**Stephen Kocsis**

**8/13/2024**

**LFOmics: A Tool for Autoencoder Neural Network Based Multiomics Integration and Function Enrichment Analysis**

**Background**

The explosion of high-throughput sequencing technologies has led to a significant increase in the volume and complexity of biological data. Multiomic datasets, which combine different types of omics data such as genomics, transcriptomics, proteomics, and metabolomics, provide a comprehensive view of biological systems. However, integrating these diverse data types poses a significant challenge due to their heterogeneous nature. To address this challenge, I have developed a web-based tool named LFOmics (Latent Feature Omics) designed to facilitate the integration and analysis of multiomic data using advanced machine learning techniques and statistical methods.

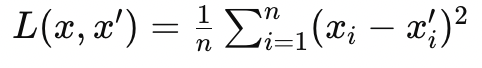
LFOmics leverages autoencoder-based machine learning models to reduce the dimensionality of multiomic data, capturing latent features that represent the underlying biological signals. These latent features are then subjected to Gene Set Enrichment Analysis (GSEA), a powerful statistical method used to identify significantly enriched biological pathways and gene sets. By combining machine learning with functional enrichment analysis, LFOmics enables researchers to uncover hidden patterns in multiomic data and gain insights into the functional relevance of these patterns.

**Math and Machine Learning Behind the Tool**

At the core of LFOmics is the autoencoder, a type of artificial neural network used for unsupervised learning. Autoencoders are particularly useful for dimensionality reduction, a process that simplifies the representation of complex data by reducing the number of variables (features) while preserving the essential information.

An autoencoder consists of two main components: an encoder and a decoder. The encoder compresses the input data into a lower-dimensional latent space, while the decoder attempts to reconstruct the original data from this compressed representation. The goal of training an autoencoder is to minimize the reconstruction error, which measures the difference between the original input and the reconstructed output. Mathematically, the autoencoder function can be represented as *h=f(Wx+b)* where *h* is the latent representation (encoded data), *W* is the weight matrix, *x* is the input data, *b* is the bias vector, and *f* is the non-linear activation function. The decoder function is given by *x′=g(Wh+b′)* where *x′* is the reconstructed data, *g* is a non-linear activation function, and *W* and *b′* are the weight matrix and bias vector, respectively.

The objective of training the autoencoder is to minimize the reconstruction loss, typically measured by the Mean Squared Error (MSE) between the original input and the reconstructed output, where *n* is the number of features:

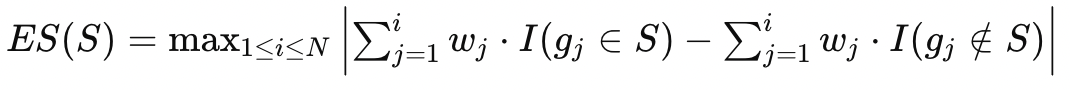


In LFOmics, the autoencoder is trained on the multiomic dataset, with rows representing genes and columns representing samples. The output of the encoder, the latent features, captures the most significant patterns in the data, allowing for a more manageable and interpretable representation of the underlying biological signals.

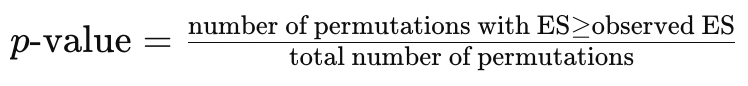
**Gene Set Enrichment Analysis (GSEA)**

After the latent features are extracted by the autoencoder, they are subjected to Gene Set Enrichment Analysis (GSEA), a statistical method used to determine whether a predefined set of genes shows statistically significant differences between two biological states.

GSEA works by ranking all genes in the dataset according to their correlation with a phenotype of interest. For a given gene set *S*, the Enrichment Score (ES) is calculated by walking down the ranked list *L*, increasing a running-sum statistic when a gene in *S* is encountered, and decreasing it when a gene not in *S* is encountered. The ES is the maximum deviation from zero encountered in this walk where *wj​* is the weight associated with the ranking of gene *j* and *I(gj​ ∈ S)* is an indicator function that is 1 if gene *gj* is in gene set *S* and 0 otherwise:



The ES reflects how much the genes in *S* are overrepresented at the extremes (top or bottom) of the ranked list *L*. To assess the statistical significance of the ES, GSEA performs a permutation test by randomly permuting the phenotype labels and recalculating the ES for each permutation. The p-value is then computed as the fraction of permutations that result in an ES as large as or larger than the observed ES:



Finally, the False Discovery Rate (FDR) is calculated to account for multiple hypothesis testing, providing a measure of confidence in the significance of the enrichment results.

**Biological Functionality Behind the Tool**

LFOmics is designed to facilitate the integration and analysis of multiomic datasets, which are essential for understanding complex biological systems. By combining different types of omics data, researchers can gain a more comprehensive view of cellular processes, disease mechanisms, and potential therapeutic targets.

The tool's biological functionality is centered around the identification of significant biological pathways and gene sets through GSEA. This analysis helps researchers understand the functional implications of the latent features extracted by the autoencoder, providing insights into the molecular mechanisms driving the observed biological phenomena.

LFOmics is applicable to all areas of biology, such as developmental biology, immunology, and neuroscience. By enabling the integration and analysis of multiomic data, the tool supports a wide range of research applications aimed at understanding the complexity of biological systems.

**How to Use the Tool**

LFOmics is designed to be user-friendly, with a web-based interface that allows researchers to upload their multiomic datasets and receive comprehensive results. The tool is accessible through a web browser and does not require any specialized software installation on the user's computer as it is a server-side analysis based application.

