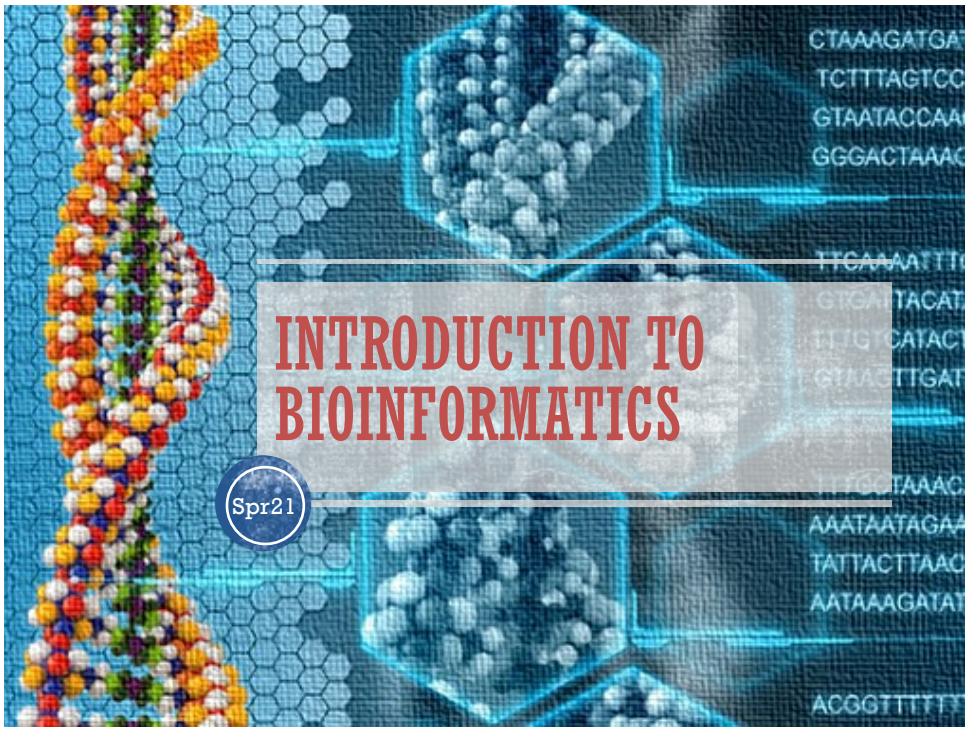


1

Understand how biological molecules are represented and stored in the computer.

Dr. Khalifa, Spr21

2

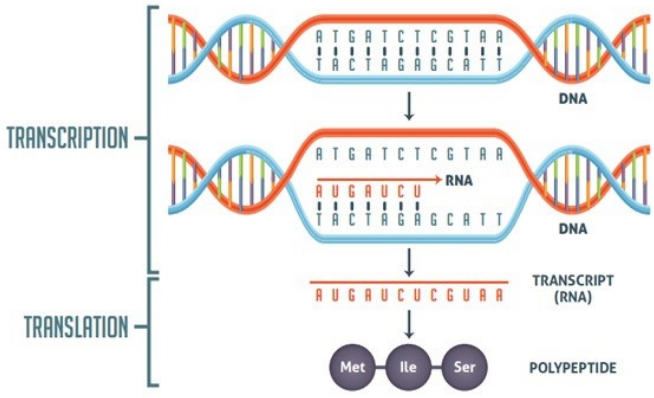
- DNA data
- Sequences as strings
- Data Storage and formats
- Database searching
- The Bioconductor package
- Reading/writing sequence data

Agenda

Sequence Data

3

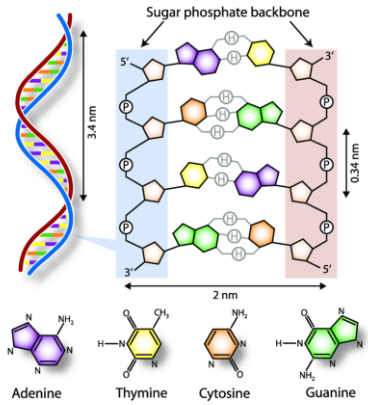
Sequence data



4

DNA sequence data

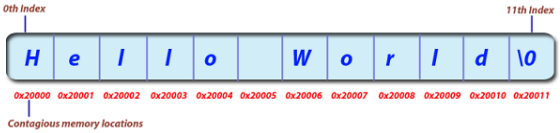
- How do we represent the crazily complex biochemical structure of DNA in the computer?
- Simplify!
  - Flatten the structure and zoom in
  - Focus on the four bases
- Use Letters for data storage
  - The whole human genome ~= 3 billion letters ~= 6 MB of data.



5

String data structure

- Σ ➔ a finite alphabet consisting of a set of characters (or symbols).
- The cardinality of the alphabet denoted by | Σ |
  - Expresses the number of distinct characters in the alphabet.
- A string or word (*w*) is an ordered list of zero or more characters drawn from the alphabet.
  - w[1 · · · n] = w[1]w[2] · · ·,w[n] , where w[i] ∈ Σ for 1 ≤ i ≤ n
  - |w| denotes the length of w.



