enbis European Network for Business and Industrial Statistics

Contribution ID: **145** Type: **not specified**

The Challenges in Building Meaningful Models with Publicly Available Omics Data

Tuesday, 12 September 2023 17:20 (20 minutes)

Omics data, derived from high-throughput technologies, is crucial in research, driving biomarker discovery, drug development, precision medicine, and systems biology. Its size and complexity require advanced computational techniques for analysis. Omics significantly contributes to our understanding of biological systems.

This project aims to construct models for Human Embryonic Kidney cells used in industry for viral vector production by incorporating five types of omics data: genomics, epigenomics, transcriptomics, metabolomics, and proteomics. With over 25 terabytes of publicly available data, the abundances of each data type vary significantly, including more than 15,000 sequence runs covering the genome, epigenome, and transcriptome, as well as approximately 300 proteomics experiments and only 6 metabolomics experiments. Skewed data availability presents challenges for integrative multi-omic approaches for meaningful machine learning.

Data generation technologies have advanced rapidly, surpassing the computational capabilities required for analysis and storage. Dealing with diverse data structures and varying database information requirements poses significant challenges. The absence of a comprehensive data warehouse incorporating multiple omics data, with standardised quality and metadata criteria, complicates information extraction from diverse sources. The persistent issue of missing or inadequate metadata continues to impact data collection, casting doubts on adherence to the FAIR principles and raising significant concerns about the reproducibility and credibility of included studies. Implementing standardised criteria and improving documentation practices across databases is crucial. Addressing these challenges and developing strategies for integrating and analysing publicly available omics data from multiple sources have immense potential to advance our understanding of complex biological systems, furthering innovation in industry.

Keywords

Omics, Big Data Integration, Machine Learning

Classification

Mainly methodology

Primary authors: Ms PRICE, Eva (University College London); Dr FEYERTAG, Felix (Oxford Biomedica); Dr DIKICIOGLU, Dugyu (UCL)

Presenter: Ms PRICE, Eva (University College London)

Session Classification: CONTRIBUTED Data Mining

Track Classification: Data mining