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

Sach Nehal

2024-07-10

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

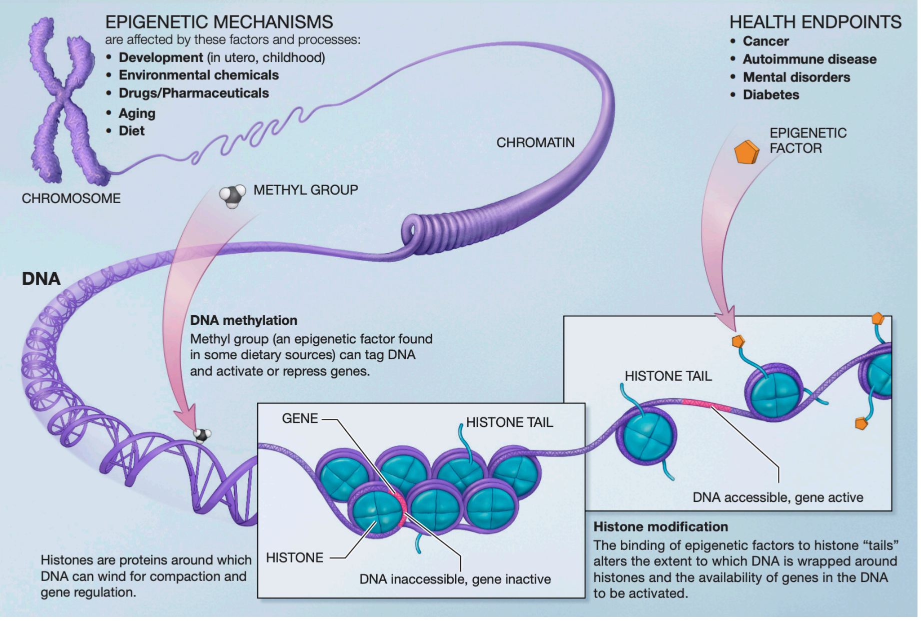
1.1 What is epigenetic data?

As you may already know, typically 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 and regulation, i.e. "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]


EPIGENETIC MECHANISMS
are affected by these factors and processes:
• Development (in utero, childhood)
• Environmental chemicals
• Drugs/Pharmaceuticals
• Aging
• Diet
HEALTH ENDPOINTS
• Cancer
• Autoimmune disease
• Mental disorders
• Diabetes
CHROMOSOME
CHROMATIN
METHYL GROUP
EPIGENETIC FACTOR
DNA
DNA methylation
Methyl group (an epigenetic factor found in some dietary sources) can tag DNA and activate or repress genes.
HISTONE TAIL
GENE
HISTONE TAIL
DNA accessible, gene active
HISTONE
DNA inaccessible, gene inactive
Histones are proteins around which DNA can wind for compaction and gene regulation.
Histone modification
The binding of epigenetic factors to histone "tails" alters the extent to which DNA is wrapped around histones and the availability of genes in the DNA to be activated.

commonfund.nih.gov

The image above shows quite simply the basics of genetic structures. Several more complex processes are involved during cell replication such as DNA transcription and translation in order to make proteins. A key takeaway in coming closer to understanding gene expression is that **Chromatin** is a complex structure made up of DNA wound around histone proteins, with some segments of DNA being accessible/inaccessible to further processes. **Euchromatin** refers to the accessible state, while **Heterochromatin** refers to a chromatin state in which DNA cannot be transcribed (inaccessible).[3] There are many different epigenetic modifications that affect chromatin accessibility.

Some common epigenetic modifications include:

1. **DNA Methylation**: Addition of methyl groups to DNA, affecting gene expression regulation[4].
2. **Histone Modifications**: Chemical changes to histone proteins that DNA wraps around, including acetylation, methylation, or phosphorylation. These changes influence chromatin structure and gene accessibility.[5]
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[6].

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

In studying gene expression and epigenetic modifications, we can more closely understand biological mechanisms that regulate development, disease, and how cells respond to epigenetic factors.

1.1.1 Common Epigenetic Techniques:

1. ***ATAC-Seq*** (Assay for Transposase-Accessible Chromatin with Sequencing): oMeasures chromatin accessibility to identify open regions of the genome where transcription factors can bind. oOutput: Peaks indicating accessible chromatin regions.
2. ***ChIP-Seq*** (Chromatin Immunoprecipitation Sequencing): oUsed to identify DNA regions bound by specific proteins (e.g., transcription factors, histones with specific modifications). oOutput: Peaks indicating binding sites or modification locations.

