YILDIZ TECHNICAL UNIVERSITY
ELECTRIC-ELECTRONIC FACULTY
COMPUTER ENGINEERING

INTRODUCTION TO BIOINFORMATICS
INSTRUCTOR: Nizamettin AYDIN

HOMEWORK #1: DNA sequence Translation to Aminoacid Sequence

Umut Güzel
18011004

1. INTRODUCTION

The main purpose of our perl program is tranlate DNA sequence into aminoacid sequence. The input file is in FASTA format. To read file, I use seek function to put cursor at the start of out DNA sequence. After that in a while loop, every DNA nucleotide translate into RNA nucleotide. In every three RNA nucleotide, I group them and put into an array named “triplet”. With a hash map that includes aminoacid equivalent of triplets, I write all aminoacids into output file. Under aminoacid sequence, I put proteins can be synthesized.

2. CODE EXPLANATION

```
open(input, "<DNA-ex.txt") or die("cannot open file");
open(out, ">result.txt") or die("cannot open file");

seek input,66,0;
```

open input and output file and put cursor into 66th character(start of DNA sequence).

```
$i=0;
$counter=1; #to count three nucleotide for every aminoacid
@triplet=(); #created an empty arrat to put RNA triplets
```

```
while(!eof(input)){ #scan file until the end of the file
    $tmp=getc(input);
    if($tmp eq 'A' or $tmp eq 'T' or $tmp eq 'G' or $tmp eq 'C'){ #check if it is a nucleotide or space
        #finding RNA equivalent of DNA nucleotide
        if($tmp eq 'A'){ #if tmp variable is not A,T,G or C; do not process that char
            $tmp='U';
        }elsif($tmp eq 'T'){
            $tmp='A';
        }elsif($tmp eq 'G'){
            $tmp='C';
        }elsif($tmp eq 'C'){
            $tmp='G';
        }
        ####################
        #combine RNA nucleotides as triplets
        if($counter==1){
            $store=$tmp;
            $counter=2;
        }elsif($counter==2){
            #$store=$triplet[i];
            $store=$store.$tmp;
            $counter=3;
        }elsif($counter==3){
            #$store=$triplet[i];
            $store=$store.$tmp;
            push(@triplet,$store);
            $counter=1;
            $i=$i+1;

        }
        ##########################
```

the while loop that scans the file and in every turn, process data taken from file to $tmp variable.

In loop, the data taken from file checked by first if statement to se if it is a DNA nucleotide or not. After that:

-There is a if..elsif statements that translates the DNA nucleotide on tmp variable to RNA nucleotide.

```
if($tmp eq 'A'){ #finding RNA equivalent of DNA nucleotide
    $tmp='U';
}elsif($tmp eq 'T'){
    $tmp='A';
}elsif($tmp eq 'G'){
    $tmp='C';
}elsif($tmp eq 'C'){
    $tmp='G';
}
#####################
```

-Another if..elsif statement to count three nucleotide and combine them on $store variable, than push it to the pre-defined @triplet array.

```
#combine RNA nucleotides as triplets
if($counter==1){
    $store=$tmp;
    $counter=2;
}elsif($counter==2){
    #$store=$triplet[i];
    $store=$store.$tmp;
    $counter=3;
}elsif($counter==3){
    #$store=$triplet[i];
    $store=$store.$tmp;
    push(@triplet,$store);
    $counter=1;
    $i=$i+1;

}
##############################
```

```
#aminoacid hash table
%aminoacids= ('UUU'=> 'F','UUC'=> 'F','UUA'=> 'L','UUG'=> 'L',
              'UCU'=> 'S','UCC'=> 'S','UCA'=> 'S','UCG'=> 'S',
              'UAU'=> 'Y','UAC'=> 'Y','UAA'=> '-STOP-','UAG'=> '-STOP-',
              'UGU'=> 'C','UGC'=> 'C','UGA'=> '-STOP-','UGG'=> 'W',
              'CUU'=> 'L','CUC'=> 'L','CUA'=> 'L','CUG'=> 'L',
              'CCU'=> 'P','CCC'=> 'P','CCA'=> 'P','CCG'=> 'P',
              'CAU'=> 'H','CAC'=> 'H','CAA'=> 'Q','CAG'=> 'Q',
              'CGU'=> 'R','CGC'=> 'R','CGA'=> 'R','CGG'=> 'R',
              'AUU'=> 'I','AUC'=> 'I','AUA'=> 'I','AUG'=> 'M',
              'ACU'=> 'T','ACC'=> 'T','ACA'=> 'T','ACG'=> 'T',
              'AAU'=> 'N','AAC'=> 'N','AAA'=> 'K','AAG'=> 'K',
              'AGU'=> 'S','AGC'=> 'S','AGA'=> 'R','AGG'=> 'R',
              'GUU'=> 'V','GUC'=> 'V','GUA'=> 'V','GUG'=> 'V',
              'GCU'=> 'A','GCC'=> 'A','GCA'=> 'A','GCG'=> 'A',
              'GAU'=> 'D','GAC'=> 'D','GAA'=> 'E','GAG'=> 'E',
              'GGU'=> 'G','GGC'=> 'G','GGA'=> 'G','GGG'=> 'G');
```

-hash map that contains all triplet combinations of RNA nucleotides and their equivalent aminoacids to find aminoacids by the values of triplet array.

```
print out "Aminoacids: ";

#print all aminoacids to file
foreach $tri (@triplet){
    print out "$aminoacids{$tri} ";
}
```

-Prints all aminoacids

```
$status=0; #when start codon occurs, status=1 ; after stop codon seen, status=0
$counter=1; #increases at every start codon after stop codon
```

```
print out "\nProteins: \n";
foreach $tri (@triplet){
    print "$aminoacids{$tri} \n";
    if($aminoacids{$tri} eq 'M'){ #check start codon
        print out "protein #$counter\n";
        $status=1;
        print out $aminoacids{$tri};
    }elsif($aminoacids{$tri} eq '-STOP-'){ #check stop codon
        $status=0;
        print out "\n";
        $counter=$counter+1;
    }elsif($status==1){ #print when the aminoacid between start and stop codon
        print out $aminoacids{$tri};
    }
}
```

-with foreach loop, we control start and stop codons and by the help of if statement and status variable, print every protein can be synthesized. In if statement when stop codon occurred, counter increased and status=0; when start codon occurred status=1.

```
close(input);
close(out);
```

Closes input and output files.

3.RESULT

Input File(DNA-ex.txt):

```
>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTC
TATTGTACCCTGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTG
CCCCCCGGGCCCGTGCCCGCCGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTC
TGAGTTGATTGAATGCAATCAGTTAAAACTTTCAACAATGGATCTCTTGGTTCCGGCATT
```

Output File(result.txt):

```
Aminoacids: L D A F L V M A H A Q E T R V G G -STOP- A Q I T W D N E A A R A A N S R R P P R G D G G P G H G R P L G L C L -STOP- Q T F A R
Q T Q L T Y V S Q F -STOP- K L L P R E P R P -STOP-

Proteins:
protein #1
MAHAQETRVGG
```

Aminoacids: L D A F L V M A H A Q E T R V G G -STOP- A Q I T W D N E A A R A A N S R R P P R G D G G P G H G R P L G L C L -STOP- Q T F A R Q T Q L T Y V S Q F -STOP- K L L P R E P R P -STOP-

Proteins:
protein #1
MAHAQETRVGG