DNA SEQUENCING

DNA Sequencing With Machine Learning

Bioinformatics

Imeh Akpan
s183179

What is DNA sequencing?

- Sequencing DNA means determining the order of the four chemical building blocks - called "bases" - that make up the DNA molecule.
- The sequence tells scientists the kind of genetic information that is carried in a particular DNA segment.
- For example, scientists can use sequence information to determine which stretches of DNA contain genes and which stretches carry regulatory instructions, turning genes on or off.
- In addition, and importantly, sequence data can highlight changes in a gene that may cause disease.

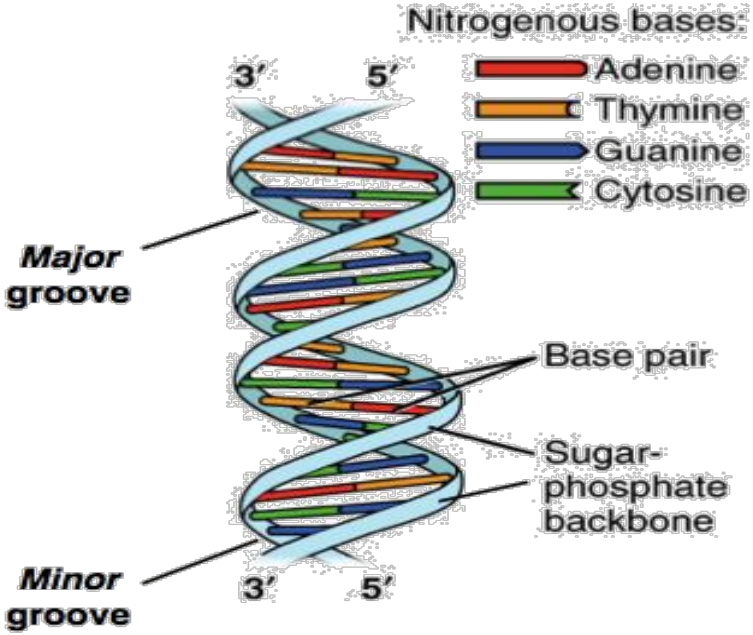
DNA BASES

- In the DNA double helix, the four chemical bases always bond with the same partner to form "base pairs."

- Adenine (A) always pairs with thymine (T);
- cytosine (C) always pairs with guanine (G).

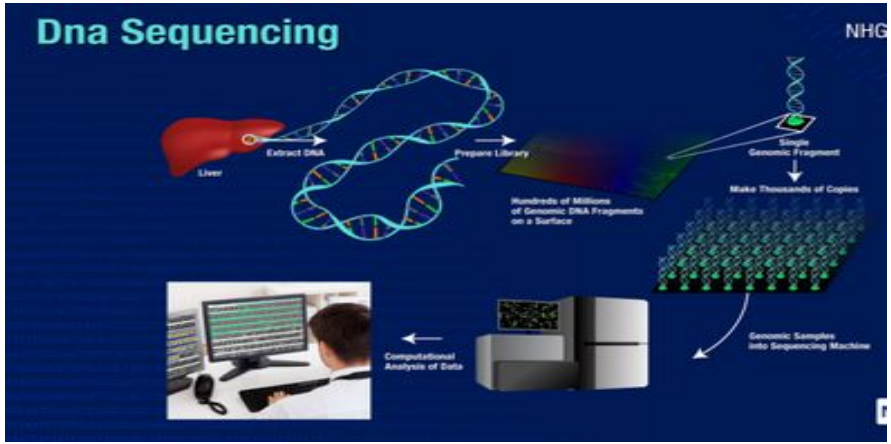
- This pairing is the basis for the mechanism by which DNA molecules are copied when cells divide, and the pairing also underlies the methods by which most DNA sequencing experiments are done.

DNA BASES


Nitrogenous bases:
Adenine
Thymine
Guanine
Cytosine
3′
5′
Major groove
Base pair
Sugar-phosphate backbone
Minor groove
3′
5′

The Human Genome Technology

The human genome contains about 3 billion base pairs that spell out the instructions for making and maintaining a human being. Since the completion of the Human Genome Project, technological improvements and automation have increased speed and lowered costs to the point where individual genes can be sequenced routinely, and some labs can sequence well over 100,000 billion bases per year, and an entire genome can be sequenced for just a few thousand dollars.



The Human Genome Technology

- One new sequencing technology involves watching DNA polymerase molecules as they copy DNA - the same molecules that make new copies of DNA in our cells - with a very fast movie camera and microscope, and incorporating different colors of bright dyes, one each for the letters A, T, C and G.

- This method provides different and very valuable information than what's provided by the instrument systems that are in most common use.

- Another new technology in development entails the use of nanopores to sequence DNA. Nanopore-based DNA sequencing involves threading single DNA strands through extremely tiny pores in a membrane.

- DNA bases are read one at a time as they squeeze through the nanopore. The bases are identified by measuring differences in their effect on ions and electrical current flowing through the pore.

Treating DNA Sequence as A Language

- Treating DNA sequence as a "language", otherwise known as k-mer counting
- A challenge that remains is that none of these above methods results in vectors of uniform length, and that is a requirement for feeding data to a classification or regression algorithm.
- So with the above methods you have to resort to things like truncating sequences or padding with "n" or "0" to get vectors of uniform length.
- DNA and protein sequences can be viewed metaphorically as the language of life. The language encodes instructions as well as function for the molecules that are found in all life forms.

Cont..

- The sequence language analogy continues with the genome as the book, subsequences (genes and gene families) are sentences and chapters,

- k-mers and peptides (motifs) are words, and nucleotide bases and amino acids are the alphabet. Since the analogy seems so apt, it stands to reason that the amazing work done in the natural language processing field should also apply to the natural language of DNA and protein sequences.

- The method I use here is simple and easy. I first take the long biological sequence and break it down into k-mer length overlapping “words”.

- For example, if I use "words" of length 6 (hexamers), “ATGCATGCA” becomes: ‘ATGCAT’, ‘TGCATG’, ‘GCATGC’, ‘CATGCA’. Hence our example sequence is broken down into 4 hexamer words.

Cont..

- Here I am using hexamer “words” but that is arbitrary and word length can be tuned to suit the particular situation.
- The word length and amount of overlap need to be determined empirically for any given application.
- In genomics, we refer to these types of manipulations as "k-mer counting", or counting the occurrences of each possible k-mer sequence.
- There are specialized tools for this, but the Python natural language processing tools make it super easy.

- **DNA Sequencing With Machine Learning**

  In this project, I will apply a classification model that can predict a gene's function based on the DNA sequence of the coding sequence alone.