Maria Luisa Santos Moreno Sanches - 111859

Algoritmos em Bioinformática: Conversão de sequências de DNA para RNA

São José dos Campos - Brasil

Maio de 2021

Maria Luisa Santos Moreno Sanches - 111859

Algoritmos em Bioinformática: Conversão de sequências de DNA para RNA

Relatório apresentado à Universidade Federal de São Paulo como parte dos requisitos para aprovação na disciplina de Algoritmos em Bioinformática.

Docente: Prof. Dr. Claudio Saburo Shida

Universidade Federal de São Paulo - UNIFESP

Instituto de Ciência e Tecnologia - Campus São José dos Campos

São José dos Campos - Brasil

Maio de 2021

Sumário

1 Assunto e Objetivos

Uma sequência no formato FASTA começa com uma descrição de uma única linha, seguida por linhas de dados de sequência. A linha de descrição é diferenciada dos dados da sequência por um símbolo de maior que ('>') no início.

Uma *string* de DNA é uma *string* formada a partir de um alfabeto contendo as letras ´A´, ´C´, ´G´ e ´T´. Dado uma *string* de DNA t correspondendo a uma fita codificadora, sua *string* de RNA traduzida u é formada pela troca de todas as ocorrências de ´T´em t por ´U´ em u.

O objetivo desta atividade é importar arquivos de nucleotídeos no formato FASTA, manipular a sequência e exportar no formato FASTA, contabilizando o número total de nucleotídeos de cada sequência.

2 Metodologia

Os 2 arquivos disponibilizados estão no formato FASTA, logo é necessário importar a biblioteca *biopython* do *Python*. Para deixar o código genérico, foi adotado a leitura dos arquivos de um diretório como dados de entrada. Abaixo mostra o código correspondente a essa leitura.

```
# All files in entrada/ directory will be used
from os import listdir
from os.path import isfile, join
path = 'entrada/'
files = [f for f in listdir(path) if isfile(join(path, f))]
```

O processo de transcrição permite a formação do RNA mensageiro com base na região codificante do DNA. Computacionalmente falando podemos analisar a transcrição como um processo de modificações em *strings*.

A leitura dos dados de cada arquivo foi realizada pelo comando *SeqIO* e a transcrição foi realizada pelo comando *transcribe()*. Observe que aplicando *sequence.transcribe( )* à variável *record.seq*, a *string* presente no objeto *SeqIO* em *sequence* é transcrita e armazenada. Para retorná-la a forma original é possível aplicar o método *back_transcribe()* a varíavel *sequence* já transcrita.

O código desenvolvido está comentado abaixo, e vale ressaltar que seu desenvolvimento foi baseado na explicação de Diego Mariano em seu livro *Introdução à Programação para Bioinformática com Biopython* (1).

```
# Import parts of Biopython
from Bio import SeqIO
from Bio.Seq import Seq

# For each file in the directory
for file in files:
    # File path to your FASTA file
    path_to_file = 'entrada/' + str(file)
    # Open file with "with" statement to avoid problems with access
    # to original file (in case computer hangs or there will be any other problem)
    with open(path_to_file, mode='r') as handle:
        # Use Biopython's parse function to process individual
        # FASTA records (thus reducing memory footprint)
        for record in SeqIO.parse(handle, 'fasta'):
            # Modifying the information to the new FASTA file
            record.description = "___RNA convertido___" + record.description
            # Original sequence
            sequence = record.seq
            # Transcribed sequence
            record.seq = sequence.transcribe()
            # Saving the file in a FASTA file
            exit_file = "saida/rna_convertido_" + str(file)
            saved_file = SeqIO.write(record, exit_file, 'fasta')
            # Checking if everything went okay
```

```
25              if saved_file!=1: print('Error while writing sequence:  ' + record.id)
26              # Printing the amount of nucleotides
27              amount_of_nucleotides = len(sequence)
28              print(record.description)
29              print('Its sequence contains {} nucleotides.'.format(amount_of_nucleotides))
```

O código na íntegra está disponível no GitHub do autor.

3 Resultados

O genoma do vírus da Dengue contabilizou o total de 10176 nucleotídeos, e o genoma do Aedes aegypti contabilizou o total de 16790 nucleotídeos. Os arquivos de saída gerados no formato FASTA contendo o RNA convertido, se encontram neste link.

Somente com esses dados é possível dizer que a sequência Aedes aegypti é a que possui um número maior de nucleotídeos.

4 Conclusão

A biblioteca *biopython* é muito útil para análises de grandes arquivos, não sendo necessário um grande conhecimento em programação para poder realizar algumas análises, o que tornou a atividade de transcrever o DNA fácil e rápida. Os resultados apresentados foram se mostraram precisos, ainda mais utilizando o método de transcrever da biblioteca *biopython*.

Referências

1 MARIANO D. C. B.; BARROSO, J. R. P. M. . C. T. S. . d. M.-M. R. C. *Introdução à Programação para Bioinformática com Biopython*. 3. ed. North Charleston, SC (EUA): CreateSpace Independent Publishing Platform, 2015. Citado na página 4.