1 Abstract

Proteins provide the functional building blocks necessary to support life. As we increase our understanding of how they work, new frontiers in medicine are opening up; we are entering an era of personalised medicine which promised to revolutionise healthcare. Yet we have really only touched the top of the iceberg and there is still much to learn.

Computer Scientists have a huge role to play in advancing this area of research. This paper seeks to understand whether, by encoding the protein sequences using common langage encoding algorithms, interesting correlations can be found that can provide direction to firhter research.

*T*O DO - Improve

2 Motivation

Within the context of an MSc dissertaion, the motivation behind this paper is multi-faceted.

1. To perform a valauable piece of research that builds upon what I've learned in my Bioinformatics module and which, in a short space of time, provides some value to the Bioinformatics research community in UCL 2. To put into practise the knowledge I've gained throughout the DSML MSc and to extend my learning to other areas of the course that I could not take (for example Statistical Natural Langauge Processing and Engineering for Data Analysis) 4. As mature student returning to a full-time job in September, to advance my knowledge set in an area that will be useful to my employer and provide further opportunites for career advancement 5. To leave in place a well-thought out and well-documented framework for my Supervisor or anyone else to take forward should he or she wish

3 Introduction

Proteins provide the functional building blocks necessary to support life. As we increase our understanding of how they work, new frontiers in medicine are opening up. We are entering an era of personalised medicine which is built upon our understadning of how proteins work. understanding of how they work and .

What is a protein? A protein is a chain of Amino Acids of varying length. In humans, proteins are assembled in the ribosomes of the cell based upon the sequences of XXX in the RNA. The genetic blueprint documents our understanding of how a set of only 20 Amino Acids are strung together according to sequences of 3 codons XXX. There is some randomness in this association which contributes to the different characterists across individuals.

Although the Amino Acid chains are formed according to the sequence of RNA XXX, as RNA is itself essentially copied from our DNA, there is a direct correlation between the nakeup of proteins and our DNA. This provides a further connection through the evolution of the oragnism - if a sequence of amino acids serve to provide an evolutionary advantage, it is more likely that the host will survive, as will their DNA, RNA and subsequently their proteins.

Through painstaking lab work, molecular biologists have been able to identify short sections of proteins that provide a particular function. This is often achieved by identifying common amino sequences across species that come from a common ancestor. These sections are found to repeat across proteins and across species - and each of these sections have been given a unique entry within a global 'protein family' library and are known as PFAM entries (pfam stabding for protein family). Proteins also provide function through their 3 dimensional structure - wherby a protein 'folds over' itself, exposing on its outside an area where other proteins can attach etc. Some areas do not fold and are categoriesed as 'DISORDER' regions.

Thus, it is possible to capture, for each protein its sequence of PFAM entries and DISORDER regions. In effect these are therefore tectual representations of the functional parts of a protein. One could consider a protein to be a sentence and the PFAM and DISORDER regiosn to be words in this sentence.

With the expolsion in interest in LLM models recently, it is intersting to see whether the encodings used for langauage models can also be used to encode protein sequences and use that encoding to identify correlations amongst proteins themselves and indeed correlations across species that may point towards areas where microbilogists can explore furter.

4 Approach

The project consisted of a number of complex tasks

- Data acquisition and data cleansing - consisting of eukaryotic protein sequences, pfam entries, disorder entries and taxonomy trees
- Preparation of a 'corpus' of 17.8 million sentences, each sentence representing a eukaryotic proteins, with the words in the sentence correspoding to to a common functional 'token' where the function of a part of the protein is known
- Creating a series of word2vec models for the corpus - each model containing a different encoding for the tokens within the protein depending upon the learning parameters provided to the model
- Cross-analysis of the encodings to identify toeksn that may be 'close' to each other in the vector space produced by the model
- Comparing the embedding spaces with other embeddings to spot correlations

4.1 Data Acquisition and preparation

This section relates to the process of acquiring data and preparing it to the point where it could be used to create a corpus for the word2vec models.

This was an extremely time-consuimg task due to the quantity of data that had to be acquired and processed and ultimately a number of different approaches were adopted before finding a scalable solution that would perform within a reasonable timeframe.

Proteins Triedn Uniref format but initially had issues processsing - lots of memory issues Eventually switched to TrEMBL as I couldn download that format from Uniprot and at the same time parse out the eukaryotic proteins proteins : 78,494,529 eukaryotic proteins loaded (TrEMBL)

This is what I used initially and all through July uniprotkb-2759$_7$8494531.dat : $EukaryoticTrEMBLproteins$ $datfilewithUniprotid, startandstop - 78,494,529entries$

In August I retried this uniref100$_t ax_2$0240801.dat : $FullUniRef100proteins$: $datfilewithUniprotid, startan$ $408,368,587entries$

pfam - 296,017,815 entries

disorder

First the raw file has over 4BN rows of xml and needed ot be chunked into smaller sizes This was done with C++ and Python These xml files were then parsed to create a .dat file (parse$_d isordered_t o_d at.py)This.d$

This was the last step in the data prep phase prior to creating the precorpus and corpus. extra.xml :

lines: 4BN 4BN lines: wc -l extra.xml 4,007,237,378

protein tags:

mobdb entries : 57MN 57MN mobidb entries (each with potentially mutliple disorder entries) grep -c "disorder$_p redictiond\ddot{b}name = \ddot{M}OBIDBLT\ddot{"}/Volumes/My\ Passport/downloads/extra.xml57,013,227$

disorder entries : 81MN

Taxonomy download from here: https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/