Type d'offre : Laboratory offer

Post date : 20.01.26

Centre National de Recherche en Génomique Humaine (CNRGH)

Internship - Transfer learning models able to handle MISSing data for the survival analysis of rare cancer from multi-OMICS data

Informations générales

Contract type : Stage

Contract length : 6 months

Education level : M2

Contact :

msavino@cnrgh.fr
agloague@cng.fr

Starting date : Sun 01/03/2026 - 12:00

Trade : IR

Topic : Autre

Centre National de Recherche en Génomique Humaine (CNRGH) :

Within the CEA, the National Center for Human Genomics Research (CNRGH), located in the Évry-Courcouronnes Genopole, is a research center dedicated to the study of the human genome. It is part of the François Jacob Institute of Biology (IBFJ), which belongs to the CEA's Fundamental Research Division (DRF).
The center provides the French and European scientific communities with the capacity to produce, store, and analyze the biological data required to carry out projects in the field of medical genomics, including research on cancer, rare diseases, and autism. It is the largest sequencing platform in France and one of the five largest in Europe.

The Mathematics and Statistics (MS) team plays a pivotal role at the National Center for Human Genomics Research. Its scope of action covers three main functions. First, the MS team is responsible for quality control of data generated by the genotyping platform. Second, the MS team acts as a reference point for the evaluation and validation of statistical analysis plans for both internal and external collaborative projects involving human genetics studies that use the CNRGH's genotyping and sequencing platforms. Finally, the MS team initiates methodological research projects on original topics of interest to the genetics community. In particular, it has developed expertise in statistical methods for studying genetic associations, including rare variants, gene–environment interactions, gene networks (pathways), and multi-omics integration.

Address :

2 rue Gaston Crémieux
91000 Evry-Courcouronnes

France

Détail de l'offre (poste, mission, profil) :

Context. In the context of cancer, accumulations of aberrations observed at multiple molecular levels are the source of the many differences observed between somatic tumor and normal cells[1]. Abnormalities on DNA may include an increased number of mutations, differentially methylated sites (epigenetic markers), or copy number variations (different numbers of copies of a chromosome segment in a cell). Such modifications have an impact on gene expression, which in turn affect proteins. Studying these molecular data (namely omics) separately is often not enough to understand the undergoing dysregulation. This led to the establishment of multi-omics studies with the hope that looking jointly at all the molecular layers would unravel the big picture. From a statistical point of view, this would result in an increase of power. Indeed, combining multiple small effects, across several omic modalities, commonly explaining the same phenomenon would increase the signal-to-noise ratio. However, to achieve this purpose, the high dimensionality of such data (more than 20.000 coding genes) has to be handled to avoid estimating spurious associations. Therefore, a tremendous number of multi-omics analysis methods have been developed[2,3].

Among the tasks addressed with multi-omics data, survival analysis consists in estimating the duration between a patient's initial diagnosis and their death. Such analysis can identify groups of patients with differential prognosis and distinguished by a molecular (omic) signature. Clinicians can further investigate such signatures for new treatments or to better adapt therapies according to the molecular specificities of a given cancer. This is one way of performing precision medicine.

Despite the promise of multi-omics data, their benefit in the field of cancer survival analysis remain limited. In an insightful study[3], 12 survival analysis methods were compared on 18 cancer data-sets analyzed separately. The aggregated results across all cancers showed that only two methods using both clinical and molecular data performed better (not statistically) than a reference model using only clinical data. Adding Deep Learning methods in a follow-up study[4] did not change the conclusions. In an ongoing work[5], we added joint Dimension Reduction (jDR) methods to the comparison. These methods estimate a reduced space representing well the commonalities between omic layers[2]. We made the hypothesis that estimating such joint reduced space, prior to survival analysis, would improve the prediction results by better dealing with the high dimensionality of the data. Preliminary results identified two jDR methods, using both clinical and omics data, statistically outperforming the reference model, using clinical data only, after aggregating the results across all cancers. Further improvement in the performance of these methodologies may be expected.

However, we are still far from identifying robust candidate multi-omics biomarkers to be further investigated by clinical trials. This could mean that the dimensionality of the data is simply too high to construct good prediction models. We identified two major ways to better handle this curse of dimensionality. First, all studies mentioned above deal with complete

data, i.e. if a subject has at least one missing omic modality, this subject is not considered in the analysis. This strategy is known to be suboptimal and can further exacerbate the curse of dimensionality[6]. Then, this issue can be also alleviated by inserting information to the targeted data-set either by (i) making use of prior knowledge or (ii) through Transfer Learning (TL). The limitation with (i) is that the model must integrate a reliable/robust prior knowledge, which is not always possible especially in the case of rare diseases, which are typically poorly characterized and supported by very limited sample sizes. Transfer Learning, on the other hand, aims at extracting this prior knowledge from a Source data-set and transfer it to the desired data-set, called the Target, to learn faster (i.e. with fewer observations) a new task out of it. A common practice is to train a model on the Source and then fine-tune it on the Target. However, in order for this transfer to work, the datasets must be related.

