Lecture 15: Genomics

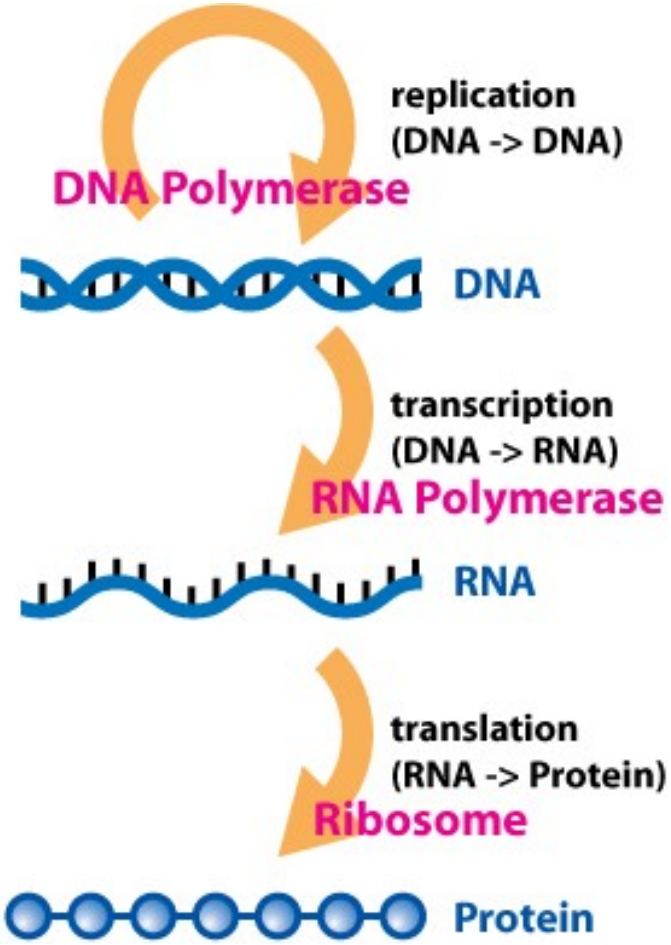
Credit

Slides are partially based on the material in Ramsundar, Bharath; Eastman, Peter; Walters, Patrick; Pande, Vijay. Deep Learning for the Life Sciences, Chapter 6.

Basic Building Blocks

- DNA is a polymer composed of four units: adenine (A), cytosine (C), guanine (G), and thymine (T).
- Proteins are polymers composed of 20 amino acids.
- DNA is responsible for recording the sequence of amino acids for an organism's proteins.
  - Each sequence of three DNA bases (called a codon) correspond to one amino acid.
    - E.g. AAA= Lysine, GCC = Alanine
- RNA is another polymer, very similar to DNA, the intermediate step in creating proteins from DNA code.
  - In place of thymine (T), it has a base called Uracil (U).

Central Dogma



The Actual Picture

- The actual transcription and translation process is much more complex.
  - Unwinding chromosome around Histones, still poorly understood.
  - Splicing: RNA needs to be spliced by removing sections and connecting back the remaining parts (called exons).
    - Many genes have multiple splice variants, i.e. a single stretch of DNA can code for multiple proteins.
  - DNA methylation makes DNA less likely to be transcribed, still poorly understood.

The Actual Picture

- Three well-known types of RNA
  - mRNA, ribosomal RNA, tRNA
- There are many other types of RNA: micro RNA (miRNA), Short Interfering RNA (siRNA). Ribozymes, Riboswitches.
- DNA is more than a string for encoding protein sequences:
  - Encoding RNA sequences, containing transcription factor binding sites, encoding splicing information, encoding histone winding instruction, …

Classical Statistical Approaches

- They struggle to represent the complex, non-linear relations in the genome.
- They are often based on simplifying assumptions:
  - Linear relations between variables, or
  - only modeling a small number of variables

Example 1: Transcription Factor Binding Prediction

Overview: Transcription Factor (TF) Binding

- Transcription Factors (TF) are proteins that influence the probability of nearby genes being transcribed.
- Every TF has a specific DNA sequence called its binding site motif that it binds to.
- Complexities
  - A TF might be able to bind to many similar sequences.
  - Some bases within the motif might be more important, it is often modeled as a position weight matrix.
  - TFs can be influenced by the physical shape of the DNA
    - E.g. how tightly the double helix is twisted,
    - TFs can only bind with motifs in unwound portions of the DNA.
  - TFs can bind to other molecules to form a different complex.

Experiment Setup

- We will use experimental data on a particular TF called JUND.
- The experiment was done to identify every place in the human genome where it binds.
- To keep it manageable, we will only include data from chromosome 22, one of the smallest (still over 50 million bases!).
- The full chromosome split into short segments, each 101 bases long.
  - Each segment has an indicator label whether or not including a JUND binding site.

Model

- Our goal is to train a model that predicts the indicator label based on the sequence.
- Sequences are coded using the one-hot encoding, where one of the four numbers is set to 1, the rest 0.
- We will use a 1D convolutional model, since we are dealing with 1D data (DNA).
  - A few convolutional layers
  - A few dense layers
  - A cross-entropy loss function

Notebook

- See the notebook on colab

Example 2: Chromatic Accessibility

Chromatin Accessibility

- Previous code can be improved by including optimizing the NN architecture, or by including more information.
- We will include information on chromatin accessibility.
  - Chromatic accessibility refers to how accessible each part of chromosome is to the outside world.
  - When the DNA is tightly wound around the histones, it cannot be accessed by TFs.
  - It depends on cell type, life cycle, environmental factors, …

Experiment

- On a particular type of cell called HepG2.
- If a region is always inaccessible, very unlikely to find JUND bond.
- Each 101-base region is associated with a number that measures accessibility.
- There are therefore two sets of features:
  - Sequence
  - Accessibility value

Example 3: RNA Interference

Overview: RNA Interference

- It was discovered in 1990s and led to a Nobel Prize in 2006.
- A short piece of RNA that is complementary to an mRNA, can bind to that mRNA and silence it.
  - Called short interfering RNA (siRNA)
  - It serves as both a mechanism for gene regulation and a defense against viruses, temporarily “turning off” genes.

Complexities

- Complex setting:
  - Some RNA molecules are more stable than others.
  - Some bind to their complementary sequences more strongly than others.
  - Some fold into shapes that make it harder to bind.
- Therefore, we need a tool for selecting siRNA sequences.

Experiments

- We will use a library of 2,431 siRNA molecules, each 21 bases long.
- Each sequence labeled with a number between 0 and 1, indicating how effective it is at silencing a gene.
- The model takes the sequence and tries to predict the effectiveness.

Code

- See code on Google Colab