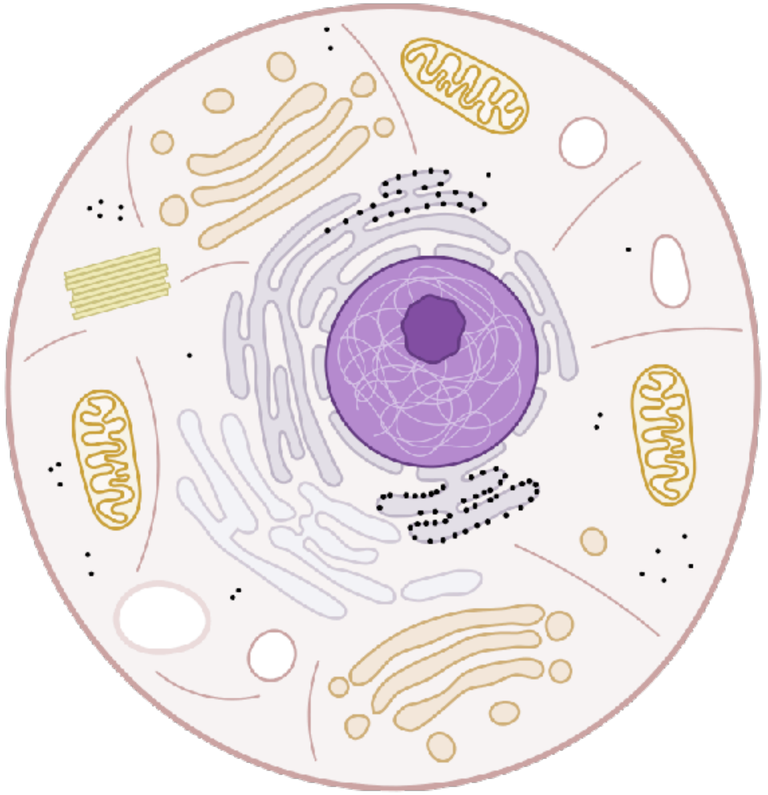
Molecular biology review
DNA, RNA, and Protein Sequences

Mark Stenglein, MIP 280A4

Typical eukaryotic cells contain nuclear and organellar dsDNA genomes

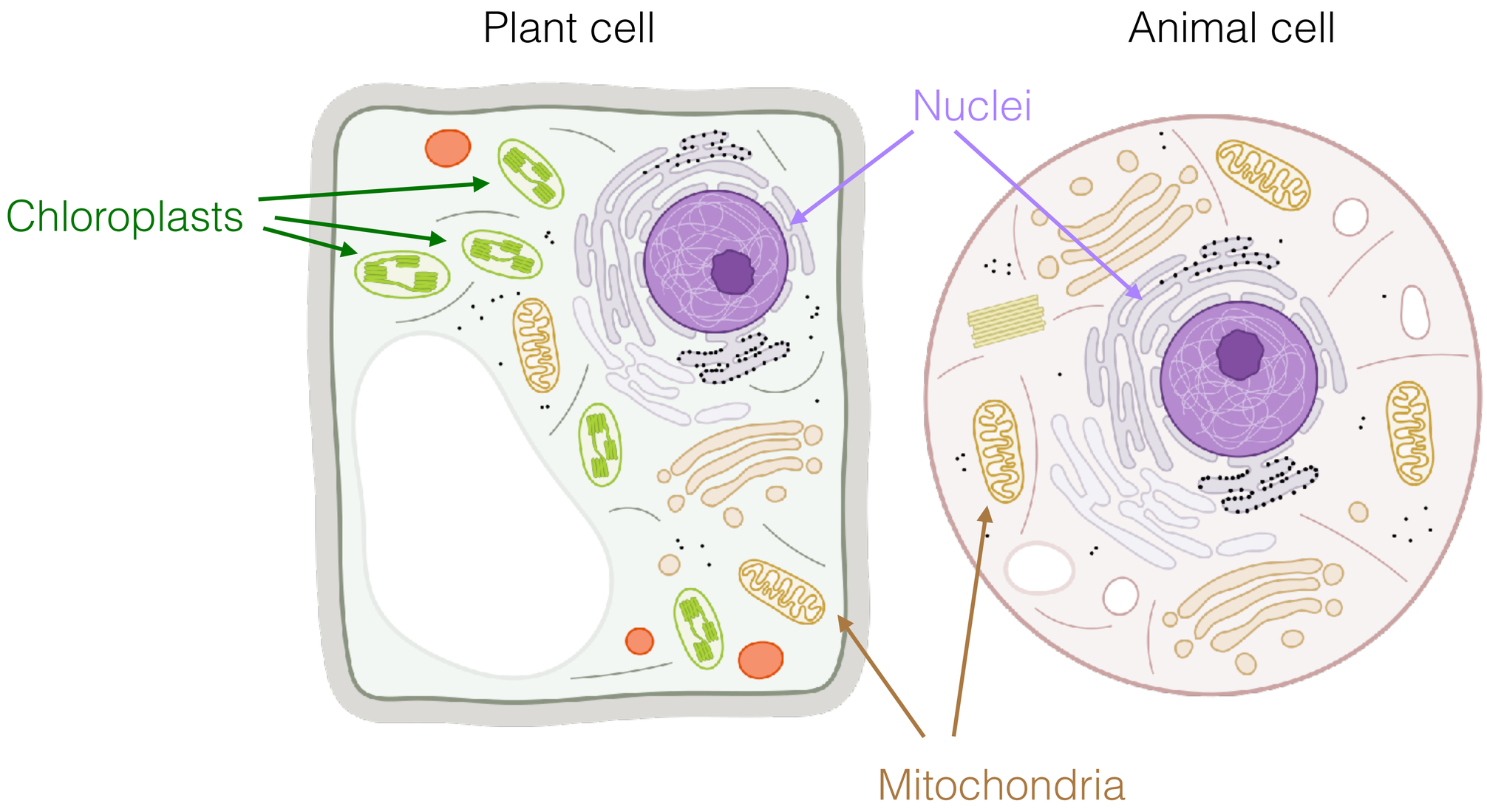
Where is the DNA in these cells?

Plant cell

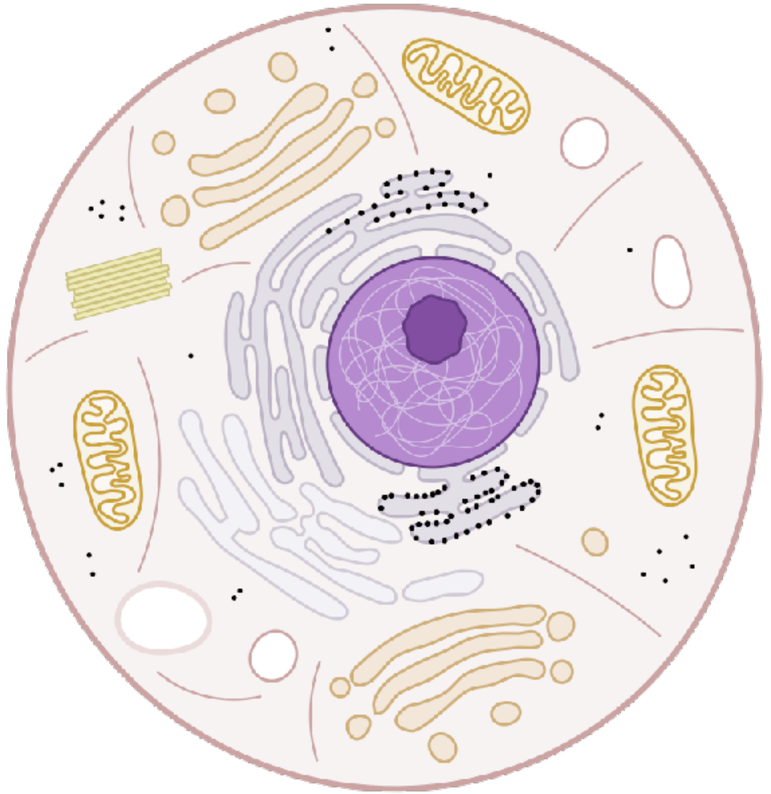


Animal cell

Typical eukaryotic cells contain nuclear and organellar dsDNA genomes



A typical eukaryotic cell also contains a ton of RNA

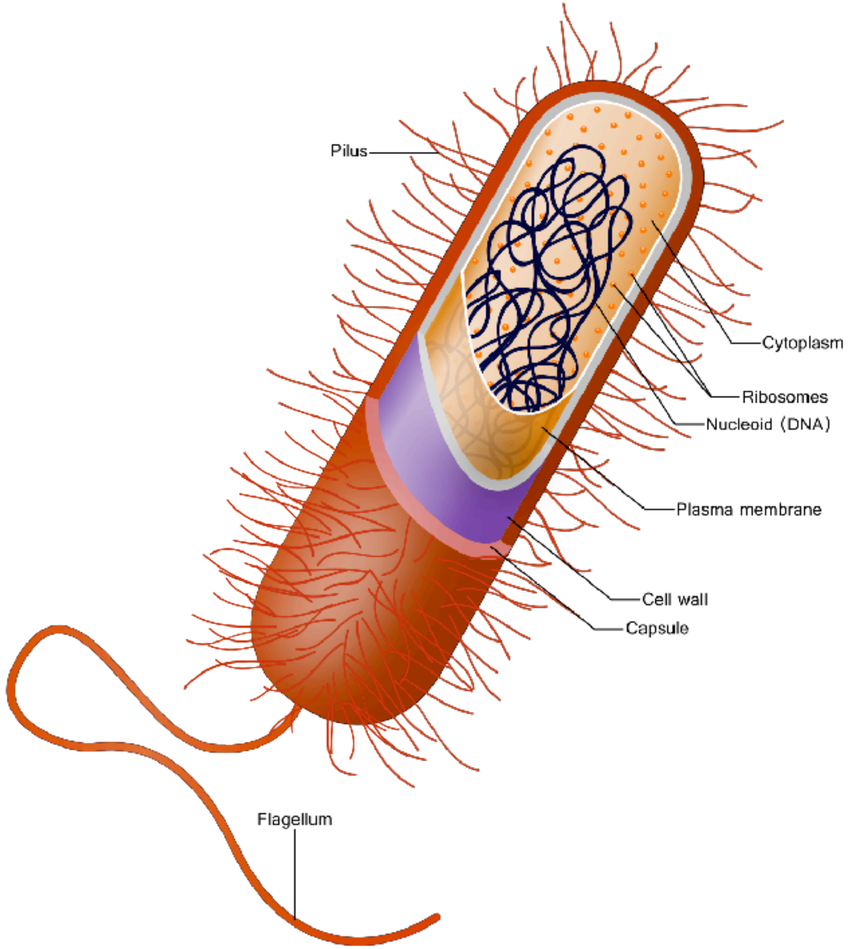


What are the main types of RNA in a cell?

