Machine Learning in Genomics: Containerised tutorials demonstrating best practises, pitfalls, and reproducibility

Sach Nehal

2024-07-08

Contents

About

Applied machine learning utilising vast amounts of data has aided in pattern identification, predictive analytics, and solving complex problems across a multitude of fields. Solving these complex problems within these fields, researchers would find differing answers to the following questions; **what machine learning techniques can we apply to the problem, how do we apply the techniques in the context of this field, and why do we need to apply them in this way?** In any case, applied machine learning requires an interdisciplinary understanding of computing techniques and the field in question.

The aim of this project is to provide you with **a set of reproducible, containerized tutorials that include all necessary data, code, and descriptions to replicate key results, along with demonstrations of common pitfalls, in the field of genomics**. It is designed for users with knowledge of machine learning but little or no background in biology as a process to learn about applying machine learning techniques in genomics.

Part I

Introduction

Chapter 1

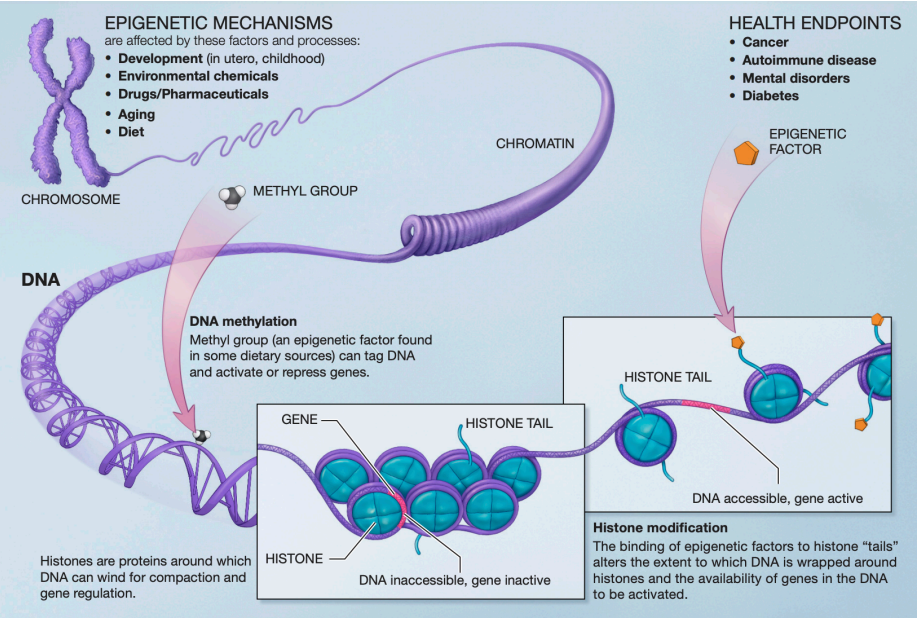
Epigenetic Data

1.1 What is epigenetic data?

As you may already know, all of the cells in your body contain the same DNA. How, then, do we have different cell types in our body? Your DNA contains a script that is able to produce the proteins required for each specific cell in your body. Which proteins, and subsequently which cells are made, depends on gene expression, "the way each cell deploys its genome."[1]

Epigenetic data arises from "the study of heritable and stable changes in gene expression that occur through alterations in the chromosome rather than in the DNA sequence."[2]

[1] Ralston and Shaw [2008]
[2] Al-Aboud et al. [2023]



commonfund.nih.gov

The key takeaways from this image: -Genetic structure of DNA, chromatin, chromosomes -Understanding histones and DNA accessibility which has implications on gene expression.

Some epigenetic alterations include:

1. **DNA Methylation**: Addition of methyl groups to DNA, affecting gene expression regulation[3].
2. **Histone Modifications**: Chemical changes to histone proteins that DNA wraps around. These changes influence chromatin structure and gene accessibility.[4]
3. **Chromatin Accessibility**: Regions of open chromatin that are accessible to transcription factors (special types of proteins that bind to DNA sequences and regulate gene expression) further dictate which regions of DNA can be expressed[5].

