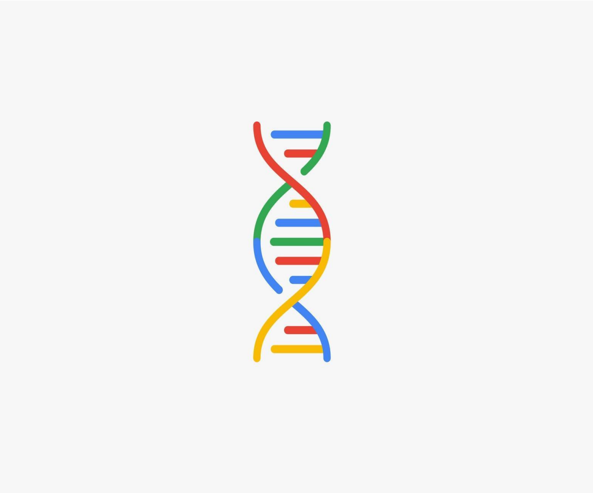


Machine Learning for DNA classification and genomics

Group members:

1803	Pratiksha
1804	Sujata
1816	Divya
1817	Hanumant
1820	Kajal

Note: Group members have written down their roll numbers on the slide they have presented.

Contents

1.Introduction to genomics using machine Learning

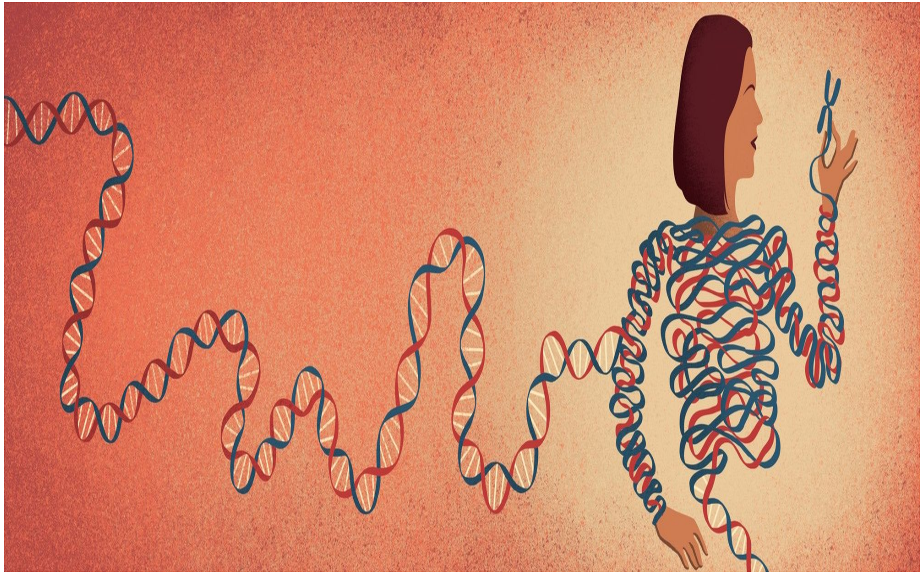
2. Genomics + DNA classification for ML

3. Application of machine learning in genomics

4. Biopython

5. Working with DNA Sequence Data

Introduction to Genomics using Machine Learning



Genome?

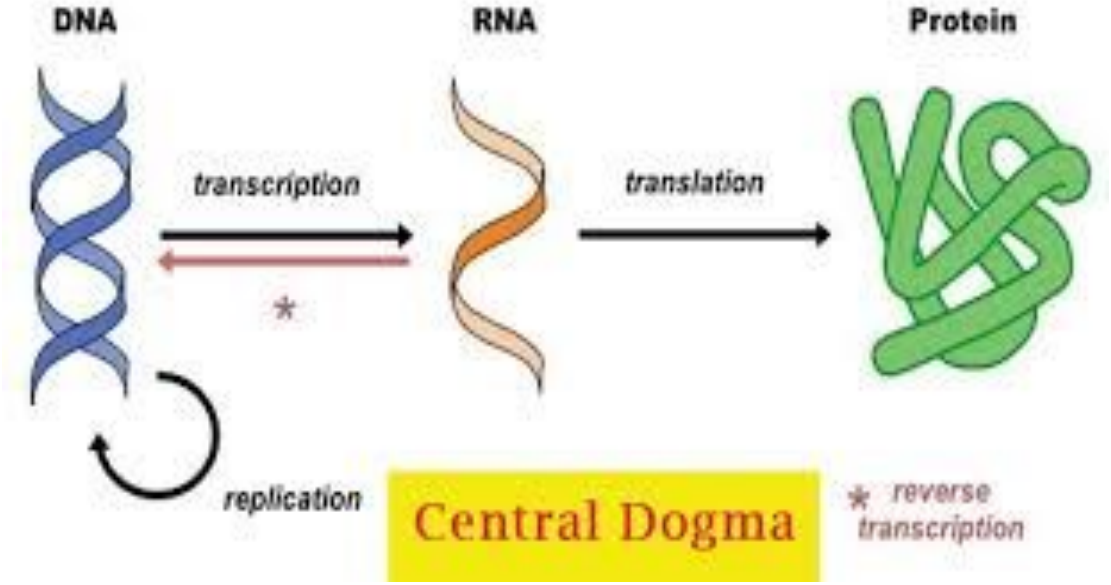
-set of genetic information in an organism

Gene?

