MT16121

Ankit Sharma

The genbank file was read using SeqIO interface and records were extracted from it. Features corresponding to each record were iterated and if the feature type equals “CDS”, then the start and stop locations were extracted into variables. The sequence was then spliced to get the coding region. Other important attributes like gene, locus_tag etc. were also extracted as they were to be written into the fasta file. The list was then written to a fasta file using Bio Python.

While writing my own parser, a list of keywords was prepared which was used as a reference to get the data from the genbank file. The data entries were processed and added to a list which was then written to fasta and genbank files.

The genbank format consists of annotations along with the sequence. It can be broken down into key value pairs and can be stored in a dictionary. It also contains a CDS tag which contains the coordinates of the coding region.

CDS region is the part that gets translated into proteins. It represents the entire coding region. The central dogma of life deals with the transfer of genetic information form gene level to the protein level which includes the process of transcription and translation. The process of transcription converts DNA sequence to RNA sequence which then gets converted to a sequence of amino acids. This is done by Ribosome. The translation is done by considering 3 nucleic acids at a time which are termed as codons.