Different types of RNA in cells

Type of RNA	Abbreviation	Function	% total RNA (approximate)
Messenger RNA	mRNA	Encodes proteins	<5%
Ribosomal RNA	rRNA	Major ribosome component	80%
Transfer RNA	tRNA	Links codons to amino acids during translation	15%
Micro RNA (and other small RNAs)	miRNA, siRNA, etc	Gene regulation, …	<5%
Other non-coding RNA	ncRNA		<5%

A typical prokaryotic cell contains one or more dsDNA chromosomes and often one or more plasmids.



What about viruses? What type of genomes do they have?

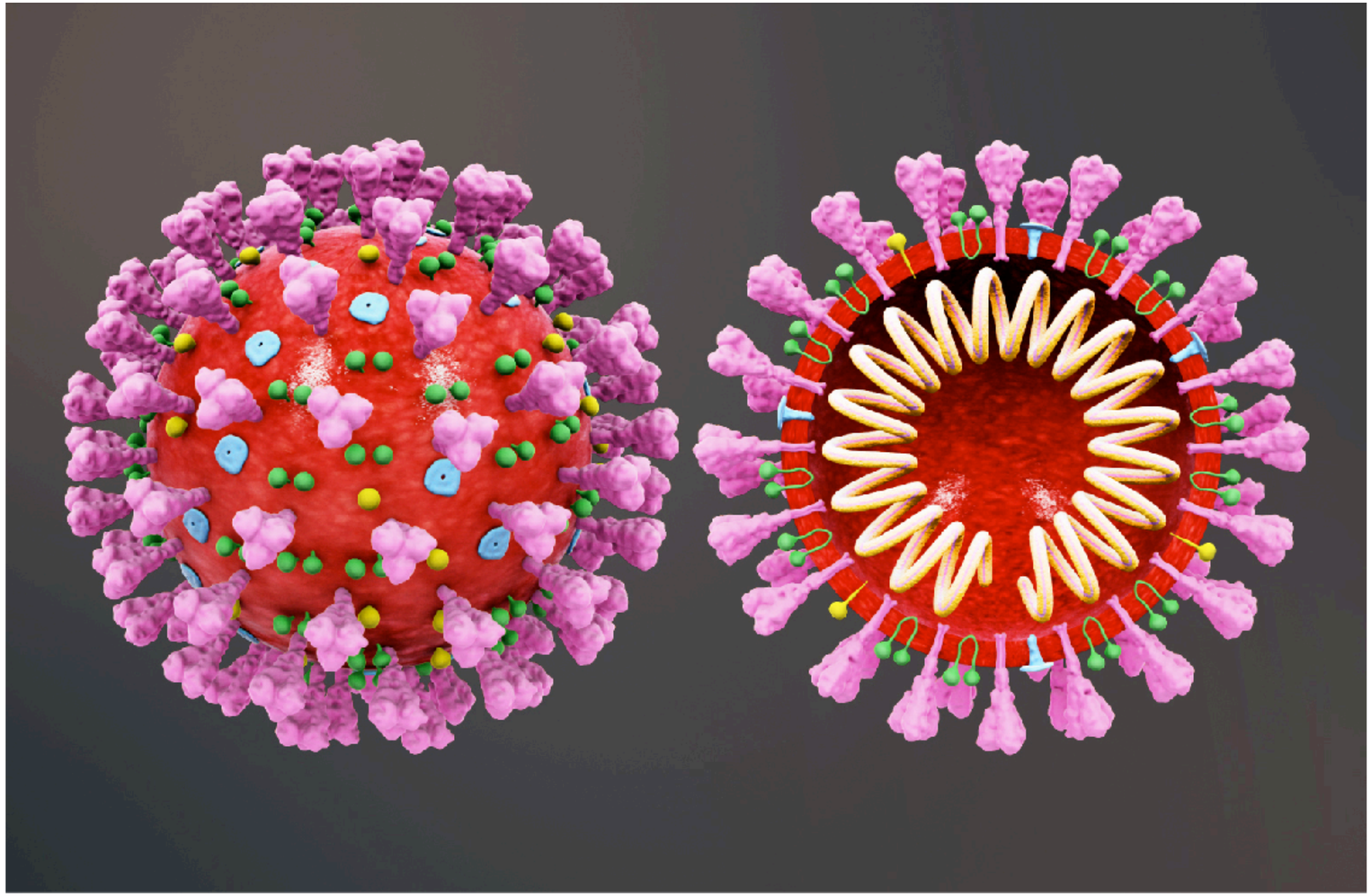


Image credit: Nagualdesign CC BY-SA 4.0, Link

Different viruses have almost every imaginable type of genomic nucleic acid

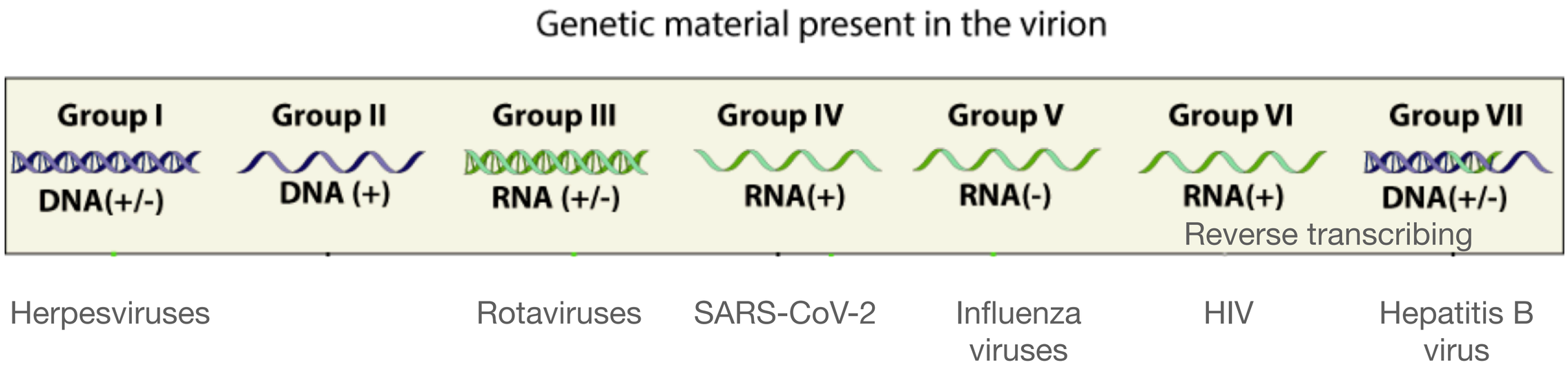


Image credit: Viral Zone Link

The central dogma of molecular biology:
DNA is transcribed into RNA, which is translated into protein

