Information storage and flow

The central dogma

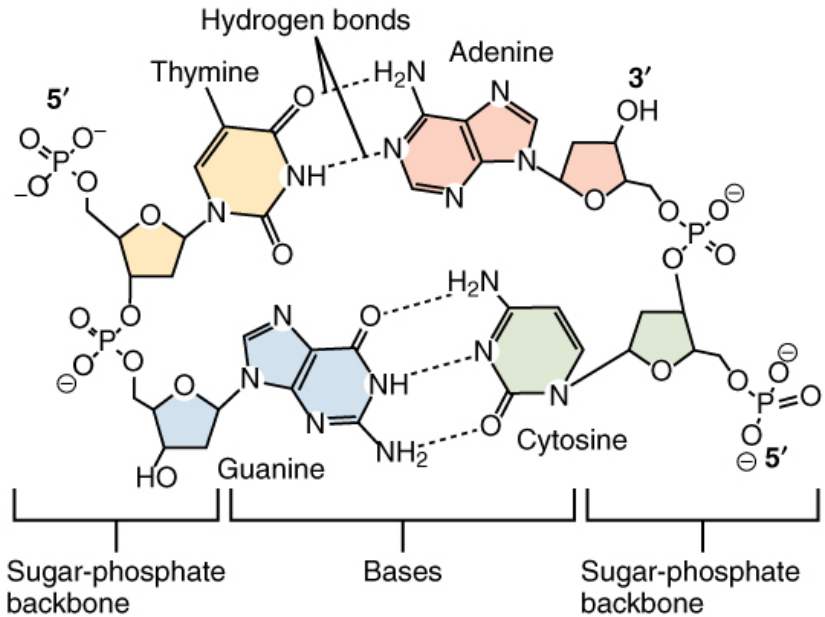
Martin Jakt

August 22, 2024

DNA

- Very long polymer made up of four different types of units (dA, dC, dT, dG).
- *Read only* memory that contains the information defining the organism.
- Double helix made up of two DNA molecules; provides a mechanism for copying.
- Maintained from generation to generation.

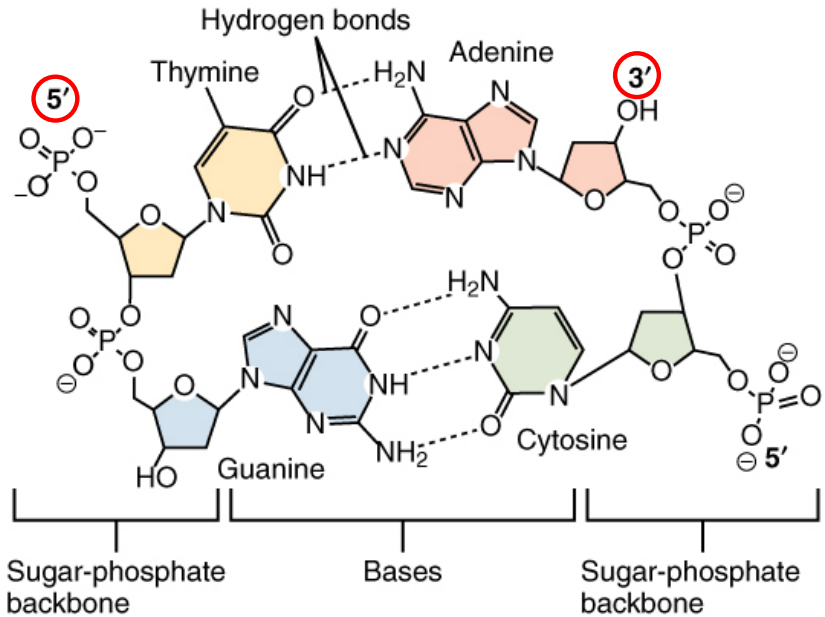
Nucleotides: the units of DNA



1

[1] Image from: "0322 DNA Nucleotides" by OpenStax College - Anatomy & Physiology, Connexions Web site. http://cnx.org/content/col11496/1.6/, Jun 19, 2013.
Licensed under CC BY 3.0 via Wikimedia Commons
https://commons.wikimedia.org/wiki/File:0322_DNA_Nucleotides.jpg#/media/File:0322_DNA_Nucleotides.jpg

Nucleotides: the units of DNA



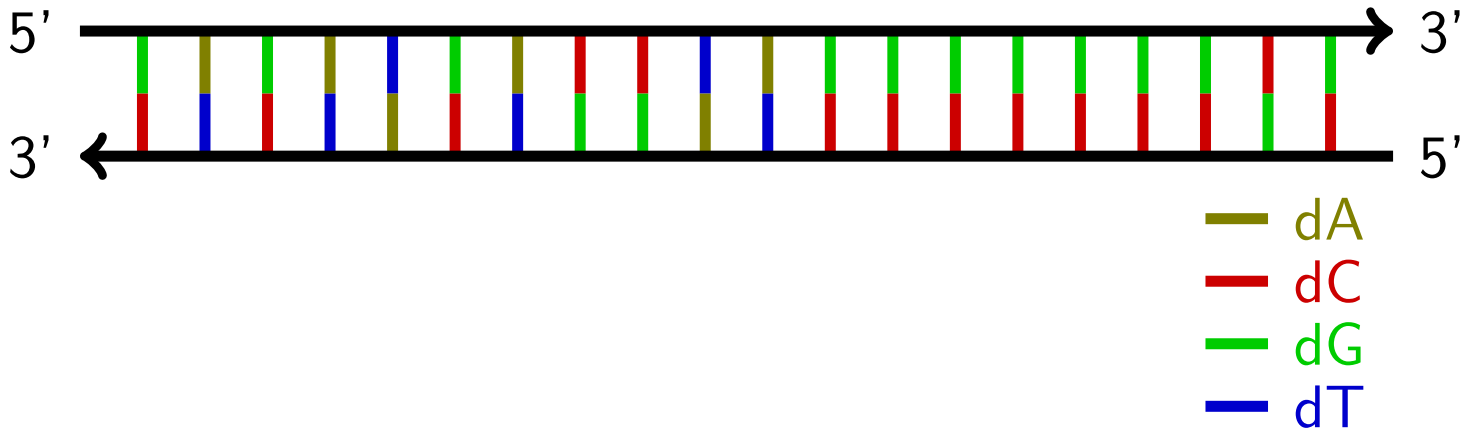
1

---

[1] Image from: "0322 DNA Nucleotides" by OpenStax College - Anatomy & Physiology, Connexions Web site. http://cnx.org/content/col11496/1.6/, Jun 19, 2013.
Licensed under CC BY 3.0 via Wikimedia Commons
https://commons.wikimedia.org/wiki/File:0322_DNA_Nucleotides.jpg#/media/File:0322_DNA_Nucleotides.jpg

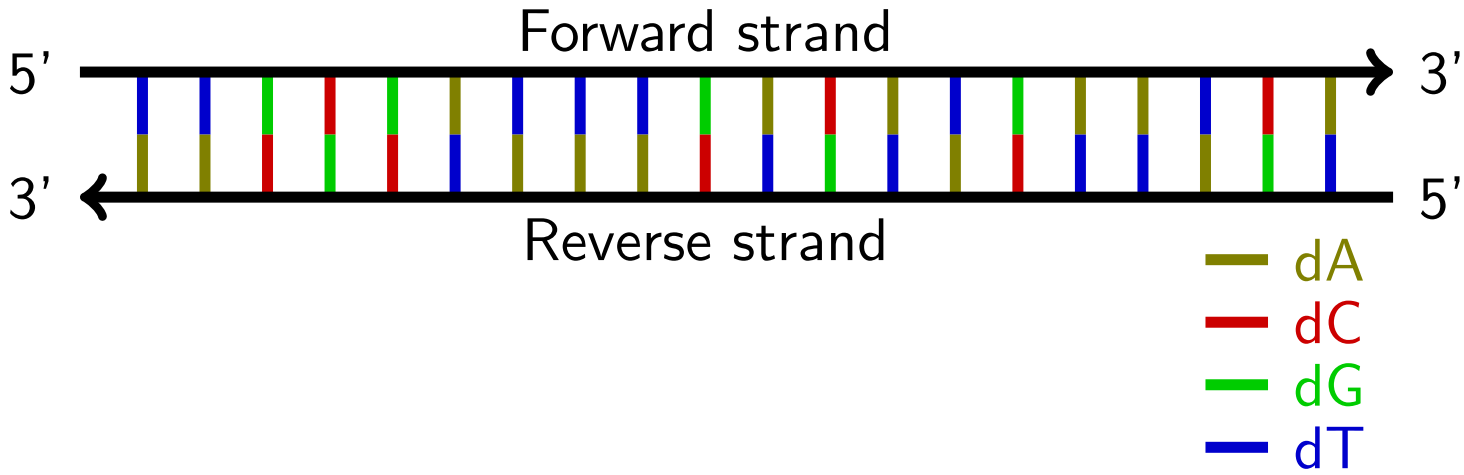
DNA Structure (1)

A linear representation



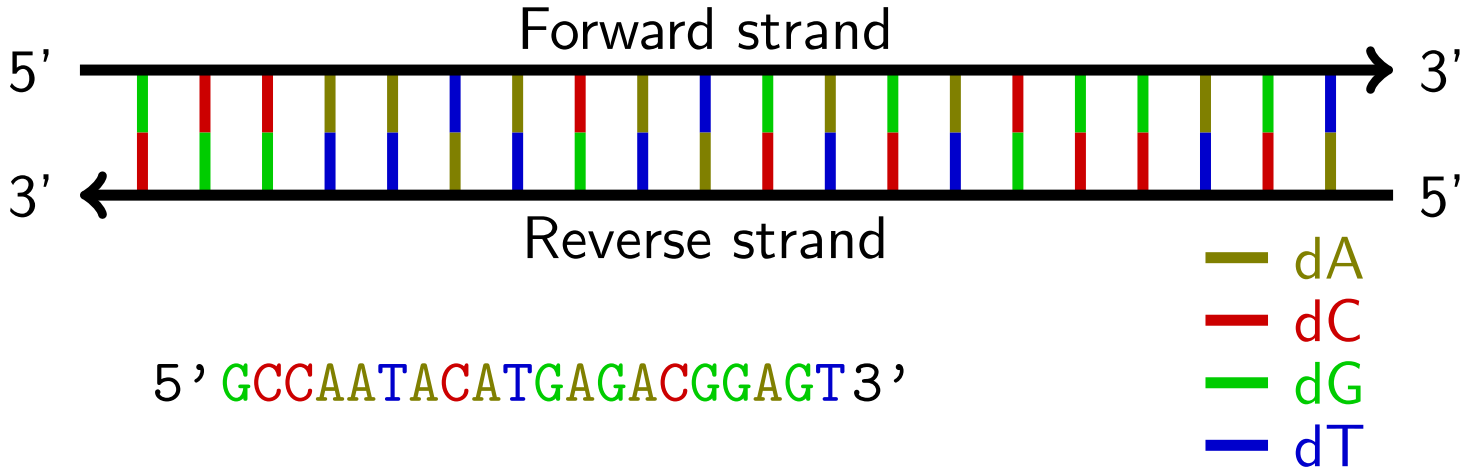
DNA Structure (1)

A linear representation



DNA Structure (1)

A linear representation



DNA Structure (2)

