Senior Design Project

Unsupervised Discovery of Disease Subtypes Using Variational Clustering of Multi-Omics Data

Team:
Herbert George-22BCE7969
Meghana V Patil-22BCE9123

Guide: Dr. Rajkumar Yesurajiu

Introduction

- Cancer shows high patient-to-patient variability
- Single-omics analysis is insufficient to capture full biology
- Multi-omics integration (RNA-seq, methylation, miRNA) builds a complete patient profile
- Reveals subtle molecular differences between patients
- Unsupervised learning identifies hidden disease subtypes
- Deep learning manages high-dimensional omics data
- Enables precise subtyping, better diagnosis, and personalized treatment

Advantages

- Explains variation in treatment response and prognosis
- Supports personalized and precision medicine
- More robust to noise and incomplete data
- Captures multiple biological layers of a patient simultaneously
- Reveals hidden patient heterogeneity missed by single-omics
- Improves disease subtype discovery

Literature Survey

Authors & Citation	Contribution	Proposed Work	Advantages	Disadvantages
Chaudhary et al., 2018, Deep learning–based multi-omics integration robustly predicts survival in liver cancer, Clinical Cancer Research [1]	Proposed a deep learning framework integrating multi-omics datasets (gene expression, methylation, and miRNA data) to predict liver cancer survival outcomes.	Develop a deep neural network that integrates heterogeneous multi-omics data to classify disease subtypes and predict patient prognosis.	Improves predictive accuracy through multi-omics integration; captures complex biological interactions; robust survival prediction.	Requires large, well-aligned multi-omics datasets; computationally intensive; interpretability challenges in deep models.
Way and Greene, 2018, Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders, Pacific Symposium on Biocomputing [2]	Used Variational Autoencoders (VAE) to learn latent representations from cancer transcriptomic data and identify biologically meaningful patterns.	Implement VAE-based feature extraction to reduce dimensionality and uncover hidden disease subtype structures from gene expression datasets.	Efficient dimensionality reduction; captures nonlinear biological relationships; enables unsupervised learning of molecular features.	Latent representations may lack direct biological interpretability; performance depends on hyperparameter tuning and dataset quality.
Xie et al., 2016, Unsupervised deep embedded clustering for representation learning, ICML [3]	Introduced Deep Embedded Clustering (DEC), which jointly learns feature representation and cluster assignments using deep neural networks.	Apply DEC to cluster patient molecular profiles to identify potential unknown disease subtypes without requiring labeled data.	Enables simultaneous feature learning and clustering; improves clustering accuracy compared to traditional methods; fully unsupervised.	Sensitive to initialization; cluster stability may vary; requires careful parameter selection.
Tian et al., 2019, Clustering single-cell RNA-seq data with a model-based deep learning approach, Nature Machine Intelligence [4]	Developed a model-based deep learning method for clustering single-cell RNA sequencing data to identify cellular heterogeneity.	Extend model-based deep clustering techniques to multi-omics and disease subtype identification at cellular resolution.	Handles high-dimensional and sparse biological datasets; captures cellular heterogeneity effectively; improves clustering reliability.	Computationally expensive; sensitive to noise and dropout events in single-cell data; requires preprocessing optimization.
Wang et al., 2014, Similarity network fusion for aggregating data types on a genomic scale, Nature Methods [5]	Proposed Similarity Network Fusion (SNF) to integrate multiple biological data types by constructing and merging similarity networks.	Combine SNF with deep learning models to enhance multi-omics data fusion before clustering or classification.	Effective integration of heterogeneous data sources; improves subtype discovery accuracy; interpretable network-based approach.	Requires careful similarity metric selection; may struggle with highly imbalanced or incomplete datasets; limited deep feature learning capability.

Research Gap

- Most existing studies rely on single-omics data, leading to incomplete insights
- Traditional clustering methods struggle with high-dimensional and noisy data
- Limited use of unsupervised deep learning for multi-omics integration
- Poor alignment between different omics layers for the same patient
- Lack of models that jointly perform representation learning and clustering

Problem Statements

Diseases such as cancer exhibit high patient-to-patient heterogeneity.