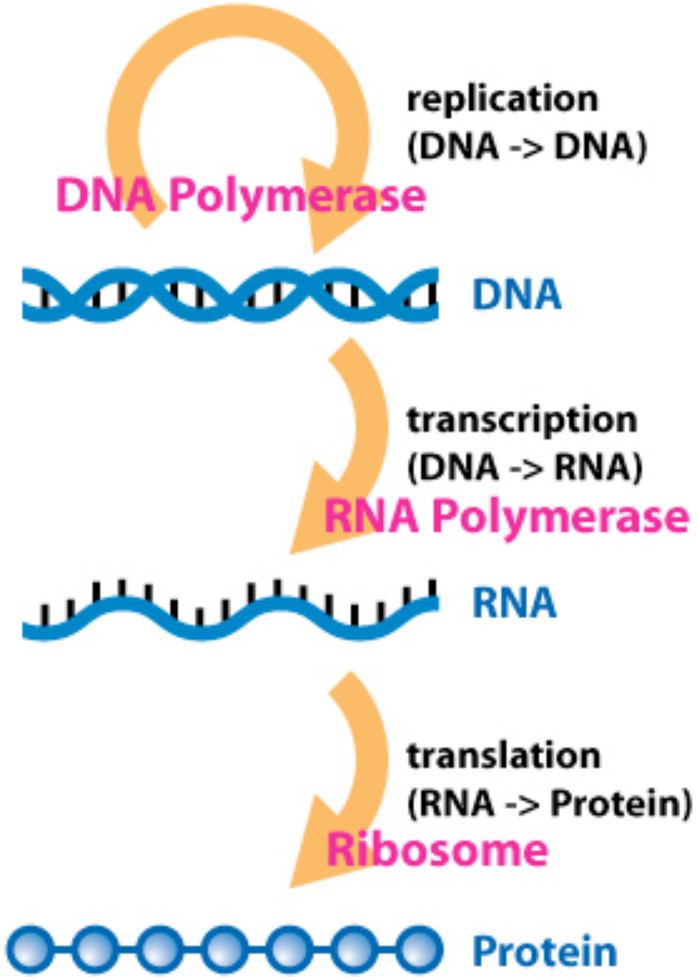


Image credit: Dhorspool CC BY-SA 3.0, Link

As usual in Biology, the full picture is somewhat more complicated

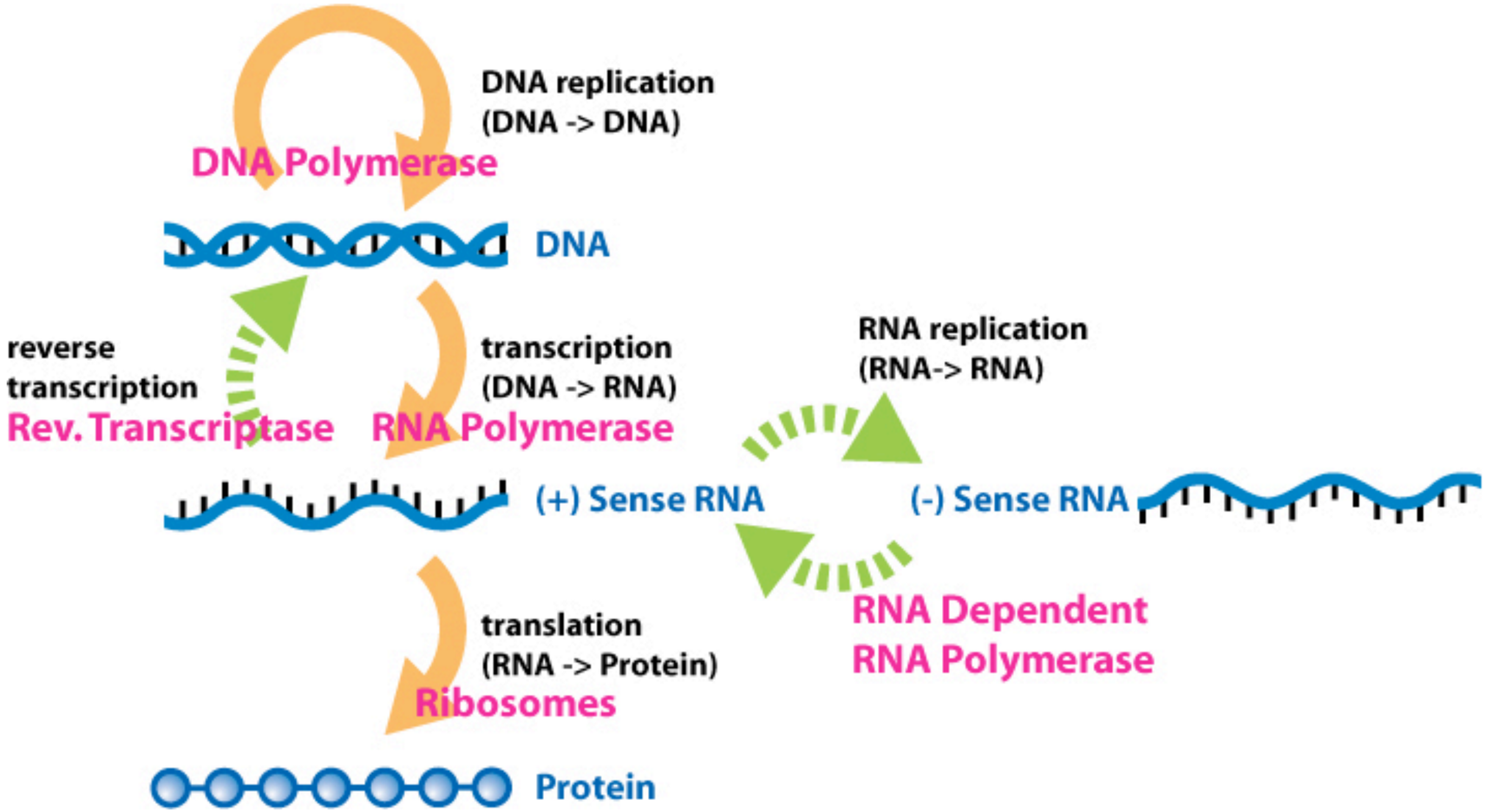


Image credit: Dhorspool CC BY-SA 3.0, Link

A DNA double helix with typical A-T and G-C base pairing

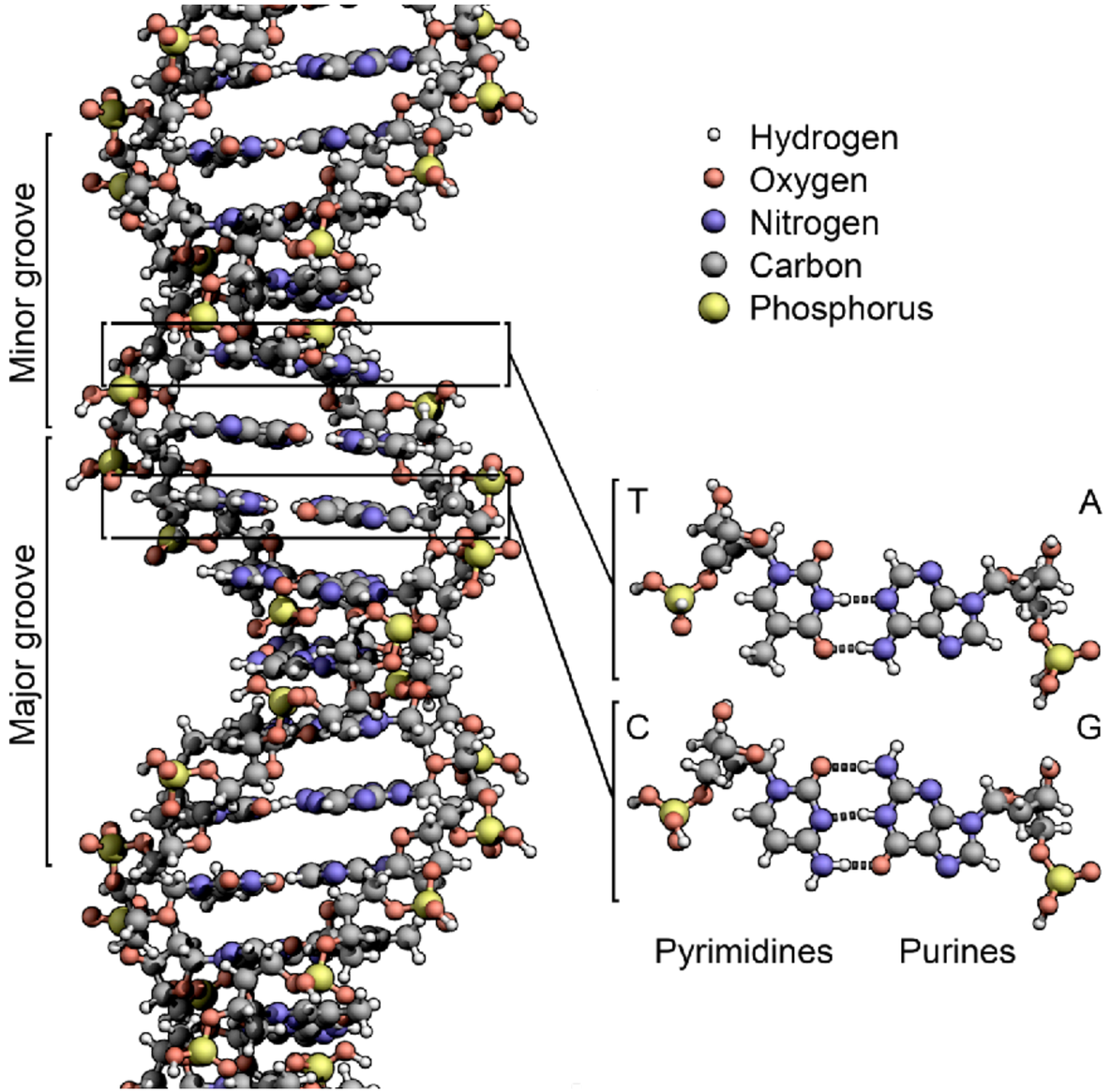
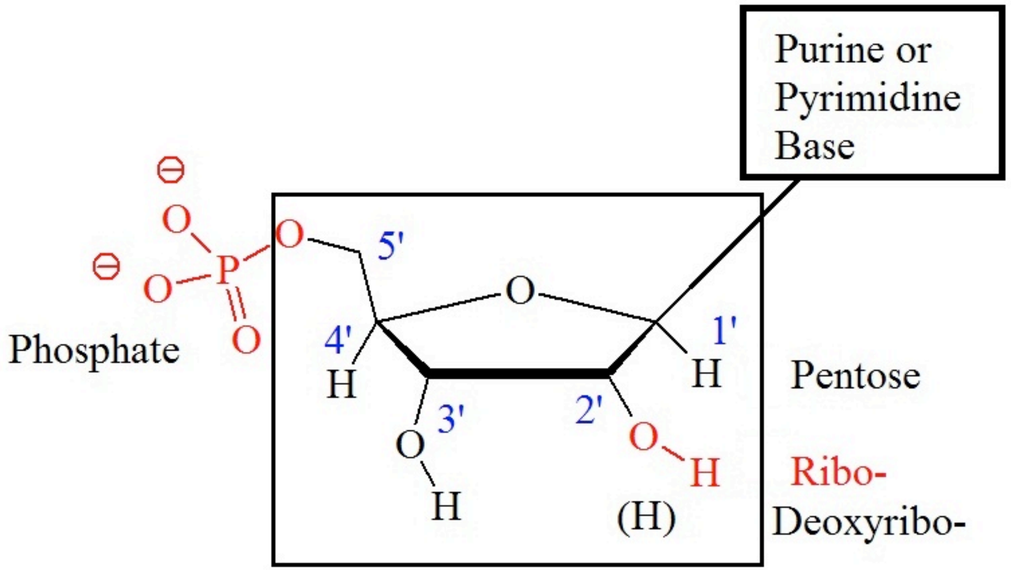


Image credit: Zephyris CC BY-SA 3.0, Link

DNA and RNA strands have 5′ to 3′ orientation



DNA strands have 5′ to 3′ orientation

5′ ACTG 3′

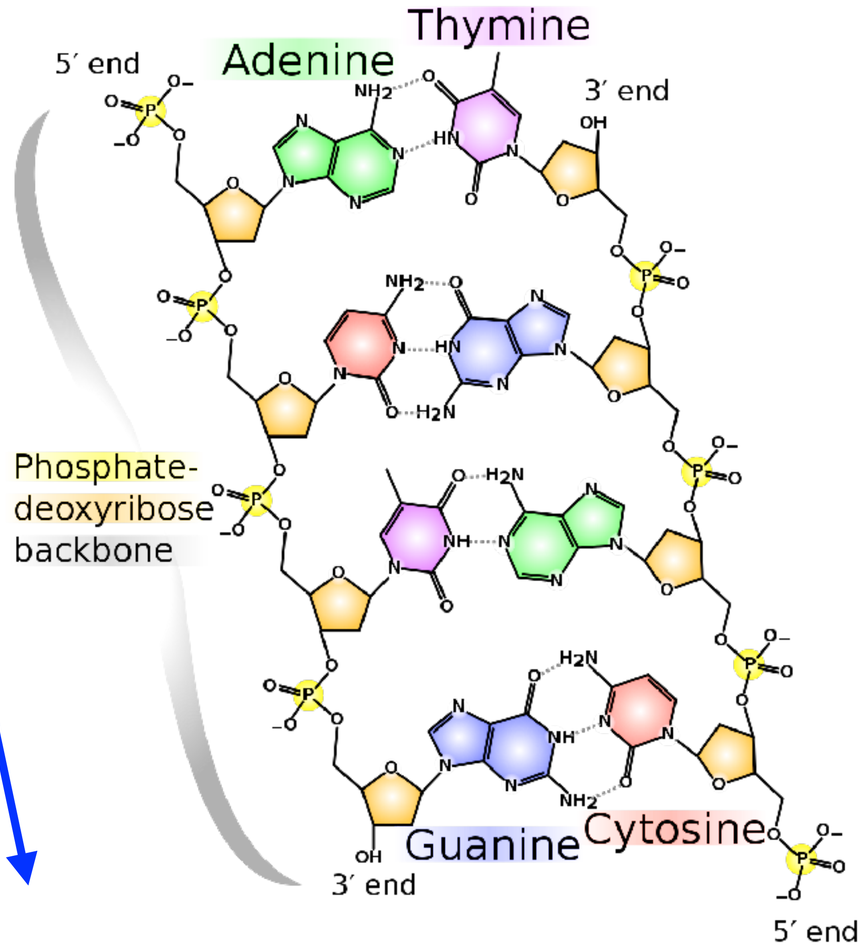
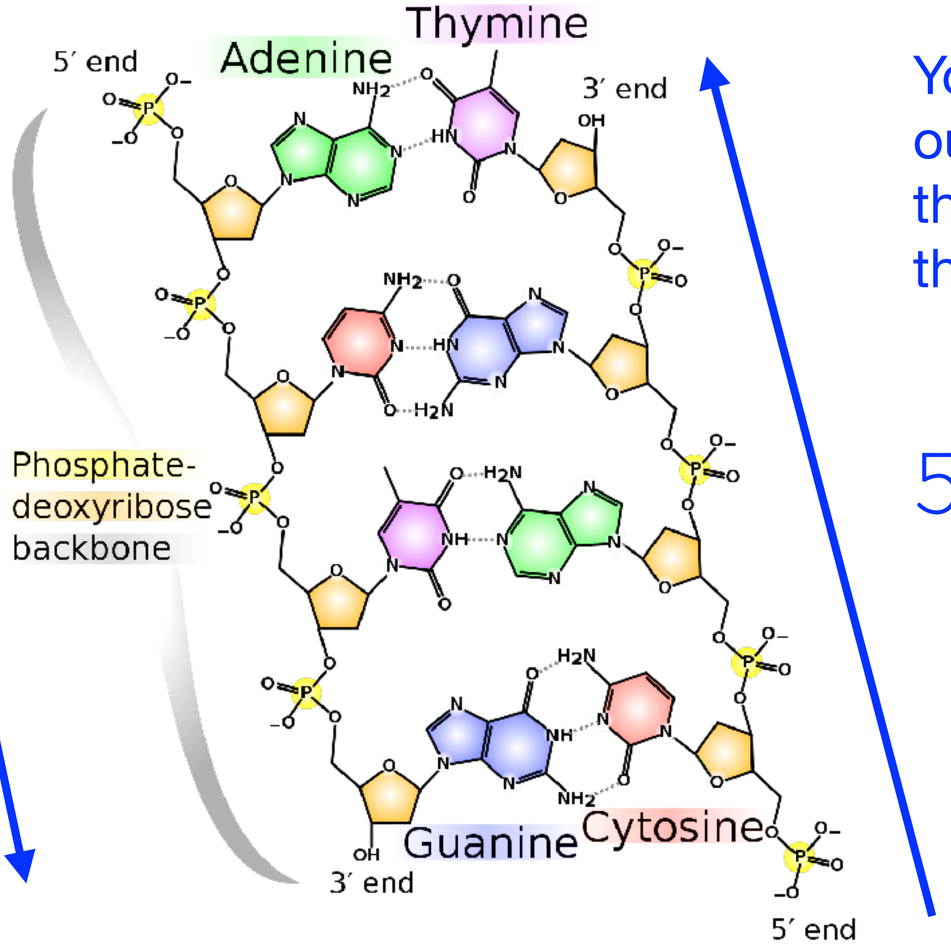


