

**M2 Internship in Biostatistics/Machine Learning
International Agency for Research on Cancer (IARC – WHO), Lyon**

Dimensionality reduction for metabolomics data in the EPIC cohort

The Biostatistics and Data Integration (<https://www.iarc.who.int/teams-nme-bdi/>) of the Nutrition and Metabolism (NME) Branch at IARC is looking for a motivated M2 student with strong background in Biostatistics and/or machine learning to apply state-of-the-art dimension reduction methods to metabolomics data in large prospective studies.

Large prospective epidemiological studies successfully identified several cancer risk factors, including tobacco smoke and obesity. Standard data collected within these large cohort studies are now increasingly supplemented with a range of omics data. By providing a snapshot of all metabolites present in a biospecimen, metabolomics data reflect individuals' health metabolic status. Their analysis allows the investigation of biological mechanisms possibly involved in cancer development, including those that might underlie established associations between lifestyle factors and cancer risk.

However, metabolomics data are high-dimensional data and exhibit complex correlation structures, which render statistical analyses challenging. The application of non-linear dimension reduction methods, including t-SNE, UMAP, autoencoders and other deep learning unsupervised approaches, could lead to the identification of metabolic signatures and, in turn, sub-groups of individuals associated with different cancer risk profiles. The integration of lifestyle and other omics data could further reflect the complex interplay between lifestyle, genetics and metabolism on cancer development, and produce even more informative signatures.

Within the EU funded project ColoMARK, the successful candidate will use R and/or Python to implement a range of machine learning methods to metabolomics data collected within the European Prospective Investigation into Cancer and nutrition (EPIC) study, a large international cohort study that recruited over 500,000 participants from 10 European countries. The NME Branch at IARC is a stimulating, international and multidisciplinary work environment.

References :

1. Van Assel et al. A probabilistic graph coupling view of dimension reduction, 2022; arXiv preprint <https://arxiv.org/abs/2201.13053>
2. Cai et al. Machine learning for multi-omics data integration in cancer, iScience, 2022; 25(2):103798. <https://pubmed.ncbi.nlm.nih.gov/35169688/>

Duration: 4 to 6 months
Required skills: Strong background in biostatistics, statistics, machine learning or related fields. Good knowledge of R and/or Python. Fluency in French and/or English.
Contact: Dr. Vivian Viallon (viallonv@iarc.who.int)
Stipend: 700€/month (net salary)
Closing date: 31/01/2023