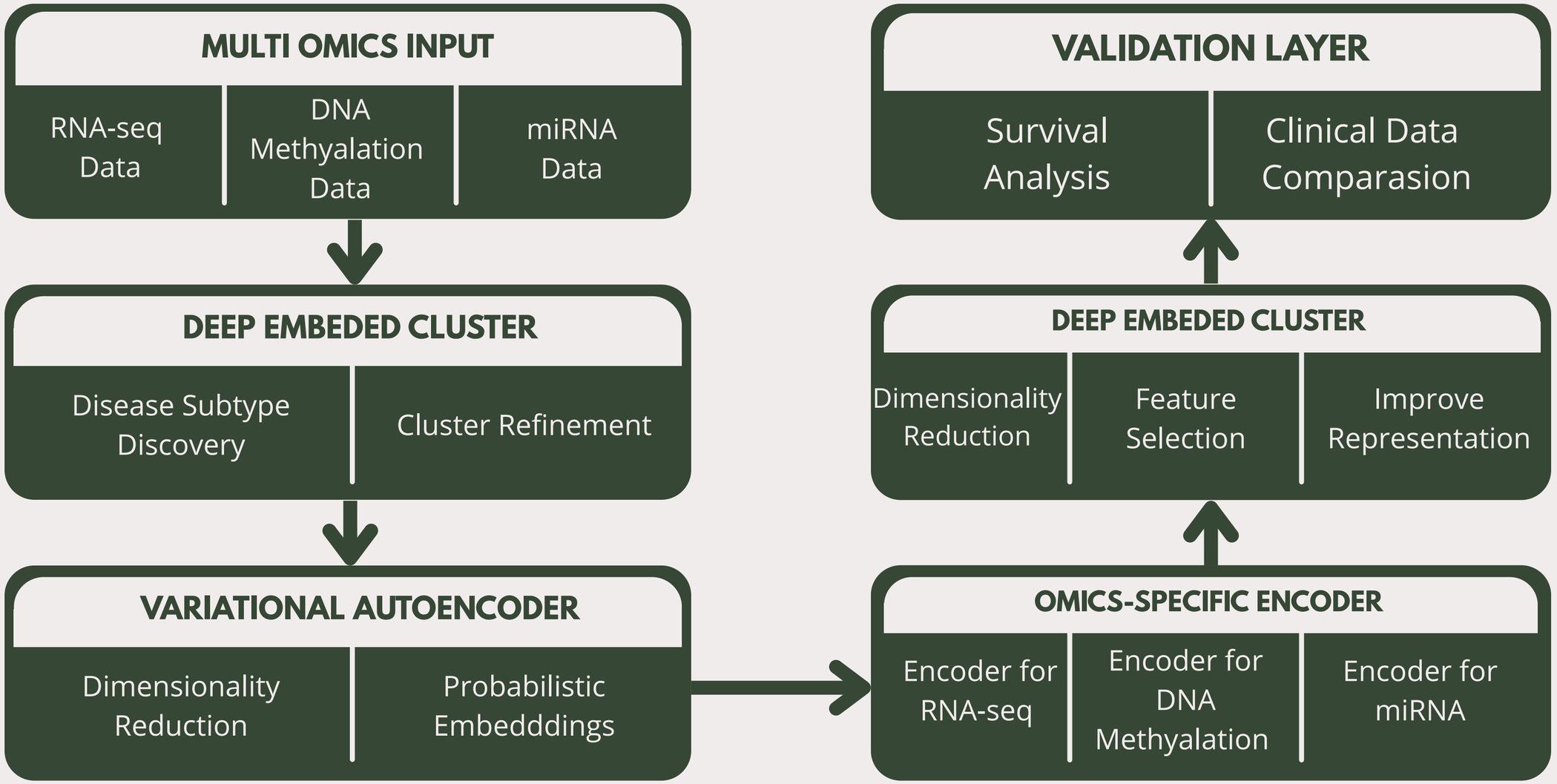
Single-omics analysis cannot capture complex cross-omics interactions.

Existing clustering methods struggle with high-dimensional, heterogeneous, unlabeled data.

An unsupervised deep learning framework is needed for effective multi-omics integration.

The goal is to identify meaningful disease subtypes automatically.

Block Diagram



Proposed Methodology

- Integrates RNA-seq, DNA Methylation, and miRNA data to capture cross-omics interactions
- Uses a Variational Autoencoder to reduce high-dimensional noise into probabilistic embeddings.
- Employs dedicated encoders for each data type to preserve unique molecular features.
- Applies DEC to jointly learn features and refine cluster assignments for subtype discovery.
- Verifies subtypes through survival analysis and clinical data comparison.

References

[1] K. CHAUDHARY, O. B. POIRION, L. LU, AND L. X. GARMIRE, “DEEP LEARNING–BASED MULTI-OMICS INTEGRATION ROBUSTLY PREDICTS SURVIVAL IN LIVER CANCER,” CLINICAL CANCER RESEARCH, 2018.
AVAILABLE: HTTPS://PUBMED.NCBI.NLM.NIH.GOV/28982688/

[2] G. P. WAY AND C. S. GREENE, “EXTRACTING A BIOLOGICALLY RELEVANT LATENT SPACE FROM CANCER TRANSCRIPTOMES WITH VARIATIONAL AUTOENCODERS,” PACIFIC SYMPOSIUM ON BIOCOMPUTING, 2018.
AVAILABLE: HTTPS://PUBMED.NCBI.NLM.NIH.GOV/29218871/

[3] J. XIE, R. GIRSHICK, AND A. FARHADI, “UNSUPERVISED DEEP EMBEDDED CLUSTERING FOR REPRESENTATION LEARNING,” INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML), 2016.
AVAILABLE: HTTPS://ARXIV.ORG/ABS/1511.06335

[4] T. TIAN, J. WAN, Q. SONG, AND Z. WEI, “CLUSTERING SINGLE-CELL RNA-SEQ DATA WITH A MODEL-BASED DEEP LEARNING APPROACH,” NATURE MACHINE INTELLIGENCE, 2019.
AVAILABLE: HTTPS://WWW.NATURE.COM/ARTICLES/S42256-019-0037-0

[5] B. WANG, A. M. MEZLINI, F. DEMIR, M. FIUME, Z. TU, M. BRUDNO, B. HAIBE-KAINS, AND A. GOLDENBERG, “SIMILARITY NETWORK FUSION FOR AGGREGATING DATA TYPES ON A GENOMIC SCALE,” NATURE METHODS, 2014.
AVAILABLE: HTTPS://PUBMED.NCBI.NLM.NIH.GOV/24464287/

Conclusion

- Multi-omics integration provides a comprehensive view of disease biology
- Unsupervised deep learning enables automatic discovery of hidden disease subtypes
- Latent representations learned from multi-omics data improve clustering quality
- The approach enhances disease understanding, prognosis, and treatment planning
- Overall, this framework has strong potential for real-world biomedical and precision medicine applications

Thank You