Version: V1.0

Date: 07.05.19

Author: Dr Georgina R Toye

Address: Department of Biological Sciences
School of Science
Birkbeck College
University of London

PARSER DOCUMENTATION

Description:

This program takes a text document containing Genbank records and sequentially returns the data for each record from various key fields. The textfile is stored locally as "genbank.txt".The data is returned as a series of consecutive entry statements in the MySQL "INSERT INTO" format for direct manual entry into the associated database. There are three separate subprograms run in tandem, which function to populate the three tables which make up the Genbank database.

Main program

This returns the following data:

acc_code = accession code;

chrom_loc = chromosomal lcation;

gene_id = gene identifier;

prot_name = protein name;

locus_span = locus DNA span;

gene_span = DNA span of the genetic unit;

CDS_span = DNA span of the entire CDS, including intervening sequences;

start_cod = start codon position

Subprograms

These return the following data:

cds_maps = the join instructions for the CDS regions, together with the appropriate accession numbers as unique labels for data coming from each record.

locus_seq = locus sequence data, together with the appropriate accession numbers as unique labels.

In each case, the data is output as a series of MySQL entry statements ready to populate the relevant table, i.e. gb_main, gb_seq and gb_cds_map, respectively.