INTRODUCTION TO BIOCONDUCTOR

Introduction to Biostrings

Paula Andrea Martinez, PhD.
Data Scientist

Biological string containers

- Memory efficient to store and manipulate sequence of characters
- Containers that can be inherited

For example:

- The BString class comes from *big string*

```
showClass("XString")

showClass("BString")

showClass("BStringSet")
```

Biostring alphabets

```
DNA_BASES  # DNA 4 bases
```

```
[1] "A" "C" "G" "T"
```

```
RNA_BASES # RNA 4 bases
```

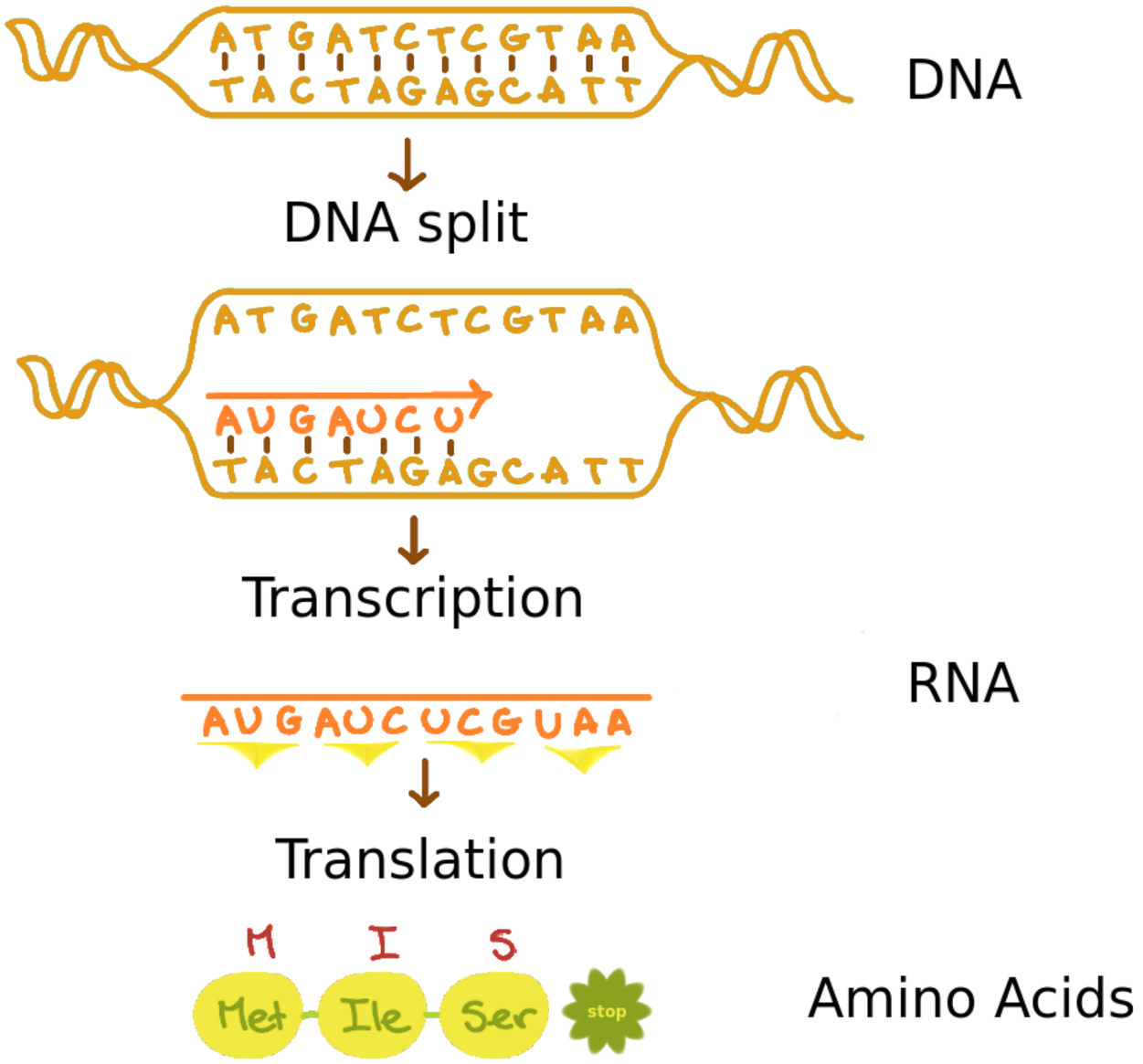
```
[1] "A" "C" "G" "U"
```

```
AA_STANDARD # 20 Amino acids
```

```
 [1] "A" "R" "N" "D" "C" "Q" "E" "G" "H" "I"
[11] "L" "K" "M" "F" "P" "S" "T" "W" "Y" "V"
```

```
DNA_ALPHABET # contains IUPAC_CODE_MAP
RNA_ALPHABET # contains IUPAC_CODE_MAP
AA_ALPHABET  # contains AMINO_ACID_CODE
```

