

Post-doc proposal: Large-scale Joint Genetic and Neuroimaging Data Analysis based on cloud computing

Duration: 12 months, with possibilities for extensions to 18/24 months.

Work place:

Parietal Team, INRIA Saclay-Île de France
CEA Saclay, Bât 145, Point Courrier 156
F-91191 GIF/YVETTE, FRANCE
<http://parietal.saclay.inria.fr/>

Introduction

Imaging genetic studies linking functional MRI data and Single Nucleotide Polymorphisms (SNPs) data face a dire multiple comparisons issue. In the genome dimension, genotyping DNA chips allow to record of several hundred thousands values per subject, while in the imaging dimension a brain image may contain 100k-1M voxels. Finding the brain and genome regions that may be involved in this link entails a huge number of hypotheses, hence a drastic correction of the statistical significance of pairwise relationships, which in turn reduces crucially the sensitivity of statistical procedures that aims at detecting the association. It is therefore desirable to set up as sensitive techniques as possible to explore where in the brain and where in the genome a significant link can be detected, while correcting for family-wise multiple comparisons (controlling for false positive rate). Another issue is the computational cost of these procedures, that need to be addressed with adequate algorithmic and computational tools.

Objectives

In this project, we will consider a unique dataset acquired in the Imagen project, an FP6 project (<http://www.imagen-europe.com>) that aims at investigating factors of addition in a population of adolescents; Imagen's database contains multi-modal neuroimaging as well as genetics and psychological data on about 2000 subjects. This database is hosted and processed at Neurospin and is available for research purpose. The candidate will be in charge of:

- Setting an analysis pipeline (based on code already available to analyze neuroimaging/genetics datasets) to extract and pre-process the relevant data for statistical analysis.
- Performing statistical analysis on simulated datasets and sub-parts of the whole database in order to set all the computational framework. These procedures will include mass-univariate linear modeling (with peak- and cluster-level tests), regularized multiple regression and a permutation-based assessment framework.
- Launch data analysis on a large scale grid and cloud environment, with the help of the Kerdata researchers (see below).
- Build the post-analytic framework to ease the interpretation of the results in both neuroimaging and genetics domains.

The analysis framework is based on algorithmic tools developed in C/Python (numpy, scipy and scikit learn). The candidate will interact i) with researchers of the Parietal team for algorithmic aspects, but also ii) with CEA researchers of Neurospin, who will provide expertise in genetics domain and iii) with the KerData team (INRIA Rennes) and the Joint MSR-INRIA Research Center, that will provide help and massive computation facilities. The project has an access to grid/cloud computing facilities to be used in collaboration with INRIA/Kerdata and MSR-INRIA partners.

The expected results is the discovery of correlation between brain activation and genetic information.

Required knowledge and background :

The candidate should have at least a basic knowledge of standard statistical concepts. He or she should have a first significant experience in parallel computation and with python language. It is important that he or she has some real interest in genetics and/or brain imaging in order to have strong interactions with specialists of these domains. He or she will benefit from the algorithmic tools developed at Parietal and of the database settings and data pre-processing tools developed by Neurospin researchers.

Relevant references

- Ge, Y., Dudoit, S., and Speed, T. (2003). *Resampling-based multiple testing for microarray data analysis*. *Test*, 12:1–77.
- Poline, J.-B., Lalanne, C., Tenenhaus, A., Duchesnay, E., Thirion, B., and Frouin, V. (2010). *Imaging genetics: Bio-informatics and bio-statistics challenges*. International Conference on Computational Statistics Paris COMPSTAT 2010, 1:101–113.
- Vounou M, Nichols TE, Montana G, Alzheimer's Disease Neuroimaging Initiative. *Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach*. *Neuroimage*. 2010 Nov 15;53(3):1147-59.

Contact

Bertrand Thirion
Parietal Team, INRIA Saclay-Île-de-France
CEA Saclay, Bât 145, Point Courrier 156
F-91191 GIF/YVETTE, FRANCE
<http://parietal.saclay.inria.fr/Members/bertrand-thirion>