Research Topic for the ParisTech/CSC PhD Program

1. ***Field :*** Information and Communication Sciences and Technologies
2. ***Subfield***: Computer Science

Title: *Graph of symbols and degeneracy for genomic data*

ParisTech School: Ecole Polytechnique

Advisor(s) Name: Mireille Régnier
Advisor(s) Email: *Mireille.regnier@polytechnique.edu*
(Lab, website): *https://www.lix.polytechnique.fr/*

Short description of possible research topics for a PhD:

Genomic data are very diverse with potentially lots of errors and uncertainty including sequencing errors and mutations.
These issues are the grounds of a research at LIX on Information Retrieval and Text Mining. Graph of words methods, defined by M. Vazirgiannnis and colleagues [1, 2] have been extremely successful in text mining that presents significant analogies to genome. We plan to employ advanced techniques based on word enumeration and combinatorics and developed by M. Régnier [3] to estimate size and also stability properties.

More specifically we will capitalize on the graph of words methods to create graphs of symbols from biological sequences data aiming at creating dense structures that employ the biological significance of the sequences. The graph of words method is very robust with regards to maintaining infomration theory related metrics (reducing the initial coding entropy in the graph structure) and the robustness of this data structure for degenerated genomic data will be studied. One should consider in turn uncertainty due to evolution, to directed evolution and to sequencing errors. Moreover we will employ and validate on such specific data the graph degeneracy technique (a computationally feasible method) to approximate the densest genome graph that apparently has significant properties.

Required background of the student:

Background should be in computer science, with strong skills in algorithms.
As an alternative, a background in combinatorics, with solid knowledge on algorithms, would fit.

A list of 5(max.) representative publications of the group: (Related to the research topic)

1. Rousseau, F. and M. Vazirgiannis. *Graph-of-word and TW-IDF: new approach to ad hoc IR*. in *22nd {ACM} International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*. 2013. {ACM}.
2. Meladianos, P., et al. *Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream*. in *Proceedings of the Ninth International Conference on Web and Social Media, {ICWSM} 2015, University of Oxford, Oxford, UK, May 26-29, 2015*. 2015. {AAAI} Press.
3. Regnier, M., et al., *A Word Counting Graph*, in *{London Algorithmics 2008: Theory and Practice (Texts in Algorithmics)}*, J.W.D. Joseph Chan and M.S. Rahman, Editors. 2009, {London College Publications}. p. 31 p.