Image credit: Madprime CC BY-SA 3.0, Link

For a dsDNA it is equally valid to write out the sequence in either orientation

Write out this DNA sequence in the 5’ to 3’ orientation of this strand

5′ ACTG 3′



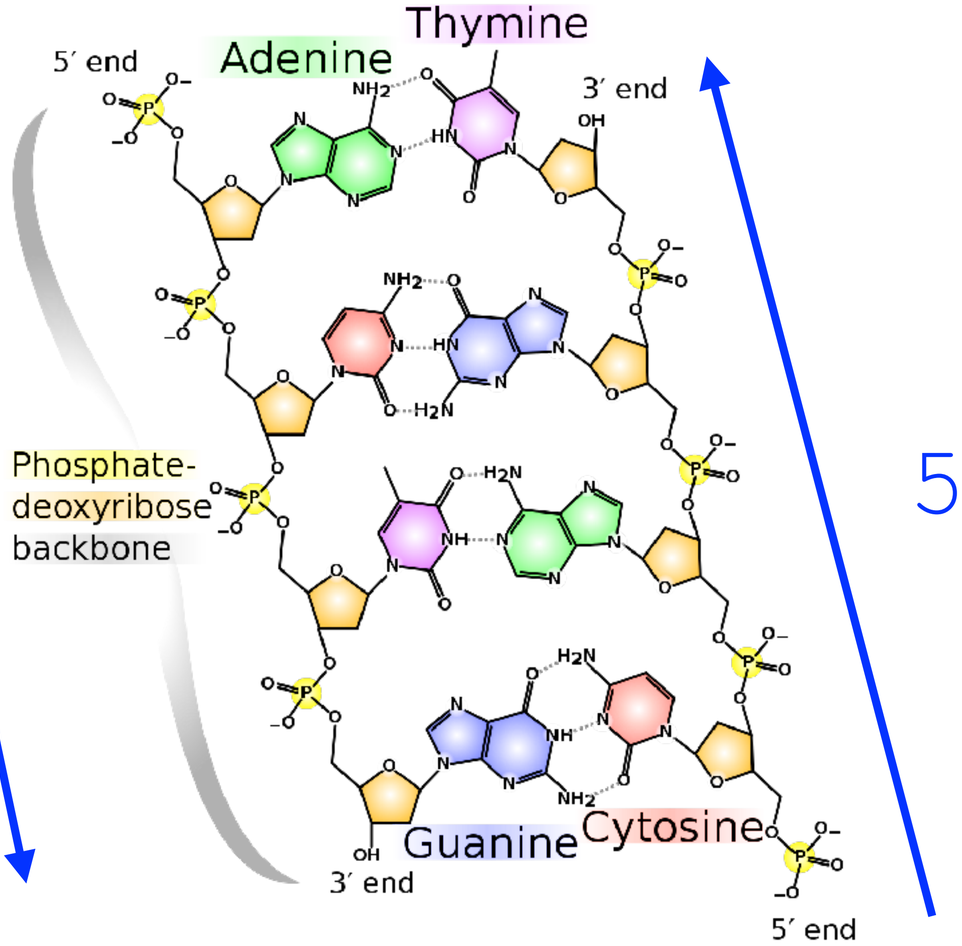
You could also write out the sequence in the orientation of this strand

5′ CAGT 3′

Image credit: Madprime CC BY-SA 3.0, Link

These 2 sequences are reverse complements of each other

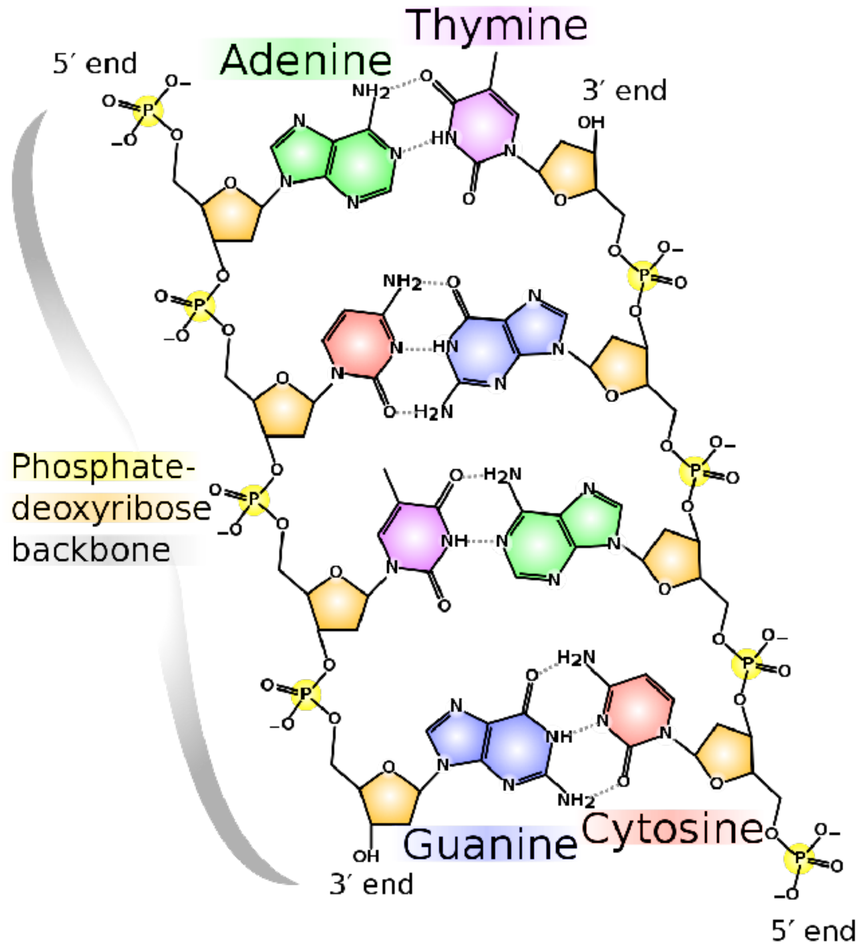
5′ ACTG 3′



5′ CAGT 3′

Image credit: Madprime CC BY-SA 3.0, Link

Sometimes reverse complement sequences are written out on 2 lines
This notation writes out both strands of a dsDNA molecule




```
5′ ACTG 3′
   ||||
3′ TGAC 5′
```

Note the opposite orientations

Image credit: Madprime CC BY-SA 3.0, Link

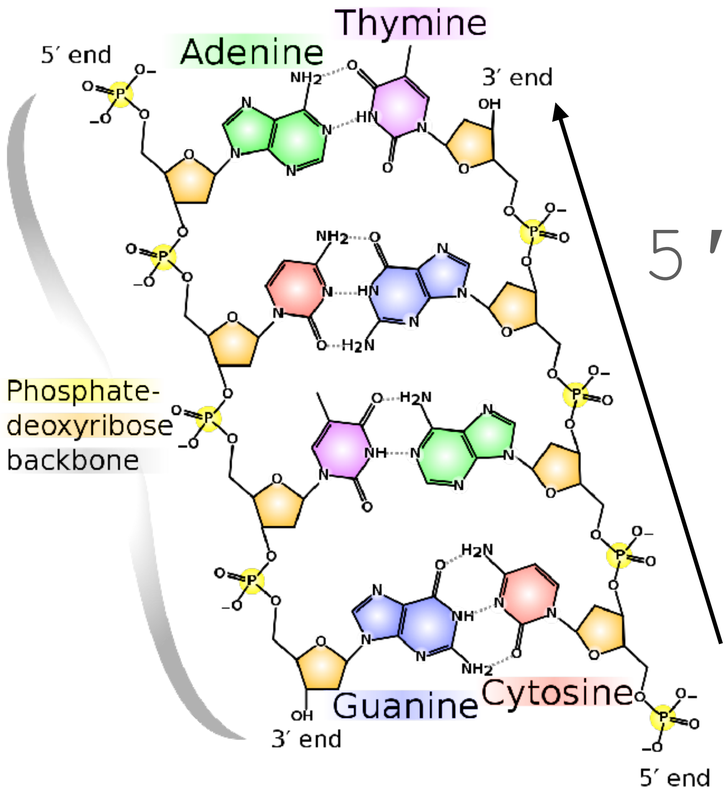
Don’t confuse this dsDNA notation with sequence alignments (e.g. from a BLAST search)

```
5′ ACTG 3′
   ||||
3′ TGAC 5′
```

```
>AB462167.1 Procyon lotor mitochondrial gene for 16S ribosomal RNA, complete
sequence, specimen_voucher: personal:Tomoharu Tokutomi:NDMC-PL-MIE18010
Length=1586

 Score = 433 bits (234),  Expect = 1e-117
 Identities = 234/234 (100%), Gaps = 0/234 (0%)
 Strand=Plus/Plus

Query  1     CAAAAACATCACCTCTAGCATTAACAGTATTAGAGGCACTGCCTGCCCAGTGACATTAGT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  840   CAAAAACATCACCTCTAGCATTAACAGTATTAGAGGCACTGCCTGCCCAGTGACATTAGT  899

Query  61    TAAACGGCCGCGGTATCCTGACCGTGCAAAGGTAGCATAATCATTTGTTCTCTAAATAGG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  900   TAAACGGCCGCGGTATCCTGACCGTGCAAAGGTAGCATAATCATTTGTTCTCTAAATAGG  959
I
Query  121   GACTTGTATGAATGGCCACACGAGGGTTTGACTGTCTCTTACTTCCAACCAGTGAAATTG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  960   GACTTGTATGAATGGCCACACGAGGGTTTGACTGTCTCTTACTTCCAACCAGTGAAATTG  1019

Query  181   ACCTTCCCGTGAAGAGGCGGGAATAAGAAAATAAGACGAGAAGACCCTATGGAG  234
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1020  ACCTTCCCGTGAAGAGGCGGGAATAAGAAAATAAGACGAGAAGACCCTATGGAG  1073
```

