README

July 5, 2019

0.0.1 Prediction of Secondary Structure of Proteins

Proteins are chain molecules built from sequences of amino acids (AA). There are 20 different amino acids that make up all of life's proteins. Protein structure can be described at three levels. 1) The primary structure is the sequence of amino acids in the chain, that is a one dimensional structure. The primary structure of millions of proteins are known today, many of them for several species. 2) The secondary structure is the result of the folding of the parts of the AA chain. The two most important secondary structures are the $\alpha-$helix and the $\beta-$sheet. 3) The tertiary structure is the real 3-D configuration of the protein under given environmental conditions (solvent, pH, temperature, etc). The tertiary structure decides the biochemical function of the protein. If the tertiary structure is changed, the protein normally loses it's ability to perform whatever function it has, since this function depends on the geometrical shape of the active site in the interior of the molecule (key-lock principle). Finding the primary structure is relatively easy, hence millions of primary structures are known nowadays. *Finding the secondary structure is difficult, and finding the tertiary structure is much more difficult*. Tertiary structures are found experimentally using X-ray crystallography and nuclear magnetic resonance (NMR). The tertiary structure of a protein usually depends on interactions between amino acids, which are far from each other in the primary structure. Hence predicting these is a tough task. The prediction of secondary structures is easier since interactions between neighboring amino acids play a larger role. In particular, the prediction of $\alpha-$ helices should be possible, since they depend mainly on the interactions between amino acids which are not more than four places in the chain away from each other. The reason is, $\alpha-$ helix makes a 100° turn per amino acid. So, 3.6 amino acids make a complete turn.The $\alpha-$helices are stabilized by the interactions between next neighbors, that is direct neighbors in the primary structure, between amino acid numbers i and $i+3$, i and $i+4$, which are neighbors in 3-D space. #### Bio-chemical structure of $\alpha-$helices i) The formation of $\alpha-$helices is a complicated process depending on many factors, not just the AA sequence, so predictions of their structure can only be approximate. ii) In the primary structure, $\alpha-$helices consist of sequences of approximately $6-20$ amino acids. iii) $\alpha-$helices are normally located near each other (often pairwise, parallel in a coiled coil structure) in three dimensions. iv) An $\alpha-$helix makes a 100° turn per amino acid (so 3.6 amino acids are a complete turn). v) Typically, there are two sides of an $\alpha-$helix. One is 'hydrophobic' (water hating) and other is 'hydrophilic'(water loving). So, former faces the interior of protein and latter faces outside (since proteins are normally sorrounded by water). vi) On the average only 10% of the amino acids of a protein are part of some $\alpha-$helix, so our input data will consist of sequences with batches of 6 to 20 "yes" interspersed between longer batches of "no". If $f(i)$ is the function that predicts α-helix at position i, then most of the time, $f(i+1) = f(i)$. The actual percentage of amino acids in α-helices differs widely from protein to protein.

In this tutorial project, we shall explore different methods of modeling the problem and also predicting the secondary structure of proteins.

In [ ]: