Due: - Written Exercises due Wednesday, June 1st, at 4 pm in Homework box in 2131 Kemper;
- Program due electronically Wednesday, June 1st, at 11:50 pm.

Written Exercises (25 pts): The written exercises should be typed and each page should have at the top your name and ID#, section, and hw#. Handwritten answers will not be graded.

J&K, 7.7.2, 7.7.4, 7.8.3, 7.10.1, 8.1.4, 8.7.2, 10.6.10, 10.8.4, 10.9.2, 11.3.2, 11.3.12.

Program (75 pts): Handin a makefile and source code files, and have your makefile produce the executable file **`translate`** (use **`all`** … at the top). The third line in the source code files must contain the author of the file, ID, and section #. Use the *handin* program for electronic submission, described in the UNIX tutorial. For this homework use:

`handin cs30 hw8 Makefile <your file1> <your file2>` …

The date and time your files are created in the cs30 directory will be counted as your submit times. If those times are later than 11:50 pm on the due date your submissions will be considered late.

From Genes to Proteins

Recall (e.g., from HW3) that DNA is a double stranded molecule composed of paired nucleotides, each of which could be `a, c, g`, or `t`, where `a` is complementary to `t` and `c` to `g`. Since they are just strings, any given sequences can be ordered alphabetically, say in an increasing order. Thus, `actg` would come after `accg,` and before `at`.

Genes are substrings of DNA which code for proteins, and carry the heritable information from our parents. Genes start with the sequence of three letters **`atg`**, called the START codon, and end with one of the three sequences **`tga`**, **`taa`**, or **`tag`**, called STOP codons. The stretch of sequence between the START and any one of the STOP codons is a potential gene. Each codon codes for an amino acid represented by a letter of the alphabet. There are 19 amino acids. Strung together amino acids form **proteins**. A substring of a DNA sequence of length a multiple of three which starts with a START and ends with one of the STOP codons can be translated into an amino acid sequence, or protein, and is called a translatable sequence. For example the sequence `AATTAAG`**`ATGGGGCTCTAA`**`AAT` contains a translatable sub-sequence at the 8th position of length 12, thus consisting of 4 codons. This sub-sequence can be translated using a **codon table** into the length three amino acid sequence `MGL`. Note that the START codon codes for an amino acid, `M,` while the STOP codons don't code for amino acids. On the other hand, the `DNA` sequence `AATGAATCTAGT` is not a translatable sequence.

In this homework you are asked to write a program that takes two command line arguments: an input file name, containing DNA sequences, and an output file name, in which you'll store the translated, protein sequences. For each sequence, the program should identify the longest possible translatable sub-sequence, if one exists, and translate it into a protein using a codon table given in the file **`codeoflife.txt`**. The translated sequences must be ordered in the sorted order of their DNA sequence. **You must store the sequences in a linked list, in alphabetical order. Using anything but a linked list to sort (e.g. qsort, etc), will amount in a 50% point reduction, even if your program works correctly otherwise.** If done properly, by adding a new sequence in the appropriate place in the list so that alphabetical order is preserved, the sorting will amount to a simple traversal trough the linked list. You can assume that the DNA sequence file will contain only proper sequences (i.e. strings over {A, C, G, T, a, c, g, t}).

Make sure the output of your program matches exactly the output below of my executable located at **`/home/cs30/public/hw8/translate`** on the csif machines. In the same directory you will find file **`codeoflife.txt`**, a tab-delimited file containing the mapping of amino acids (column 1) to codons (column 2).

```
[cs30@pc50 hw8]$ translate seqs.txt prots.txt

[cs30@pc50 hw8]$ more seqs.txt
aaATttaTggattagcaagcag
ACGATGATGATGGGGCCCTAATAGTGATAAAAAACT
AAAATAATTTGGA
ATGAAATGGTAGATGAAACCCGGGATATGATAG

[cs30@pc50 hw8]$ more prots.txt
none
MD
MMMGP
MKPGI

[cs30@pc50 hw8]$
```