Week 01

Work done

Spent most of the week studying genetics concepts as these were completely and largely unknown to the researcher. While proteins themselves will be the central object of study, understanding them requires also understanding other parts of the genetic complex.A short summary of the knowledge collected is as follows.

The collected genetic material of an organism is called a genome or genotype. The way it is expressed in an organism as it develops, especially considering external factors, is the phenotype. At a base level, the DNA is a backed up genetic code base with four key elements(nucleotides) in various arrangements. The double spiral splits into halves and unfolds into RNA, which bonds with the free organic molecules in the cell nucleus. Codons, units of three nucleotides, work to program aminoacid compounds. Multiple codons called create a chain of such amino acids which is known as a protein. In an organism, a protein serves multiple purposes, from chemical signals to support structures.

In short, we can create analogies of DNA as a factory spitting out structures based on programmed code. Although the process is likely to be a lot more imprecise and random due to the inherent organic and evolutionary nature of the elements, hopefully this analogy will be useful going forward.

On the other hand, a bit less time has been spent reviewing natural language processing methods, as the researcher is familiar with these methods already.

Word embedding are methods where strings in a text are mapped out to mathematical vectors based on various methods. A common one is the contextual word vector, where each element in the vector measures the frequency of all other common words surrounding the target word. This is known as the bag of words or n-grams(where in is the number of surrounuding word) and relies on methods such Tf-idf to counts the frequency of words in a corpus of text and decide which is the most important one. Other methods include unsupervised learning methodologies. Finally, we have refinement techniques such as stemming(a raw cutting of the word's terminations such as -ind, -ed) and lemmatization(a more fine tuned version of stemming which can instead return the infinitive of a verb).

Putting the two areas together, it seems like proteins, being amino acid chains, can be represented as word vectors, allowing us to use natural language processing methods to classify proteins(classifiers), study interactions and bonds(vector similarities), and produce new protein sequences(LSTM).

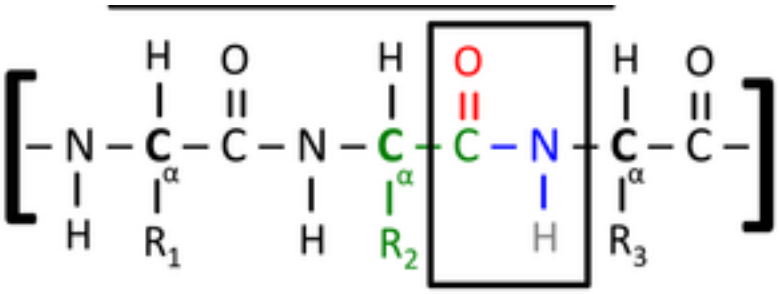
Problems

1. Some conceptual problems regarding genetics remain. For example, the following statement:

It is essential in humans, meaning the body cannot synthesize it; it must be obtained from the diet. Tryptophan is also a precursor to the neurotransmitter serotonin, the hormone melatonin and vitamin B3.[3] It is encoded by the codon UGG.

Raises the following question: If the human body can't synthesize it, why would it be encoded in a codon that should logically and statistically occurs often enough in a DNA of 4 billion nucleotides.

2. From an initiate point of view, word vector embeddings of proteins seem mind boggling. Unlike classical strings, proteins seem to have a 2-d structure, represented a main chain and side chains arising from it. This is before the protein folds in 3 dimensions, but even so, the representations will prove challenging.



To do for next week:

Reading the provided article and LSTMs. Considering the researcher's complete lack of a background in genetics, approaching a scientific paper loaded with specialist terms would not have been possible before the work above.