Manipulate Biological Data Using Biostrings Package Exercises(Part 4)

Bioinformatics is an amalgamation Biology and Computer science. Biological Data is manipulated using Computers and Computer software's in Bioinformatics. Biological Data includes DNA; RNA & Proteins. DNA & RNA is made of Nucleotide which are our genetic material in which we are coded.Our Structure and Functions are done by protein, which are build of Amino acids
Here in this we try to manipulate DNA, RNA, Protein strings using Biostring Package
Install Packages
Biostrings

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1

Create an RNA string and find palindromes in the sequence

Exercise 2

Create a DNA string and find palindromes in the sequence

Exercise 3

Create a DNA string and find the dinucleotide frequency of the sequences

Exercise 4

Create an RNA string and find the dinucleotide frequency of the sequences

Learn more about Data Pre-Processing in the online course R Data Pre-Processing & Data Management – Shape your Data!. In this course you will learn how to:

- import data into R in several ways while also beeing able to identify a suitable import tool
- use SQL code within R
- And much more

Exercise 5

Create a DNA string and find the oligonucleotide frequency in the sequences

Exercise 6

Create an RNA string and find the oligonucleotide frequency in the sequences

Exercise 7

Create a DNA string and find the trinucleotide frequency in the sequences

Exercise 8

Create an RNA string and find the trinucleotide frequency in the sequences

Exercise 9

Print amino acid alphabets

Exercise 10

Create an Amino acid string and print the frequency of the amino acid strings in the sequence

---

Manipulate Biological Data Using Biostrings Package Exercises(Part 3)

Bioinformatics is an amalgamation Biology and Computer science.Biological Data is manipulated using Computers and Computer software's in Bioinformatics. Biological Data includes DNA; RNA & Proteins. DNA & RNA is made of Nucleotide which are our genetic material in which we are coded.Our Structure and Functions are done by protein, which are build of Amino acids
In this exercise we try compare between DNAs, RNAs & Amino Acid Sequences to fund out the relationships.
Comparison is done using sequence alignment or sequence comparison techniques.
There are two types of sequence alignment exists.
1.Pairwise alignment
2.Multiple Sequence Alignment

Pairwise alignment refers to comparison between two sequences, where as Multiple Sequence Alignment refers to comparing more than two sequences.

In the exercises below we cover how we can do pairwise alignment using Biostrings package in Bioconductor.

Install Packages

Biostrings

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1

Create two DNA strings and do pairwise alignment using local, global and overlap alignment techniques and print the score.

Exercise 2

Create two DNA strings and do pairwise alignment and write the alignment to an .aln file.

Exercise 3

Create two Amino acid strings and do pairwise alignment

Exercise 4

Create two Amino acid strings and do pairwise alignment using BLOSUM62 substitution matrix.

Exercise 5

Create two Amino acid strings and do pairwise alignment using BLOSUM100 substitution matrix

Exercise 6

Create two Amino acid strings and do pairwise alignment using PAM250 substitution matrix

Exercise 7

Compare between BLOSUM62 substitution matrix of R and that of the NCBI Database using any two amino acid of your choice.

Exercise 8

Do pairwise alignment using Needlemann Wunch Alignment algorithm and print the score, suppress any warnings.

Exercise 9

Create two DNA Strings and translate the same to amino acids and do pairwise alignment between the amino acid sequences

Exercise 10

Create two RNA Strings and translate the same to amino acids and do pairwise alignment between the amino acid sequences

---

Manipulate Biological Data Using Biostrings Package Exercises(Part 2)

Bioinformatics is an amalgamation Biology and Computer science.Biological Data is manipulated using Computers and Computer software's in Bioinformatics. Biological Data includes DNA; RNA & Proteins. DNA & RNA is made of Nucleotide which are our genetic material in which we are coded.Our Structure and Functions are done by protein, which are build of Amino acids

In this exercise we try correlate the relation between DNA, RNA & Protein.

Conversion of DNA to RNA is known as Transcription. DNA/RNA to protein is known as Translation.

Here we also discuss Sequence Alignment Techniques. Sequence Alignment is comparing the similarity between the sequences to check how much the DNA,RNA or Protein are related to each other.
Three are three types of Sequence Alignment
1. Global Alignment
2. Local Alignment
3. Over lap Alignment

In the exercises below we cover how we can Manipulate Biological Data using Biostrings package in Bioconductor.