Note the use of fixed-width fonts like Courier to write sequences:

Almost always dsDNA sequences are written out for one strand only



5′ CAGT 3′

or just

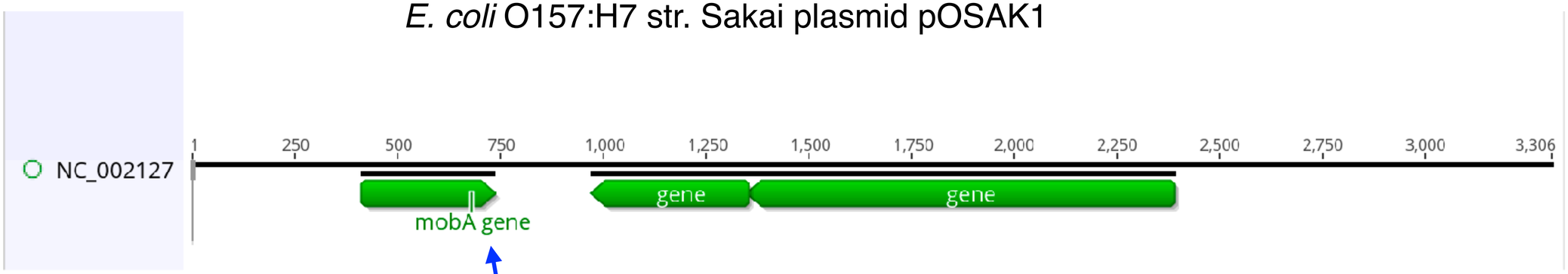
CAGT

- The choice of strand is arbitrary
- Either orientation is acceptable
- For a particular organism there is usually a customary orientation
- Always written 5′ to 3′
- The sequence of the other stand is implied.

Image credit: Madprime CC BY-SA 3.0, Link

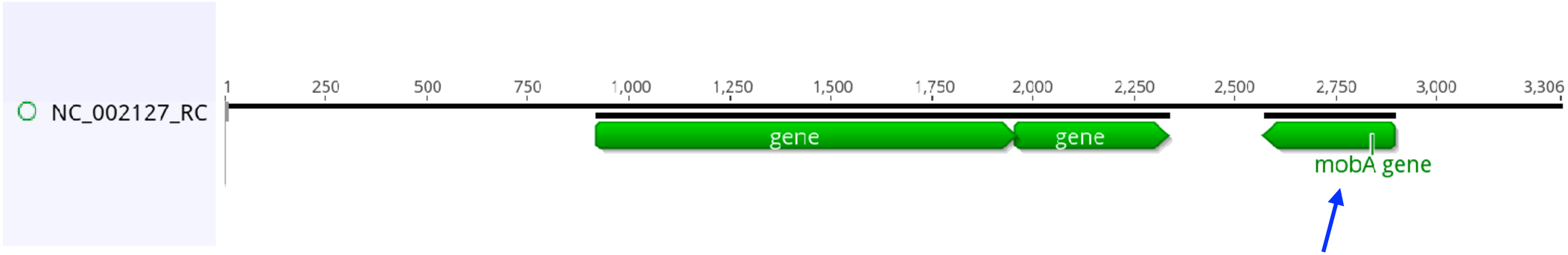
Although both orientations are valid there are different biological meanings for the two orientations

E. coli O157:H7 str. Sakai plasmid pOSAK1



The mobA gene here is in its coding orientation: the sequence will match the mobA mRNA

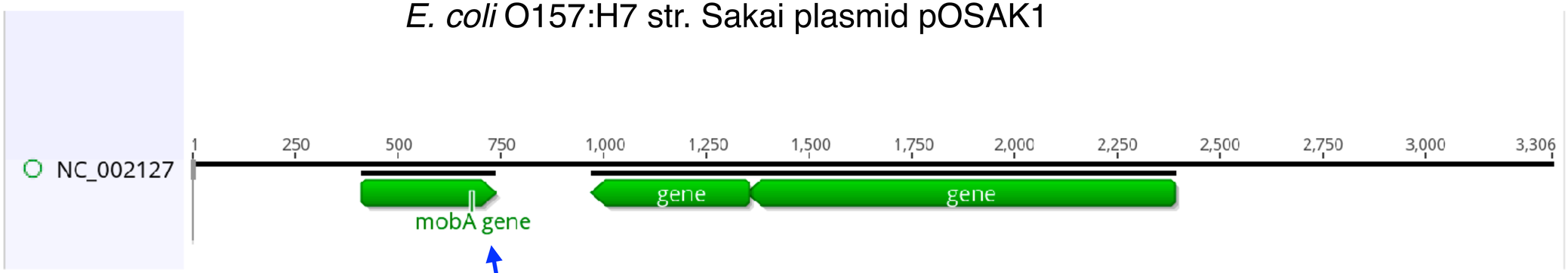
E. coli O157:H7 str. Sakai plasmid pOSAK1 reverse complement



The mobA gene now is in anti-coding sense: the mobA mRNA will be the reverse complement of this sequence

Annotation, information that describes features of sequences, takes into account the orientation of the sequence.

E. coli O157:H7 str. Sakai plasmid pOSAK1



The mobA gene is in the sense of the sequence between positions 413 and 736

Annotation from NCBI

```
gene            413..736
                /gene="mobA"
                /locus_tag="pOSAK1_01"
                /db_xref="GeneID:1789662"
CDS             413..736
                /gene="mobA"
                /locus_tag="pOSAK1_01"
                /codon_start=1
                /transl_table=11
                /product="plasmid mobilization"
                /protein_id="NP_052604.1"
                /db_xref="GI:10955263"
                /db_xref="GeneID:1789662"
                /translation="MTKRSGSNTRRRAISRPVRLTAEEDQEIRKRAAECGKTVSGFLR
                AAALGKKVNSLTDDRVLKEVMRLGALQKKLFIDGKRVGDREYAEVLIAITEYHRALLS
                RLMAD"
```

The good news: Since you almost always download a sequence and its annotation together this is not something you really need to worry about.

Exercise: write the reverse complement of these sequences
The reverse complements should be written written 5′ to 3′

1) 5′ CAAAGGT 3′

2) 5′ GGGCCCAAATTT 3′

Exercise: download and install Geneious software

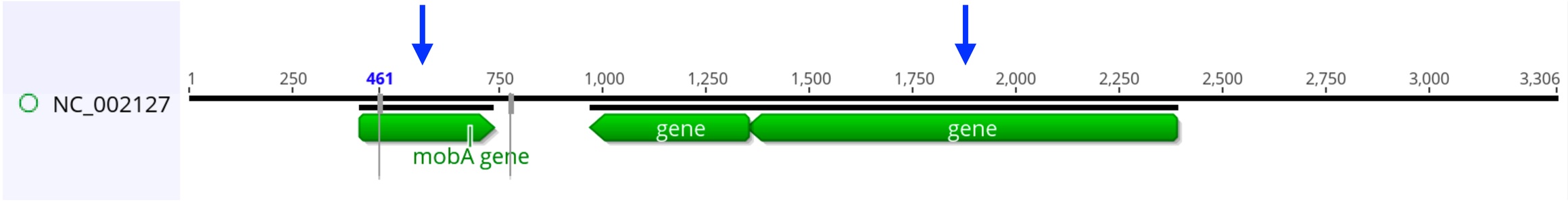
1. Google Geneious: go to https://www.geneious.com/
2. Download from Resources -> Download
3. Activate using your email and license key:

   PRIME-KE7XN-RHAMT-CSJQK

Exercise: download the *E. coli* pOSAK1 plasmid sequence in Geneious

1. Open Geneious
2. Create a new folder in the top-level Local folder named E_coli or something: File->New->Folder
3. In the NCBI/Nucleotide section at the bottom left, search for NC_002127
4. After the sequence is found, click on Download Full Sequence
5. Drag this sequence to your new folder
6. Answer the relevant exercise questions

Exercise: extract coding sequences from pOSAK1



1. Extract the mobA gene coding sequence.
   A) Click on the mobA annotation
   B) Click the extract button.
   C) Answer the relevant exercise questions.

2. Extract the coding sequence *in the coding orientation* of the largest gene on this plasmid (the gene between positions 1348 and 2388)
   A) Click on the annotation for this gene
   B) Click the extract button. Answer the pop-up question in Geneious so that you get this sequence in the coding orientation.
   C) Answer the relevant exercise questions.

What are the three major molecular properties that distinguish RNA from DNA?

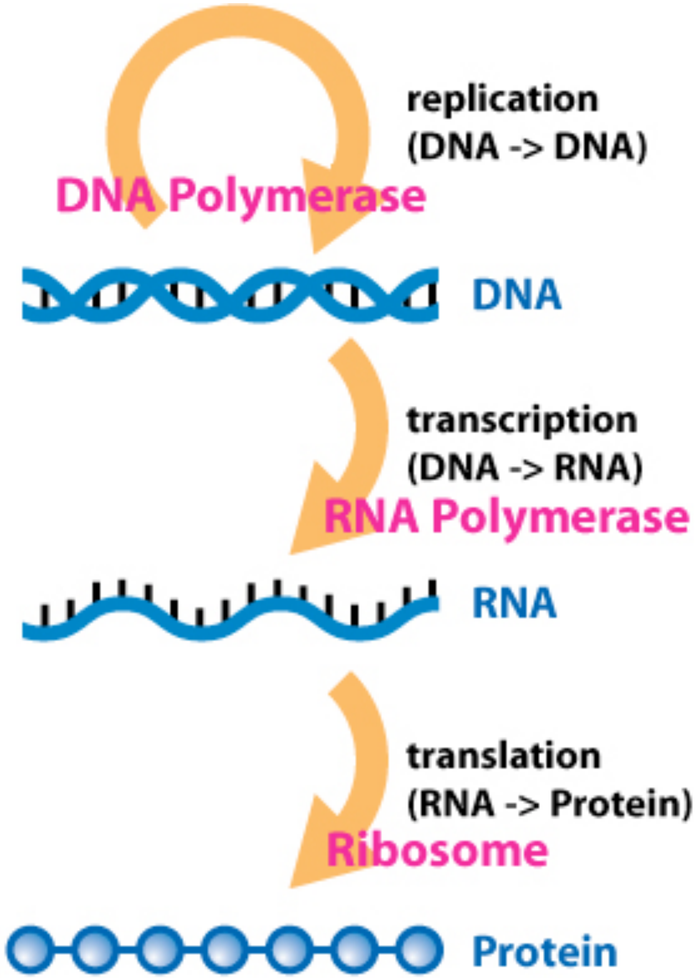


Image credit: Dhorspool CC BY-SA 3.0, Link

RNA vs. DNA: (1) ribose not deoxyribose, (2) usually single stranded, (3) Us not Ts

