LABORATOIRE DE BIOLOGIE ET MODELISATION DE LA CELLULE

ENS - CNRS UCBL UMR 5239 - INSERM U 1210
46, allée d'Italie - 69364 Lyon cedex 07 – France

Tél : +33 4 72 72 81 71 - http://www.ens-lyon.fr/LBMC

Enriching Kernel-Based testing for single cell transcriptomics

Single-Cell transcriptomics now allows the quantification of gene expression at the scale of individual cells, encoded in count matrices countaining thousands observations (cells) and tens of thousands features (gene expression values). The analysis of such data requires new methodological frameworks, dedicated to their complexity and size. A major challenge consists in comparing the distribution of gene expression between conditions (ex: control vs treatment). In a recent work we developed a non-parametric test based on supervised kernel-based classification. Our procedure belongs to the family of Maximum Mean Discrepancy tests (MMD), that rely on a distance between the expectation of distributions embeddings in a Reproducing Kernel Hilbert Space (RKHS). This strategy can be enriched by considering the dependency structure of the data in this RKHS, which appears central in the field of single cell transcriptomics, to better account for biological variability. This test is then restated in a test based on a kernel Fisher Discriminant Analysis (kFDA). The application of our procedure to experimental data is very promising, and raises new research challenges in machine learning.

One main challenge is to quantify the importance of features (genes) that explain the discrimination between populations. This is a very general quiestion for kernel-based methods (non linear), for which there is no consensus framework to assess features importance. A possible strategy would be to use permutations to quantify each feature's importance, which is computationnally greedy, but could be simple to implement and interpret for biologists. Another strategy could be to perform feature selection using penalized methods like the lasso.

Another very promising aspect will be to incorporate the spatial positionning of cells in the tissue, thanks to the so-called spatial-transcriptomics technology. This new challenge will consists in integrating the spatial component to the kernel-based test that compares distributions.

The candidate will be co-supervised by Franck Picard (CNRS, ENS Lyon) and Bertrand Michel (EC Nantes), experts in computational statistics and statistical learning. The candidate will work at the ENS de Lyon, in an interdisciplinary environment, between mathematics, computer science and biology. Moreover, the candidate will benefit from the SingleStatOmics ANR project that gathers an interdisciplinary consortium in machine learning / IA dedicated to single cell genomics, with experts in machine learning, optimal transport and statistics.References :

- A kernel two-sample test; Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, Alexander Smola, The Journal of Machine Learning Research Volume 133/1/2012 pp 723–773
- Testing for Homogeneity with Kernel Fisher Discriminant Analysis, NIPS 2007, Moulines Eric, Francis Bach, Zaïd Harchaoui
- Regev A, Teichmann SA, Lander ES, et al. The Human Cell Atlas. Elife. 2017; 6:e27041.
- Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31 (2020).