Install Packages
Biostrings

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1

Create a DNA String and find out the complement of the DNA

Exercise 2

Create a RNA String and find out the complement of the RNA

Exercise 3

Create a DNA string and find the reverse complement of the same.

Exercise 4

Create a RNA string and find the reverse complement of the same.

Exercise 5

Create a DNA string and translate the same into Amino Acids using Standard Genetic codon and print the three letter codon of the amino acids

Exercise 6

Create a DNA string and translate the same into Amino Acids using Standard Genetic codon and print the three letter codon of the amino acids

Exercise 7

Create two DNA Strings and align the sequence using Global Alignment technique and print the score of the alignment

Exercise 8

Create two DNA Strings and align the sequence using Global Alignment technique and print the score of the alignment after specifying your own score for match and mismatch among the sequence

Exercise 9

Create two DNA Strings and align the sequence using Local Alignment technique and print the score of the alignment

Exercise 10

Create two DNA Strings and align the sequence using Overlap Alignment technique and print the score of the alignment

---

Manipulate Biological Data Using Biostrings Package Exercises (Part 1)

Bioinformatics is an amalgamation Biology and Computer science.Biological Data is manipulated using Computers and Computer software's in Bioinformatics. Biological Data includes DNA; RNA & Protiens. DNA & RNA is made of Nucleotides which are our genetic material in which we are coded.Our Structure and Functions are done by protein, which are build of Amino acids

In the exercises below we cover how we can Manipulate Biological Data using Biostrings package in Bioconductor.

Install Packages
Biostrings

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed

on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1

Print out the standard Genetic Code table using Biostrings Package

Exercise 2

Print the first codon in the standard genetic code

Exercise 3

Print out the Standard RNA Genetic Code table using Biostrings package

Exercise 4

Print out the Standard RNA Genetic Code of Stop codon using Biostrings package

Exercise 5

Print out the standard Amino acid codon table using Biostrings

Exercise 6

Print the code of the start codon from the standard genetic code

Exercise 7

Print the three letter code of Amino acid Methionine using Biostrings

Exercise 8

Create a DNA string and print the length and dinucleotide frequency of the string

Exercise 9

Create RNA string and print the length and dinucleotide frequency of the string

Exercise 10

Create a Protein string and print the length of the protein

Accessing and Manipulating Biological Databases Exercises (Part-3)

In the exercises below we cover how we can Manipulate Biological Data using Seqinr packages

Install Packages

seqinr
Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1
Read a Fasta File in your current directory and print the complement of the sequence
Exercise 2
List the Standard codon tables available for translation of the nucleotide sequences
Exercise 3
List the Vertebrate Mitochondrial codon tables available for translation of the nucleotide sequences
Exercise 4
Find out the synonym codons of a any three letter codon in the standard codon table

Exercise 5
Read a Fasta File in your current directory and translate into amino acids using Vertebrate Mitochondrial codon table
Exercise 6
Read a Fasta File in your current directory and do frame 1 translation into amino acid sequences.
Exercise 7
Read a Fasta File in your current directory and do frame 1

translation into amino acid sequences and create a string of amino acid sequences from the vector

Exercise 8

Read a Fasta File in your current directory and do translation into amino acid sequences and find the Iso Electric Point of the translated sequence

Exercise 9

Read a Fasta File in your current directory and do frame 1 translation into amino acid sequences and find the Molecular Weight of the translated sequence

Exercise 10

Open the Nucleotide Fasta file and translate the sequences to Amino acids and plot a graph based on different categories of amino acids in the sequence.

Accessing and Manipulating Biological Databases Exercises (Part 2)

In the exercises below we cover how we can Access and Manipulate Biological Data bases through rentrez & Seqinr packages

Install Packages
rentrez

seqinr

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1
Read a Fasta File in your current directory and print the sequence

Exercise 2

Read a Fasta File in your current directory and print the length of the sequences

Exercise 3

Read a Fasta File in your current directory and count each nucleotide in the file.

Exercise 4

Read a Fasta File in your current directory and print the details of the sequences
Exercise 5

Read a Fasta File in your current directory and count all dinucleotides in the file.

Exercise 6

Read a Fasta File in your current directory and print the GC contents

Exercise 7

Read a Fasta File in your current directory and print sequences as characters.