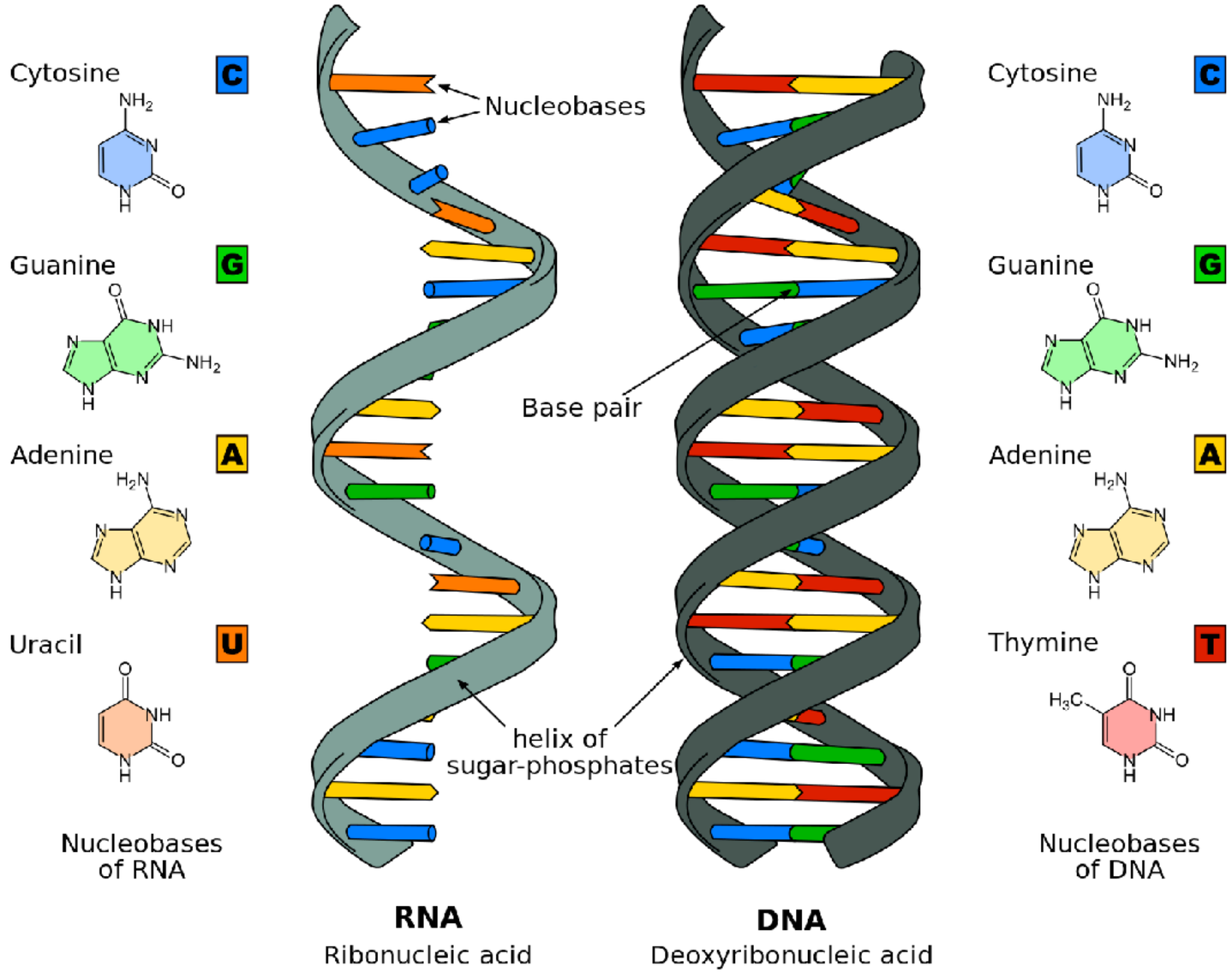


Image credit: Sponk CC BY-SA 3.0, Link

In practice in computational biology, RNA sequences are usually written with Ts not Us even though this is inaccurate biologically

The SARS-CoV-2 reference sequence on NCBI (NC_045512)
The genome actually has Us not Ts

```
  1 attaaaggtt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct
 61 gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact
121 cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctcgtctatc
181 ttctgcaggc tgcttacggt ttcgtccgtg ttgcagccga tcatcagcac atctaggttt
241 cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac
301 acacgtccaa ctcagtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggctttgg
361 agactccgtg gaggaggtct tatcagaggc acgtcaacat cttaaagatg gcacttgtgg
421 cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagccctatg tgttcatcaa
481 acgttcggat gctcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact
541 cgaaggcatt cagtacggtc gtagtggtga gacacttggt gtccttgtcc ctcatgtggg
601 cgaaatacca gtggcttacc gcaaggttct tcttcgtaag aacggtaata aaggagctgg
661 tggccatagt tacggcgccg atctaaagtc atttgactta ggcgacgagc ttggcactga
```

Note that lower case vs. upper case bases generally doesn't have any special meaning

RNA sequences are also almost always written in the 5' to 3' orientation.

Sequences on public databases are identified by “accessions”. These are unique identifiers for each sequence. The accession of this sequence is: NC_045512

Genbank accessions also have versions, which are defined after the accession and a period.

ACCESSION.VERSION

Versions increase by 1 if a sequence is ever updated.

What version is the SARS-CoV-2 reference sequence (NC_045512)?