General Goal. The objective of this internship is to study models able to both deal with missing data and perform Transfer Learning to tackle the curse of dimensionality in cancer survival multi-omics studies. These methodologies will be especially evaluated in the context of rare cancers (less than 6-15 new cases per 100.000 people per year; though 22-27% of cancer diagnosed and 25% of cancer mortality[7]) that could benefit the most from these approaches.

Tasks. To achieve this goal, the first task of this internship will be to perform a benchmark study on the biggest public multi-omic cancer data-set, The Cancer Genome Atlas (TCGA), gathering 33 cancer types for more than 11.000 patients across 8 modalities. Following previous TL studies working with complete data[8,9], all types of cancer but one will compose the Source data-set and the remaining one will act as the Target rare cancer. This setting is built upon the fact that preliminary studies have shown that information are shared across cancers through multiple omics data[10,11]. This would allow to learn a general “cancer knowledge” transferable to a targeted cancer. This setting will be repeated for several Target cancer to draw robust conclusions[3]. Furthermore, different missing data situations will be manually generated from the Source, the Target or both. Despite that jDR methods have already proven to outperform the others when a Target cancer is analyzed alone[5], in this study, the Source data-set will be composed of enough observations so that classical Machine Learning and Deep Learning methods are expected to be comparable[12]. Hence, both jDR[13] and Variational Auto-Encoders [14,15], their equivalent within a Deep-Learning framework, will be evaluated in this benchmark.

In a second time, such analysis will be applied on an adult rare cancer multi-omic cohort provided by Dr. Agusti ALENTORN. This cohort gathers 147 clinical, 123 transcriptomic, 115 Whole Exome Sequencing and only 64 methylation profiling data on Primary Central Nervous System Lymphoma[16].

Send CV and motivation letter to Mary SAVINO and Arnaud GLOAGUEN

Mary SAVINO : msavino@cnrgh.fr
Arnaud GLOAGUEN : agloague@cng.fr

Requirement :

M2 or last year of engineer school with specialty/knowledge in Computer Science / Statistics / Machine Learning / Deep Learning / BioStatistics.

Working knowledge in programming (R / Python, ...).

Previous experience with applications to genomics will be a plus.

1. Sun, W. *et al.* The association between copy number aberration, DNA methylation and gene expression in tumor samples. *Nucleic Acids Research* **46**, 3009–3018 (2018).

2. Cantini, L. *et al.* Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat Commun* **12**, 124 (2021).

3. Herrmann, M. *et al.* Large-scale benchmark study of survival prediction methods using multi-omics data. *Brief Bioinform***22**, bbaa167 (2021).

4. Wissel, D. *et al.* Systematic comparison of multi-omics survival models reveals a widespread lack of noise resistance. *Cell Rep Methods* **3**, 100461 (2023).

5. Goff, V. *et al.* Impact of joint Dimension Reduction methods for survival prediction - Extension of a multi-omics benchmark study. in (2024).

6. Flores, J. E. *et al.* Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front Artif Intell* **6**, 1098308 (2023).

7. Boyd, N. *et al.* Rare cancers: a sea of opportunity. *Lancet Oncol***17**, e52–e61 (2016).

8. Chai, H. *et al.* Predicting bladder cancer prognosis by integrating multi-omics data through a transfer learning-based Cox proportional hazards network. *CCF Trans. HPC* **3**, 311–319 (2021).

9. Li, Y. *et al.* Transfer Learning for Survival Analysis via Efficient L2,1-Norm Regularized Cox Regression. in *2016 IEEE 16th International Conference on Data Mining (ICDM)* 231–240.

10. Sato, G. *et al.* Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis. *Nat Commun* **14**, 3671 (2023).

11. Li, Y. *et al.* Pan-cancer proteogenomics connects oncogenic drivers to functional states. *Cell* (2023)

12. Hanczar, B. *et al.* Assessment of deep learning and transfer learning for cancer prediction based on gene expression data. *BMC Bioinformatics* **23**, 262 (2022).

13. Hirst, D. P. *et al.* MOTL: enhancing multi-omics matrix factorization with transfer learning. *Genome Biology* **26**, 224 (2025).

14. Benkirane, H. *et al.* Multimodal CustOmics: A unified and interpretable multi-task deep learning framework for multimodal integrative data analysis in oncology. *PLOS Computational Biology* (2025)

15. Ranjbari, S. *et al.* Integration of incomplete multi-omics data using Knowledge Distillation and Supervised Variational Autoencoders for disease progression prediction. *Journal of Biomedical Informatics* **147**, 104512 (2023).

16. Hernández-Verdin, I. *et al.* Molecular and clinical diversity in primary central nervous system lymphoma. *Ann Oncol* **34**, 186–199 (2023).

Lien vers l'offre sur le site dataia.eu :https://da-cor-dev.peppercube.org/node/1510