

University of Cincinnati
Department of Electrical & Computer Engineering and Computer Science

20 ENFD 112 – Fundamentals of Programming

LABORATORY 5: DNA GENE IDENTIFICATION, PATTERN MATCHING

Spring 2008

1. Objective

The objectives of this assignment are to a) understand rudimentary pattern matching algorithms; b) understand how important it is to design code that runs efficiently in the presence of very large data files.

2. Background

The Human Genome Project

The Human Genome Project was an international effort to discover all of the approximate 30,000-100,000 human genes (the human genome) and to determine the complete sequence of the 3 billion DNA subunits (called bases). The results of the project have been made publicly available in an effort to lower the barriers to effective biomedical research. The project was formally begun in October 1990 and was planned to last 15 years. However, rapid technological advances resulted in the completion of the project by 2003, two years ahead of schedule.

As part of the project, parallel studies were carried out on selected model organisms, such as the bacterium *E. coli* and the *Drosophila*, to help develop the technology to quickly sequence genes and interpret human gene function. The U.S. Department of Energy's Human Genome Program (operated by Lawrence Berkley Laboratory) and the National Institutes of Health's National Human Genome Research Institute (NHGRI) together make up the U.S. Human Genome Project.

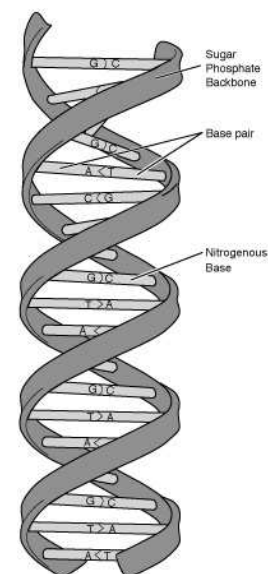
Private companies are also racing to be the first to completely map the genome. Companies like Celera, Affymetrix, and Gene Logic, are hoping to patent their information before the government can discover it and release it freely. If you are more interested in the HGP then check out the website at <http://www.ornl.gov/hgmis> or USA Today.com's set of stories on genomics.

Computer Modeling of DNA

DNA, the building block of all forms of life, is a double helix molecule consisting of a sequence of *base-pairs* or rungs (see the figure on the right). Each rung is formed from either an A-T or G-C pairing where A, T, G, and C correspond to the four DNA nucleotides of adenine, thymine, cytosine, and guanine, respectively. Because of this you only need to specify one side of the sequence to know the entire double helix. Thus, DNA is modeled in a computer program as a sequence of symbols, each taking one of four values: characters 'A', 'T', 'C', and 'G'. For example, the following represents a possible DNA subsequence:

CGTGACAGTCCTCTCCTTTACCGAAAGGGAAGAATAAAAGTGGCGTGATGCATTACGC

A DNA sequence such as the one modeled above is read from left to right.



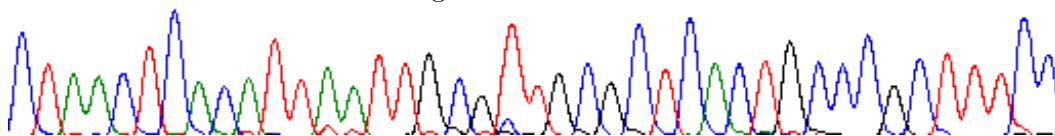
A *gene* is a subsequence of a DNA molecule and is one of 20 nucleic acids upon which all proteins in the body are built. A DNA molecule is partitioned into sections, called *codons*, containing three base-pairs each and represented or typed as three labels corresponding to the nucleotides of those base-pairs (for example, *CGA*). A gene subsequence begins at a condon and ends at a condon. Therefore, the number of base-pairs in a gene is always a multiple of three. There is only one condon type that can begin a gene: *TAC*. There are three condon types that can end a gene: *ACT*, *ATT*, and *ATC*. For example, the DNA sequence shown above contains the gene:

TACCGAAAGGGAAGAATAAAAGTGGCGTGATGCATT

which starts at position 19 (7th condon) in the sequence, from the left, and has 36 base-pairs (12 condons). Genes do not overlap: this implies the end condon of a gene is the first one that is encountered after the start condon in a sequence. In humans, the average gene is about 20,000 base-pairs long.

Sequencing

Scientists usually send off a chunk of DNA to be sequenced by an automatic sequencing machine. The results come back as shown in the figure below.



where the colors indicate the probability that one of the four nucleotides is at a particular spot. This type of figure is converted into a sequence of letters, like *ACGT...* It is up to the scientists to analyze the string of letters to determine what the sequence actually does.

3. Problem

Your assignment is to read a file containing a genetic sequence and identify the number of genes, their starting position relative to the beginning of the file, and their length. Your program will output a report to the screen and to a file indicating who generated the report and the information listed above. To do this, you will have to write a function that automatically searches for and finds the start and end codons. The input file format is just a string of 'A', 'C', 'T', and 'G' characters, all on one line.

Write the program to solve this problem as a function called `pattern` which takes two arguments: the first, named `filename`, specifies the name of the input file and the second, named `identity`, specifies who is using the program. When run, the user specifies the name of the file to be read. For example,

```
> pattern('short.dat', 'John Franco');
```

is how to execute your code from Matlab assuming that a file named `short.dat` exists in the current directory and the user is the instructor.

4. Analysis and Coding

Read the given file into a variable `genome` as follows:

```
fid = fopen(filename, 'rt');          % open the file
if fid < 0
    error(['File ' filename ' not found']);
```

```
end
genome = fgetl(fid);
```

where `filename` was obtained as an argument to the `pattern` function.

Establish a variable, say `strt_condon`, as an index into `genome` that points to the current start condon for a gene. Initially, `strt_condon` is 1. Establish a variable, say `end_condon`, as an index into `genome` that points to a condon that is to be checked to see if it is one of three end condon patterns. This variable is not set until a start condon has been located. When that happens, `end_condon` takes the value of `strt_condon` plus 3. The `end_condon` cursor advances and tests are made until an end condon is found. At that point, statistics are recorded about the gene that has been discovered and the `strt_condon` variable is set to `end_condon` plus 3. The `strt_condon` cursor then advances and tests are made to locate a start condon. When one is found, the above process repeats.

5. Submission

Submit a single file called `pattern.m` on or before May 4 using blackboard. See the course webpage at <http://gauss.ececs.uc.edu/Courses/HTML/E112.html> for instructions.