-sequence of DNA that decodes the protien

1820

Central Dogma of Molecular Biology



1820

2. Genomics :

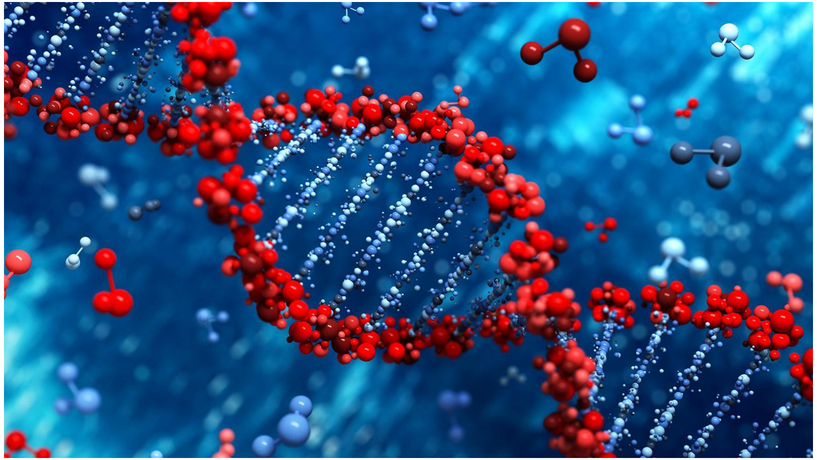
Genomics is an interdisciplinary field of biology that concentrates on studying the structure, function, mapping, and editing of genomes. A genome is a complete set of DNA of an organism; it includes all of the genes.

Genomics is the sub discipline of genetics devoted to the

- Mapping
- Sequencing
- Functional analysis of genomics

We can divide genomics into several subsets:

1. Regulatory genomics
2. Structural genomics, and
3. Functional genomics.



Applications of Genomics:

1. Identify comparison for new nucleic acid sequences.
2. Analysis of gene expression profile.
3. Database of model organism.
4. Hunting for disease-related genes.
5. Screening of poisonous side effect genes.

Today, genomics is a powerful field for innovation encompassing technologies such as deep learning, computer vision, and natural language processing.

DNA Classification for ML

Used Google Colab,

Github link:
https://github.com/DivyaGazinkar/MCA--2021-machine-learning-

3. Applications Of Machine Learning in Genomics :

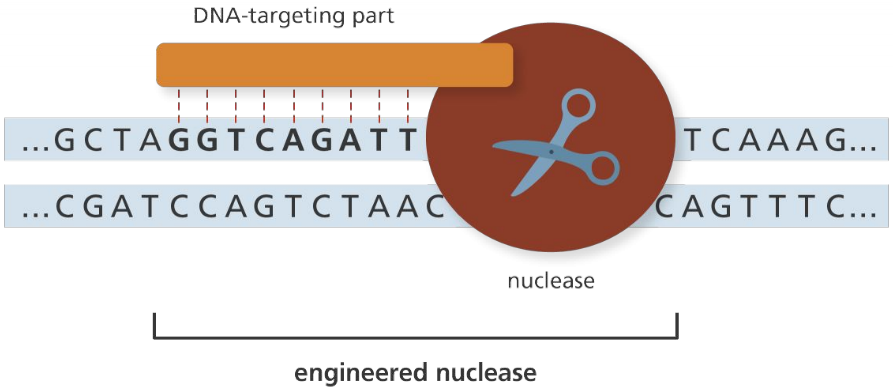
1) Genome Sequencing :

Area where machine learning is causing disruptions is **genome sequencing**, a recent field of interest in medical diagnostics. It includes modern DNA sequencing techniques that allow researchers to sequence the entire human genome in one day. The classic sequencing technology required more than a decade for completion when the human genome was sequenced first.

Companies like Deep Genomics are now operating on the market and using machine learning to help researchers interpret genetic variation**.** In particular, development teams design algorithms based on patterns identified in large genetic data sets. These patterns are then translated to computer models that help researchers to interpret how genetic variation affects critical cellular processes like metabolism, cell growth, or DNA repair. Disruption to the normal functioning of these processes can potentially cause diseases like cancer. That’s why using *machine learning in genomics* research is so important.