1.1.1 Key Epigenetic Techniques:

1. ***ATAC-Seq*** (Assay for Transposase-Accessible Chromatin with Sequencing): `o Measures chromatin accessibility to identify open regions of the genome where transcription factors can bind. o Output: Peaks indicating accessible chromatin regions.`

[3] Al-Aboud et al. [2023]
[4] T. [2007]
[5] Kappelmann-Fenzl [2021]

2. ***ChIP-Seq*** (Chromatin Immunoprecipitation Sequencing): o Used to identify DNA regions bound by specific proteins (e.g., transcription factors, histones with specific modifications). o Output: Peaks indicating binding sites or modification locations.

1.2 What does epigenetic data look like?

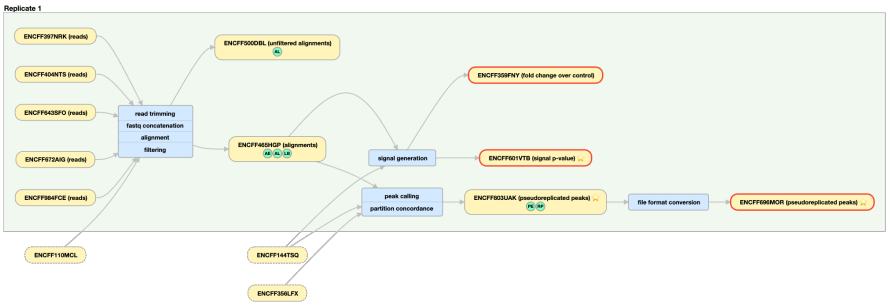
Epigenetic data can be represented in various forms, depending on the type of modification being studied and the methods used to gather the data. **ATAC-Seq** and **ChIP-Seq** are the common methods I will focus on, but there are others that may produce different forms of data.

1.2.1 Representing epigenetic data

1. ***Raw Sequence Reads:*** o These are the basic output of sequencing experiments, such as those from ChIP-Seq or ATAC-Seq. o Reads are processed and aligned to a reference genome before undergoing peak calling.
2. ***Peak Calling:*** o A method used to identify regions in the genome where there is significant enrichment of sequencing reads. This indicates the presence of DNA-protein interactions (e.g., transcription factor binding sites) or accessible chromatin regions. o Peaks represent areas where epigenetic marks or chromatin accessibility are concentrated.

Representing Peaks: - ***P-value or Fold-change:*** P-value: Indicates the statistical significance of the peak, helping to distinguish true peaks from background noise. Fold-change: Represents the difference in read density between treated and control samples, indicating the strength of the signal. - ***Types of Peaks:*** Categorical Peaks: Simple yes/no indication of a peak's presence. Continuous Peaks: More nuanced representation that includes the intensity or enrichment level of the peak, often visualized as a signal track.

Example Data Pipeline



encodeproject.org

EXPLANATION OF PIPELINE + WHAT DATA WE NORMALLY USE IN ML

1.2.2 Example Data Representations:

The following is an example of what chIP-Seq data looks like using UCSC's Genome Browser. The experiment data comes from the encodeproject.org. EXPLANATION OF EXPERIMENT + EXPLAIN GRAPH

UCSC Genome Browser

encodeproject.org

1.2.3 Transformations to stop extreme p-values

Arcsinh-transformation

1.2.4 Epigenetic Data and Gene Expression:

While epigenetic data provides crucial insights into gene regulation, it is not the same as direct measurements of gene expression (such as RNA-Seq data). Epigenetic modifications can influence gene expression, but they do so by altering the chromatin state and regulatory landscape rather than directly measuring mRNA levels.