Exercise 8

Open the Nucleotide Fasta file and translate the sequences to Amino acids in Forward Translation

Exercise 9

Open the Nucleotide Fasta file and translate the sequences to Amino acids in Reverse Translation

Exercise 10

Open the Nucleotide Fasta file and translate the sequences to Amino acids and print the three letter codons of the translated amino acids

---

Accessing and Manipulating Biological Databases Exercises (Part 1)

In the exercises below we cover how we can Access and Manipulate Biological Data bases through rentrez & seqinr packages

Install Packages
rentrez
seqinr

Answers to the exercises are available here

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1

Print all the available data bases which you can access through rentrez package

Exercise 2

Print all the searchable terms in a database

Exercise 3

Display the details of any database of your choice

Exercise 4

Retrieve and print 10 ids of nucleotide sequences from nuccore database about Human.

Exercise 5

Retrieve and print 20 ids of protein sequences from protein database about Human.

Exercise 6

Create a Fasta File for a particular human protein sequence from the listed ids.

Exercise 7

Create a Fasta File for a particular human nucleotide sequence from the listed ids.

Exercise 8

Open the Nucleotide Fasta file and print the details using seqinr package.

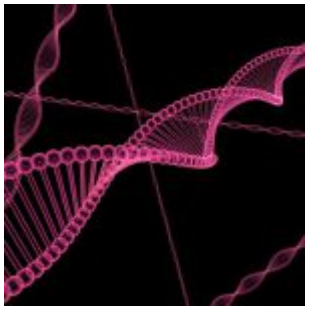
Exercise 9

Open the Protein Fasta file and print the details using seqinr package

Exercise 10

Open the Nucleotide Fasta file and print only sequence from the created Fasta file striping all other information.

Bioinformatics Tutorial with Exercises in R (part 1)



Bioinformatics is an interdisciplinary field of study that combines the field of biology with computer science to understand biological data. Bioinformatics is generally used in laboratories as an initial or final step to get the information. This information can subsequently be utilized for the wet lab practices. However, it can also be used as a standalone computational method to get the answers to relevant biological questions. Bioinformatics has greatly accelerated the discovery process and is used in both academics as well as industry to understand genomics & proteomics data, determine the protein structure, RNA structures, analyzing sequence data for molecular biological work, gene identification, to name a few applications. Over the years it has grown into a complex field of discipline requiring specific education and professional training.

We begin with a simple introduction in sequence data. This tutorial assumes that readers have basic knowledge about the central dogma and biological molecules such as protein, DNA and RNA; and also understand that they are biopolymers naturally or synthetically produced for a specific role. Wikipedia is a great resource to get the basic introduction on these. However, I will also write a few sentences in my exercises to refresh the knowledge of the readers. Numerous programming languages have been used for the development of Bioinformatics, ranging from low end languages such as Java, C/C++ to high end scripting languages such as perl, python and the R statistical language. Bioinformatics also involves

extensive database management implementation for storage, query and updating the sequence and numerical data. Most of the Bioinformatics software can be implemented either on a Windows, Mac or Linux platform. This tutorial also assumes that the reader has some understanding about R programming, RStudio and installation of packages. We will use numerous packages both common as well as strictly developed for Bioinformatics.

The open source community known as Bioconductor specifically develops the Bioinformatics tools using R for the analysis and comprehension of high-throughput genomic data. It boasts to have two releases each year, 1296 software packages, and an active user community. It has packages developed for application ranging from basic sequence alignment to recent years' NextGen sequencing machines. We hope that we will able to discuss more about it in later chapters, when not only readers would be able to appreciate its relevance but also utilize it in their daily examples. Let us begin with some little steps by installing a few packages on our machines. I personally prefer to begin with a fresh installation to make sure their are no conflicts from the preinstalled packages.

Learn more about Bioconductor in the (Free!) online courses Introduction to Bioconductor: Annotation and Analysis of Genomes and Genomics Assays and Case Studies in Functional Genomics offered by Harvard University.
To begin with let us first install a package known as Biostrings by running the following command on your RStudio.

```
source("https://bioconductor.org/biocLite.R")
biocLite("Biostrings")
```

This would take up to 5 minutes depending upon your internet connection and computer. Note that typing install.packages("biostrings") in RStudio might not result

into success because of version issues. As usual use library(Biostrings) command to load the package.

Deoxyribonucleic acid, or DNA, stores biological information which is expressed in form of RNA inside the cells. The two antiparallel DNA strands are termed polynucleotides and they are composed of simpler monomer units called nucleotides. Each nucleotide is composed of one of four nitrogen-containing nucleobases—either cytosine (C), guanine (G), adenine (A), or thymine (T). The randomly arranged nucleotide finally write up the code of life in the form of strict vocabulary.