[1] For more information IUPAC DNA codes http://genome.ucsc.edu/goldenPath/help/iupac.html


ATGATCTCGTAA
TACTAGAGCATT
DNA
DNA split
ATGATCTCGTAA
AUGAUCU
TACTAGAGCATT
Transcription
AUGAUCUCGUAA
RNA
Translation
M
I
S
Met
Ile
Ser
stop
Amino Acids

Transcription DNA to RNA

```
# DNA single string
dna_seq <- DNAString("ATGATCTCGTAA")
dna_seq
```

```
  12-letter "DNAString" instance
seq: ATGATCTCGTAA
```

```
# Transcription DNA to RNA string
rna_seq <- RNAString(dna_seq)
rna_seq
```

```
  12-letter "RNAString" instance
seq: AUGAUCUCGUAA
```

Translation RNA to amino acids

```
RNA_GENETIC_CODE
```

```
rna_seq
```

```
  12-letter "RNAString" instance
seq: AUGAUCUCGUAA
```

```
# Translation RNA to AA
aa_seq <- translate(rna_seq)
aa_seq
```

Three RNA bases form one AA: `AUG = M, AUC = I, UCG = S, UAA = *`

```
 4-letter "AAString" instance
seq: MIS*
```

Shortcut translate DNA to amino acids

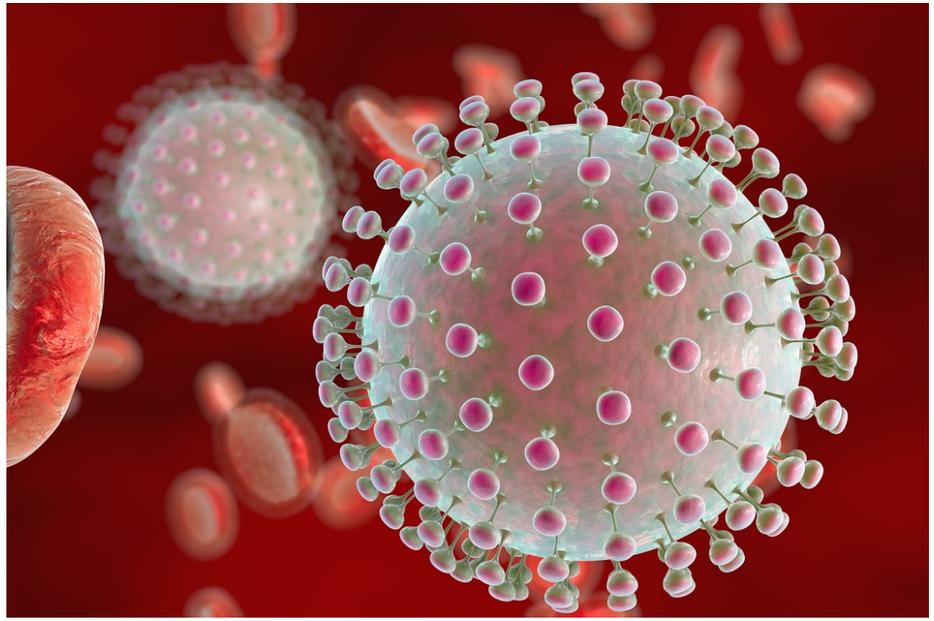
```
dna_seq
```

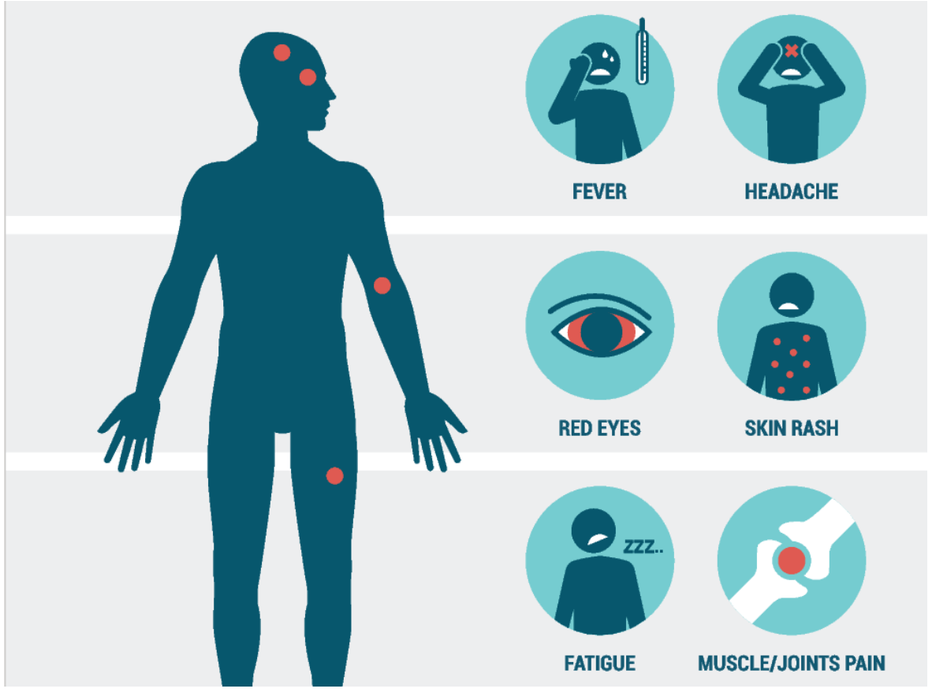
```
 12-letter "DNAString" instance
seq: ATGATCTCGTAA
```

```
# translate() also goes directly from DNA to AA
translate(dna_seq)
```

```
 4-letter "AAString" instance
seq: MIS* # Same result as before
```

The Zika virus





INTRODUCTION TO BIOCONDUCTOR

Let's practice with the Zika virus!

INTRODUCTION TO BIOCONDUCTOR

