**Institute of Biochemistry and Biophysics, Department of Bioinformatics**

# Project 2

## Predicting Transcript Expression in Mouse Embryonic Stem Cells Using Integrated ChIP-Seq and ATAC-Seq Data

### Background:

This project focuses on analyzing a dataset that integrates chromatin accessibility (ATAC-Seq) and histone modification signals (ChIP-Seq) to predict transcript expression levels in mouse embryonic stem cells (ESCs). The dataset comprises thousands of transcripts, where each row represents a specific transcript ID and each column corresponds to signal values from ChIP-Seq or ATAC-Seq assays around transcription start sites (TSS) or transcription termination sites (TTS). The primary objective is to understand how chromatin features influence gene expression and develop predictive models that utilize these features to estimate transcript expression levels.

### 1. Exploratory Data Analysis (EDA)

The first step in any machine learning project is to explore and understand the dataset. In this task, you will perform exploratory data analysis (EDA) on the datasets and get ready for training machine learning models.

### Tasks:

1. Descriptive Statistics:
   - Calculate basic statistics (mean, median, variance) for each feature and the expression levels.
   - Identify any skewness or trends in the data.
2. Data Visualization:
   - Create histograms and box plots to visualize the distribution of ChIP-Seq and ATAC-Seq mean signals separately.
   - For each histone modification, extract the five features with the highest mean signal values and their corresponding distance from TSS or TTS.

  - Group these features based on their distances from TSS or TTS. Do you observe any patterns? Are these patterns biologically meaningful? Why or why not?

---

**2. Transcript Expression Prediction**

In this task, you will build regression models to predict transcript expression levels based on chromatin features. You will train multiple regression models and evaluate their performance.

**Tasks:**

1. Data Splitting:
   - Split the dataset into training and testing sets. Implement cross-validation (e.g., K-fold cross-validation) to ensure robust evaluation.
   - To avoid data leakage, ensure that all transcripts of each gene are in either the training set or the testing set (not in both).
2. Model Training:
   - Select 4 different machine learning algorithms (e.g., Linear Regression, Ridge Regression, Lasso Regression, Regression Tree, etc.).
   - Train each model on the training data to predict expression values. You may need to train models on log (1+TPM) values for better results.
3. Model Evaluation:
   - Evaluate each model using the following performance metrics:
     - Mean Absolute Error (MAE)
     - Mean Squared Error (MSE)
     - Root Mean Squared Error (RMSE)
     - $R^2$ Score
     - Spearman Correlation
   - Compare the regression models based on these performance metrics. Discuss which model provides the best regression results and why it is most appropriate for this dataset.
4. Model Optimization:
   - Use either Grid Search or Random Search in combination with nested Cross-Validation to tune the hyperparameters of one model of interest.
   - After optimization, reevaluate the model on the testing set and compare its performance with the non-tuned model.

---

### 3. Feature Selection and Importance Analysis

In this task, you will predict transcript expression levels using selected features. Follow a similar approach as in the previous task and report your results.

**Additional Tasks:**

1. Feature Correlation Analysis:
   - Compute pairwise correlations between ChIP-Seq/ATAC-Seq features and transcript expression.
   - Visualize the correlation matrix to detect highly correlated features and discuss potential multicollinearity issues.
2. Feature Selection Methods:
   - Apply techniques such as Recursive Feature Elimination (RFE) or Lasso-based feature selection.
   - Analyze which features are selected most frequently and discuss their biological relevance.
   - **Retrain your models using the selected features**.
   - Evaluate the impact of reducing the number of features on model performance.
3. Feature Importance Visualization:
   - Use techniques like SHAP (SHapley Additive exPlanations) or feature importance from tree-based models (e.g., Random Forest) to interpret which features contribute most to the predictions.
   - Visualize and explain the significance of the top-ranked features in the context of gene regulation and chromatin dynamics.

---

**Keynotes:**

The required data can be downloaded from here.

You are required to create a comprehensive Jupyter Notebook report that includes your code, results, and answers to the provided questions. The report should also contain brief explanations of your code and any relevant analysis. Send your answers as a zip file via email to Mahdi.Anvari7@ut.ac.ir by **December 24, 2024**. Make sure to include your name and student number in the email subject line.

You can research your answers online or using textbooks and you can discuss your solutions with your classmates, but you need to disclose all the resources that you used in your report. If you use tools like ChatGPT, include your prompt and the answer in your report.

Good luck