Answers to the exercises are available here.

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

Exercise 1
Using the following command, create a list of DNA exactly the same as the one given below, then list the outcome by writing the vector myDNA.

```
myDNA <- c("ATGTTGCATTCATTAATTAAGAACGACCCAATACA")
myDNA
```

Exercise 2
In recent years there has been exponential advancement in the field of DNA sequencing. High throughput methods have been developed to speed up the sequencing project. Still the basic DNA sequencing is a technology that reads each nucleotide step-by-step by chemical methods to decipher what order of letters A, T, G and C were placed that resulted into the specific DNA sequence. However, for this problem we have given you a short result from sequencing data. In a typical laboratory sequencing results, some of the DNA sequence could look something like this “CTGATTT-GATGGTC-NAT” where apart from ATGC we could see some dashes (skipped) and N (unknown) nucleotide. Copy these into mySequencing and print the result.

Use the length() command to find out the length. Which of the following is the length

A [1] 1
B [2] 19
C [1] 0

Exercise 3
Now use the package specific command called DNAString to copy the seq

myDNASeq <- DNAString("CTGATTT-GATGGTC-NAT")
Now use length command to find the length of the myDNASeq. Compare the difference between mySequencing and myDNASeq .

Exercise 4
Use the class() command to analyze the datatype of the myDNA and myDNASeq. Which of the following represents the result

A
class(myDNASeq)
[1] "DNAString"
attr(,"package")
[1] "Biostrings"
> class(myDNA)
[1] "integr"
B

> class(myDNASeq)
[1] "DNAString"
attr(,"package")
[1] "Biostrings"
> class(myDNA)
[1] "character"

C None of the above
D Both

Exercise 5

In this problem let us attempt to paste the myDNA and myDNASeq sequences using paste and analyse the results. Try to understand the syntax of paste command using ?paste command. Copy the result into pastedDNA and print it by typing pastedDNA .
Try to analyze the pastedDNA using class command.

Exercise 6
Did you notice that final result was a character class and not the Biostring as expected? This would make pastedDNA not usable for biostring for any purpose. What happens is that Biostrings introduces a new data structure hierarchy which is different than the vector datatype of R. It has few sequence constructors such as DNAString(), RNAString(), AAString() for type of biomolecules (DNA, RNA and Amino Acids). This results into the fact that Biostrings objects cannot be compared with R strings (myDNA == myDNASeq is an invalid command). You will always end up with FALSE upon attempting to compare.
Next we would look at some basic transformations in Biostrings that can be implemented on DNA data. These transformations are reverse() , complement() , reverseComplement() and translate() . Run the myDNASeq <- DNAString("CTGATTT-GATGGTC-NAT") again. Run the complement(myDNASeq) . Analyze the data.

Exercise 7
Did you find out what as happened in previous problem? This just created the complement of each nucleotide. The DNA usually exists as a double helix with both strands running antiparallel. Each base of ATGC is paired with some base on complementary strand. A has preference for T (and vise versa), G has preference for C (and vise versa). So the complementary strand is going to have the nucleotide that is pairs preferentially. However, the unknown nucleotide N gets written as N because the sequencer could not tell what it was.
In next problem run reverse(myDNASeq)

Exercise 8
Did you notice what has happened? Did nucletide in DNA

sequences got read from back to front and not front to back? Or not. Each complementary strand is usually written from back to front because the complementary strands are anti-parallel. This because the sign of each strand is opposite. This is basically because the strand which basically is a polymer of nucleotide either has 5’ or 3’ end (more about it later). So the +ve strand runs from 5’-3’ and -ve strand runs from 3’-5’. In the next problem run reverseComplement(myDNASeq)

Exercise 9

Did you notice what happened. The DNA sequences not only got complementary but also reversed. This is how double stranded DNA exists in nature.
Run alphabetFrequency(myDNASeq)

Exercise 10

The values in problem 9 gave you the frequency of occurrence of each nucleotide in this DNA. This is an important thing to know when analyzing the DNA.
Finally run translate(myDNASeq) . This would yield the hypothetical protein sequence that myDNASeq would produce. Afterall that is one of the important role of DNA, to code for the protein. What did you get?

In our next exercises we will work little more with Biostrings to analyze the DNA at little more. Happy learning.