Sequence handling

Paula Andrea Martinez, PhD.
Data Scientist

Single vs set

- **XString** to store a **single** sequence
  - BString for any string
  - DNAString for DNA
  - RNAString for RNA
  - AAString for amino acids
- **XStringSet** for **many** sequences
  - BStringSet
  - DNAStringSet
  - RNAStringSet
  - AAStringSet

Create a stringSet and collate it

```
# read the sequence as a set
zikaVirus <- readDNAStringSet("data/zika.fa")
```

```
length(zikaVirus)  # the set contains only one sequence
[1] 1
width(zikaVirus)   # and width 10794 bases
[1] 10794
```

```
# to collate the sequence use unlist
zikaVirus_seq <- unlist(zikaVirus)

length(zikaVirus_seq) # A 10794-letter "DNAString" instance
[1] 10794

width(zikaVirus_seq)
# Error unable to find width for "DNAString"
```

From a single sequence to a set

```
# to create a new set from a single sequence
zikaSet <- DNAStringSet(zikaVirus_seq,
                        start = c(1, 101, 201),
                        end = c(100, 200, 300))
zikaSet
```

```
  A DNAStringSet instance of length 3
    width seq
[1]   100 AGTTGTTGATCTGTGTGAGTCAGACT...AATTTGGATTTGGAAACGAGAGTTT
[2]   100 CTGGTCATGAAAAACCCCAAAGAAGA...GTAAACCCCTTGGGAGGTTTGAAGA
[3]   100 GGTTGCCAGCCGGACTTCTGCTGGGT...CAGCAATCAAGCCATCACTGGGCCT
```

```
length(zikaSet)
[1] 3
width(zikaSet)
[1] 100 100 100
```

