Master internship in bioinformatics / applied mathematics:

Gene prognostic biomarker for cancer patients using data mining and network based *a priori*

Project:

Correlated to the fast decreasing of DNA/mRNA sequencing costs, there has been in the last years an increase of public available data for cancer. As an example, in the public American database TCGA, there are more than 1000 patients diagnose with invasive breast cancer for which both whole genome mRNA sequencing data are available for tumor tissues and also associated clinical data including survival. In the past, important genes were discovered in different laboratories, and further validated as prognostic gene markers in a validation cohort. Actually, both group of genes and patients are clustered separately through resemblance of their profile, and the prognostic values for each group of patients are inferred through Kaplan-Meier approach. Here we propose a data driven learning approach to take both sequencing and clinical data into account to select genes, and to validate the approach with another cohort. To further improve gene selection, we will take into account known biological information in the form of a gene network in which edges correspond to gene interaction.

The work will be done mainly in a biological laboratory with a close collaboration with Florent Chatelain in Gipsa-lab. It will be in direct link with the PhD study of Rémy Jardillier, and on previous work with network based *a priori*. Discriminative models such as Cox regression will be considered to perform feature selection, namely prognostic gene marker identification, in high dimension, with penalization adapted to the network organization.

Supervisors: Laurent Guyon; laurent.guyon@cea.fr (contact email)
Rémy Jardillier; remy.jardillier@student.ecp.fr
Florent Chatelain; florent.chatelain@gipsa-lab.grenoble-inp.fr

Web links: http://laurent.guyon.phd.free.fr/
http://www.gipsa-lab.fr/page_pro.php?vid=727

Keywords: data analysis, network, bioinformatics, survival data, cox regression, cancer, High-dimensional statistics

Education required: Engineering schools or Master in bioinformatics, computer science or applied mathematics (in progress), with a clear interest in biological applications.

Duration: A least 3 months