MSIAM/MoSIG/SIGMA – 2nd year research internship

Structural complexity and compression of tree-structured data

Supervisors: Marion Revolle(1), François Cayre(1) and Jean-Baptiste Durand(2)
E-mail: Marion.Revolle@gipsa-lab.grenoble-inp.fr, Francois.Cayre@gipsa-lab.grenoble-inp.fr, Jean-Baptiste.Durand@imag.fr
Location of internship: GIPSA-lab

(1) GIPSA-lab, 11 rue des Mathématiques, Saint Martin d'Hères. +33(0) 4 76 82 63 78
(2) Laboratoire Jean Kuntzmann & Inria, Equipe Projet Mistis, Saint Martin d'Hères. +33(0)4 76 63 57 09

Context:

To assess the structural complexity of a collection of elements, approaches based of the concepts of entropy or Kolmogorov complexity have been developed. However, computing entropies requires either strong statistical assumptions or a huge amount of data, while computing Kolmogorov complexities requires a Turing machine. A convenient proxy, referred to as the algorithmic complexity, is obtained by approximating the Kolmogorov complexity by a compression rate, using a recursive encoder for example in the Lempel-Ziv family. Extensions of this theory allow the definition of measures of relative complexity of an object x with respect to y, which is the compression rate of x when y is known. This leads to the definition of a distance $d(x,y)$ (Revolle *et al.*, 2016). Up to now, only sequences have been considered; however extending this approach to elements with more complex structures would allow us to define the structural complexity and similarity of, for example, trees and graphs. This requires recursive encoders to be developed for such structures.
Chen & Reif (1996) developed a compression algorithm for trees and claimed it to belong to the family of Lempel-Ziv compression algorithms. However the dictionary built while traversing the tree must be transmitted, which contradicts the Lempel-Ziv principle of online reconstruction of the dictionary while uncompressing the tree.

Tasks:

This project consists in conceiving and implementing some encoder for unordered trees proved to satisfy a Lempel-Ziv-like property. Based on such compression scheme, an extension of the algorithmic complexity and relative complexity will be proposed, as well as the associated distance.
The encoder and distance will be used to analyze the structure of plants. We expect to obtain low structural complexity for self-similar plants (Godin & Ferraro, 2010) and higher complexities for arbitrary plants. The distance will be used to perform plant clustering.

Prerequisites:

Basic knowledge in lossless compression theory and advanced programming skills are required.

This work is intended to be continued as a PhD thesis.

References:

Y. Chen and W. Reif. Efficient lossless compression of trees and graphs. In: *Proceedings of Data Compression Conference* (1996).
G. Godin and P. Ferraro. Quantifying the degree of self-nestedness of trees: application to the structural analysis of plants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**(4), 688-703 (2010).
M. Revolle, F. Cayre and N. Le Bihan. SALZA: Soft algorithmic complexity estimates for clustering and causality inference (2016). http://arxiv.org/abs/1607.05144