Complement sequence


```
a_seq <- DNAString("ATGATCTCGTAA")
a_seq

  12-letter "DNAString" instance
seq: ATGATCTCGTAA
```

```
complement(a_seq)

  12-letter "DNAString" instance
seq: TACTAGAGCATT
```

Rev a sequence

```
zikaShortSet
```

```
 A DNAStringSet instance of length 2
width seq                       names
[1]    18 AGTTGTTGATCTGTGTGA     seq1
[2]    18 CTGGTCATGAAAAACCCC     seq2
```

```
rev(zikaShortSet)
```

```
 A DNAStringSet instance of length 2
width seq                       names
[1]    18 CTGGTCATGAAAAACCCC     seq2
[2]    18 AGTTGTTGATCTGTGTGA     seq1
```

Reverse a sequence

```
zikaShortSet
```

```
 A DNAStringSet instance of length 2
width seq                         names
[1]    18 AGTTGTTGATCTGTGTGA        seq1
[2]    18 CTGGTCATGAAAAACCCC        seq2
```

```
reverse(zikaShortSet)
```

```
 A DNAStringSet instance of length 2
width seq                         names
[1]    18 AGTGTGTCTAGTTGTTGA        seq1
[2]    18 CCCCAAAAAGTACTGGTC        seq2
```

Reverse complement

```
# Original rna_seq sequence
 8-letter "RNAString" instance
seq: AGUUGUUG
```

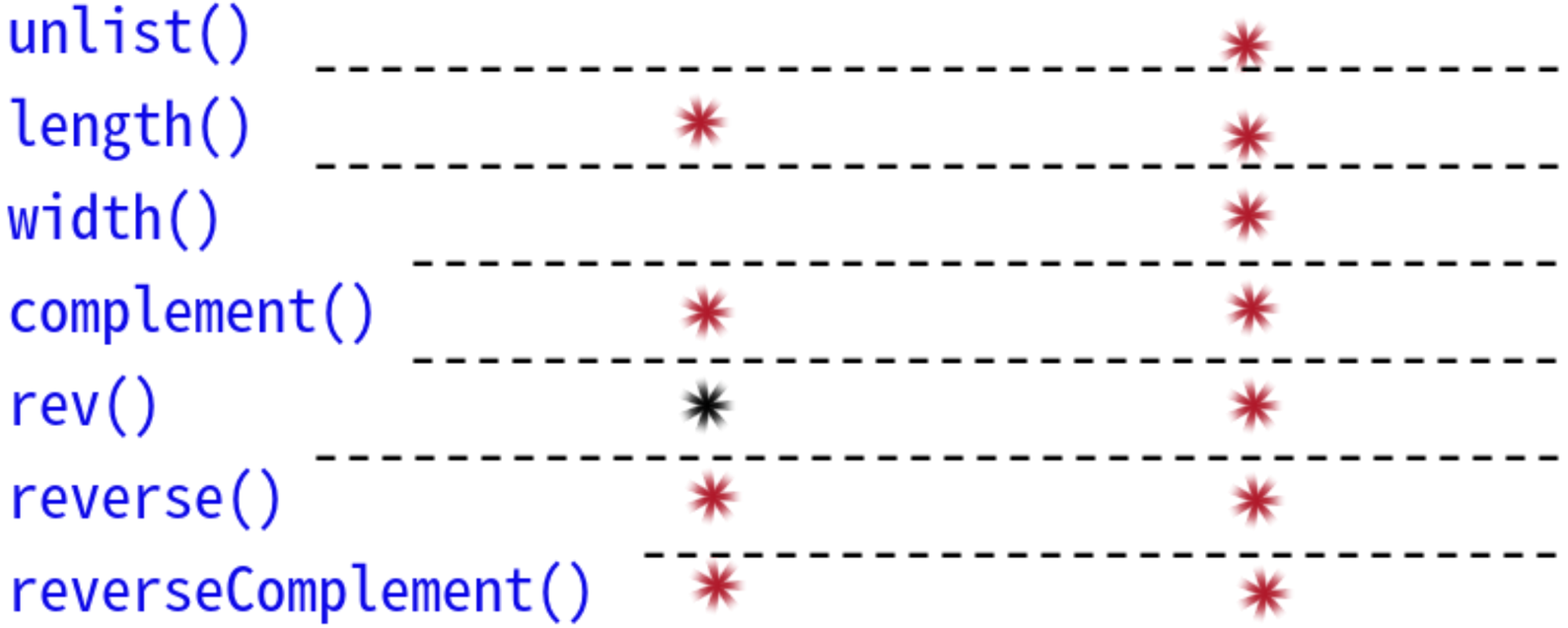
```
reverseComplement(rna_seq)
```

```
  8-letter "RNAString" instance
seq: CAACAACU
```

```
# Using two functions together
reverse(complement(rna_seq))
```

```
  8-letter "RNAString" instance
seq: CAACAACU
```