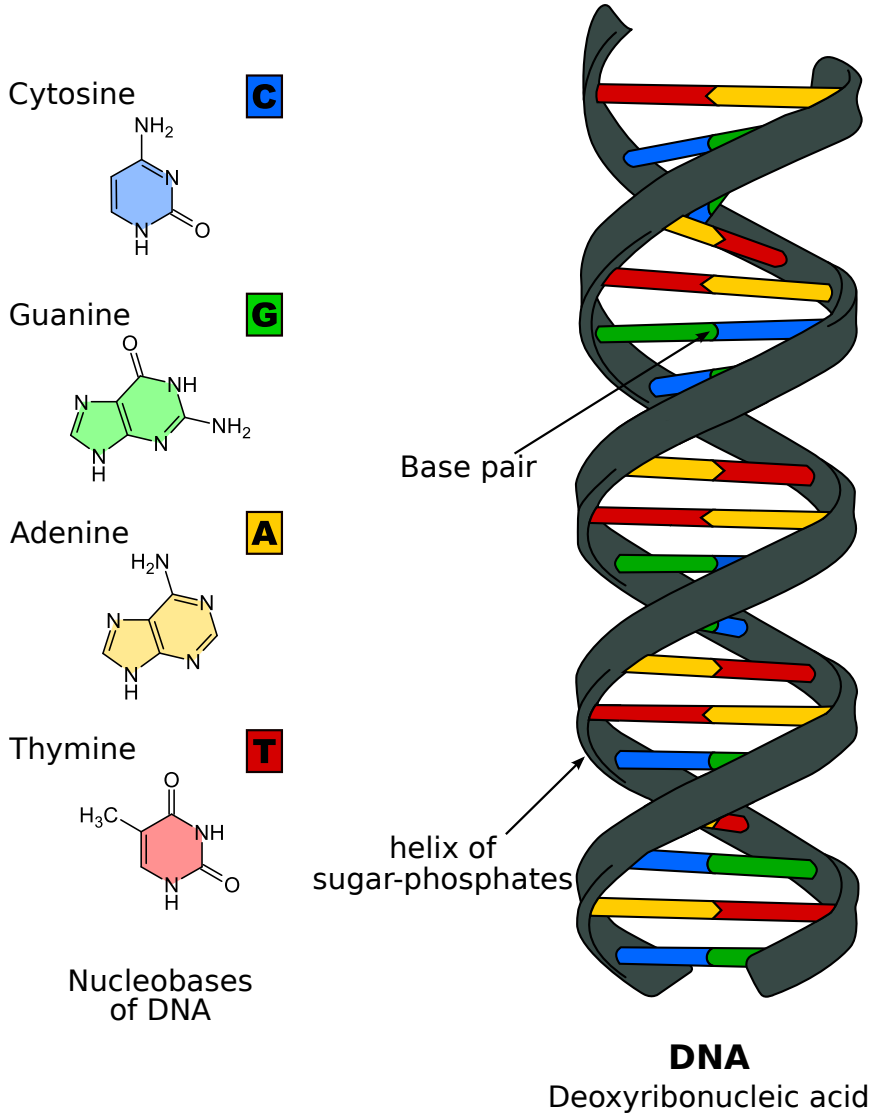
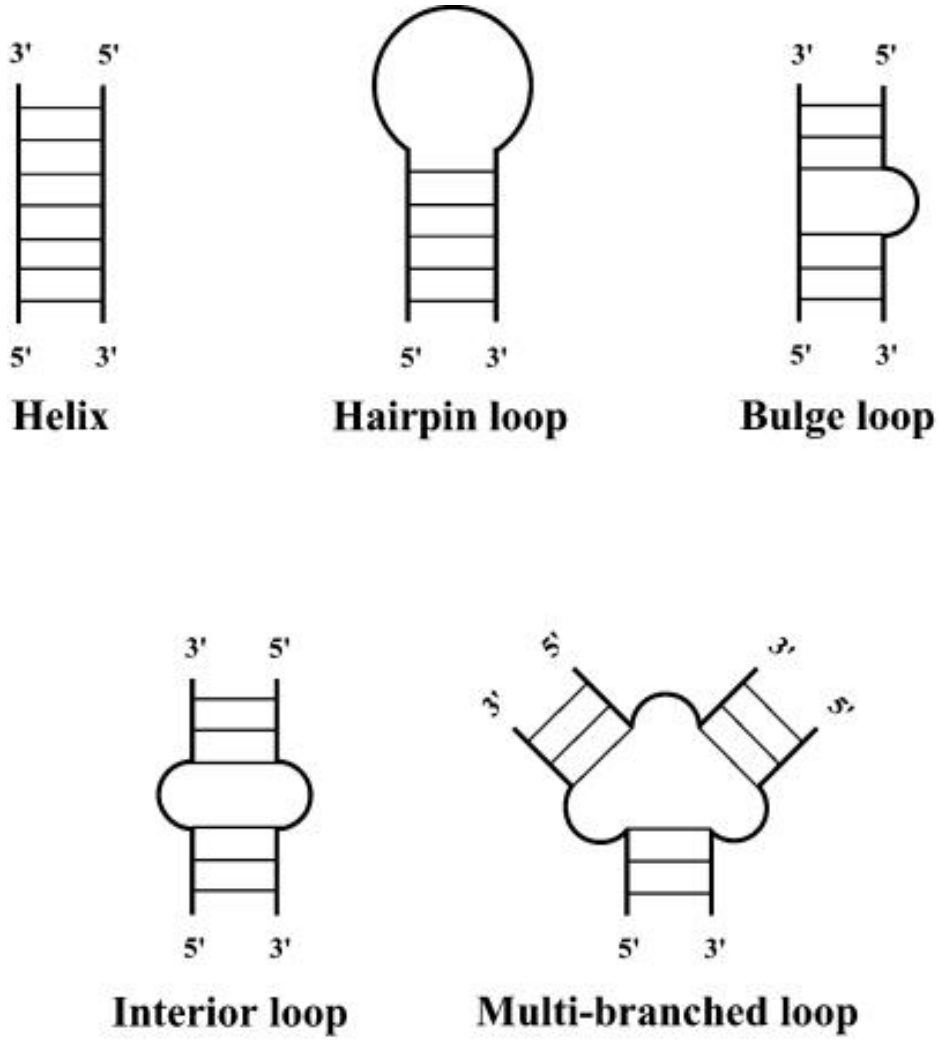


Image modified from:
chemical structures of nucleobases by Roland1952. Licensed under CC BY-SA 3.0 via Wikimedia Commons
https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg

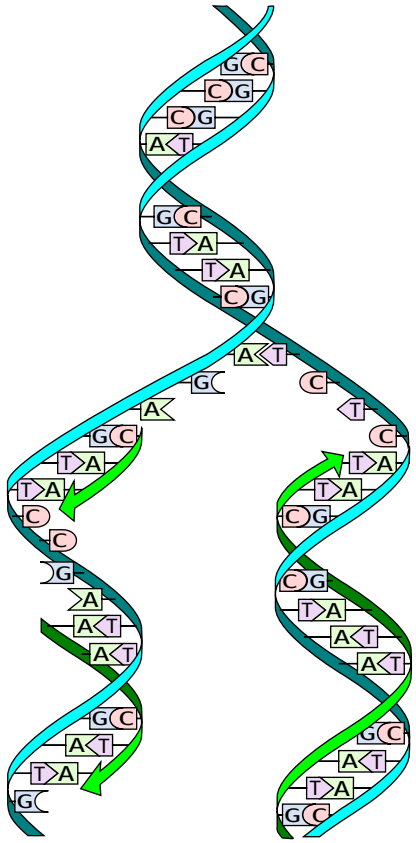
Base pairing to structure



Single stranded RNA / DNA molecules can form complex structures.

Figure 1, from:
Nucleic Acids Res. 2003 Dec 15; 31(24):7280-7301

Copying DNA

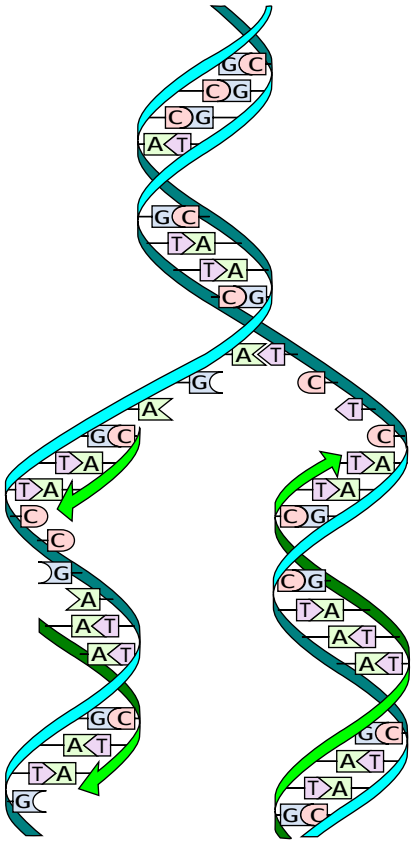


2

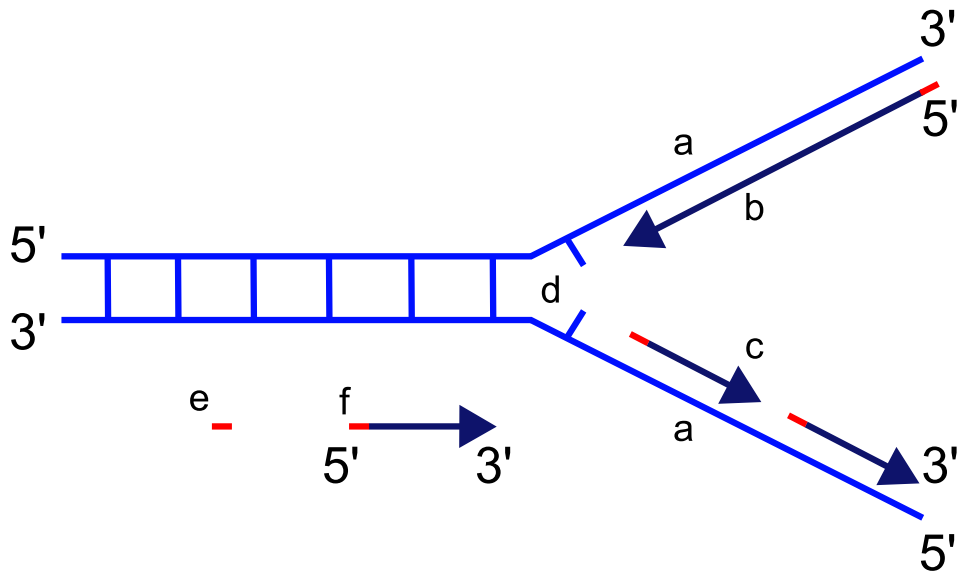
[1] https://commons.wikimedia.org/wiki/File:DNA_Replication_split.svg

