CS100M Spring 2008 Project 5 due Thursday 4/10 at 6pm

2 From DNA to protein

The human genome has three billion pairs of nucleotides, also called bases: Adenine, Cytosine, Guanine, and Thymine. We will refer to the four bases by their first letters, T, C, A, and G. When a gene is active, the corresponding DNA is transcribed to RNA, and the RNA prescribes the production of proteins, which are chains of amino acids. There are 20 amino acids used by living cells to encode proteins. Since the DNA "alphabet" has only four letters, scientists first postulated and later confirmed that triplets of bases encode the amino acids. (3 is the smallest n such that 4^n is at least 20.)

The $4^3 = 64$ possible triplets of DNA, called codons, and the names of the amino acids that they encode are shown in the translation table below. Since a triplet of bases encode one amino acid, an amino acid sequence (using the abbreviation letters) is exactly 1/3 the length of the DNA sequence. (Here we simplify the problem by ignoring the RNA. In terms of coding, the RNA differs from the DNA only in one letter: a T in DNA is a U in RNA.)

Codon	Amino acid	Abbreviation	Codon	Amino acid	Abbreviation
ttt	Phenylalanine	F	att	Isoleucine	I
ttc	Phenylalanine	F	atc	Isoleucine	I
tta	Leucine	L	ata	Isoleucine	I
ttg	Leucine	L	atg	Methionine	M
tct	Serine	S	act	Threonine	T
tcc	Serine	S	acc	Threonine	T
tca	Serine	S	aca	Threonine	T
tcg	Serine	S	acg	Threonine	T
tat	Tyrosine	T	aat	Asparagine	N
tac	Tyrosine	T	aac	Asparagine	N
taa	Ochre	Y*	aaa	Lysine	K
tag	Amber	Y*	aag	Lysine	K
tgt	Cysteine	C	agt	Serine	S
tgc	Cysteine	C	agc	Serine	S
tga	Opal	Y*	aga	Arginine	R
tgg	Tryptophan	W	agg	Arginine	R
ctt	Leucine	L	gtt	Valine	V
ctc	Leucine	L	gtc	Valine	V
cta	Leucine	L	gta	Valine	V
ctg	Leucine	L	gtg	Valine	V
cct	Proline	P	gct	Alanine	A
ccc	Proline	P	gcc	Alanine	A
cca	Proline	P	gca	Alanine	A
ccg	Proline	P	gcg	Alanine	A
cat	Histidine	H	gat	Aspartic acid	D
cac	Histidine	H	gac	Aspartic acid	D
caa	Glutamine	Q	gaa	Glutamic acid	E
cag	Glutamine	Q	gag	Glutamic acid	E
cgt	Arginine	R	ggt	Glycine	G
cgc	Arginine	R	ggc	Glycine	G
cga	Arginine	R	gga	Glycine	G
cgg	Arginine	R	ggg	Glycine	G

* "stop codon" for indicating the end of an active gene's coding sequence--not an actual amino acid

You will write a function to read a protein data file, extract some information including the DNA sequence, translate the DNA into the sequence of amino acids, and create and return a MATLAB structure for that protein. You can download your favorite protein from various data banks! The provided data files were downloaded from NCBI (National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov/`). Your function has the following specification:

```
function Pro= MakeProtein(fname)
% Make a STRUCTURE for a protein given its summary data file.
% fname is the string file name
% Pro is a protein with
%   Pro.id assigned the LOCUS identifier (line 1, col 13-25)
%   Pro.def assigned the DEFINITION
%     (start at line 2 col 13; variable number of lines)
%   Pro.dna assigned the dna sequence
%     (start after the line ORIGIN; end before the line //)
%     A string containing only the letters--no spaces or digits
%   Pro.amino assigned the amino acid sequence, translated from dna
```

The given function `conversionTable` returns a cell array representing the "table" for translating from DNA to amino acids as shown above. Read the function and its comments to learn the layout of the table. You must use `conversionTable` in your solution.

Several data files are provided so that you can test your function on different data. The shortest one is `pdata8.txt`, so you may want to use that one as the starting point.

Hint: Create helper functions (*subfunctions*) to break down the problem! If you can name a specific (sub)task, then chances are good that that task can be separated out as a specific subfunction. Good candidates are "read data from file" and "translate to amino acid."