Indo/US Collaborative Research Grants

National Science Foundation of US and Technology Innovation Hubs of India

Title: Integration of Multiomics Data using Deep Neural Networks: Feature Extraction, Association Mining, Big Data Realization and Privacy Preservation
Indian PI: Professor Rajat K. De, Technology Innovation Hub at Indian Statistical Institute, Kolkata, India
US PI: Professor Yevgeniy Vorobeychik, Computer Science & Engineering, Washington University in St. Louis, Saint Louis, USA

Integration of data originating from different sources, also called multimodal data, is a challenging task as the data types and distribution of data over various sources are different. Thus the problem being considered in the present proposal is on integration of multimodal data using deep neural networks. Here we will consider multiomics data as multimodal data. Multiomics studies have enabled us to understand the mechanistic drivers behind complex disease states and progressions, thereby providing novel and actionable biological insights into health status. However, integrating data from multiple omics modalities is challenging due to the high dimensionality of data and noise associated with each platform. Non-overlapping features and technical batch effects in the data make the task of learning more complicated. Conventional machine learning tools are not as effective in such data integration tasks. In addition, existing methods for single cell multiomics integration are computationally expensive. This has encouraged the development of novel deep neural network models for integration of high-dimensional multiomics data, which would be capable of learning meaningful features for further downstream analysis. One such deep neural model, called UMINT, that we have developed for extracting features from multiomics data, is depicted in Figure 1. Some other models are now under development for visualization of multiomics single cell data. Thus, the system will facilitate seamless analysis of the data with a diverse range of applications, and practices in precision medicine as well as healthcare management. In order to reduce computational complexity as well as overall cost of processing, we will develop a distributed computing platform under Map-Reduce paradigm to deploy the aforesaid deep neural network models using commodity hardware. Additionally, one of the key consideration in analyzing high-granularity individual-level data of this kind is to ensure privacy of the individuals who have contributed it. To this end, we are developing effective algorithmic approaches for preserving privacy of multiomic data release, focusing for the moment on the genomic summary statistics data. In particular, our recent focus is on the use of deep generative models in adversarial training to develop general techniques for privacy-preserving data sharing.

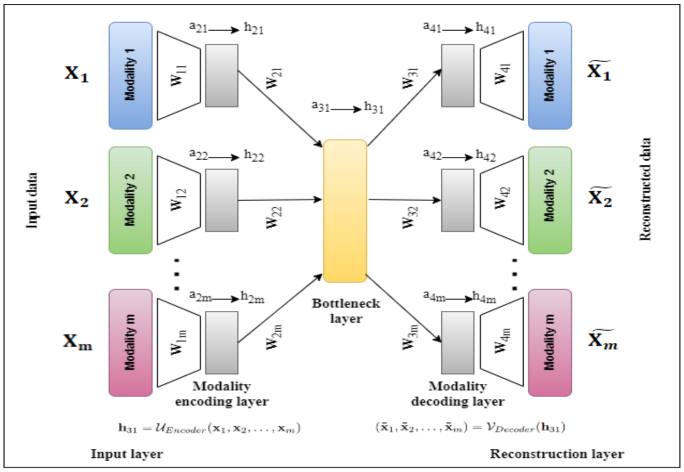
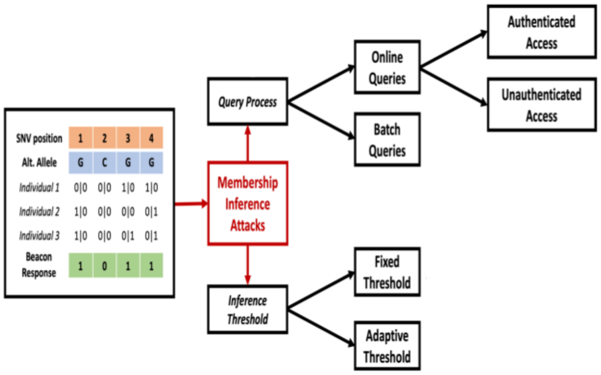

Input data
$\mathbf{X_1}$
$\mathbf{X_2}$
$\mathbf{X_m}$
Modality 1
Modality 2
Modality m
W_{11}
W_{12}
W_{1m}
$a_{21} \rightarrow h_{21}$
$a_{22} \rightarrow h_{22}$
$a_{2m} \rightarrow h_{2m}$
W_{21}
W_{22}
W_{2m}
$a_{31} \rightarrow h_{31}$
Bottleneck layer
W_{31}
W_{32}
W_{3m}
$a_{41} \rightarrow h_{41}$
$a_{42} \rightarrow h_{42}$
$a_{4m} \rightarrow h_{4m}$
W_{41}
W_{42}
W_{4m}
Modality 1
Modality 2
Modality m
$\widetilde{\mathbf{X_1}}$
$\widetilde{\mathbf{X_2}}$
$\widetilde{\mathbf{X}_m}$
Reconstructed data
Modality encoding layer
Modality decoding layer
$\mathbf{h}_{31} = \mathcal{U}_{Encoder}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$
$(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_m) = \mathcal{V}_{Decoder}(\mathbf{h}_{31})$
Input layer
Reconstruction layer

Figure 1: UMINT



SNV position	1	2	3	4
Alt. Allele	G	C	G	G
Individual 1	0|0	0|0	1|0	1|0
Individual 2	1|0	0|0	0|0	0|1
Individual 3	1|0	0|0	0|1	0|1
Beacon Response	1	0	1	1

Figure 2. Genomic data access models and privacy threats.