1.3 Sources of epigenetic data

Blueprint Roadmap Encode (Main Focus)

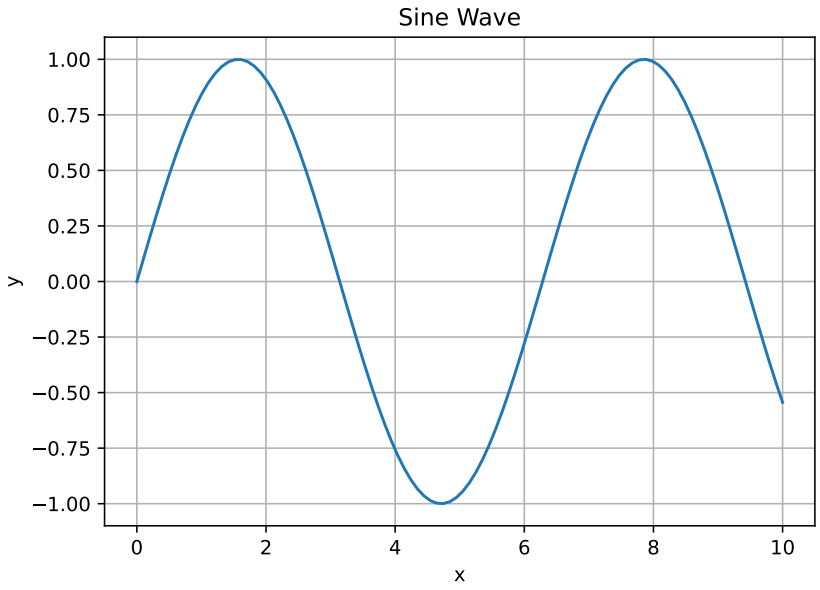
Handling bigWig files Data loaders and pre-processing Dealing with missing data (oversampling, undersampling, weighting)

##testpython code

```
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0, 10, 100)
y = np.sin(x)

plt.plot(x, y)
plt.xlabel('x')
plt.ylabel('y')
plt.title('Sine Wave')
plt.grid(True)
plt.savefig('images/sine_wave.png')
plt.show()
```



Part II

Training models with DNA input

Chapter 2

Loss functions, and peak metrics

Chapter 3

Base pair averaging

Chapter 4

Training tricks

Chapter 5

Choosing which genomic regions to train on

Chapter 6

Effect of differences in sequencing depths

Chapter 7

Reproducibility of machine learning models

7.1 Seeding

7.2 Dashboarding

Chapter 8

Testing

Part III

Software libraries for model building

Chapter 9

gReLU

Chapter 10

Kipoi

Chapter 11

Weights and Biases

Part IV

ML pitfalls in genomics

Chapter 12

Pitfalls overview

Part V

Model interpretability

Chapter 13

Creating and visualising a simple model

Chapter 14

TF mo-Disco

Part VI

Using existing models

Chapter 15

Using the gReLU model zoo

Chapter 16

Fine tuning of Enformer

Part VII

Predicting in novel cell types

Chapter 17

Incorporating ATAC-seq info

Chapter 18

Use of cell type averages

Part VIII

More complex models

Chapter 19

Training multi-headed models

Chapter 20

Training siamese twin models

Bibliography

Nora M. Al-Aboud, Connor Tupper, and Ishwarlal Jialal. *Genetics, Epigenetic Mechanism*. National Library of Medicine, 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK532999/#article-22137.r1.

Melanie Kappelmann-Fenzl. *Design and Analysis of Epigenetics and ChIP-Sequencing Data*. Springer, 2021. URL https://doi.org/10.1007/978-3-030-62490-3_12. ISBN 978-3-030-62490-3.

Amy Ralston and Kenna Shaw. *Gene Expression Regulates Cell Differentiation*. Nature Education, 2008. URL https://www.nature.com/scitable/topicpage/gene-expression-regulates-cell-differentiation-931/#:~:text=All%20of%20the%20cells%20within,each%20cell%20deploys%20its%20genome. Nature Education 1(1):127.

Kouzarides T. *Chromatin modifications and their function*. National Library of Medicine, 2007. URL https://doi.org/10.1016/j.cell.2007.02.005. PMID: 17320507.