[2] https://commons.wikimedia.org/wiki/File:Replication_fork.svg

Copying DNA



3



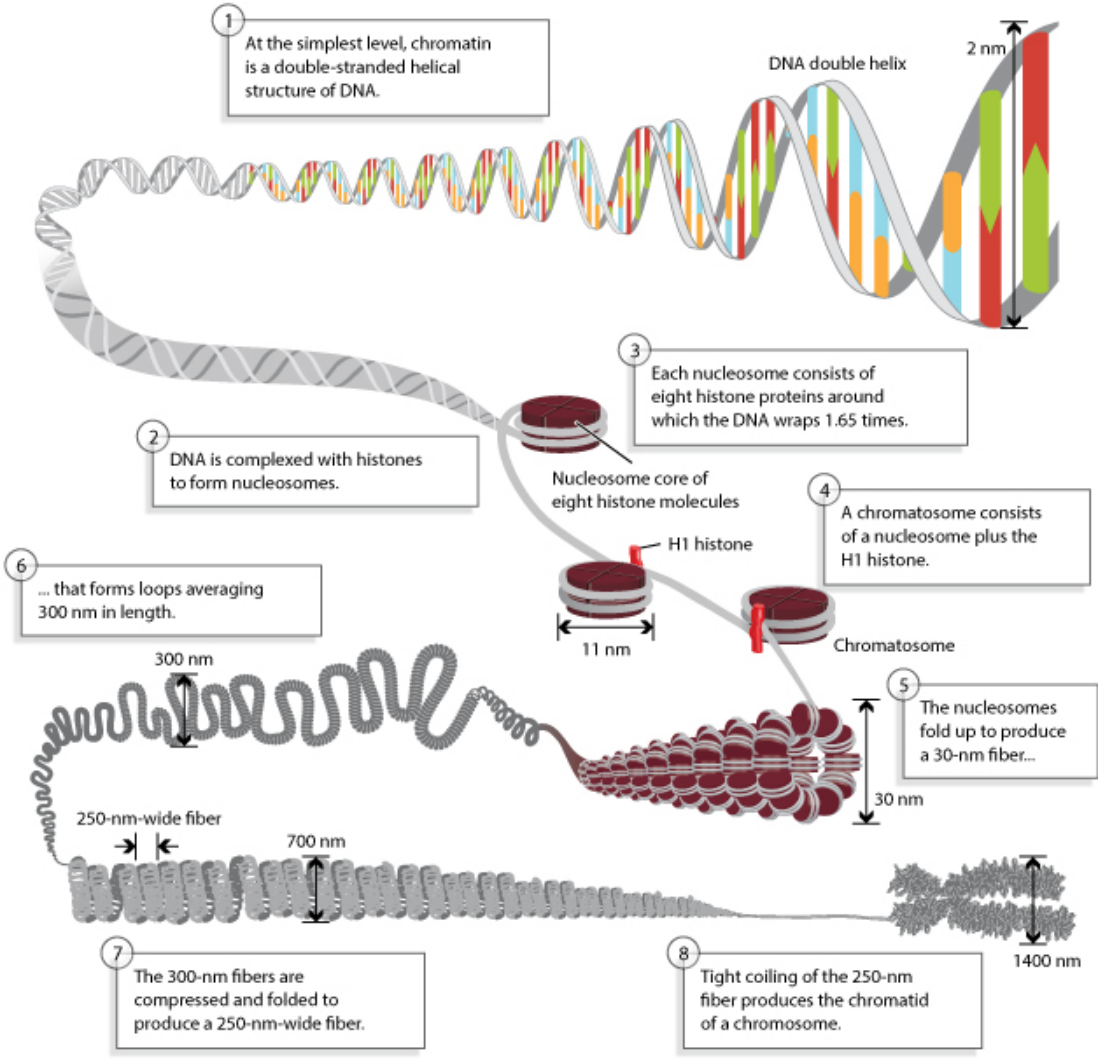
2

[1] https://commons.wikimedia.org/wiki/File:DNA_Replication_split.svg

[2] https://commons.wikimedia.org/wiki/File:Replication_fork.svg

Packaging of DNA in the Nucleus

1 bp $\sim$0.34nm
6×10^9 bp / diploid genome
$\sim$2m



DNA Packaging: Nucleosomes and Chromatin
Anthony T. Annunziato, Ph.D. (Biology Department, Boston College) © 2008 Nature Education
http://www.nature.com/scitable/topicpage/dna-packaging-nucleosomes-and-chromatin-310

Packaging and gene regulation

- Histones can be modified by methylation, acetylation, sumoylation, phosphorylation, biotinilation, ubiquitination at specific residues.
- Specific modifications are correlated with:
  - Active transcription
  - Repressed state
  - Gene features and their states (eg. promoters / enhancers / splice sites(?))
- Histone modifications both set and read by transcription factors.
- Large number ($\sim$70) of histone modifications known with a potentially huge combinatorial code.
- Specific histone modifications are also associated with DNA methylation.

For more details:

`https://www.cellsignal.com/common/content/content.jsp?id=science-tables-histone`

`http://www.activemotif.com/documents/1815.pdf`

The informatician's representation

```
ACTGATAGA
|||||||||
TGACTATCT
```

The informatician's representation

```
ACTGATAGA
|||||||||
TGACTATCT
```

or more simply as:

```
5' ACTGATAGA 3'
```

The informatician's representation

```
ACTGATAGA
|||||||||
TGACTATCT
```

or more simply as:

```
5' ACTGATAGA 3'
```

Simply the sequence of the nucleotides
For human about 3×10^9 letters.

How to read DNA sequence

The Gene

- A functional unit of DNA.
- Contains a region that is transcribed into RNA and encodes the amino acid sequence of a protein or a functional RNA molecule.
- Includes the regions that determine when the gene is active (RNA is transcribed).

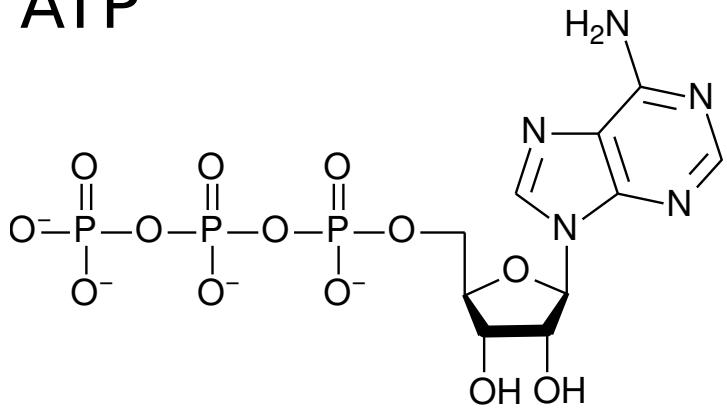
As always, this is a bit of an over-simplification.

What is RNA?

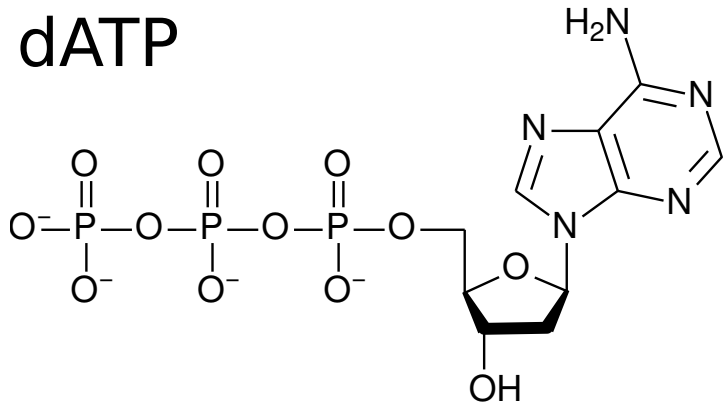
- Like DNA but made up of (A, C, U, G)
- Chemically very similar to the DNA but less stable
- The units contain three of the same bases as DNA and can base-pair with DNA molecules (forming hybrid double stranded molecules).
- Uracil (U) base instead of T found in DNA molecules. Functionally equivalent and can base pair with A.
- RNA molecules transmit genetic information from the DNA, and are used either as functional RNA molecules or to encode protein structures.

DNA and RNA monomers

ATP



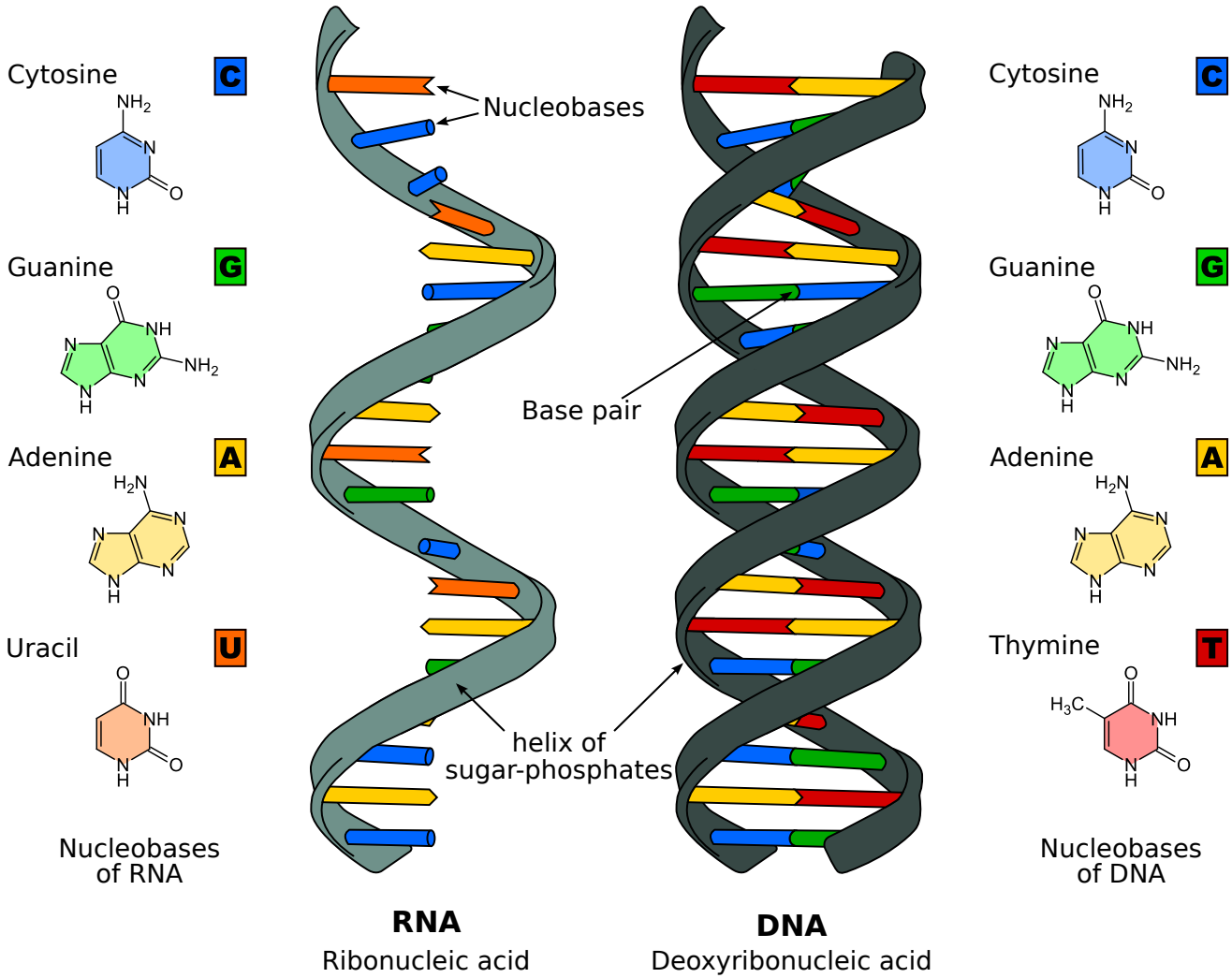
dATP



Spot the difference!

image modified from:
http://www.wikidoc.org/index.php/Nucleotide

DNA & RNA structure



RNA can also be double stranded.

image from:
chemical structures of nucleobases by Roland1952. Licensed under CC BY-SA 3.0 via Wikimedia Commons
https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg

What is Protein?

- Polymer molecules made up of chains of 20 different amino acids.
- Make up both structural (e.g. cytoskeleton) and functional (eg. enzymes) components of the cells.
- Can form an almost infinite variety of shapes that are determined by their amino acid sequence (primary structure).
- The amino acid sequence determines how the protein molecule folds into specific shape and also it's chemical properties.

Amino Acids

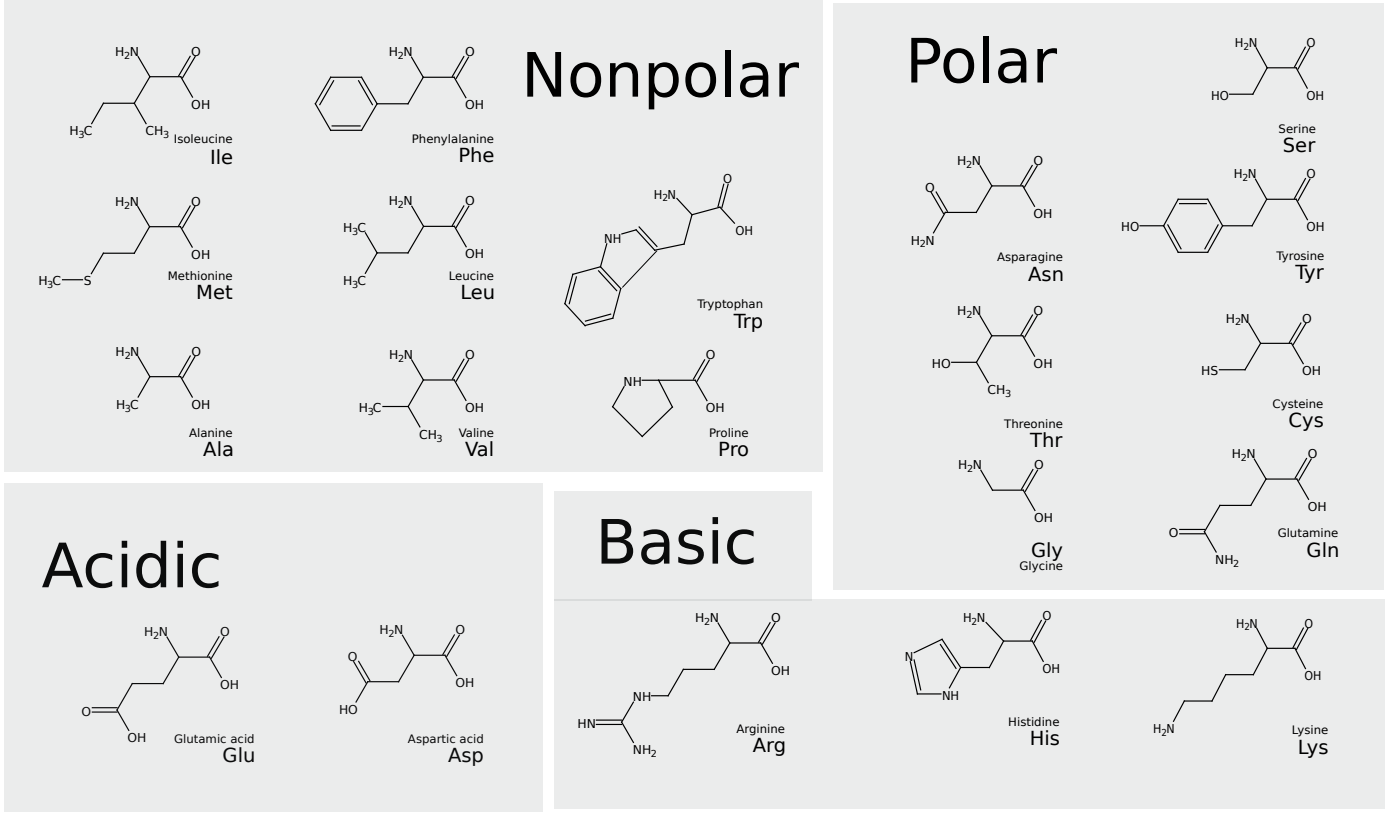
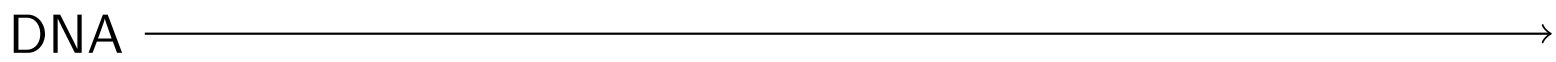