Protein sequences are the other main type of biological sequence

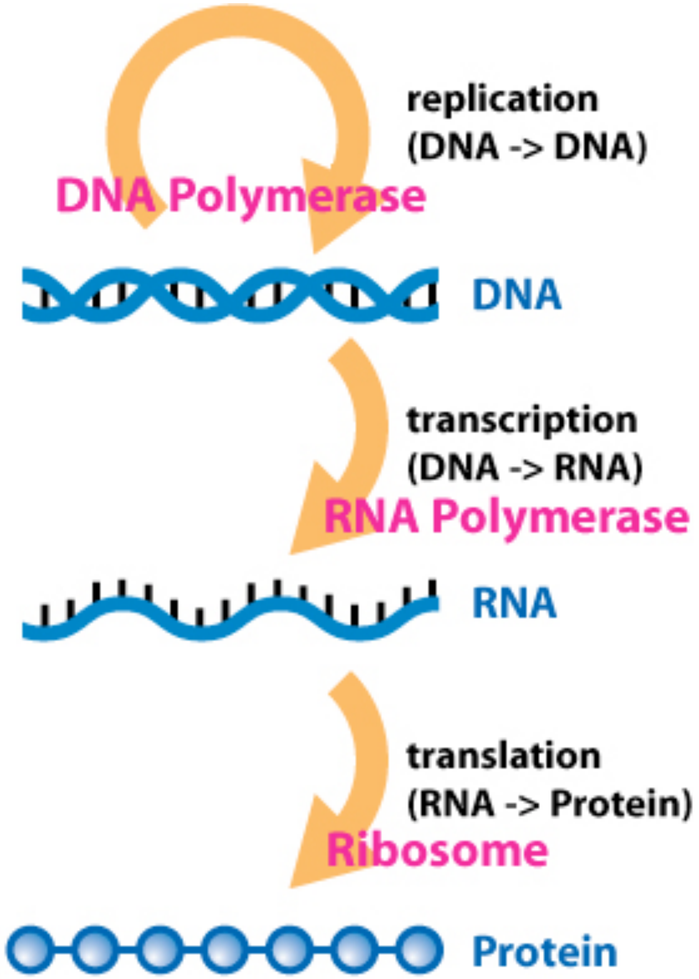


Image credit: Dhorspool CC BY-SA 3.0, Link

Proteins are translated from mRNAs

Protein sequences are always oriented amino terminal to carboy terminal

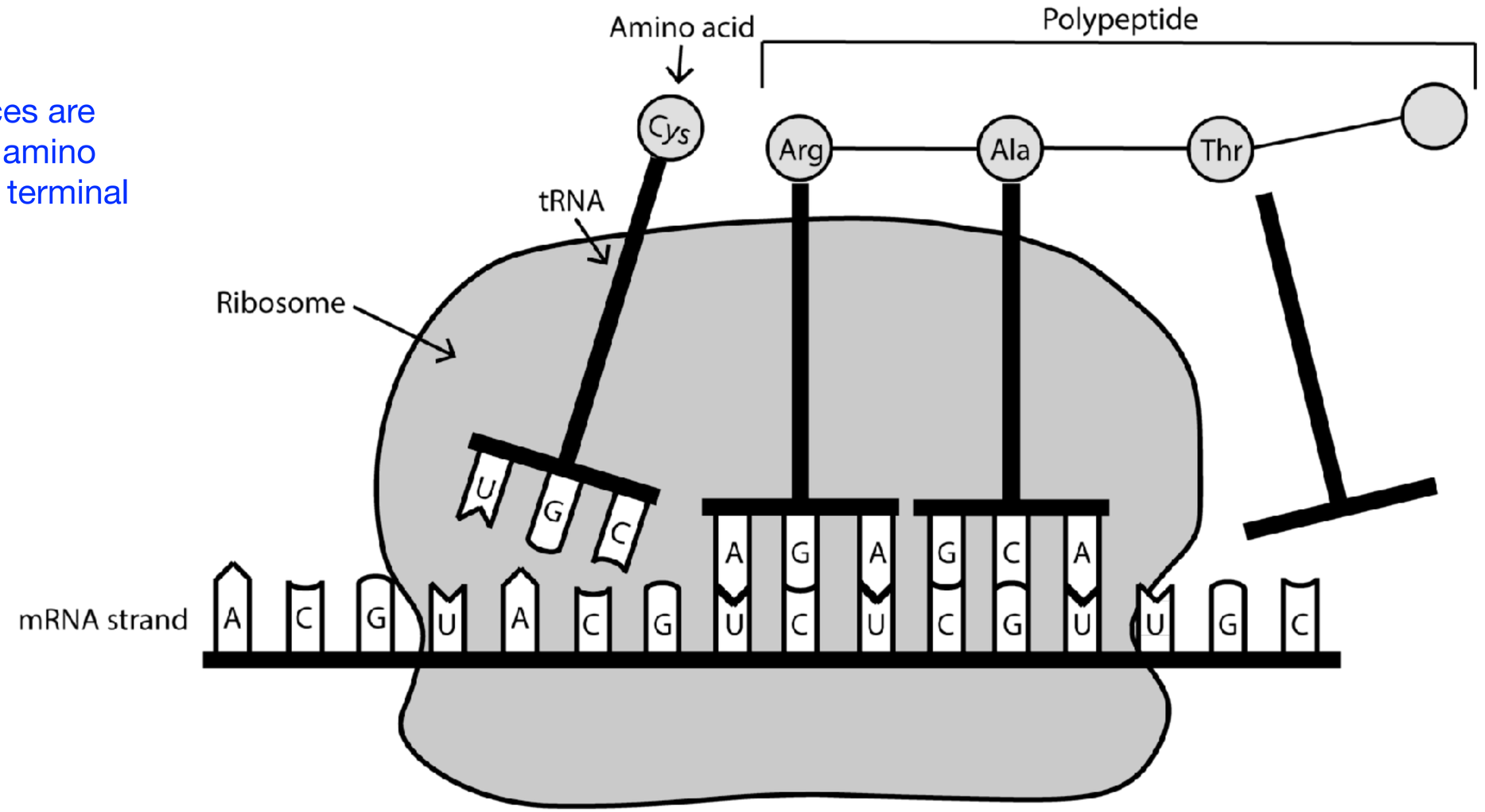


Image credit: Sarah Greenwood CC BY-SA 4.0, Link

Proteins are translated from mRNAs according to the genetic code

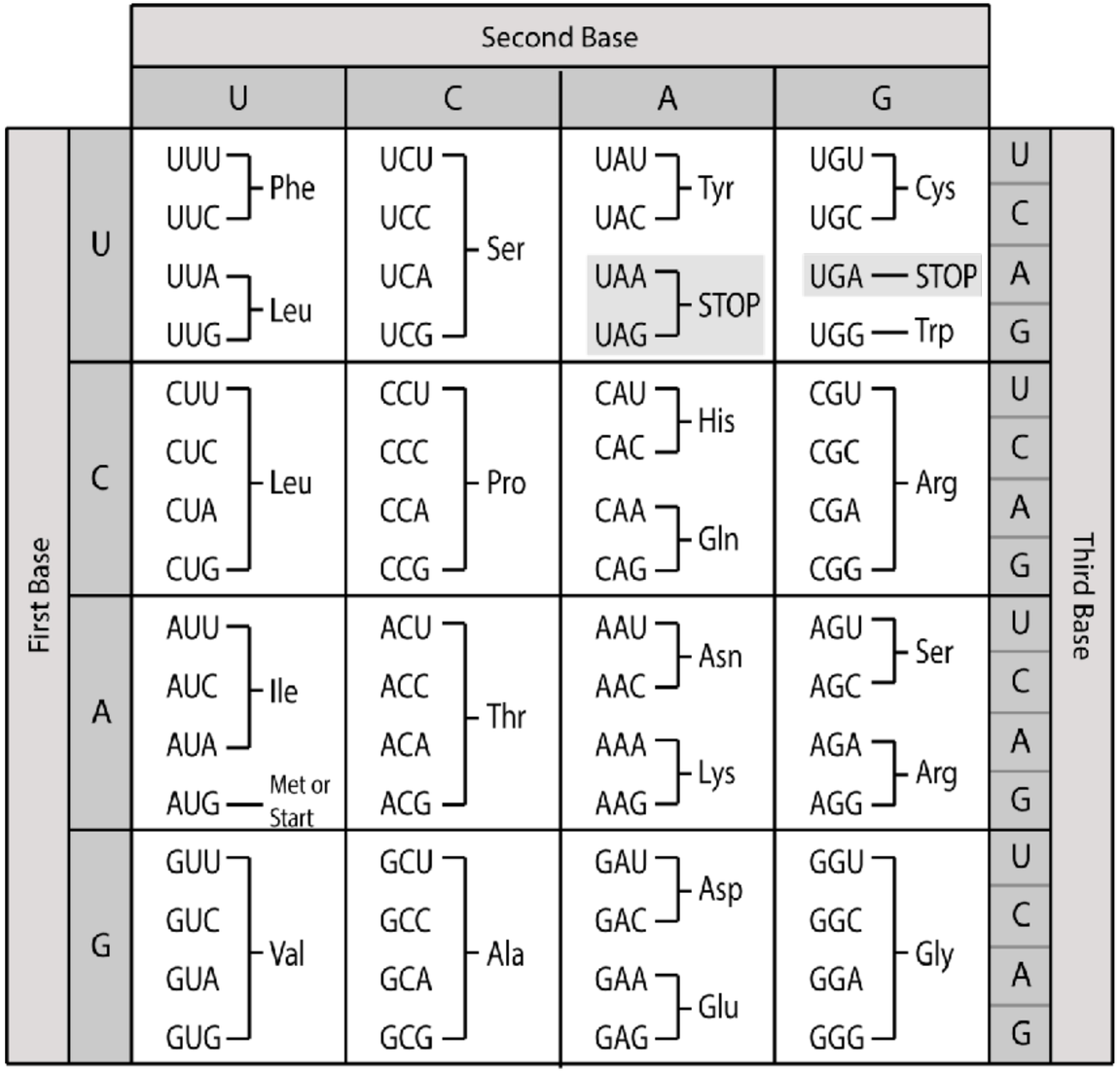
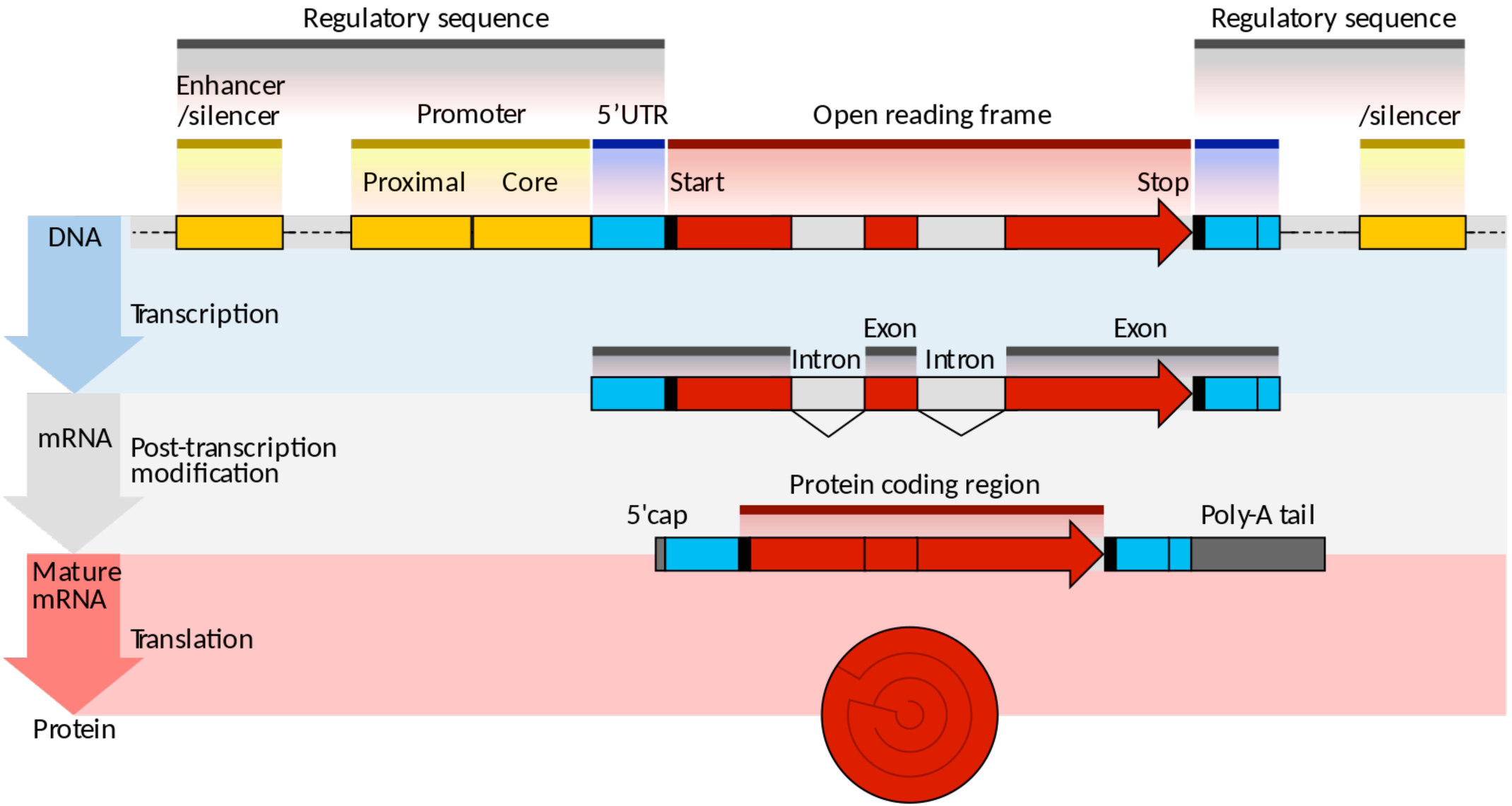


Image credit: Sarah Greenwood CC BY-SA 4.0, Link

Coding sequences in real life

How are eukaryotic coding sequences structured in genomes?

Eukaryotic coding sequences are typically on exons separated by introns



Coding sequences in real life

How are *prokaryotic* coding sequences structured in genomes?