6

Sequences as Strings

- The basic types of DNA, RNA, and protein molecules can be represented as strings
- DNA are strings over the alphabet {*A,C, G, T*}
  - four bases adenine, cytosine, guanine, and thymine
- RNA are strings over the alphabet {*A,C, G,U*}
  - uracil replacing thymine
- Proteins are strings over an alphabet of the 20 amino acids

20 natural amino acid notation

Amino Acid	3-Letter[4]	1-Letter[4]
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Data Storage and formats

- Types of text-based format for representing biological sequences (DNA, RNA and protein):
  - Raw Sequence: Data without description.
  - FASTA Format: One line of description, then sequence.
  - GenBank Record: Lots of detailed description about the sequence.

FASTA Example

- Simple and widely used!
- begins with a single-line description, followed by lines of sequence data.
- The description starts with a greater-than (">") symbol in the first column.
  - Multiple Entries
- It is recommended that all lines of text be shorter than 80 characters in length.

>PF00181|NF01243182 50S ribosomal protein L2 [Coxiella burnetii]
MALVKTKPTSPGRRFVVKVVHPELHKGDPYAPLVESKNR
INSRNNQGRITVRRRGGGHKRNYRIIDFKRDKEGIEGKVE
RLEYDPNRSAHIALVLYPDGERRYIIAPKGVHKGSKVVSG
REAPIRPGNCLPLQNIPLGATIHNIELKPGKGAQLVRSAGA
SAQLAAKEGIYAIIRMRSGETRKILAVCRACIGEVSNSEHN
LRSLGKAGAKRWRGRRPTVRGVAMNPVDHPHGGGEGK
TSGGRHPVSPTGKPTKGYKTRRNKRTSNMIIRDRRKK

https://zhanglab.ccmb.med.umich.edu/FASTA/

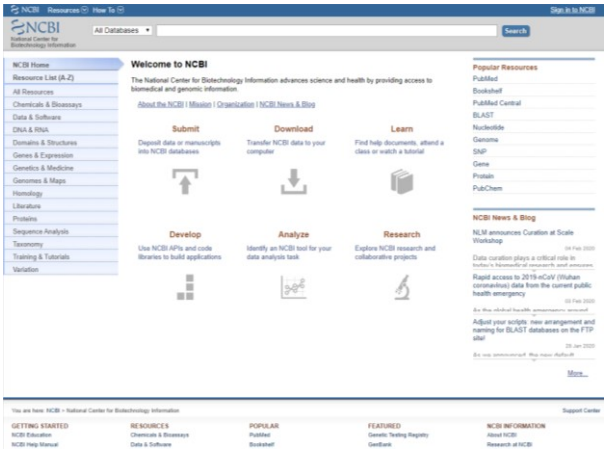
9

Biological Databases

- US, The National Centre for Biotechnology Information (NCBI) sequence database www.ncbi.nlm.nih.gov
- Europe, the European Molecular Biology Laboratory (EMBL) Sequence Database www.ebi.ac.uk/embl
- Japan, the DNA Data Bank of Japan (DDBJ) www.ddbj.nig.ac.jp.
- These three databases exchange data every night, so at any one point in time, they contain almost identical data.
- Each sequence is stored in a separate record and is assigned a unique identifier that can be used to refer to that sequence record.
  - The identifier is known as an accession and consists of a mixture of numbers and letters.
  - Different databases have different accessions, as they each use their own numbering systems for referring to their own sequence records.

10

The NCBI Database



https://www.kelleybioinfo.org/algorithms/basics/databases/ncbi.pdf

GenBank Record

Dengue virus 1, complete genome

NCBI Reference Sequence: NC_001477.1

FASTA Graphics

Go to:

```
LOCUS       NC_001477              10179 bp ss-RNA     linear   VRL 03-MAY-2019
DEFINITION  Dengue virus 1, complete genome.
ACCESSION   NC_001477 REGION: 95..10273
VERSION     NC_001477.1
DBLINK      BioProject: PRJNA485481
KEYWORDS    RefSeq.
SOURCE      Dengue virus 1
  ORGANISM  Dengue virus 1
            Viruses; Riboviria; Orthornavirae; Kitrinoviricota; Flasuviricetes;
            Amarillovirales; Flaviviridae; Flavivirus.
REFERENCE   1  (bases 1 to 10179)
  AUTHORS   Puri,B., Nelson,W.M., Henchal,E.A., Hoke,C.H., Eckels,K.H.,
            Dubois,D.R., Porter,K.R. and Hayes,C.G.
  TITLE     Molecular analysis of dengue virus attenuation after serial passage
            in primary dog kidney cells
  JOURNAL   J. Gen. Virol. 78 (PT 9), 2287-2291 (1997)
   PUBMED   9292016
REFERENCE   2  (bases 1 to 10179)
  AUTHORS   McKee,K.T. Jr., Bancroft,W.H., Eckels,K.H., Redfield,R.R.,
            Summers,P.L. and Russell,P.K.
  TITLE     Lack of attenuation of a candidate dengue 1 vaccine (45AZ5) in
            human volunteers
  JOURNAL   Am. J. Trop. Med. Hyg. 36 (2), 435-442 (1987)
   PUBMED   3826504
```

