[1]irene.garcia@uib.es, Mathematics and Computer Science Department, University of the Balearic Islands and Health Research Institute of the Balearic Islands (IdISBa),

[2]bessem.chouaia@unive.it, Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari Venezia

[3]merce.llabres@uib.es, Mathematics and Computer Science Department, University of the Balearic Islands and Health Research Institute of the Balearic Islands (IdISBa)

[4]simeoni@unive.it, Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari and European Centre for Living Technology,Venice, Italy

The metabolism involves chemical reactions that link one to the other, creating a complex network of reactions. In this work, we analysed the potential of a simplified representation of the metabolism as a graph, where nodes are metabolic pathways, and there is an edge between two nodes if their corresponding pathways share one or more compounds. We call it an Abstract Metabolic Network (AMN). Our goal was to investigate the extent to which AMNs help discern the different taxonomic groups and capture evolutionary steps. We considered the metabolism of 7141 species stored in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database and performed a large-scale comparison of their AMNs using graph kernels. Specifically, we employed the Vertex Histogram, Shortest Path, Weisfeiler-Lehman and Pyramid Match graph kernels. We performed various experiments, first by considering the whole set of selected species, then by considering all the Eukaryotes and, within Eukaryotes, their kingdoms separately, and finally focusing on Prokaryotes. The comparison results were investigated through exploratory data analysis (heatmaps, multi-dimensional scaling and clustering techniques) as well as machine learning techniques (support vector machine) for prediction analysis. Looking at the experiments, we observe that all the results turned out to be biologically and evolutionary meaningful. Moreover, although performing differently, all the considered kernels reported a similar clustering pattern at a higher taxonomic level, thus suggesting that such patterns are clear and robust. This allows us to state that AMNs can reflect key evolutionary processes within the metabolism. However, it is also clear that, in general, they fail to capture fine-grain differences between species due to the need for more information on the reactions within each metabolic pathway. This result gives rise to new research questions that we would like to address for future work: How to use AMNs to highlight the relevant topological similarities and differences among the metabolism of different species? Is it possible to explore the pathways dependencies of two or more symbiotic species?

Keywords: Graph kernels, multivariate data analysis, support vector machine