# Introduction

# Background

## Biology Background

**DNA**

DNA is the molecule that is the hereditary material in all living cells.

DNA molecules are sequences of Nucleotides (ATCG).

**NUCLEOTIDS –AMINO ACIDS**

ATCG

**GENOME**

A genome is the genetic material of an organism. More practically, it is a long chain (the DNA) of 4molecules of length N, which vary from species to species. This 4 molecules are represented as{A,G,T,C} and are what we called the genetic code.The genome contains the information of cellular agents – proteins – that carry out functions in thecell; e.g., transforming one product into another. Thus, the cellular machinery reads parts of thegenome – genes – and translate this information into proteins

Genome is all of the information stored in the DNA of a specific species.

The complete list of nucleotides that forms the DNA is called the genome sequence.

Bacterial genomes are generally smaller and less variant in size among species when compared with genomes of eukaryotes. Bacterial genomes can range in size from 130 kbp to over 14 Mbp.

The total length of the human genome is over 3Bbp (billion base pairs).

\*\*bp: Genome size is the total amount of DNA contained within one copy of a single complete genome. It is typically measured in terms of mass in picograms or as the total number of nucleotide base pairs, usually in megabases (millions of base pairs, abbreviated Mb or Mbp).

We focus on bacterial genome analysis.

**GENES**

Genome sequences show that there are 500-1200 genes in parasitic bacteria, 1500-7500 genes in free-living bacteria, and 1500-2700 genes in archaea.

The size of the genome of E. coli is in the middle of the range. The common laboratory strain has 4,288 genes, with an average length ~950 bp, and an average separation between genes of 118 bp . But there can be quite significant differences between strains. The known extremes of E. coli are from the smallest strain that has 4.6 Mb with 4249 genes to the largest strain that has 5.5 Mb bp with 5361 genes

We still do not know the functions of all the genes. In most of these genomes, ~60% of the genes can be identified on the basis of homology with known genes in other species.

## Deep Learning-NLP Background

**SUPERVISED LEARNING**

**UNSUPERVISED LEARNING**

**CLASSIFICATION**

**ANALOGY BETWEEN BIO-ENG and NLP**

**E.g.**

Amino-acids (A,C,T,G) = letters

recurrent combination of amino acids (ACT,ATG…) = words

proteins = sentences

genes = text paragraphs composed of sentences

global protein classification tasks( e.g. enzyme class prediction) = text classification tasks (e.g. sentiment analysis)

# Methods

## Problem analysis

### Data

Focus on : prokaryotes

Source of the data:

Format of the data: .fasta (check library <https://pypi.org/project/pyfasta/> )

### Task 1: Train a model to predict the presence or absence of a gene in a given DNA sequence.

We want to train our model to identify portions of the genome with any biological significance and therefore to be able to predict the presence of a gene.

It a supervised classification task.

In prokaryotes etcetc

### Task 2: Train a model to predict the start and end of a gene in a given DNA sequence.

Description and challenges

### Task 3: Train a model to predict the start and end of the protein coding a DNA sequence.

Description and challenges

# Pre-processing

DL algorithms performances improves significantly using big datasets. The larger a dataset is, the more pre-processing it needs.

## Choice of dataset and cleaning

We choose data xx from dataset xx available at xx.

Data can be incomplete (missing features), incompatible (size), inconsistent (wrong format) or noisy (wrong labels, outliers etc).

## Data Sampling

How first task is a classification problem with two classes: coding region vs non-coding region.

We might argue that we need to select a good number of samples from both classes.

In eukaryotes genes are only a very small portion of the DNA sequence.

In the Escherichia coli (and prokaryotes) on the contrary almost all the DNA contains genes:

* Genome size = 4.6 Mb
* Average number of genes = 4250
* Coding-region average length = 950 bp
* Non coding region average lenght = 118 bp

## Word2Vec – CBOW, let’s get the Embeddings

Commonly in NLP “embedding”, is a distributed representation of a word, which retains the meaning and the relationships with other words in the vocabulary.

Similar words are mapped by the model (Word2vec in this case) into vectors that are close to each other in the embedding space.

We want to have a good generalized representation of the DNA string, long list of ACGT, using embedding and NLP models.

For DNA embeddings k-mers are commonly used.

K-mers are the words that compose the proteins language, and we want to translate them in numerical vector, feedable to Neural Network, while capturing their “meaning” and how they relate to each other in the sequence.

The k-mers can reveal hidden patterns in that sequence population. k is a integer and can range from 2 to several dozens, depending on the application.

Implementation Process of Cbow for DNA

* Generate K-mers from DNA sequence, those are the words in our corpus
* Data Preparation — Clean, normalise and tokenise words from the corpus
* Hyperparameters — Learning rate, epochs, window size, embedding size
  + [window\_size]:  context words are words that are neighbouring the target word
  + [n]: This is the dimension of the word embedding and it typically ranges from 100 to 300 depending on your vocabulary size. Dimension size beyond 300 tends to [have diminishing benefit](http://www.aclweb.org/anthology/D14-1162). **Do note that the dimension is also the size of the hidden layer.**
  + [epochs]: This is the number of training epochs. In each epoch, we cycle through all training samples.
  + [learning\_rate]: The learning rate controls the amount of adjustment made to the weights with respect to the loss gradient.
* Generate Training Data — Build vocabulary, one-hot encoding for words, build dictionaries that map id to word and vice versa
* Model Training — Pass encoded words through forward pass, calculate error rate, adjust weights using backpropagation and compute loss
* Inference — Get word vector and find similar words
* Further improvements — Speeding up training time with Skip-gram Negative Sampling (SGNS) and Hierarchical Softmax

## Features reduction (Data compression)

Shall we consider using PCA to reduce the number of features of the embedded vectors?

## Normalization

# Methods

## Find the Gene

## AWD-LSTM Model

## Cross-validation

## SVM

# Conclusions

# References