2) Gene Editing :

Gene editing is defined as a method of making specific alterations to DNA at the cellular or organism level. CRISPR is a gene editing technology that offers a faster and less expensive way of conducting gene editing. In order to use CRISPR, researchers must first select an appropriate target sequence. This can be a daunting process involving many choices and unpredictable outcomes. Machine learning offers the capability to significantly reduce the time, cost and effort necessary to identify an appropriate target sequence.



How does genome editing work?

- Genome editing uses a type of enzyme called an ‘engineered nuclease’ which cuts the genome in a specific place.
- Engineered nucleases are made up of two parts:
    - A nuclease part that cuts the DNA.
    - A DNA-targeting part that is designed to guide the nuclease to a specific sequence of DNA.
- After cutting the DNA in a specific place, the cell will naturally repair the cut.
- We can manipulate this repair process to make changes (or ‘edits’) to the DNA in that location in the genome.

3) Clinical Workflow :

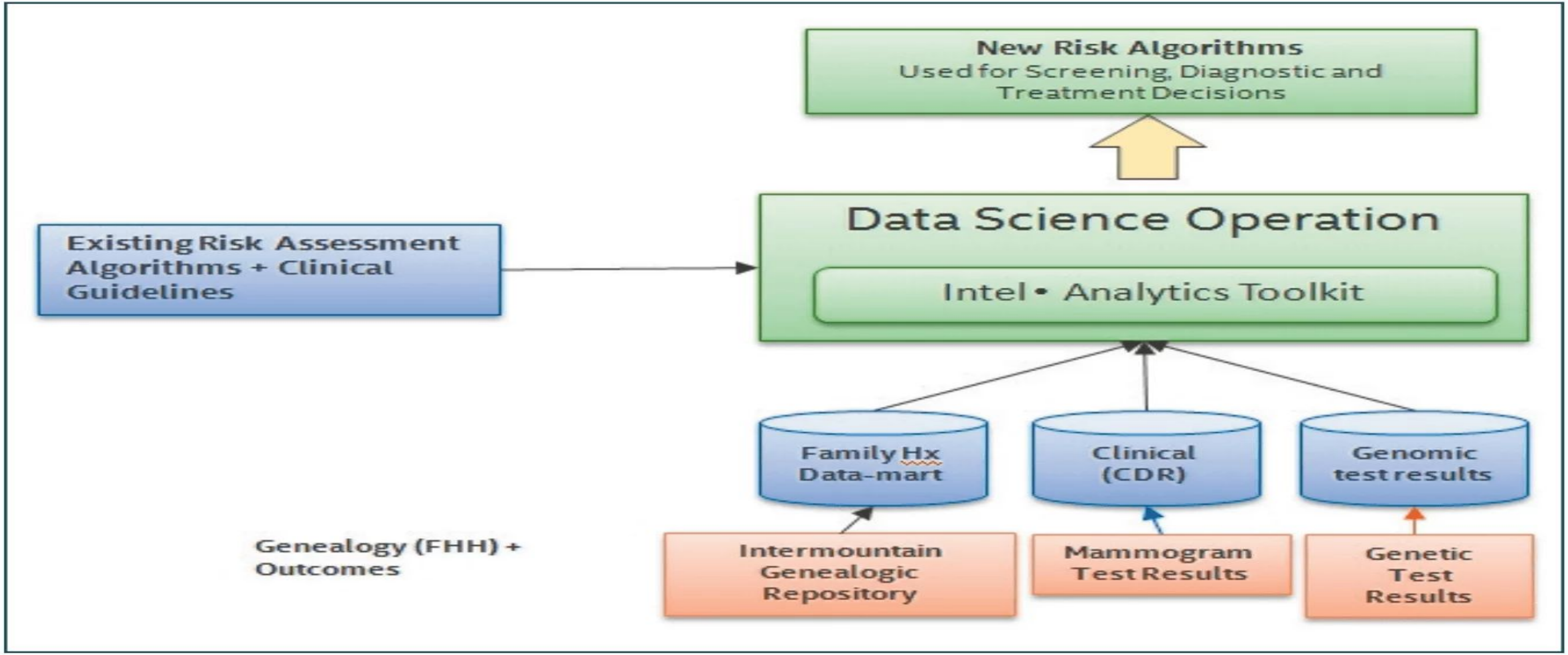
There are often gaps in the patient data available to the different members of a healthcare team serving a patient. This challenge has sparked an interest in using machine learning to improve the efficiency of the clinical workflow process. Intel has designed an Analytics Toolkit which integrates machine learning capabilities into the clinical workflow process.