Single sequence	Set of sequences
XString	XStringSet
ATCGGTAC	ATCGGTAC CCGTAACTT CTTATCGAA



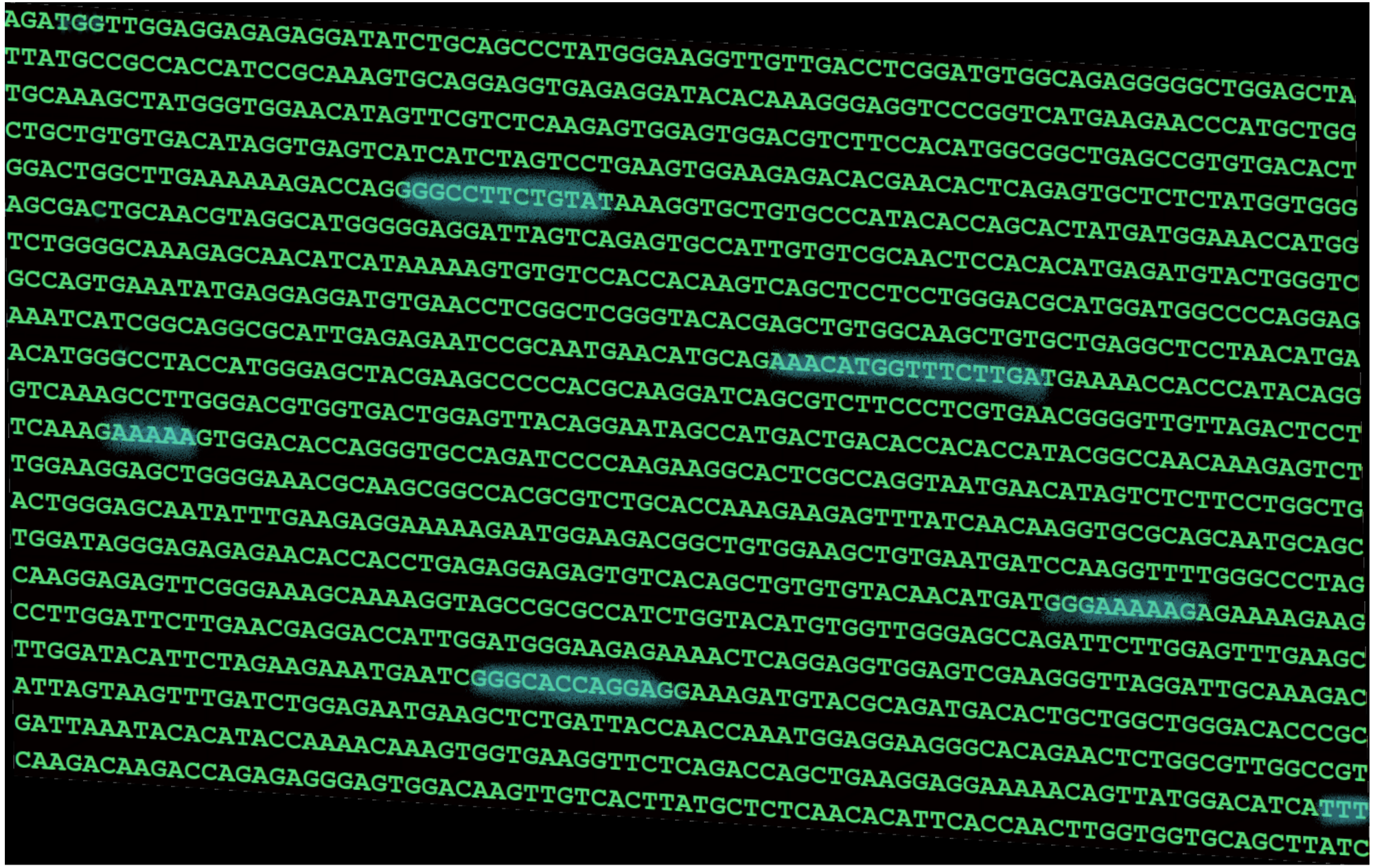
INTRODUCTION TO BIOCONDUCTOR

Let's practice sequence handling!

INTRODUCTION TO BIOCONDUCTOR

Why are we interested in patterns?

Paula Andrea Martinez, PhD.
Data Scientist


AGATGGTTGGAGGAGAGAGGATATCTGCAGCCCTATGGGAAGGTTGTTGACCTCGGATGTGGCAGAGGGGGCTGGAGCTA
TTATGCCGCCACCATCCGCAAAGTGCAGGAGGTGAGAGGATACACAAAGGGAGGTCCCGGTCATGAAGAACCCATGCTGG
TGCAAAGCTATGGGTGGAACATAGTTCGTCTCAAGAGTGGAGTGGACGTCTTCCACATGGCGGCTGAGCCGTGTGACACT
CTGCTGTGTGACATAGGTGAGTCATCATCTAGTCCTGAAGTGGAAGAGACACGAACACTCAGAGTGCTCTCTATGGTGGG
GGACTGGCTTGAAAAAAGACCAGGGGCCTTCTGTATAAAGGTGCTGTGCCCATACACCAGCACTATGATGGAAACCATGG
AGCGACTGCAACGTAGGCATGGGGGAGGATTAGTCAGAGTGCCATTGTGTCGCAACTCCACACATGAGATGTACTGGGTC
TCTGGGGCAAAGAGCAACATCATAAAAAGTGTGTCCACCACAAGTCAGCTCCTCCTGGGACGCATGGATGGCCCCAGGAG
GCCAGTGAAATATGAGGAGGATGTGAACCTCGGCTCGGGTACACGAGCTGTGGCAAGCTGTGCTGAGGCTCCTAACATGA
AAATCATCGGCAGGCGCATTGAGAGAATCCGCAATGAACATGCAGAAACATGGTTTCTTGATGAAAACCACCCATACAGG
