A protein is a large molecule manufactured in the cell of a living organism to carry out essential functions within the cell. The primary structure of a protein is a sequence of amino acids. There are 20 common amino acids, each of which has a chemical name (e.g., "Glycine"), a three-letter abbreviation (e.g., "Gly"), and a one-letter code (e.g., "G"). See

`http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html`

for a table about the chemistry of the amino acids and

`http://courses.cs.vt.edu/~algnbio/genetic_code/index.php`

for information about how the amino acids fit into the genetic code. The twenty one-letter codes are:

```
A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
```

For the purpose of representing and manipulating the primary structure of a protein, it suffices to use the one-letter codes in a string. For example,

```
MLQSIIKNIWIPMKPYYTKVYQEIWIGMGLMGFIVYKIRAADKRSKALKASAPAPGHH
```

is the amino acid sequence for a human protein called "6.8 kDa mitochondrial proteolipid". In this project, amino acid sequences will always be upper-case, with no white space inserted. There are many online databases from which protein sequences can be obtained. One is SWISS-PROT, which can be found at `http://ca.expasy.org/sprot/`. As of March 27, 2004, SWISS-PROT contained database entries for 146,720 amino acid sequences. Each database entry has much more information about a protein than its sequence, as can be seen by going to

`http://ca.expasy.org/cgi-bin/get-full-entry?[SWISS_PROT-ID:'68MP_HUMAN']`

This is the entry for the protein whose sequence was given above.

A complete protein record may contain a fairly large number of logical fields. These are flagged with two-character sequences occurring at the beginning of each line. A full listing of the possible fields is given in Table 1 on page 2 of this specification. It is important to note that some protein records will contain only a proper subset of the possible fields. In addition the amount of data for each field can vary considerably.

For our purposes in this assignment, we will use a text file of complete SWISS-PROT entries. You do not need to be concerned with validating the correctness of the database entries.

A full description of the logical significance of the various fields, and any format constraints, is given in the UniProt User Manual, which is available at `http://us.expasy.org/sprot/userman.html`.

Note: some of the sequence data files may contain multiple entries corresponding to the same accession code. In such a case, your implementation should recognize if an accession code is already in the index structure, and if so simply reject the duplicate entries.

It is also possible that a protein record may list more than one accession code. In that case, you should index the record using the first code that is listed (the primary accession code) and disregard the rest.

Table 1 Protein record field specifications:

Each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry, are shown in the table below.

Line code	Content	Occurrence in an entry	Comments
ID	Identification	Once; starts the entry	
AC	Accession number(s)	Once or more	[O,P,Q][0-9][A-Z,0-9]^3[0-9]
DT	Date	Three times	
DE	Description	Once or more	
GN	Gene name(s)	Optional	
OS	Organism species	Once or more	
OG	Organelle	Optional	
OC	Organism classification	Once or more	
OX	Taxonomy cross-reference(s)	Once or more	
RN	Reference number	Once or more	
RP	Reference position	Once or more	
RC	Reference comment(s)	Optional	
RX	Reference cross-reference(s)	Optional	
RG	Reference group	Once or more (Optional if RA line)	
RA	Reference authors	Once or more (Optional if RG line)	
RT	Reference title	Optional	
RL	Reference location	Once or more	
CC	Comments or notes	Optional	
DR	Database cross-references	Optional	
KW	Keywords	Optional	
FT	Feature table data	Optional	
SQ	Sequence header	Once	
(blanks)	Sequence data	Once or more	
//	Termination line	Once; ends the entry	

As shown in the table, some line types are found in all entries, others are optional. Some line types occur many times in a single entry. Each entry must begin with an identification line (ID) and end with a terminator line (//).

Note that some formatting details must be inferred from the sample data files provided on the course website, and the detailed documentation available online in the UniProt User Manual.

Figure 1 Sample Protein Record:

```
ID   143S_MOUSE     STANDARD;      PRT;   248 AA.
AC   O70456;
DT   16-OCT-2001 (Rel. 40, Created)
DT   16-OCT-2001 (Rel. 40, Last sequence update)
DT   28-FEB-2003 (Rel. 41, Last annotation update)
DE   14-3-3 protein sigma (Stratifin).
GN   SFN OR MKRN3.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=FVB/N;
RA   Karpitskiy V.V., Shaw A.S.;
RL   Submitted (APR-1998) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: P53-regulated inhibitor of G2/M progression (By
CC       similarity).
CC   -!- SUBUNIT: Homodimer (By similarity).
CC   -!- SUBCELLULAR LOCATION: Cytoplasmic or may be secreted by a non-
CC       classical secretory pathway (By similarity).
CC   -!- SIMILARITY: Belongs to the 14-3-3 family.
DR   EMBL; AF058798; AAC14344.1; -.
DR   HSSP; P29312; 1A38.
DR   MGD; MGI:1891831; Sfn.
DR   GO; GO:0005737; C:cytoplasm; IDA.
DR   GO; GO:0000079; P:regulation of CDK activity; IDA.
DR   InterPro; IPR000308; 14-3-3.
DR   Pfam; PF00244; 14-3-3; 1.
DR   PRINTS; PR00305; 1433ZETA.
DR   ProDom; PD000600; 14-3-3; 1.
DR   SMART; SM00101; 14_3_3; 1.
DR   PROSITE; PS00796; 1433_1; 1.
DR   PROSITE; PS00797; 1433_2; 1.
KW   Multigene family.
SQ   SEQUENCE   248 AA;  27713 MW;  D433390433FB3F48 CRC64;
     MERASLIQKA KLAEQAERYE DMAAFMKSAV EKGEELSCEE RNLLSVAYKN VVGGQRAAWR
     VLSSIEQKSN EEGSEEKGPE VKEYREKVET ELRGVCDTVL GLLDSHLIKG AGDAESRVFY
     LKMKGDYYRY LAEVATGDDK KRIIDSARSA YQEAMDISKK EMPPTNPIRL GLALNFSVFH
     YEIANSPEEA ISLAKTTFDE AMADLHTLSE DSYKDSTLIM QLLRDNLTLW TADSAGEEGG
     EAPDDPHI
//
```

Note that there are absolutely no stated limits on the lengths of the strings that occur in the protein records.