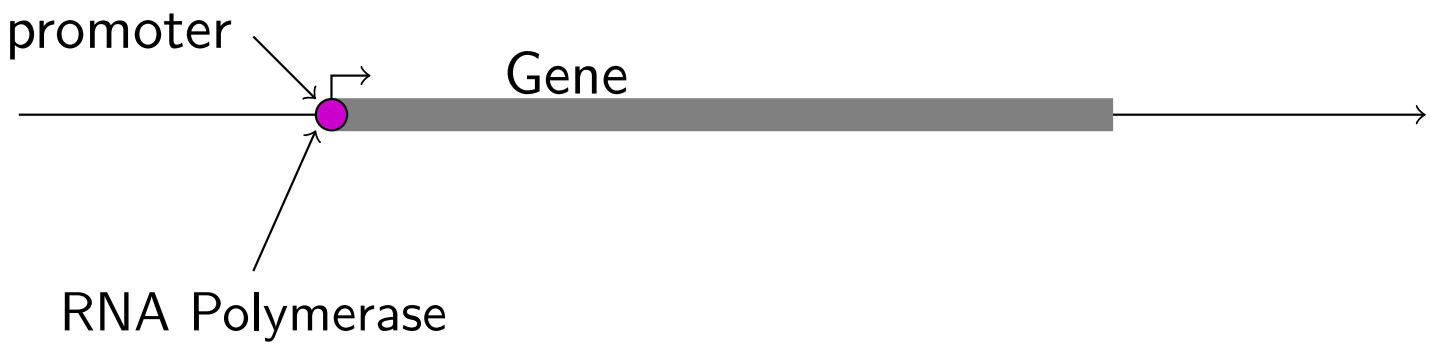
Image elements taken from:
https://en.wikipedia.org/wiki/Genetic_code#/media/File:GeneticCode21-version-2.svg
originally by Abgent.