* **Step 1 (uploading your data):** To use LFOmics, the first step is to prepare your multiomic dataset in a CSV file format. The file should have genes as rows and samples as columns, with the first column containing the gene symbols and the remaining columns containing the expression levels for each sample. Once your data is prepared, navigate to the LFOmics web interface and click on the "Upload Your Multiomics Data" section. Use the "Choose CSV file" button to select your file and then click "Upload and Process" to initiate the analysis.
* **Step 2 (data processing and autoencoder training):** After the file is uploaded, LFOmics automatically processes the input data by standardizing the expression levels across samples. This step ensures that the data is normalized and ready for autoencoder training. The autoencoder is then trained on the processed data to extract latent features that represent the most significant patterns in the dataset. The training process involves multiple epochs, with the model learning to minimize the reconstruction error between the original input and the reconstructed output. The training loss is plotted and saved as a PNG file, providing a visual representation of the model's performance.
* **Step 3 (gene set enrichment analysis):** Once the autoencoder training is complete, the latent features are subjected to GSEA to identify significantly enriched biological pathways and gene sets. The top 1000 genes contributing to the latent features are selected for enrichment analysis using the KEGG 2019 Human gene set database. The results of the GSEA include enrichment scores, p-values, and FDR values for each gene set, as well as a bar plot showing the top 10 enrichment scores. These results are saved as CSV and PNG files, respectively, and can be downloaded from the results page.
* **Step 4 (view and download results):** After the analysis is complete, you will be redirected to the results page, where you can view and download the results. The page provides links to download the latent feature matrix, decoded latent features, enrichment results, and visualizations. The results are also saved in a SQL database, allowing for easy retrieval and further analysis. Each result is associated with a unique identifier, making it easy to track and manage multiple analyses.

**Discussion**

LFOmics represents a powerful and versatile tool for the integration and analysis of multiomic data. By combining advanced machine learning techniques with statistical methods for functional enrichment analysis, the tool enables researchers to uncover hidden patterns in complex biological datasets and gain insights into the underlying molecular mechanisms.

One of the key strengths of LFOmics is its ability to handle large and complex datasets, making it suitable for a wide range of research applications. The use of autoencoders for dimensionality reduction allows the tool to capture the most significant patterns in the data, while GSEA provides a rigorous statistical framework for identifying biologically relevant pathways and gene sets.

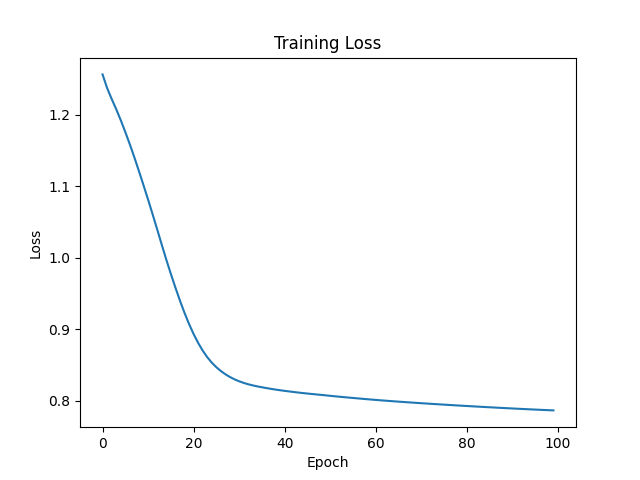
The web-based interface of LFOmics makes it accessible to researchers without requiring specialized computational skills. The tool is designed to be user-friendly, with a straightforward workflow that guides users through the process of data upload, processing, analysis, and results retrieval.

However, there are some limitations to the current implementation of LFOmics. For example, the tool currently supports only the KEGG 2019 Human gene set database for GSEA. Expanding the range of available gene sets would increase the tool's applicability to different research contexts. Additionally, while the autoencoder model used in LFOmics is effective for many types of data, it may not capture all relevant patterns in datasets with highly complex or non-linear relationships. Future versions of the tool could explore the use of more advanced neural network architectures, such as variational autoencoders or generative adversarial networks, to address these limitations.

Overall, LFOmics provides a valuable resource for researchers seeking to integrate and analyze multiomic data. The tool's combination of machine learning and functional enrichment analysis offers a powerful approach to understanding the complexity of biological systems and uncovering the molecular mechanisms driving disease and other biological processes. With further development and expansion, LFOmics has the potential to become an essential tool in the field of multiomic data analysis.

**Example Outputs:**

This example was generated by using a small subset of multiomics data including RNA-seq and single cell RNA-seq data. The bulk RNA-seq data used in this project was sourced from the study titled "ESR1 and p53 Interactome Defines Mechanisms of Therapeutic Response to Tamoxifen Therapy in Luminal Breast Cancer Patients" under the accession number [GSE263089](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE263089). This dataset provided valuable insights into gene expression patterns across different biological conditions. The single-cell RNA-seq (scRNA-seq) data was obtained from the study titled "Proteogenomic integration of single-cell RNA and protein analysis identifies novel tumour-infiltrating lymphocyte phenotypes in breast cancer" under the accession number [GSE199219](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE199219). This dataset allowed for the exploration of gene expression at the single-cell level, contributing to the understanding of cellular heterogeneity. These datasets were instrumental in demonstrating the capabilities of LFOmics in integrating and analyzing complex multiomic data.

****

**A graph of a bar graph

Description automatically generated with medium confidence**