Intel partnered with the Transformation Lab at the Intermountain Healthcare in Salt Lake City, Utah, to efficiently integrate genomics in the institution’s breast cancer treatments and patient care.

four major components:

1. A centralized database of genomic data that is linked to “clinical and patient data”
2. All clinicians and genetic counselors have access to Electronic Health Records (EHRs)
3. All data from genetic tests are integrated into EHRs
4. Clinical Decision Support tools (CDS) are operational and accessible. Examples of clinical decision support include family health histories, screenings and past clinical data.

A visual representation of the phases – image from Intel


New Risk Algorithms
Used for Screening, Diagnostic and Treatment Decisions
Existing Risk Assessment Algorithms + Clinical Guidelines
Data Science Operation
Intel • Analytics Toolkit
Family Hx Data-mart
Clinical (CDR)
Genomic test results
Genealogy (FHH) + Outcomes
Intermountain Genealogic Repository
Mammogram Test Results
Genetic Test Results

4) Consumer genomics products :

Genetic testing and **consumer genomics** are becoming an increasingly important market for innovation. The anticipated market expansion of these areas is powered by the growing awareness among societies of how genetic tests can be used to determine the likelihood of developing a particular disease. Companies such as 23AndMe or Ancestry.com are becoming household names among consumers.

For example, 23AndMe offers a Genetic Weight based on combining data from 600,000 research participants with the use of machine learning. The report is capable of **delivering insights** into how unique factors such as age and genotype impact one’s weight.

Future Applications of Machine Learning in Genomics

Pharmacy genomics

Pharmacy genomics is an emerging field within precision medicine that examines the role of genomics in the context of an individual response to particular drugs. This area is a quickly developing one but still relatively new.

However, researchers are already experimenting with machine learning techniques. For example machine learning models were applied to **determine a stable dose of a particular drug** in renal transplant patients

In the future, researchers will be using machine learning models to better **understand the individual response** to particular treatments and, as a result, create more personalized treatments.

Genetic screening of newborns

Some experts believe that newborn **genetic screening** might become a standard practice during the next decade. The idea is to collect data birth and then integrate it into the individual EHR (Electronic health Record)profile. Another fact of this trend is making noninvasive screening capabilities available to women during pregnancy. These would be geared at identifying particular diseases such as Down syndrome.

The Newborn Screening Center at the National Taiwan University Hospital implemented machine learning to improve the accuracy its web-based newborn screening system for metabolism defects. Results of the study showed that instances of false positives were reduced “from 21 to 2 for phenylketonuria (PKU), from 30 to 10 for hypermethioninemia, and 209 to 46 for 3-methylcrotonyl-CoA-carboxylase (3-MCC) deficiency.”

Agriculture

Genomics is a discipline relevant to our food production industry. Experts imagine that in the future machine learning will be helping farmers to **improve soil quality and crop yield.** The California-based startup PathoGn combines genomics and machine learning to create diagnostic tools for preventing and predicting diseases and crops. Today the startup is called Trace Genomics and focuses more on soil health.

However, we can easily imagine using genetic data to predict the health of crops so that farmers can better predict and optimize yields. On a large scale, such innovations could lead to significant global improvements in crops and solve world problems such as hunger.

Biopython

Biopython is the largest and most popular bioinformatics package for Python.

It contains a number of different sub-modules for common bioinformatics tasks.

It also contains C code to optimize the complex computation part of the software.

It runs on Windows, Linux, Mac OS X, etc.

Features

Biopython is portable, clear and has easy to learn syntax.

- Interpreted, interactive and object oriented.
- Supports FASTA, PDB, GenBank, Blast, SCOP, PubMed/Medline, ExPASy-related formats.
- Option to deal with sequence formats.
- Tools to manage protein structures.
- BioSQL − Standard set of SQL tables for storing sequences plus features and annotations.
- Access to online services and database, including NCBI services (Blast, Entrez, PubMed) and ExPASY services (SwissProt, Prosite).
- Access to local services, including Blast, Clustalw, EMBOSS.

Advantages

Biopython requires very less code

- Provides microarray data type used in clustering.
- Reads and writes Tree-View type files.
- Supports structure data used for PDB parsing, representation and analysis.
- Supports journal data used in Medline applications.
- Supports BioSQL database, which is widely used standard database amongst all bioinformatics projects.
- Supports parser development by providing modules to parse a bioinformatics file into a *format specific record object or a generic class of sequence plus features.*
- Clear documentation based on cookbook-style.

5.Working with DNA Sequence Data

Using Google colab :

Classifying Human , Chimpanzee and Dog DNA and checking how models performs with respect each DNA

Data Source : All data can be downloaded from kaggle

1817

References:

For Genomics + DNA classification for ML:

- https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/what-genomics#:~:text=Genomics%20is%20the%20study%20of,structure%20and%20function%20of%20genomes.
- https://www.youtube.com/watch?v=eg8DJYwdMyg
- https://codete.com/blog/machine-learning-genomics/
- https://scikit-learn.org/stable/supervised_learning.html#supervised-learning