CSI5126. Algorithms in bioinformatics
Fall 2018

Assignment 1

Deadline: October 1, 2018, 18:00

[PDF]

Solution

- a1.zip

Learning outcomes

- Through the development of simple computer programs, familiarize yourself with DNA, RNA, and protein sequences, as well as the genetic code.

In the work place, one would use an existing application or API to perform the tasks of this assignment — see the Resources Section. However, I believe that writing simple programs by yourselves to carry out these tasks can help you learn more easily the the biology.

Instructions

For all the questions, assume that the information is stored in FASTA format [1]. I am also expecting to run your program from the command line:

```
$ java A1Q1 input.fa
```

Here, **java** refers to the Java Virtual Machine, **A1Q1** is a file containing the byte-code of the java program (A1Q1.java was compiled to produce A1A1.class). Finally, **input.fa** is a file containing some input encoded using the FASTA format.

- You must do this assignment individually.
- Assignments must be submitted using Brightspace (information will be communicated soon).
- Each program must be documented.
- For the programming questions, I should be able to run your programs **without downloading additional libraries** (your program should run on any operating system, I will use macOS to test your program).
- Clearly identify yourself on every file that you hand in; this includes your name and student id.

1 Transcription (5 marks)

Write a simple program taking as input a DNA sequence stored into a file. The program must transcribe the input to RNA. The result is displayed on the standard output. For instance, given a file with the following DNA content:

```
> Unknown
ACTGTTGTTCGGTGATCATCAGTTGTACAACGTCCTAACAACATCACATGCAATGCTTATGATATTCTTC
```

Your program would display the following information on the output:

```
ACUGUUGUUCGGUGAUCAUCAGUUGUACAACGUCCUAACAACAUCACAUGCAAUGCUUAUGAUAUUCUUC
```

[1] https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp

2 Reverse complement (5 marks)

Write a simple program taking as input a DNA sequence stored into a file. The program must display the reverse complement sequence. For instance, given a file with the following DNA content:

```
> Unknown
ACTGTTGTTCGGTGATCATCAGTTGTACAACGTCCTAACAACATCACATGCAATGCTTATGATATTCTTC
```

Your program would display the following information on the output:

```
GAAGAATATCATAAGCATTGCATGTGATGTTGTTAGGACGTTGTACAACTGATGATCACCGAACAACAGT
```

3 All six reading frames (5 marks)

Write a simple program taking as input a DNA sequence stored into a file. The program must display all six translation reading frames. For example, given the follow DNA content:

```
> Unknown
ACTGTTGTTCGGTGATCATCAGTTGTACAACGTCCTAACAACATCACATGCAATGCTTATGATATTCTTC
```

Your program would produce the following output. Here the star is used to represent the stop codon.

```
> 5'3' Frame 1
T V V R * S S V V Q R P N N I T C N A Y D I L

> 5'3' Frame 2
L L F G D H Q L Y N V L T T S H A M L M I F F

> 5'3' Frame 3
C C S V I I S C T T S * Q H H M Q C L * Y S

> 3'5' Frame 1
E E Y H K H C M * C C * D V V Q L M I T E Q Q

> 3'5' Frame 2
K N I I S I A C D V V R T L Y N * * S P N N S

> 3'5' Frame 3
R I S * A L H V M L L G R C T T D D H R T T
```

4 Database search (5 marks)

One of our life science colleagues has just sequenced this DNA fragment. We would like to know if it corresponds to a protein coding sequence. If so, does it match a known protein sequence. To solve this problem, you must translate this DNA sequence into all six possible reading frames, and search each one of using the resources available at the National Center for Biotechnology Information (NCBI).

For the online search.

1. Go to the NCBI Web site: https://www.ncbi.nlm.nih.gov.
2. Go to the Sequence Analysis section of the Web site (hint: consult the menu on the left-hand side of the page).
3. In the tools section, you will find a link entitled **Basic Local Alignment Search Tool (BLAST)**. BLAST is a well known application "[to find] regions of similarity between biological sequences".
4. Since our inputs are protein sequences, go to the **Protein BLAST** Web page.
5. We will be using the database **RefSeq** (reference proteins, refseq_proteins), which is a curated database.
6. For all six reading frames, paste the sequence in the appropriate box and perform a search.

Here is the input DNA sequence.

```
> Unknown
ACTGTTGTTCGGTGATCATCAGTTGTACAACGTCCTAACAACATCACATGCAATGCTTATGATATTCTTC
TTCATCATGCCAGGCACGATGGCAGGACTAGGCAACTTACTAGTGCCATTCCAGATGAGTGTACCGGAGT
TAGTATTCCCAAAGATTAATAACATCGGTATATGATTTTTAGTATGTGGTCTACTTTTGATTACGGGTTC
ATCTTGGATGGAGGAAGGTTCAGGAACGGCCTGAACCGTCTATCCACCACTAGCGCTCACTGCAAGTCAT
AGCGGACTTGCTGTAGATACGTTCATTATCGCATTGCACATGGCCGGTGCAAGCTCCCTTACAGGAAGCA
TCAACCTTATATGTACAATCGCCTATGCCCGCCGTTCACTCATGGCGATGCTGCAGTCATCACTTTATCC
CTGATCCATTACAATCACTGCAGCGTTACTCATAGGAGTTGTGCCTGTGCTAGCAGGTGCTATCACGATG
CTACTCACTGATAGAAGTTGGAGTACCAGCTTCTATGACAGTTCGGCAGGCGGTGATCCTATGTTGTATC
AGCACTTATTCTGGGTGTTTGGGCATCCAGAAGTCTATATCATCATACTTCCAGTATTCGGTATAGTCAG
```

Answer the following questions:

- What is the likely identity of the protein?
- What is the accession number of that protein?
- What is the name of the organism?
- From which kingdom of life is the organism from?
- What is the E-value of the resulting alignment?
- Did you encounter any specific problem translating the input sequence? If so, tell me about it.

5 Genetic Code (5 marks)

Since its discovery 50 years ago, the genetic code [2] has never ceased to amaze. For instance, we now know that biases in codon usage play key roles in the subtle regulation of gene expression.

For this question, write a simple program to analyze the genetic code. In particular, your program must output the following information:

- For each of the 20 naturally occurring amino acids, as well as the stop codon, print the associated codons.
- The (Hamming) distance between any two pairs of codon is 0 (if the codons are identical), 1, 2, or 3 (if all three positions are distinct). In the first part of this question, we have seen that each amino acid is encoded by at least one, but generally many codons. Let's define $D(i, j)$ as the minimum number mutations transforming a codon for amino acid i into a codon for amino acid j. Over all possible codons encoding the amino acid i and over all the codons encoding for amino acid j, your program must find the pair with the minimum (Hamming) distance. The program must display the information for all possible pairs of amino acids.

Resources

- http://emboss.sourceforge.net
- http://biopython.org
- http://biojava.org
- http://bioperl.org
- http://bioruby.org
- http://www.bioconductor.org
- See also: https://links.bioinformatics.ca

[2] https://www.ncbi.nlm.nih.gov/books/NBK21950/

References

[1] Christina E Brule and Elizabeth J Grayhack. Synonymous Codons: Choose Wisely for Expression. *Trends in genetics : TIG*, 33(4):283–297, April 2017.

[2] Tessa E F Quax, Nico J Claassens, Dieter Söll, and John van der Oost. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular cell*, 59(2):149–161, July 2015.

A Frequently Asked Questions [FAQ]

1. **“None.”**

   For now!

Modified October 15, 2018