ACATGGGCCTACCATGGGAGCTACGAAGCCCCCACGCAAGGATCAGCGTCTTCCCTCGTGAACGGGGTTGTTAGACTCCT
GTCAAAGCCTTGGGACGTGGTGACTGGAGTTACAGGAATAGCCATGACTGACACCACACCATACGGCCAACAAAGAGTCT
TCAAAGAAAAAGTGGACACCAGGGTGCCAGATCCCCAAGAAGGCACTCGCCAGGTAATGAACATAGTCTCTTCCTGGCTG
TGGAAGGAGCTGGGGAAACGCAAGCGGCCACGCGTCTGCACCAAAGAAGAGTTTATCAACAAGGTGCGCAGCAATGCAGC
ACTGGGAGCAATATTTGAAGAGGAAAAAGAATGGAAGACGGCTGTGGAAGCTGTGAATGATCCAAGGTTTTGGGCCCTAG
TGGATAGGGAGAGAGAACACCACCTGAGAGGAGAGTGTCACAGCTGTGTGTACAACATGATGGGAAAAAGAGAAAAGAAG
CAAGGAGAGTTCGGGAAAGCAAAAGGTAGCCGCGCCATCTGGTACATGTGGTTGGGAGCCAGATTCTTGGAGTTTGAAGC
CCTTGGATTCTTGAACGAGGACCATTGGATGGGAAGAGAAAACTCAGGAGGTGGAGTCGAAGGGTTAGGATTGCAAAGAC
TTGGATACATTCTAGAAGAAATGAATCGGGCACCAGGAGGAAAGATGTACGCAGATGACACTGCTGGCTGGGACACCCGC
ATTAGTAAGTTTGATCTGGAGAATGAAGCTCTGATTACCAACCAAATGGAGGAAGGGCACAGAACTCTGGCGTTGGCCGT
GATTAAATACACATACCAAAACAAAGTGGTGAAGGTTCTCAGACCAGCTGAAGGAGGAAAAACAGTTATGGACATCATTT
CAAGACAAGACCAGAGAGGGAGTGGACAAGTTGTCACTTATGCTCTCAACACATTCACCAACTTGGTGGTGCAGCTTATC

What can we find with patterns?

