Abstract

Introduction

Change Proteins provide the functional building blocks necessary to support life. A protein is a chain of Amino Acids of varying length.

The amino acids that form proteins can ultimately be traced back to our DNA, thus there is a strong link between the consituent parts of a protein and evolution. If a sequence of amino acids serve to provide an evolutionary advantageous capability, that sequence will survive through the generations.

Throgh parinstaking lab work, molecular biologists have been able to identify short sections of proteins that provide a particular function. This is often achieved by identifying commone 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.

Motivation Withih 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