The Central Dogma

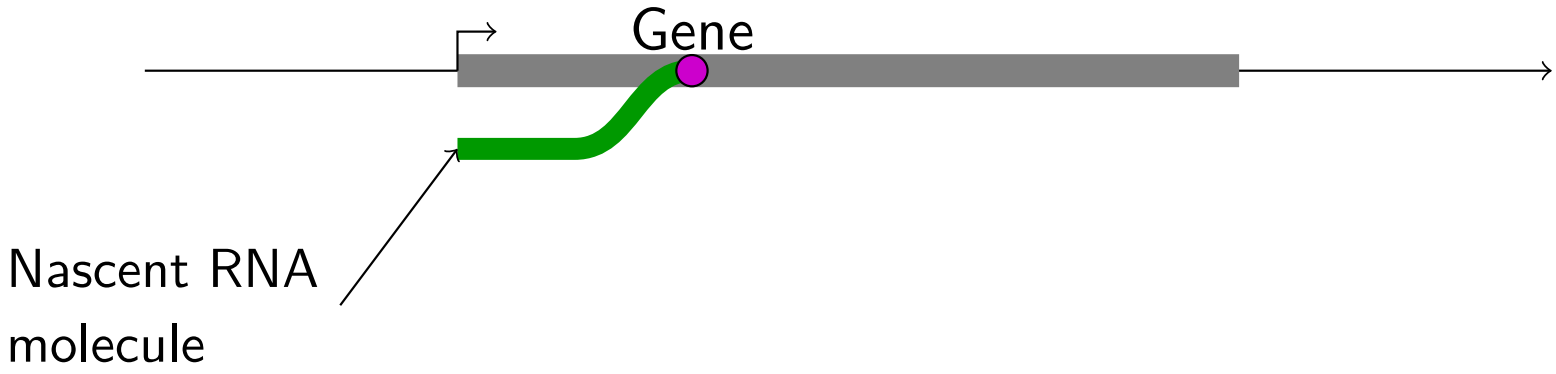


The Central Dogma

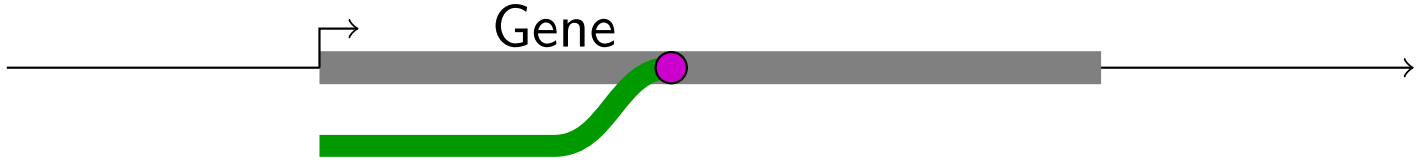
The Central Dogma



The Central Dogma

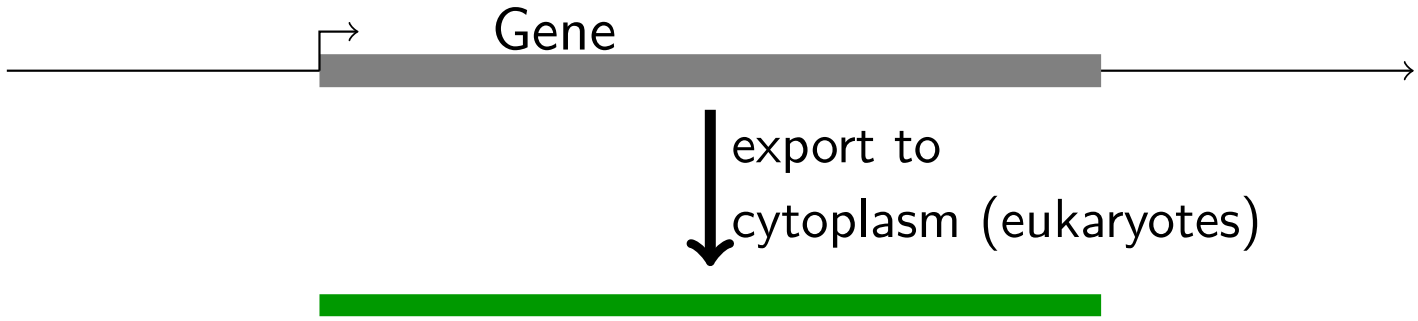


The Central Dogma

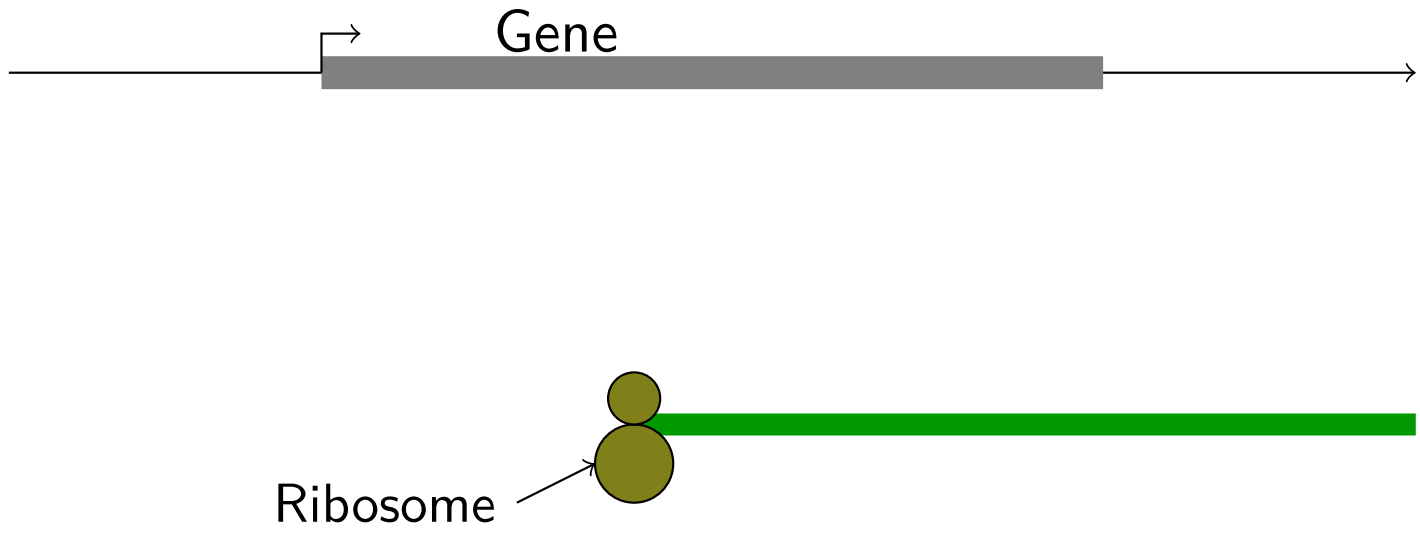


The Central Dogma

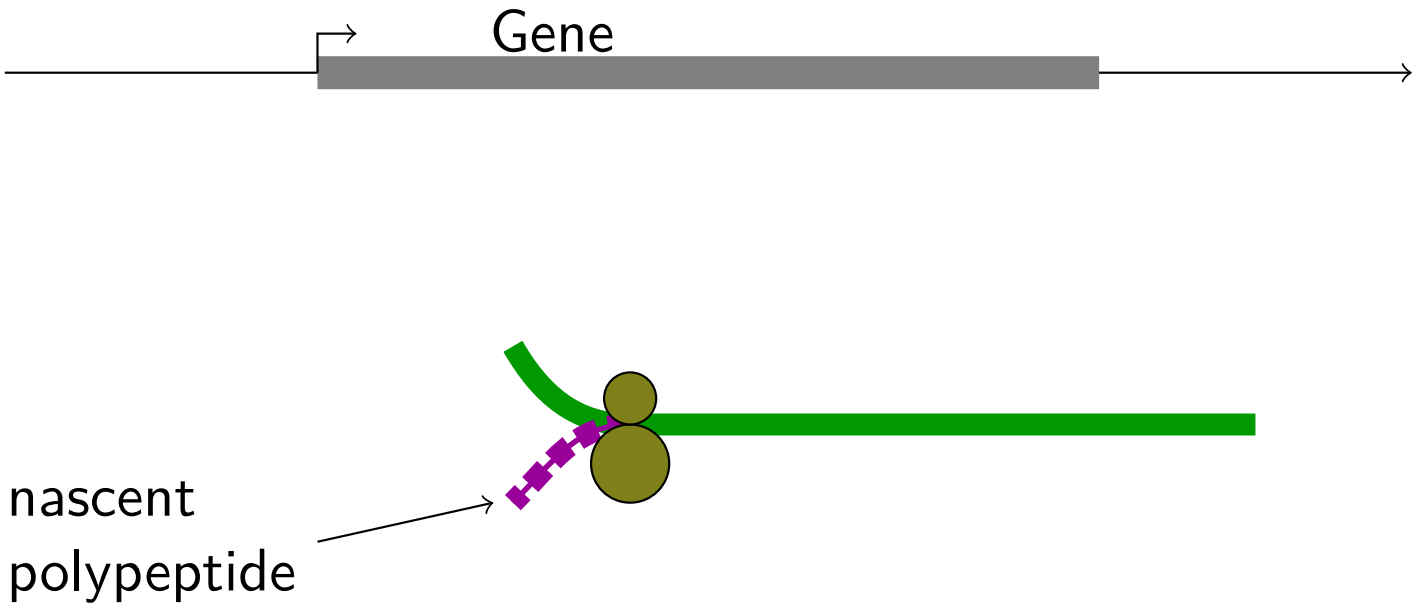
The Central Dogma



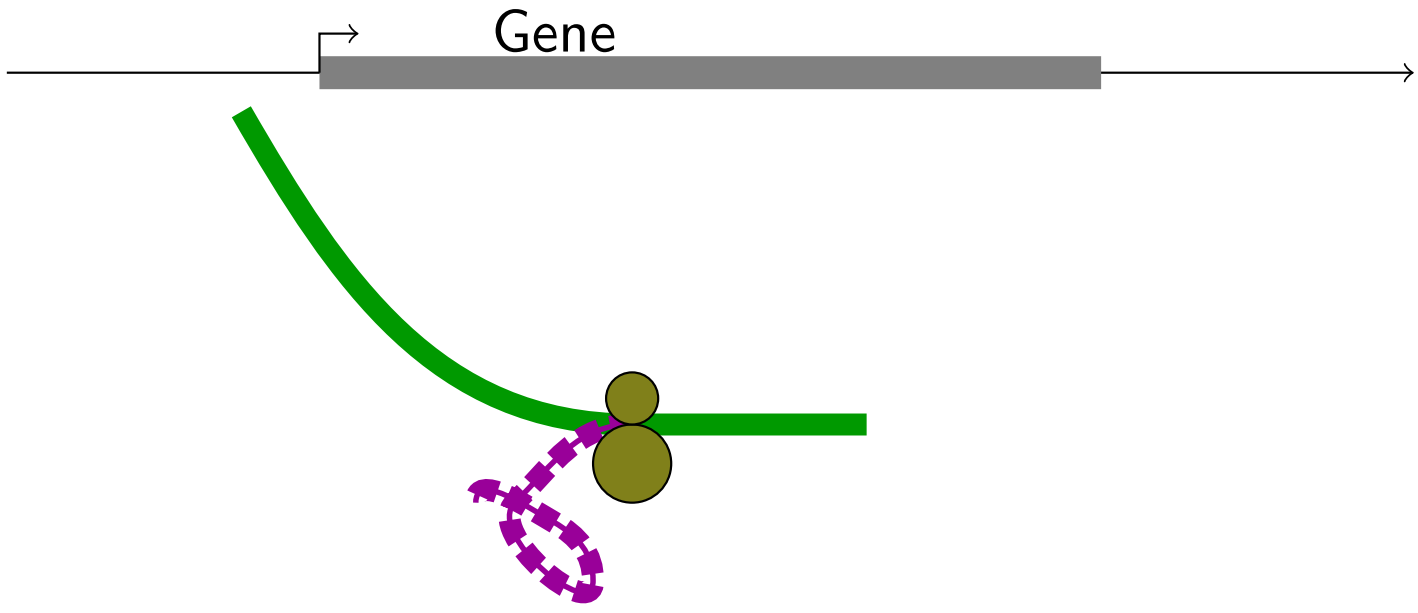
The Central Dogma



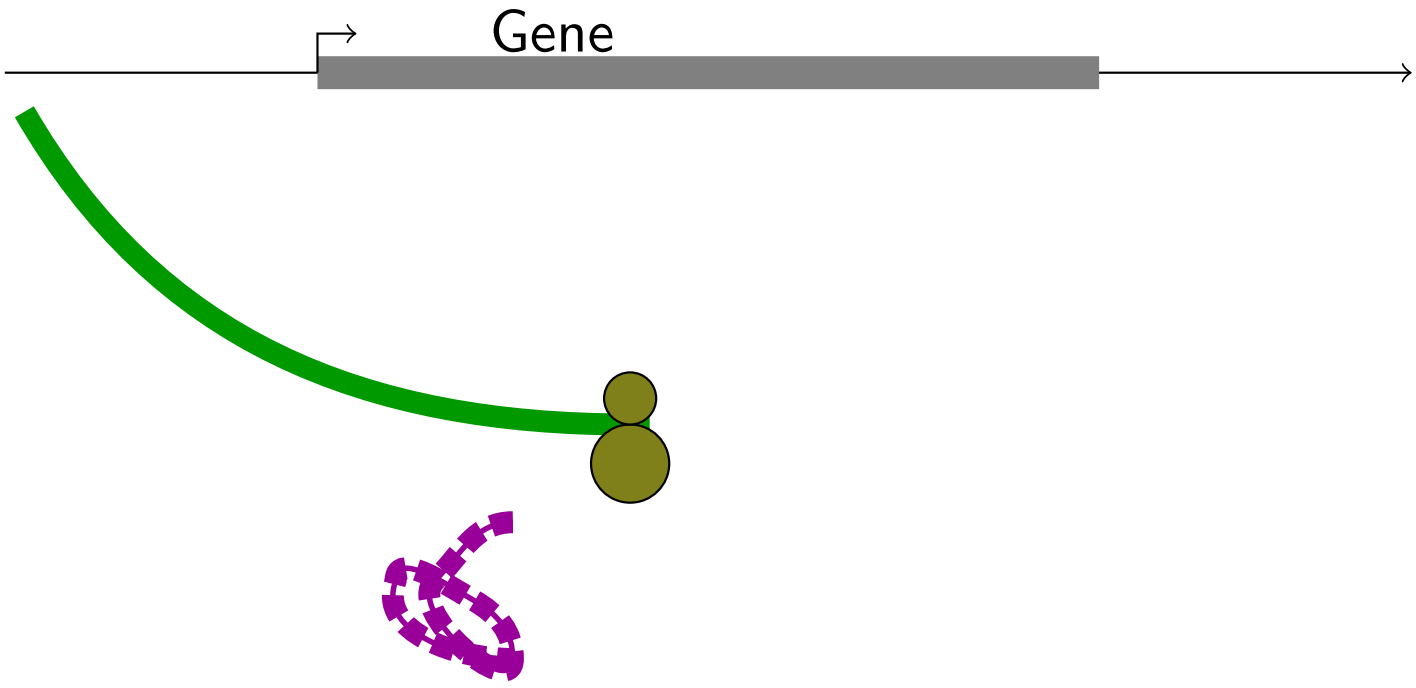
The Central Dogma



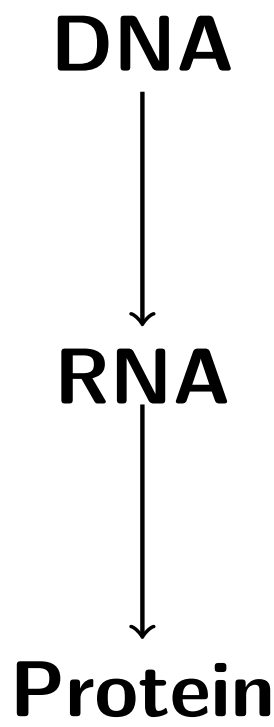
The Central Dogma



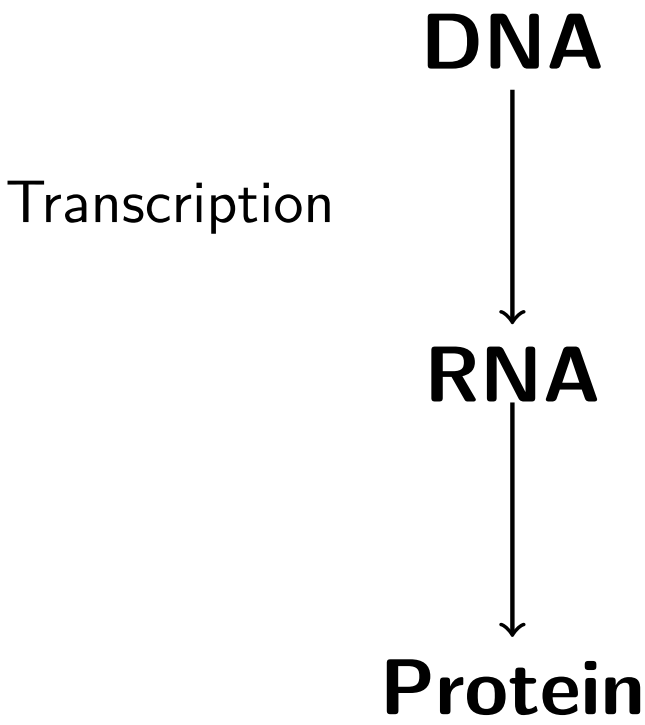
The Central Dogma



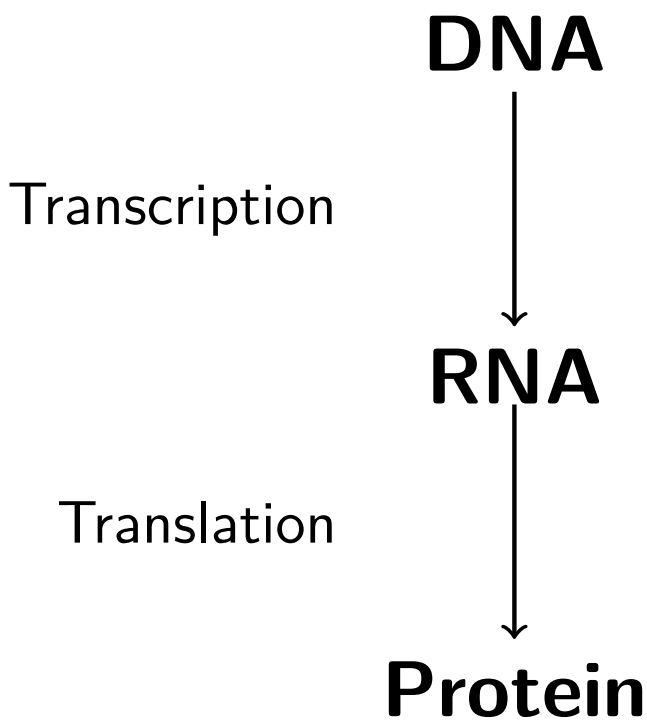
The Central Dogma



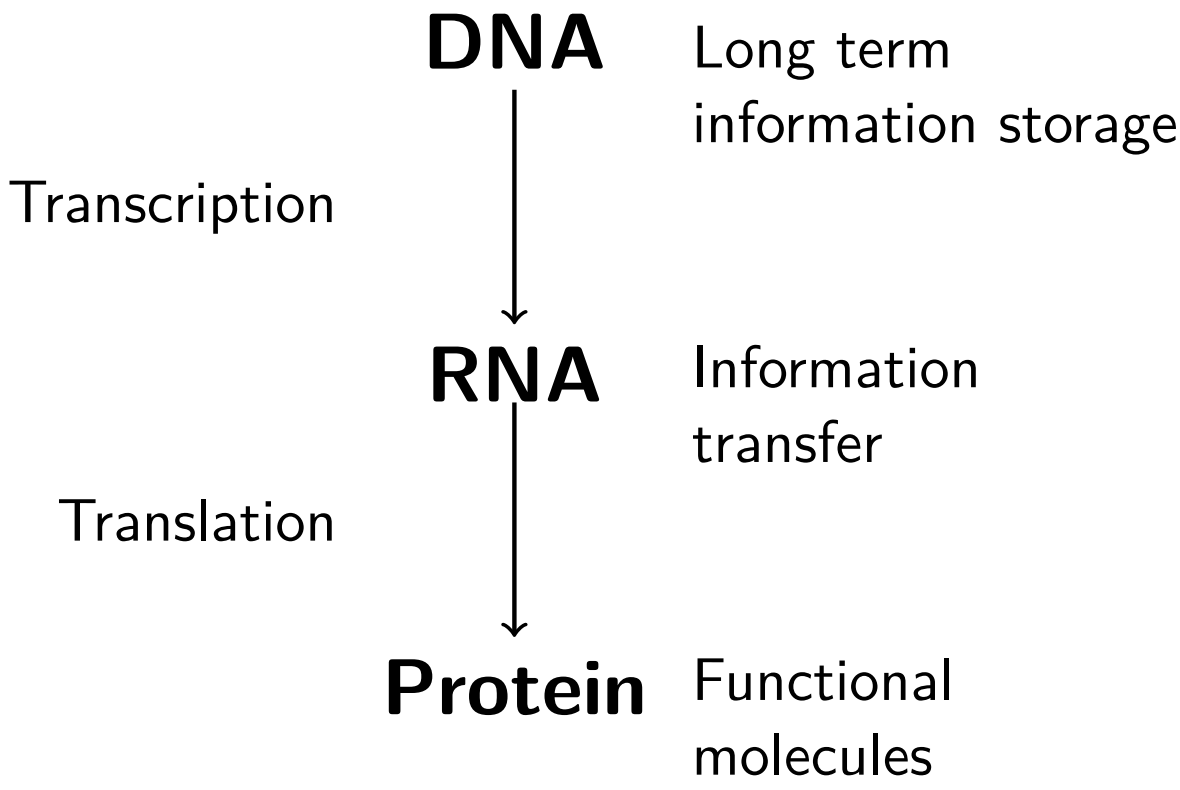
The Central Dogma



The Central Dogma

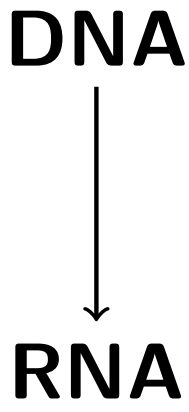


The Central Dogma



