HW 3: Reading & Writing FASTA Files

For this assignment, you will make programs that can read and write FASTA files. The notebook you turn in should be called `fasta-reader.ipynb`.

The FASTA format is a text-based format made to hold either nucleotide sequences (in which each letter represents a DNA base) or peptide sequences (in which each letter represents an amino acid). The first line of a FASTA file begins with a `>` symbol, followed by a description of the data to come. Then, every line after that holds the long sequence of data. The difficult part is that many of these sequences are hundreds or even thousands of characters long. Thus, data may be split over many lines. Here's an example of a peptide sequence that comes from the Sri Lankan elephant:

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

The single sequence starts with `LCL` and extends all the way to `ENY`.

Like for last week's homework, you must make both a reader and a writer for `.fasta` files. The reader will ask the user for the name of a file. (You do not need to modify it this time.) The reader must ask the user for a file's name. Given this, it must open the file, and then output both the descriptor (removing the `>`) and the sequence, like this:

```
Please enter a FASTA file: elephant.fasta

Descriptor: gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus
maximus]
Sequence: LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIP
YIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIK
DFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVI
LGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGXIE
NY
```

The writer will be more difficult. It will first ask the user for the name of a new file, and the descriptor and sequence to place into that file. It will prepend the descriptor with the necessary `>`, and it will split the sequence into lines of 70 characters (except the last one, which will probably have fewer):

```
Please enter the name for a new FASTA file: prolactin-precursor.fasta
Enter its descriptor: LCBO - Prolactin precursor - Bovine
Enter its sequence: MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAV
```

```
MVSHYIHDLSSEMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLY
HLVTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDEDARYSAFY
NLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

Thank you.  The file "prolactin-precursor.fasta" has been created.
```

(The above sequence has 230 characters. Therefore, in the saved file, the sequence will be split into three lines of 70, followed by one line of 20.)

As always, style is important. Be sure to give your program adequate comments, so that someone else could read your code easily.