ON STUDYING DUPLICATIONS IN GENOMES BY USING RANDOM GRAPHS

Nadia Pisanti[1,2] *pisanti@di.unipi.it*

Alain Viari[1] *Alain.Viari@snv.jussieu.fr*

[1] Department of Computer Science, University of Pisa
[2] Atelier de Bionformatique, University of Paris VI

Abstract. Investigating similarities among genes within a genome is useful for studying its evolution because families of similar genes may correspond to proteins having a common ancestor (homologs). In the usual approaches, one computes a distance for every pair of genes, and similarities are detected as pairs whose distance is smaller than a given threshold, considered as significant according to a statistical model. The resulting data, which depends upon the threshold, can be represented as a graph. In this work, we suggest a different way for detecting similarities and genes families, which does not rely upon fixing a single threshold nor evaluating directly the significance of the distance values. The principle is to compare the graphs obtained for all possible thresholds to suitable random graphs. Our idea is that the properties of a given graph that represents the biological data are interesting wherever they differ from those of the random graph. This approach requires to search for a suitable model of random graph and to identify suitable properties to be analysed. Here we propose a first model and apply the method to the complete genome of *Mycoplasma genitalium*. Despite the many approximations we made, the results revealed interesting data that could not have been retrieved with traditional techniques.