1.2 What does epigenetic data look like?

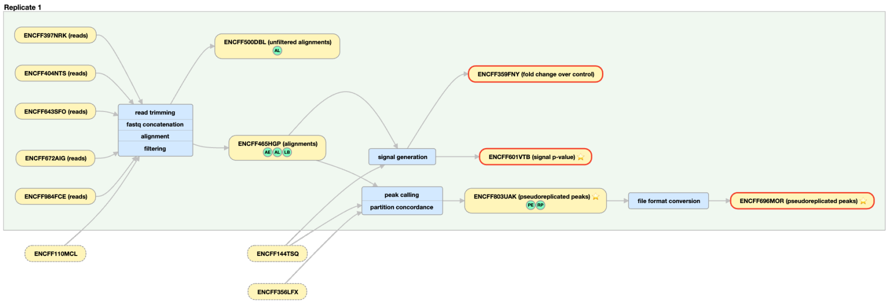
Epigenetic data can be represented in various forms, depending on the type of modification being studied and the methods used to gather the data. **ATAC-Seq** and **ChIP-Seq** are the common methods I will focus on, but there are others that may produce different forms of data, such as WGS (whole-genome sequencing) which produces nucleotide sequencing data, or Bisulfite conversion of DNA producing data on methylation levels across the genome.

1.2.1 Representing epigenetic data

1. ***Raw Sequence Reads:*** oThese are the basic output of sequencing experiments, such as those from ChIP-Seq or ATAC-Seq. oReads are processed and aligned to a reference genome before undergoing peak calling.
2. ***Peak Calling:*** oA 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. oPeaks represent areas where epigenetic marks or chromatin accessibility are concentrated.

Representing Peaks: o ***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. o ***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. Thresholded/Pseudoreplicated Peaks: Usually categorical, these peaks are of high confidence regions from multiple replicates (experiments) or pseudoreplicates (artificial data splits), to ensure reliability and reproducibility.

Example Data Pipeline



encodeproject.org

This example data pipeline originates from a ChIP-seq experiment targeting the H3K36me3 histone modification in brain tissue. The aim of the experiement is to map the locations where the H3K36me3 histone modification is present along the genome. The yellow bubbles represent downloadable data sets of different types, while the blue boxes represent step types (e.g. peak calling). In the left column are multiple data sets of raw sequence reads, which then undergo data quality steps before being aligned (first blue box) to the reference human genome GRCh38 (denoted by ENCFF110MCL below the reads). The next steps include Peak calling (categorical peaks) and signal generation (continuous peaks) to produce the data we normally use in our machine learning models. This data pipeline process aids in normalisation, noise reduction, and dimensionality reduction of the data.

1.2.2 Example Data Representations:

The following is an example of what this same chIP-Seq data targeting the H3K36me3 histone modification in brain tissue looks like using UCSC's Genome Browser. The pseudoreplicated peaks represent categorically, the significant locations along the genome where the H3K36me3 histone modification is present.

UCSC Genome Browser

encodeproject.org

1.2.3 Transformations to stop extreme p-values

Arcsinh-transformation

1.3 Sources of epigenetic data

Blueprint Roadmap Encode (Main Focus)

1.4 Handling bigWig files

1.5 Data loaders and pre-processing

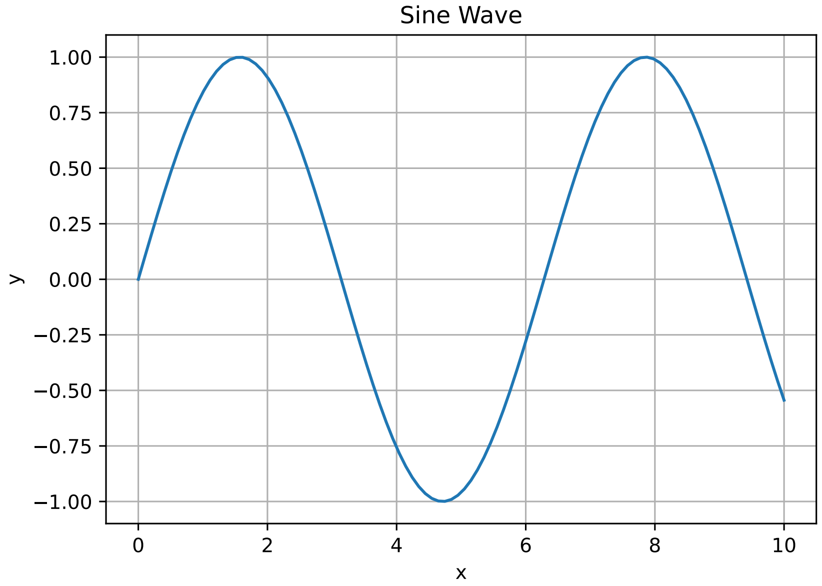
1.6 Dealing with missing data (oversampling, undersampling, weighting)

##testpython code

```python
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()
```


Sine Wave
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
y
0
2
4
6
8
10
x

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

12.1 Distributional differences

12.2 Dependent examples

12.3 Confounding

12.4 Leaky pre-processing

12.5 Unbalanced classes

12.6 Balancing the proportion of peaks / no-peaks in validation sets

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.

Zainab Shahid, Brittany Simpson, Kathleen H. Miao, and Gurdeep Singh. *Genetics, Histone Code.* StatPearls Publishing LLC, 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK538477/. PMID: 30860712.

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.