Big Bio-Data Analysis (Artificial Intelligence and Machine Learning)

12 May 2023

# Machine Learning & BioInformatics Case Studies

By

**Richard Sserunjogi**
Department of Computer Science,
Makerere University, Uganda
sserurich@gmail.com

AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS & DATA INTENSIVE SCIENCE

# House Keeping

**Resources**

- Introduction to Machine Learning and Bioinformatics

**https://github.com/atwine/big-bio-data-class-2023/tree/main/Resources**

**TODO: Expected Outcome**

- Read about the selected topics from the book
- Prepare personal notes from your reading

# Introduction

- Informative yet accessible ways in which the two increasingly intertwined areas of **bioinformatics** and **machine learning** borrow strength or motivation from each other.
- **Bioinformatics** is an emerging field of science growing from an application of mathematics, statistics and information science to the study and analysis of very large biological datasets with the help of powerful computers.

- Among the natural sciences, **biology** is the one that studies highly complex systems known as **living organisms**.
- Before the era of modern technology, it was primarily a **descriptive science** that involved careful observation and detailed documentation of various aspects of a living being (e.g., its appearance, behavior, interaction with the surrounding environment, etc.).
- These led to a reasonably **accurate classification of all visible forms of life** under the sun (the binomial nomenclature by Carolas Linnaeus)

- And to a theory of how the lifeforms we see all around us came into being over billions of years (**the theory of evolution by Charles Darwin**).
- However, the **technology available in those days was not good enough** for probing the internal mechanisms that made life possible and sustained it i.e.the complex biochemistry of metabolism, growth and self-replication.
- **The biological datasets** that were available for statistical analysis in those days (including clinical and epidemiological data) were **relatively small and manageable**, and standard classical procedures (such as two sample t-tests, ANOVA and linear regression) were adequate to handle them.

- But starting from the **middle of the twentieth century**, key breakthroughs in the **biomedical sciences and rapid technological development changed everything**.
- Not only did they enable us to probe the inner sanctum of a living organism at the molecular and genetic levels, but also brought a sea-change in our concept of medicine.
- A series of new discoveries gave us unprecedented insight into the modus operandi of living systems (the most famous one being
- the Franklin-Watson-Creek double helix model for DNA) and the ultimate goal of medical scientists **became personalized medicine**.

- New experiments powered by advanced technology were now generating enormous amounts of data on various features of life and increasingly efficient computing machines were making it possible to create and query gigantic databases.
-

# Classification Techniques

- **Applications of Classification Techniques to Bioinformatics  Problems**
  - Most active area has been the class prediction problem (e.g.different stage of cancer of patients) using primarily gene but more recently protein microarray data
  - Peptide and protein identification in mass spectrometry

# Connections between Machine Learning and Bioinformatics

Direct applications of standard machine learning algorithms or specializations of them to particular contexts.
Focus on three problem areas:

- DNA and amino acid sequence analysis,
- gene expression analysis
- network inference.

# Machine Learning in Structural Biology: Interpreting 3D Protein Images

# Statistical Methods for Classifying Mass Spectrometry Database Search Results

# Modeling and Analysis of Quantitative Proteomics Data Obtained from iTRAQ Experiments

# References

Introduction to Machine Learning and Bioinformatics

# Thank you!

If you have any questions feel free to email me:

sserurich@gmail.com