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 shorter sample is given in Figure 4 on page 9 of this specification.

Besides the amino acid sequence (found in a formatted form under the heading "Sequence information" near the bottom of the page) other important information in the SWISS-PROT entry includes:

- The primary accession code (e.g. "P56378") that is always a unique 6-character identifier for the entry;
- A protein name (e.g., "6.8 kDa mitochondrial proteolipid") that is an official name for the protein that indicates its function in a cell; and
- One or more source organism fields (e.g., "Homo sapiens (Human)") that give the species where the protein is found.
- A molecular weight given to the nearest mass unit.

For our purposes in this assignment, rather than use a text file of SWISS-PROT entries, the program will manipulate a binary database file that contains condensed protein records consisting of a primary accession number, a protein name, a source organism, the amino acid sequence, and the molecular weight. More precisely, each database record will be stored in the following format:

Significance	**Type**	**Comments**
Accession code	sequence of characters	alphanumeric, always 6 characters
Protein name length	unsigned short	
Protein name	sequence of characters	no meaningful restriction on contents
Number of source organisms	unsigned short	The following pair of entries will be repeated the specified number of times
Organism name length	unsigned short	
Organism name	sequence of characters	no meaningful restriction on contents
Amino acid sequence length	unsigned short	
Amino acid sequence	sequence of characters	upper-case, from code set given above
Molecular weight of protein	unsigned int	