- Gene start
- Protein end
- Regions that enhance or silence gene expression
- Conserved regions between organisms
- Genetic variation

Pattern matching

- `matchPattern(pattern, subject)`
  - 1 string to 1 string
- `vmatchPattern(pattern, subject)`
  - 1 set of strings to 1 string
  - 1 string to a set of strings

Palindromes


```
findPalindromes() # find palindromic regions in a single sequence
```

Not new biology

- The Genetic code was first described by Nirenberg in 1963 On the coding of genetic information Nirenberg, Marshall et al. Cold Spring Harb Symp Quant Biol 1963, 28

- How translation might differ according to the reading frame, was first described by Streisinger in 1966 Frameshift Mutations and the Genetic Code Streisinger, George et al. Cold Spring Harb Symp Quant Biol 1966, 31: 77-84

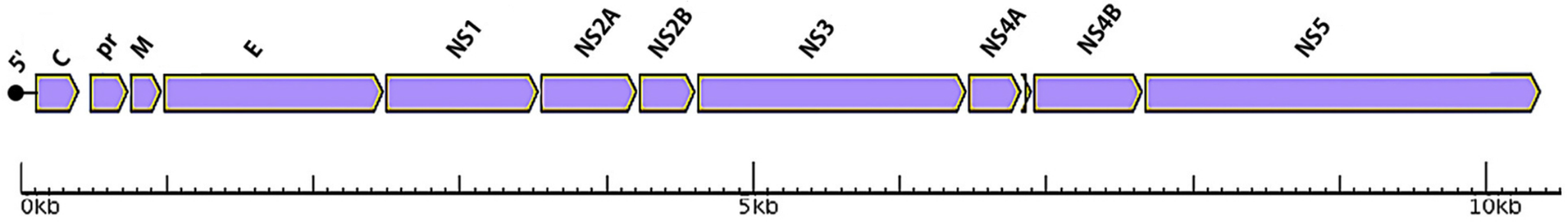
Translation has six possibilities

```
# Original dna sequence
[1]    30 ACATGGGCCTACCATGGGAGCTACGAAGCC
```

```
# 6 possible reading frames, DNAStringSet
[1]    30 ACATGGGCCTACCATGGGAGCTACGAAGCC             + 1
[2]    30 GGCTTCGTAGCTCCCATGGTAGGCCCATGT             - 1
[3]    29  CATGGGCCTACCATGGGAGCTACGAAGCC             + 2
[4]    29  GCTTCGTAGCTCCCATGGTAGGCCCATGT             - 2
[5]    28   ATGGGCCTACCATGGGAGCTACGAAGCC             + 3
[6]    28   CTTCGTAGCTCCCATGGTAGGCCCATGT             - 3
```

```
# 6 possible translations, AAStringSet
[1]    10 TWAYHGSYEA                                 + 1
[2]    10 GFVAPMVGPC                                 - 1
[3]     9 HGPTMGATK                                  + 2
[4]     9 AS*LPW*AH                                  - 2
[5]     9 MGLPWELRS                                  + 3
[6]     9 LRSSHGRPM                                  - 3
```

Conserved regions in the Zika virus



Adapted figure From Mosquitos to Humans: Genetic Evolution of Zika Virus Wang, Lulan et al. Cell Host & Microbe 2016, Vol 19 5: 561-565

Facts

- The Zika Virus has a positive strand genome.
- It lives in humans, monkeys and mosquitoes.
- The Flaviviruses family and share 11 conserved proteins.

INTRODUCTION TO BIOCONDUCTOR

Let's practice finding patterns!