Open NCBI database website

Search and download the DNA sequence for the DEN-1 with accession: NC_001477

Search and download the DNA sequence for the DEN-2 with accession : NC_001474

Search and download the DNA sequence for the DEN-3 with accession NC_001475

Exercise 1

The Dengue virus causes Dengue fever, a neglected tropical disease

13

BioConductor

- the Bioconductor set of R packages (www.bioconductor.org) contains several packages with many R functions for analyzing biological data sets.
- Bioconductor has a particular approach to making packages available.
- Each six months, in spring and fall, the current 'devel' version of packages is branched to become the next 'release'.
  - Packages within a release are tested with one another, so it is important to install packages from the same release.
- The first step to package installation is to make sure that the BiocManager package has been installed using standard R procedures.

```
if (!require(BiocManager)) install.packages("BiocManager",
repos = "https://cran.r-project.org")

BiocManager::install(c("Biostrings", "GenomicRanges"))
```


14

Reading a sequence File in R

- the SeqinR package contains R functions for obtaining sequences from DNA and protein sequence databases, and for analyzing DNA and protein sequences.
- Steps:
  - Download the FASTA file from database website
  - Load the seqinr package
  - Read the FASTA file from its location

```
mysequence <- read.fasta(file = "myfasta.fasta")
```

  - Access sequence data using `getSequence` command
    - the first element of the list object contains the DNA sequence

15

Load and `seqinR` package

Read the DEN-1 sequence

Use getSequence function

Print the first 10 bases in the sequence

Exercise 2

The Dengue virus causes Dengue fever, a neglected tropical disease

16

Retrieving a sequence in R

- Retrieving sequences from databases requires the following:
  - The `seqinr` library installed and loaded in the R session
  - A sequence ID or keyword for searching
  - Access to the database via the Internet
- R uses the ACNUC database:
  - it brings together data from various different sources
  - organized into various different ACNUC (sub)-databases
    - To List all sub-databases

```
choosebank()
```


Retrieving a sequence in R - Steps

1. Choose the data bank

```
choosebank( source )
```

2. query the bank with the search of your interest, using accession number

```
Q<-query( "Q", "AC=########")
attributes(Q)
```

3. Fetch a specific sequence from the query object

```
Seq<- getSequence(Q$req[[#]] )
```

4. Find more info.

```
Access <- getName(Q$req[[#]])
Annot <- getAnnot(Q$req[[#]])
```

5. close the data bank

```
closebank()
```

Connect to the refseqViruses NCBI database

Search and download the DNA sequence for the DEN-1 with accession: NC_001477

How many sequences were retrieved?

prints out the first 50 nucleotides

Exercise 3

The Dengue virus causes Dengue fever, a neglected tropical disease

19

All in one function!

- A little book of R, Page#15

```
> getncbiseq <- function(accession)
  {
     require("seqinr") # this function requires the SeqinR R package
     # first find which ACNUC database the accession is stored in:
     dbs <- c("genbank","refseq","refseqViruses","bacterial")
     numdbs <- length(dbs)
     for (i in 1:numdbs)
     {
        db <- dbs[i]
        choosebank(db)
        # check if the sequence is in ACNUC database 'db':
        resquery <- try(query(".tmpquery", paste("AC=", accession)), silent = TRUE)
        if (!(inherits(resquery, "try-error")))
        {
           queryname <- "query2"
           thequery <- paste("AC=",accession,sep="")
           query(`queryname`,`thequery`)
           # see if a sequence was retrieved:
           seq <- getSequence(query2$req[[1]])
           closebank()
           return(seq)
        }
        closebank()
     }
     print(paste("ERROR: accession",accession,"was not found"))
  }
```

Error!
Find it, fix it!

20

Complex queries

- Search for a sequence by a particular :
  - NCBI accession → "AC=" argument
  - Type (DNA or mRNA) → "M=" argument
  - organism or taxon → "SP=" argument
  - Journal publication → "R=Jor/vol/page"
  - Find more on "`query()`" function help page

Connect to the ACNUC "genbank"

search for mRNA sequences from the parasitic worm "Schistosoma mansoni"

How many sequences were retrieved?

What are the accession numbers of the first tow?

write the 10th sequence data to a FASTA-format file

Exercise 4

The Dengue virus causes Dengue fever, a neglected tropical disease

Export data to FASTA

- You can write out a sequence to a FASTA-format file in R by using the `write.fasta()` function from the SeqinR R package.

```
write.fasta(myseqs, mynames, file.out = "Myfile.fasta")
```

- The arguments are:
  - the name of the output file using the "file.out" argument
  - the R variable that contains the sequence using the "sequences" argument

    `myseqs <- getSequence(Q)`
  - the name that you want to give to the sequence using the "names" argument.

    `mynames <- getName(Q)`

23

Questions?

- A Little Book of R For Bioinformatics : P#11 -17
- SeqinR manual: https://cran.r-project.org/web/packages/seqinr/seqinr.pdf

24