Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Machine Learning for Genomics
Fall Semester 2024

Project 1: Prediction of Gene Expression from Chromatin Landscape

Assigned on: **5:00pm on 09.10.2024** Due by: **12:00pm on 30.10.2024**

Overview

With this exercise sheet, we will present the first practical project of the course. The topic of this project is to use chromatin information to predict gene expression. The goal is to build a simple (or more sophisticated) machine learning model that uses different available chromatin tracks to accurately predict expression of genes in an unseen cell line. The closer your predictions are to the ground truth, the better your model captured the biology of the cell machinery that determines the level of gene expression.

In particular, in this exercise you will:

a) explore various chromatin (a.k.a. epigenetic) datasets, such as histone marks and chromatin accessibility,

b) choose and implement a prediction model that takes various epigenetic datasets as inputs and predicts gene expression level,

c) evaluate the performance of your method.

We have split the work into several work packages. There is no need for you to follow this division of work exactly, but it should serve as a guide to which steps are deemed to be important.

Data

During the lectures, you learned about such cellular processes as transcription factor binding and histone modification, and their roles in regulating transcription. These processes are closely linked to the "openness" of the DNA, also known as chromatin accessibility. Overall, transcription factor binding, histone modifications, and chromatin accessibility all play a role in up-regulation or down-regulation of nearby genes. For this exercise, you will focus on using genomic profiles of histone modifications and/or chromatin accessibility to accurately predict gene expression. To test the robustness of your prediction model, you will be working with three different cell lines – X_1, X_2, X_3. **Please note that all files correspond to the hg 38/GRCh 38 version of the human genome.** In the context of this project, you will have available the following datasets for the three cell lines:

- H3K27me3 (ChIP-seq)
- H3K4me1 (ChIP-seq)
- H3K4me3 (ChIP-seq)

- H3K27ac (ChIP-seq)
- H3K36me3 (ChIP-seq)
- H3K9me3 (ChIP-seq)
- Chromatin Accessibility (DNase-seq)
- Gene Expression (obtained using the CAGE experiment)
- Gene Information (obtained using CAGE and RefSeq)

Let chr denote the list of all non-sex chromosomes (i.e., autosomes). We define A, B and C as disjoint subsets of chr such that $A + B + C = chr$. Specifically, we define these sets as:

- $A = \{chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr15, chr16, chr17, chr18, chr20, chr21, chr22\}$
- $B = \{chr14, chr19\}$
- $C = \{chr1\}$

You are expected to train your predictive model on set A of cell lines X_1 and X_2 and validate on set B of cell lines X_1 and X_2; you are also free to use different splits of $A + B$ for training and validation. Finally, your model will be tested on the gene expression values of C of X_3 which are unknown to you.

To make the feature and target values comparable across different cell lines, the datasets have been already normalized using the CHIPIN method (outside of the scope of this exercise).

Gene expression files will consist of gene names along with the expression values of those genes. The locations of gene Transcription Start Sites (TSS)s and gene body along with the strand are given in the gene information file. The feature datasets (ChIP-seq and DNase-seq) will represent signals along the genome. For each feature, you will be provided with a BED narrowpeak (`.bed`) file and a BigWig (`.bw` or `.bigwig`) file. You are free to use either or both of them. Please note that BED narrowpeak files follow the headers of narrowpeak files. Further details about these file formats can be found at https://genome.ucsc.edu/FAQ/FAQformat.html. You are encouraged to visualize the data using *Integrative Genomics Viewer (IGV)* or *USCS Genome Browser*.