Prokaryotic coding sequences are typically continuous

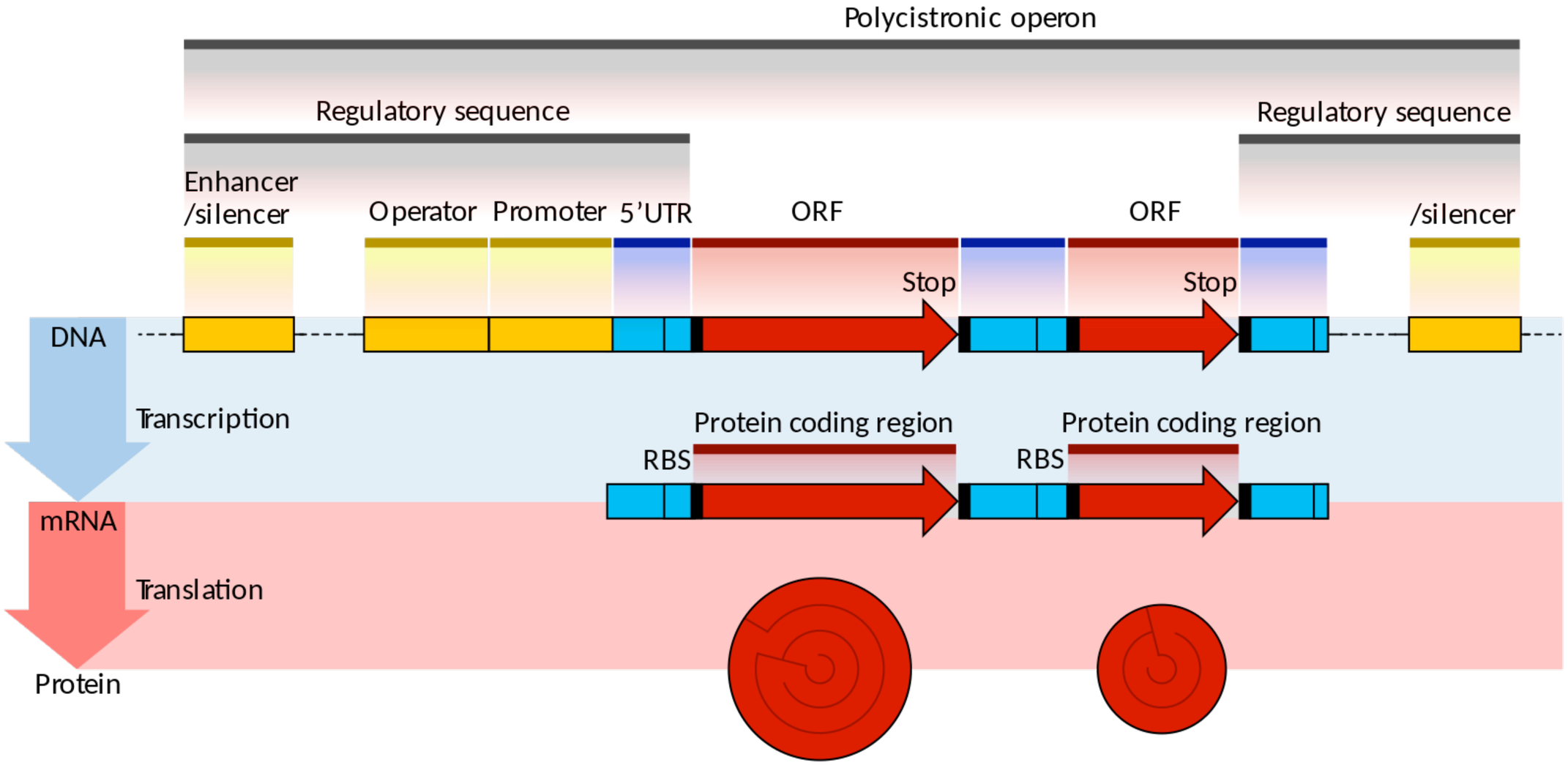


Image credit: Thomas Shafee CC BY 4.0, Link

What about viruses?

Viruses use a variety of strategies to produce different proteins from compact genomes

HIV uses splicing and ribosomal frame shifting

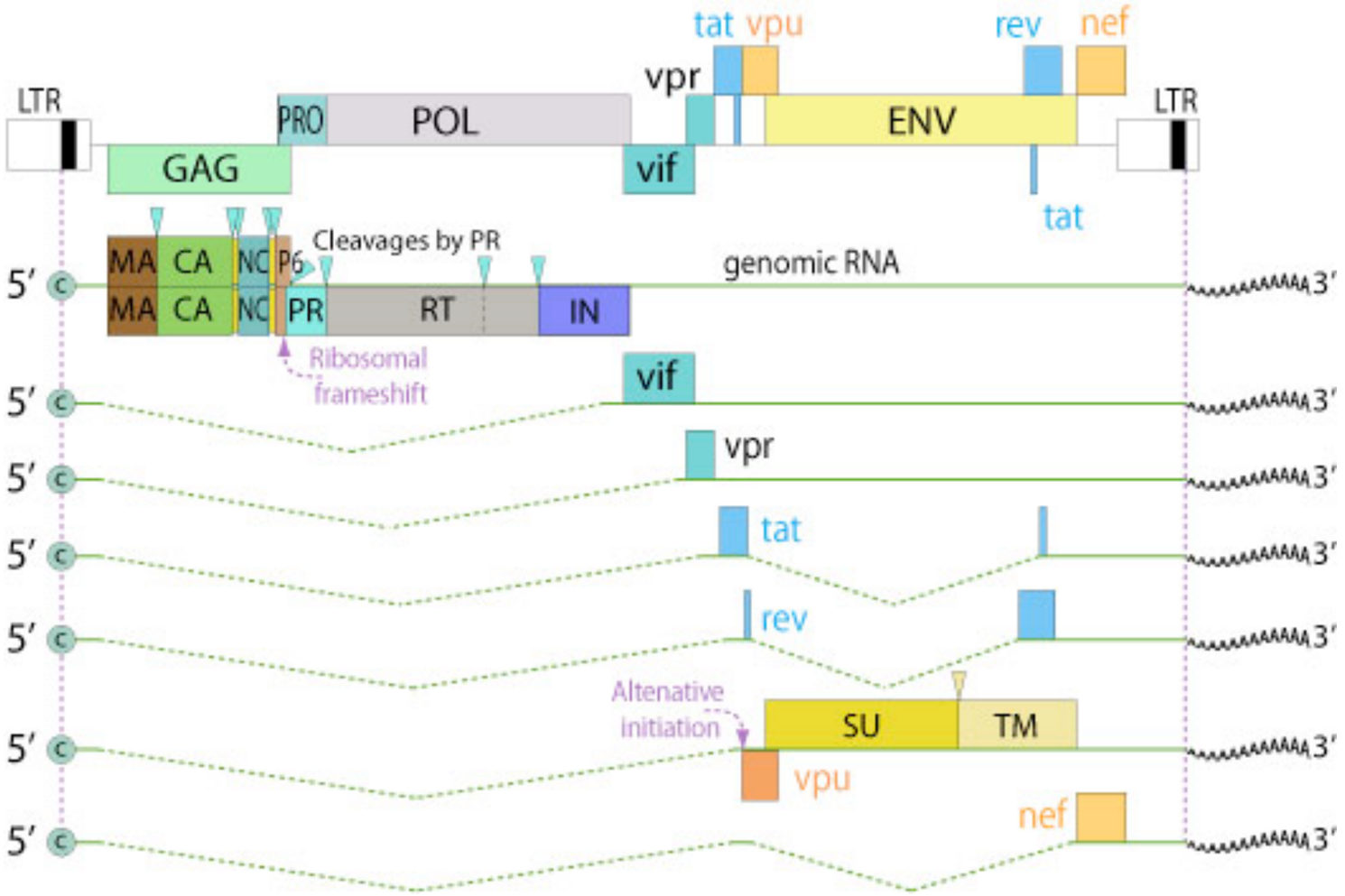


Image credit: Viralzone

Viruses use a variety of strategies to produce different proteins from compact genomes

Influenza viruses encode genes on different RNA molecules (genome segments) (And splicing and ribosomal frameshifting, etc.)

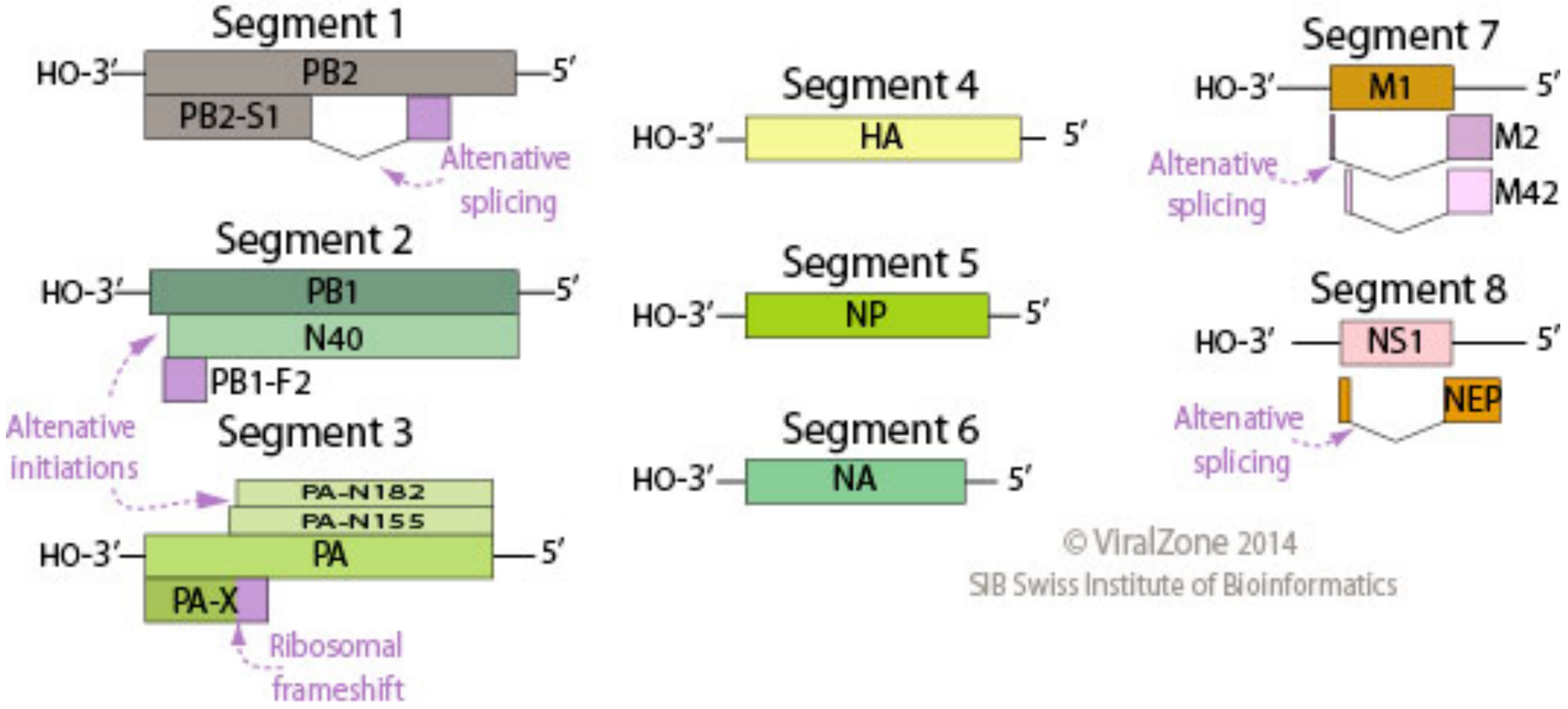
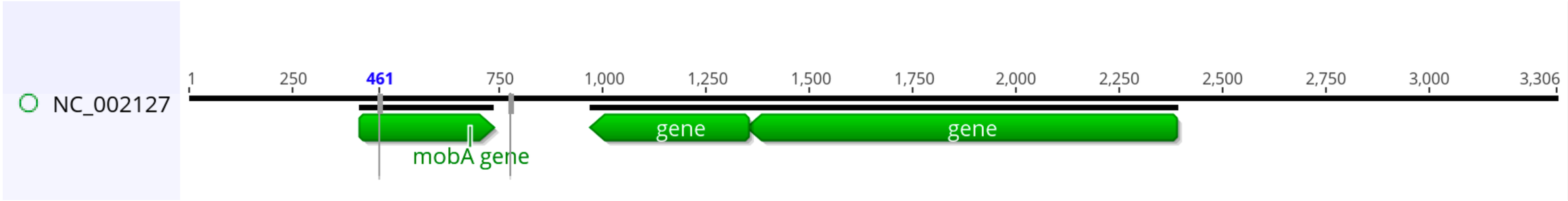


Image credit: Viralzone

Exercise: extract the protein sequence encoded by the mopA gene on pOSAK1



1. Extract the mobA gene protein sequence.
   A) Click on the mobA annotation
   B) Click the Translate button
   C) Select “Translate selected annotation” to create a new protein sequence in Geneious.
   D) Answer the relevant exercise questions.