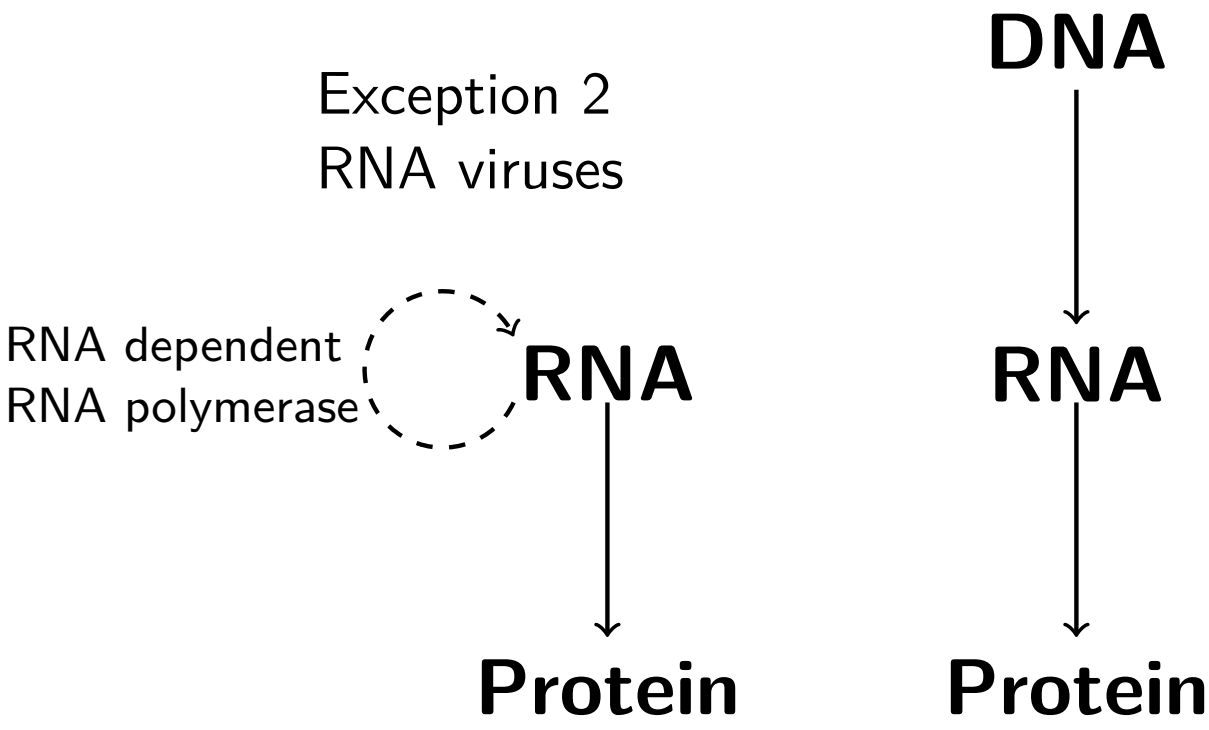
The Central Dogma

Exception 1

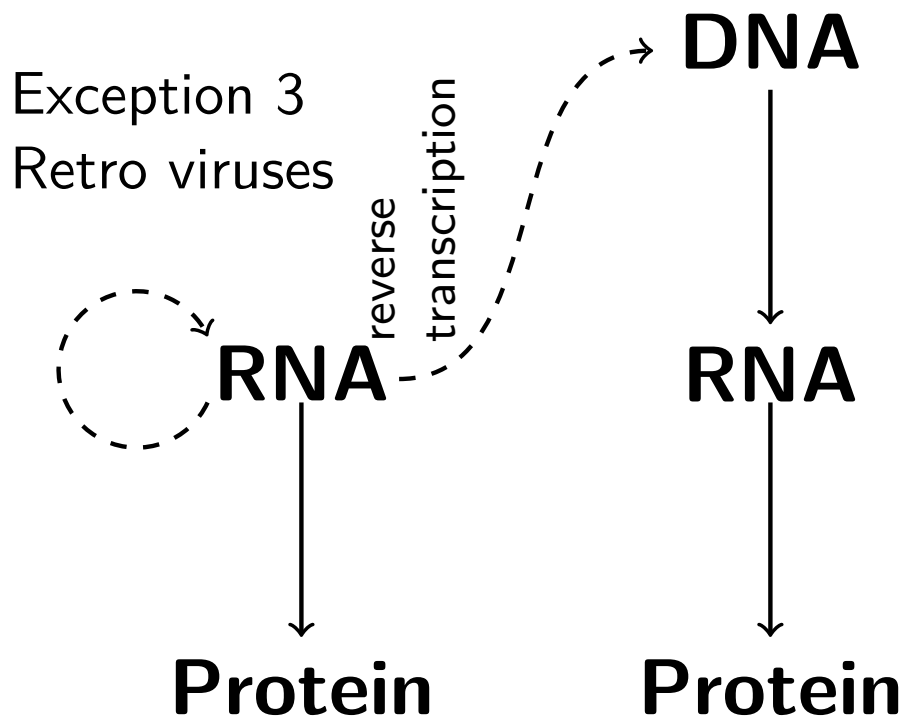


Functional
RNA molecules

The Central Dogma

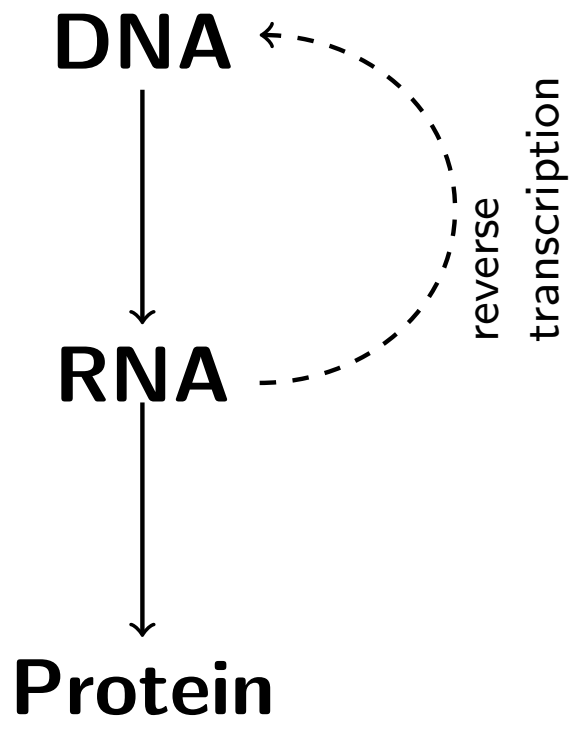


The Central Dogma

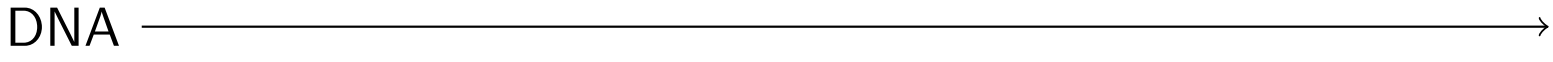


The Central Dogma

Exception 4
Retro-transposons



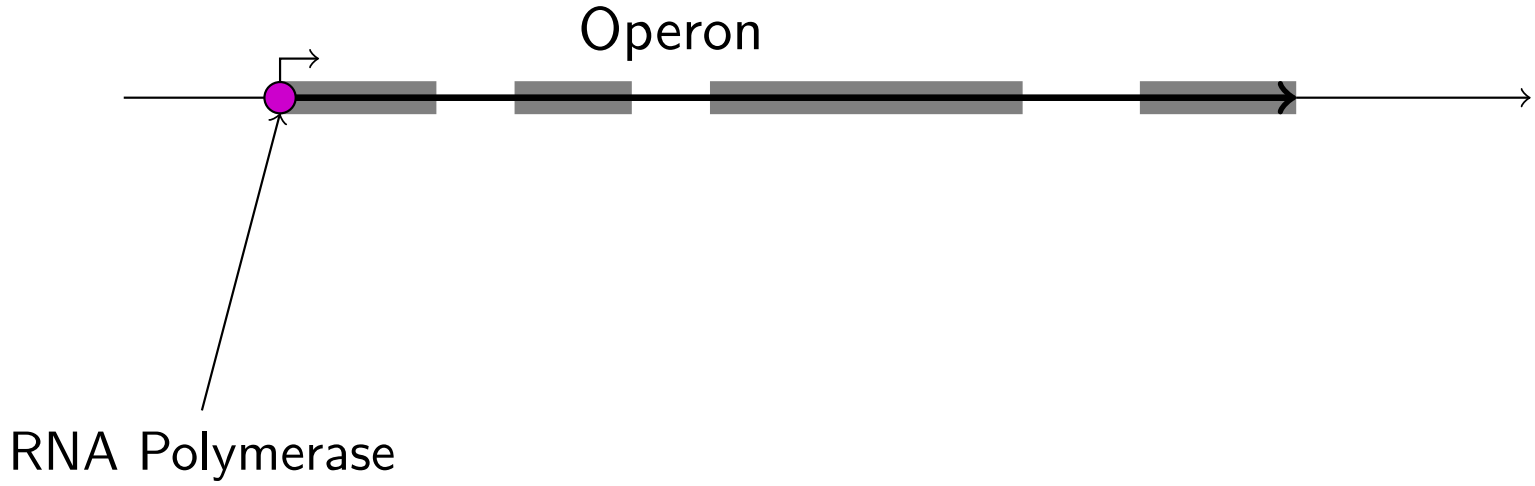
Operons and polycistronic messages



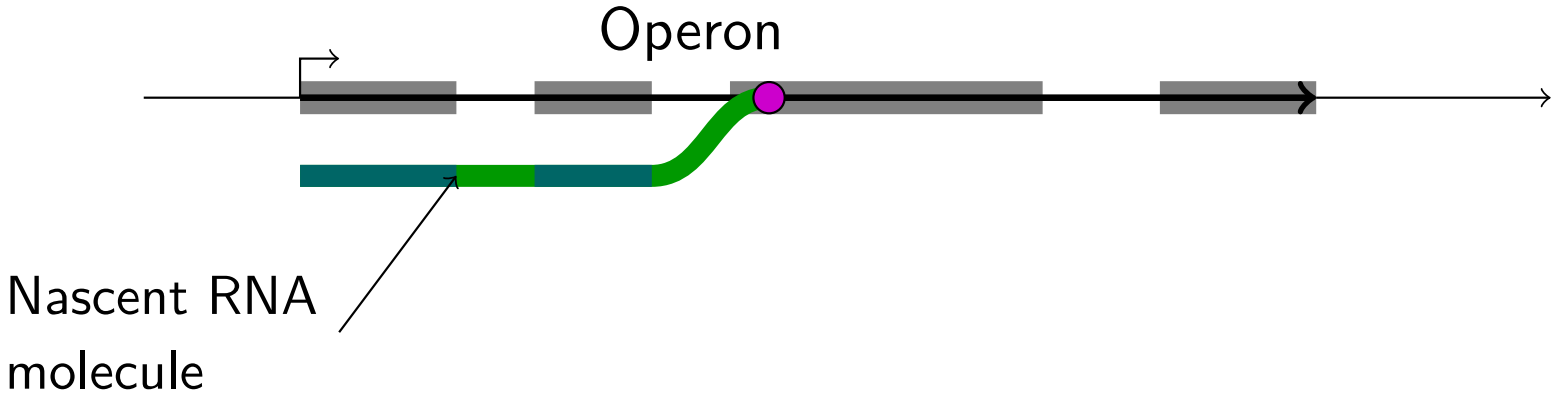
Operons and polycistronic messages

Genes in prokaryotes

Operons and polycistronic messages

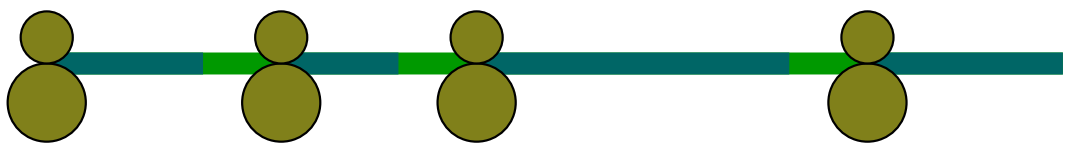


Operons and polycistronic messages

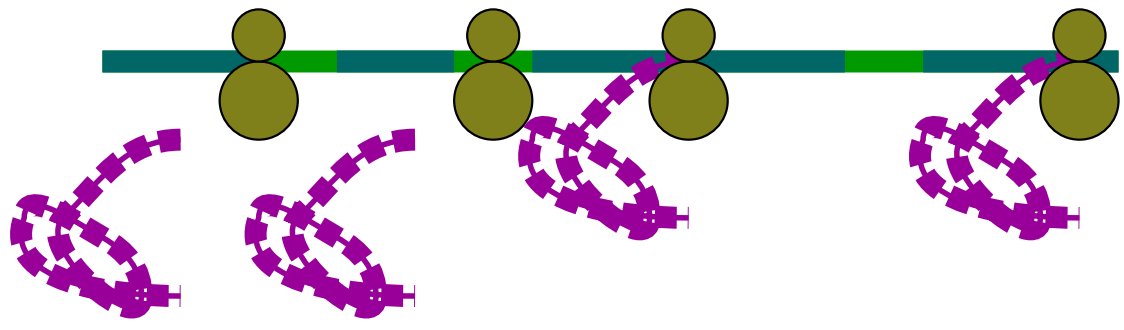


Operons and polycistronic messages

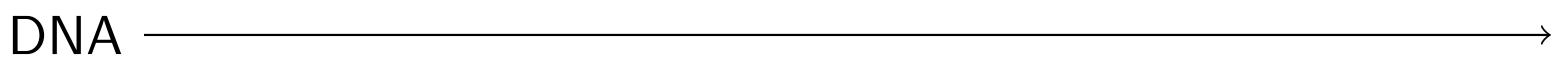
Operons and polycistronic messages



Operons and polycistronic messages