Work Package 1.1 – Modelling Choices & Data Pre-processing

Using the input data defined above, you will now be required to make some important choices:

a) Which histone marks should I use?

b) Should I use the chromatin accessibility data?

c) Should I use `.bed` files or `.bw` files or a combination of both?

d) Should I use the signal coming only from specific genomic regions?

e) Should the signal be further modified/normalized?

f) Should I use DNA sequence (reference genome) as an additional input? (Please note that this data are not provided to you but can easily be downloaded using `fastaFromBed` of the *bedtools* Python package)

The goal of this work package is to decide which pre-processing methods are relevant to answer the above questions and ensure that you can achieve the best result in the next work package. It is also crucial to determine which part of the genomic region can be responsible to predict the expression of a particular gene. To provide a perspective, it is known that DNA segments representing distant regulatory elements (a.k.a. distant enhancers) that are 1 million base pairs away from the gene TSS can nevertheless contribute to the regulation of gene expression. However, the probability of such events is low; the likelihood for a DNA segment to have an impact on regulation of a gene decreases with the distance between the segment and the gene (47% information is encoded in the 40 kb region around the TSS and 84% information is encoded in the 200 kb.). **We expect you to systematically determine the features that will be the best input to the model.** You can try to visualize the datasets in IGV or UCSC Genome Browser to help answer some of the above questions.

Do not hesitate to do a web search and read papers of the ENCODE consortium on this topic.

Work Package 1.2 – Model Building

You are now ready to build a model that will accurately predict gene expression of any given gene using chromatin information. But which model should you use? Should you build a linear model or a non-linear model (like a Decision Tree) or try both? Should I try to use a more complicated model? To answer these questions, you are encouraged to do a literature search on this topic.

As mentioned before, the model is to be trained and validated using the chromosome sets A and B of cell lines X_1 and X_2. **In this work package, you are expected to decide and build the best model for this task.**

Work Package 1.3 – Evaluation Metrics

Since the end goal is to be able to identify highly and lowly expressed genes, we will be using **Spearman's correlation** to evaluate your model's performance on the holdout chromosome C of cell line X_3. Spearman's correlation (also known as Spearman's ρ) evaluates the monotonic relationship between the ranked values and is independent of the raw values. If we have n observations, let d_i denote difference between the two ranks of observation i, then Spearman's correlation between the two ranks is:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \tag{1}$$

In this exercise, the Spearman's correlation will capture the correlation between the model-based gene rank list (based on predicted expression) and CAGE-based gene rank list (based on experimental expression).

You can also make use of other metrics like Pearson's correlation and R-squared. However, these will not be used for grading.

Pearson's correlation (also known as Pearson's r) evaluates the linear relationship between the raw values. If a and b are two sets of raw values with μ_a and μ_b as mean raw values respectively, then Pearson's correlation between a and b is:

$$r = \frac{\sum_i (a_i - \mu_a)(b_i - \mu_b)}{\sqrt{\sum_i (a_i - \mu_a)^2 \sum_i (b_i - \mu_b)^2}} \tag{2}$$

R-squared (also known as R^2) is defined as the percentage of variance in the true values that can be explained by the model. It determines the goodness of fit of your regression model. *Please note that this interpretation is only true for linear models.* Assume you have n observations and f_i and y_i are predicted and true values respectively for observation i. The mean true value is denoted by μ. Then, R-squared is:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \mu)^2} \tag{3}$$

Submission

For the implementation of your project, we provide a Jupyter notebook with an implementation of the evaluation metrics. We recommend using Python, however, students more comfortable with R can use an R kernel in a Jupyter notebook. In this case, we suggest following the structure provided in the original notebook and adapting the code where needed. We also provide you with a base conda environment to help you start working. You are free to use any libraries. Please submit your final conda environment along with the code. If there is any additional knowledge needed for us to conduct the grading, please add a README file containing the a brief description of your scripts.

Each submission needs to be accompanied by a brief summary of the group's solution and briefly outline the parts you worked on yourself. Everybody should be submitting this independently, but exchanging ideas within and across groups is encouraged. For discussion across groups, please use Moodle.

Grading

In order to pass, you need to meet the baseline as defined below:

- In Work Package 1.3, obtain the Spearman's correlation of 68.5 %.

A bonus of +0.25 on the semester grade will be given to the 10% students who perform the best in terms of Spearman's correlation on the holdout data set.