HW 6: Analyzing the Kraken Genome

Now it is your job to analyze some genes from the kraken genome. This notebook should be called `kraken-genetics.ipynb`.

Download the file `kraken.gene`. It contains ten lines of 5000 bases each. You must locate the gene located within each line, and output some basic information on it. This will require a combination of loops, if/else blocks, and regular expressions.

For purposes of this assignment, consider a gene to be a stretch of DNA that begins with a start codeon (the bases `ATG`) and continues by threes until the earliest stop codon (`TAA`, `TAG`, or `TGA`). Note that the stop codon must be "in frame". That is, the number of bases between the start codon and the stop codon must be a multiple of three, or else it is not really a stop codon.

Here is the standard genetic code, by which the information in a gene is used to create a new polypeptide chain, which may be incorporated into a protein. Each codon codes for a different amino acid, which forms one of the "links" in the new chain.

1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	Phenylalanine (nonpolar)	TCT	Serine (polar)	TAT	Tyrosine (polar)	TGT	Cysteine (polar)	T
	TTC		TCC		TAC		TGC		C
	TTA	Leucine (nonpolar)	TCA		TAA	Stop	TGA	Stop	A
	TTG		TCG		TAG		TGG	Tryptophan (n)	G
C	CTT		CCT	Proline (nonpolar)	CAT	Histidine (basic)	CGT	Arginine (basic)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Glutamine (polar)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Isoleucine (nonpolar)	ACT	Threonine (polar)	AAT	Asparagine (polar)	AGT	Serine (polar)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lysine (basic)	AGA	Arginine (basic)	A
	ATG	Methionine (n)	ACG		AAG		AGG		G
G	GTT	Valine (nonpolar)	GCT	Alanine (nonpolar)	GAT	Aspartic acid (acidic)	GGT	Glycine (nonpolar)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glutamic acid (acidic)	GGA		A
	GTG		GCG		GAG		GGG		G

Thus, if the gene starts with `ATGCAG`, the growing polypeptide chain starts with methionine, which is followed by a glutamine. The polypeptide chain continues until the gene reaches a stop codon, which doesn't code for any amino acid.

Your program will ask the user for the file to use. For each line, you must output:

- Where the gene begins (using 1-based counting).

- How many amino acids will be in the created polypeptide chain.
- The percentage of these that are polar, nonpolar, acidic, and basic.

For example, using the given data, your output should start like this:

```
Please enter the file to analyze: kraken.gene
Line 1:
  Gene starts at position 923, and codes for a chain of 892 amino acids.
  Polar: 34.98%, Nonpolar: 42.04%, Acidic: 5.83%, Basic: 17.15%
⋮
```

Be sure to use functions as much as you can. The main block of code (the block outside of any function) should just be a high-level outline of your program. Keep all the functions at the beginning of your code. The only thing that should be before your functions is the line importing the `re` module, and possibly some constants.

You may find this assignment easier if you make your own small `.gene` file first. Debug your code on the small file, before you attempt the larger one.

Finally, note that I will use a different source file, with a different length, when I test your program. Be sure that your program is general enough to work in this case!