Introns and Exons

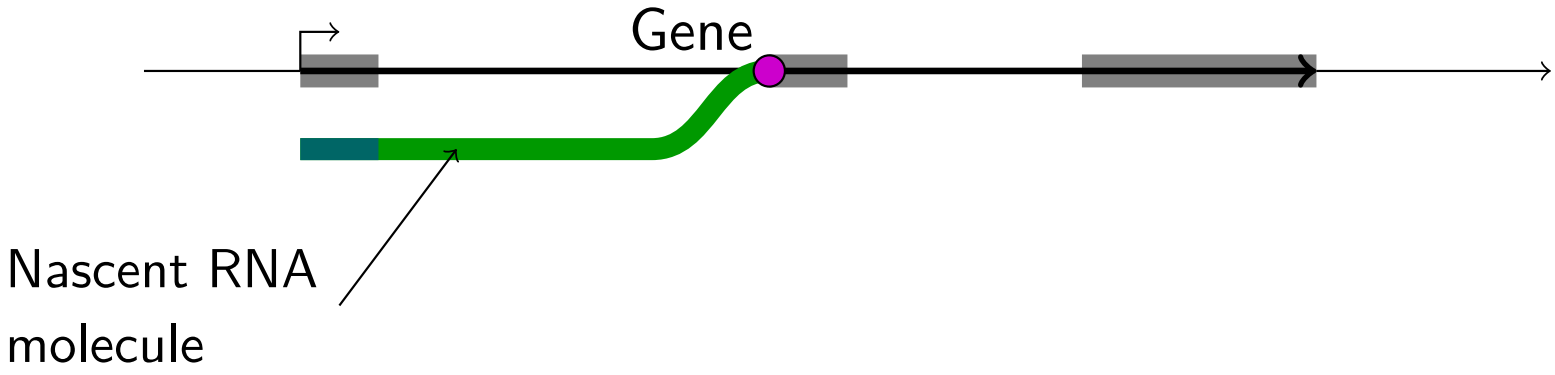


Genes in eukaryotes

Introns and Exons

Genes in eukaryotes

Introns and Exons

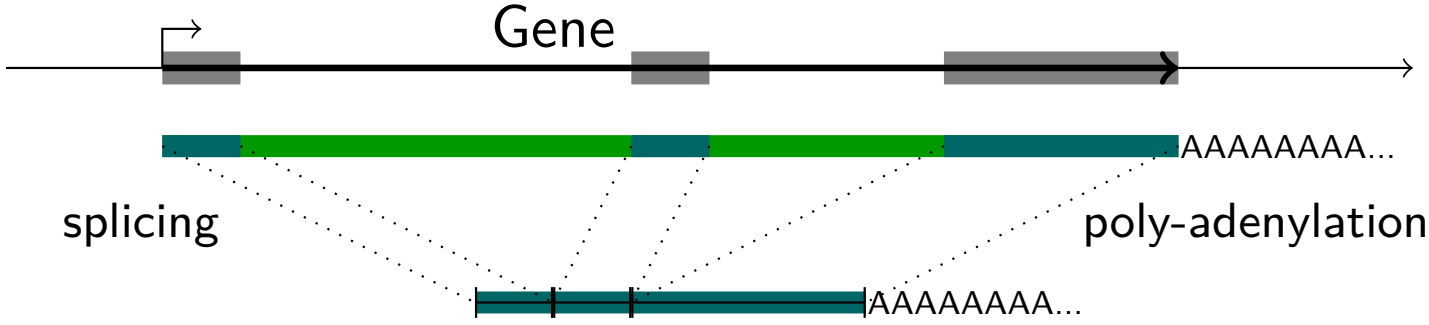


Genes in eukaryotes

Introns and Exons

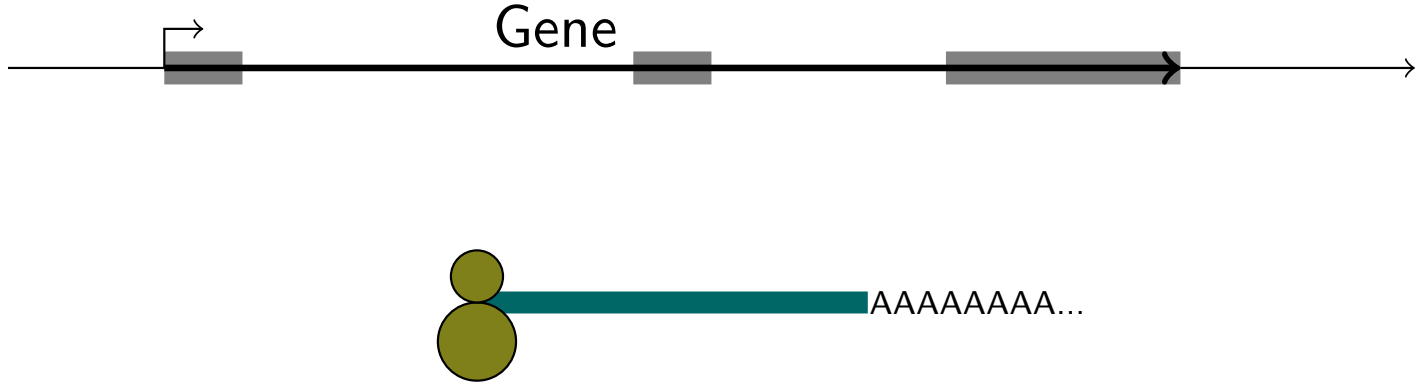
Genes in eukaryotes

Introns and Exons



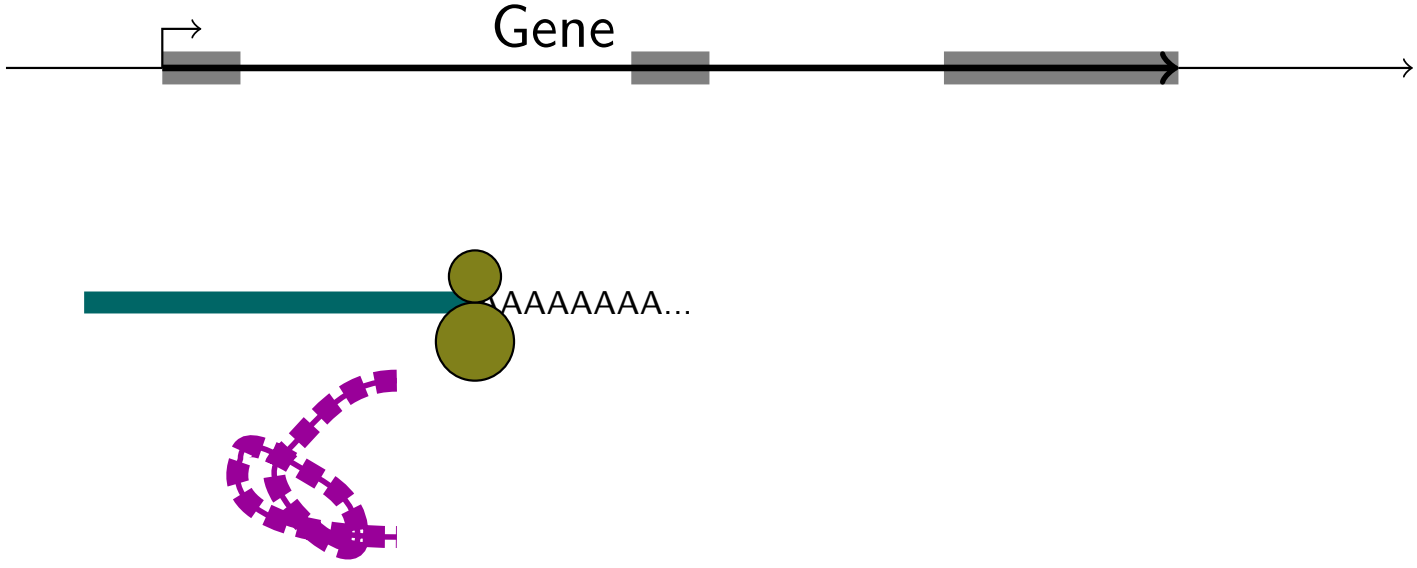
Genes in eukaryotes

Introns and Exons



Genes in eukaryotes

Introns and Exons



Eukaryotes & Prokaryotes

Eukaryotes

- One transcript containing introns and exons
- Introns removed and exons combined by splicing
- Enzymatic (non-template) addition of As at 3' end
- One protein produced

Prokaryotes

- One transcript containing several open reading frames
- Each open reading frame translated seperately
- Several proteins produced

Eukaryotes & Prokaryotes

Eukaryotes

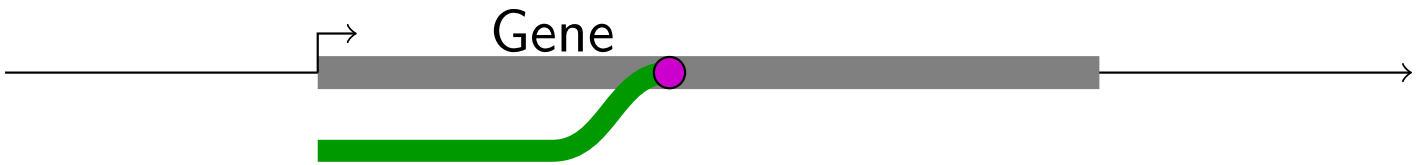
- One transcript containing introns and exons
- Introns removed and exons combined by splicing
- Enzymatic (non-template) addition of As at 3' end
- One protein produced

Prokaryotes

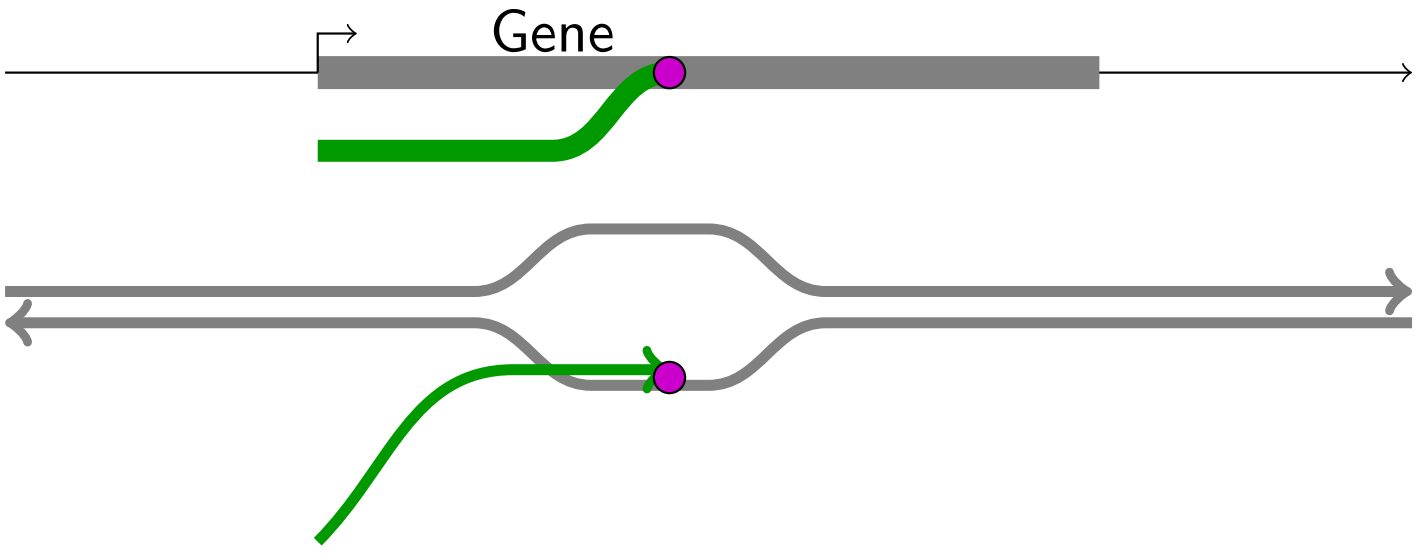
- One transcript containing several open reading frames
- Each open reading frame translated seperately
- Several proteins produced

- Some polycistronic messages identified in eukaryotes.
- Introns can be found in prokaryotes (but *very rare*)

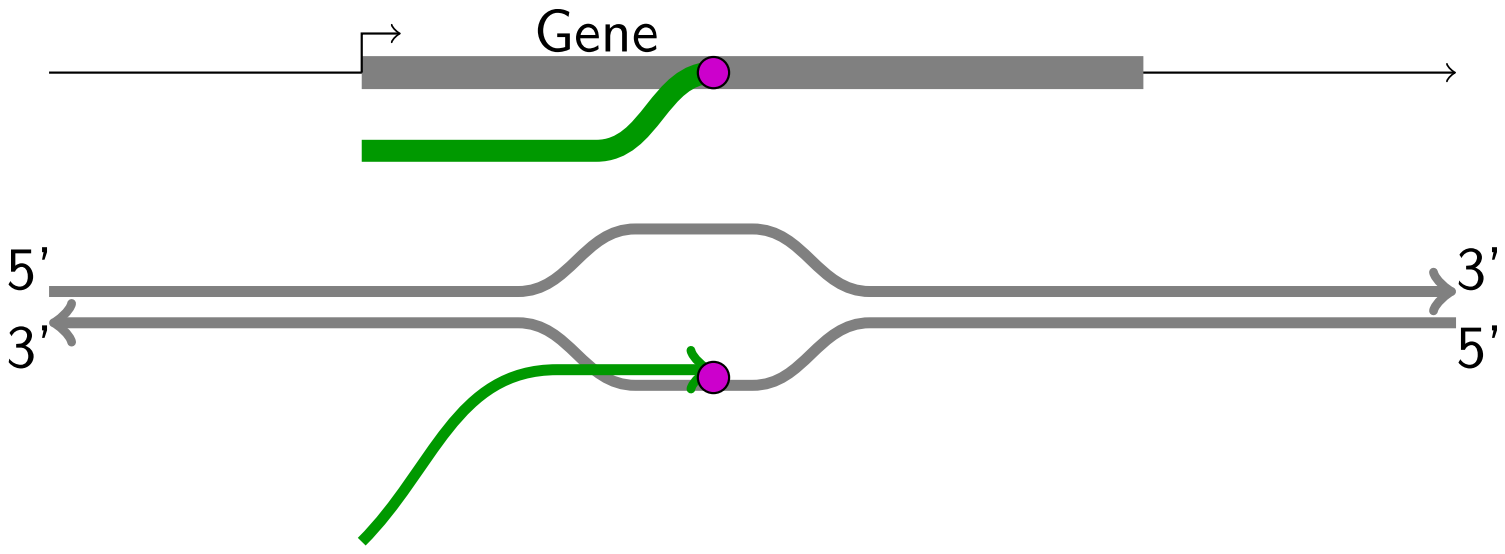
Strandedness!!



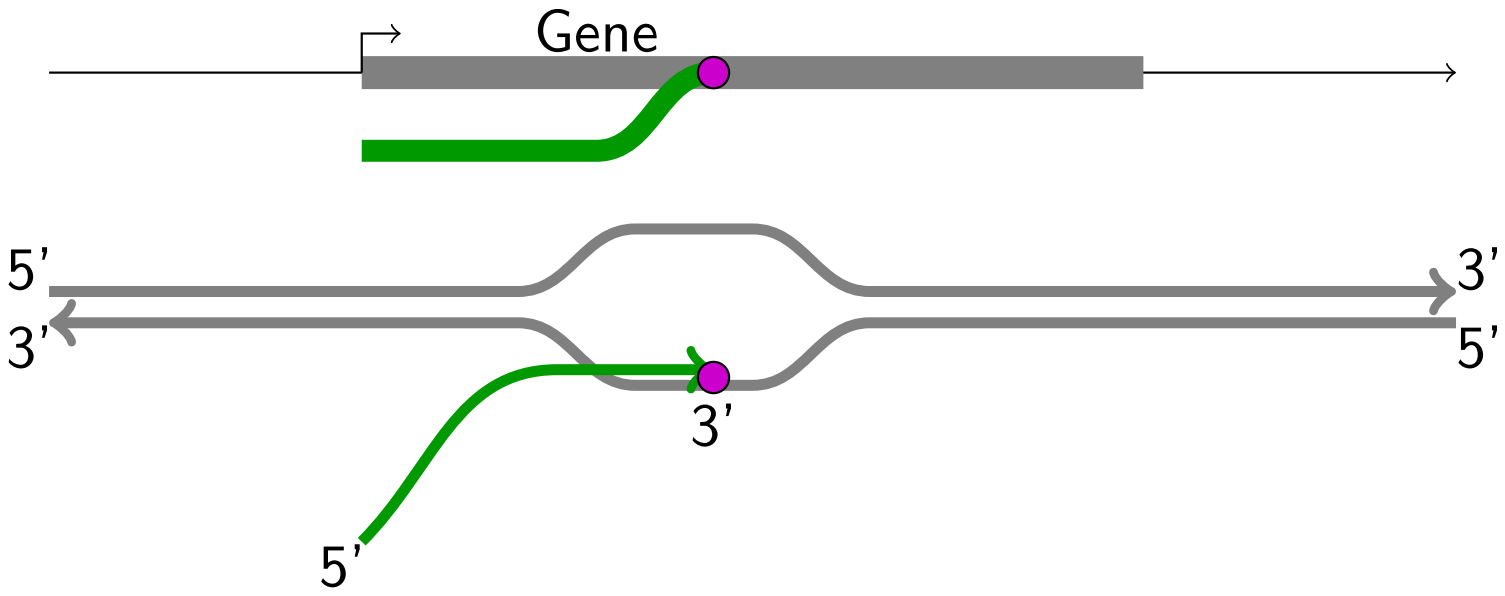
Strandedness!!



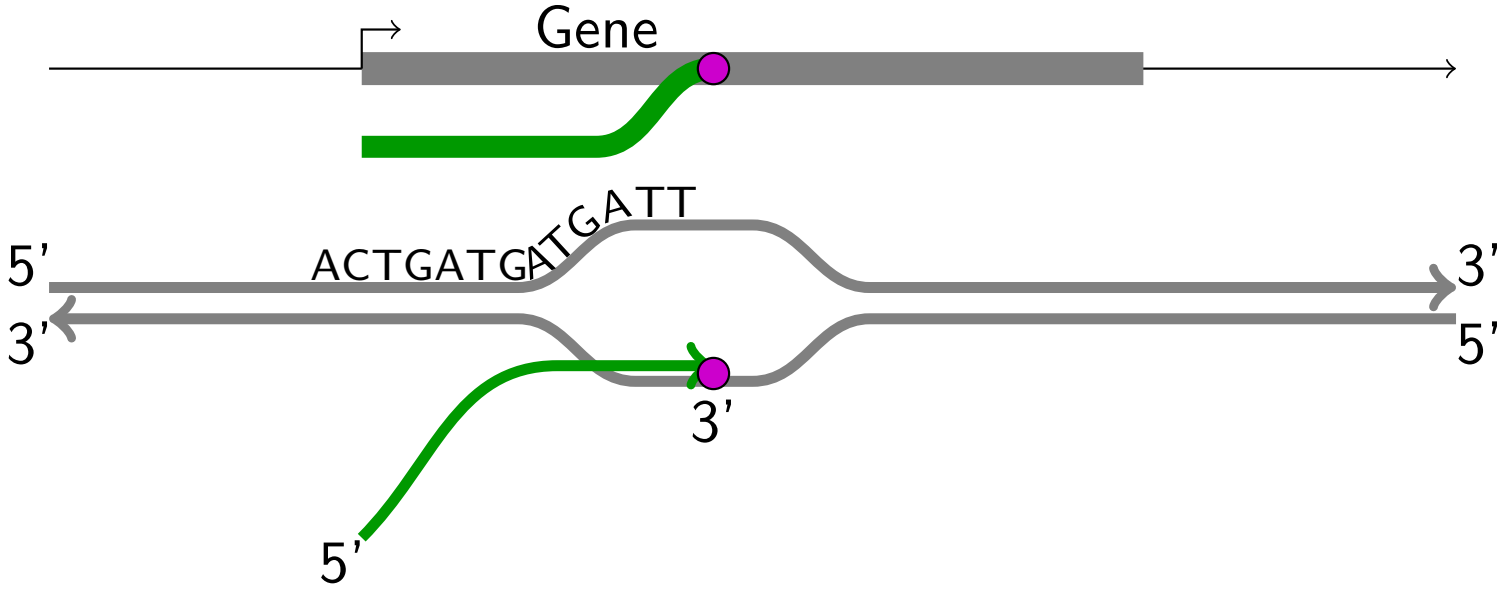
Strandedness!!



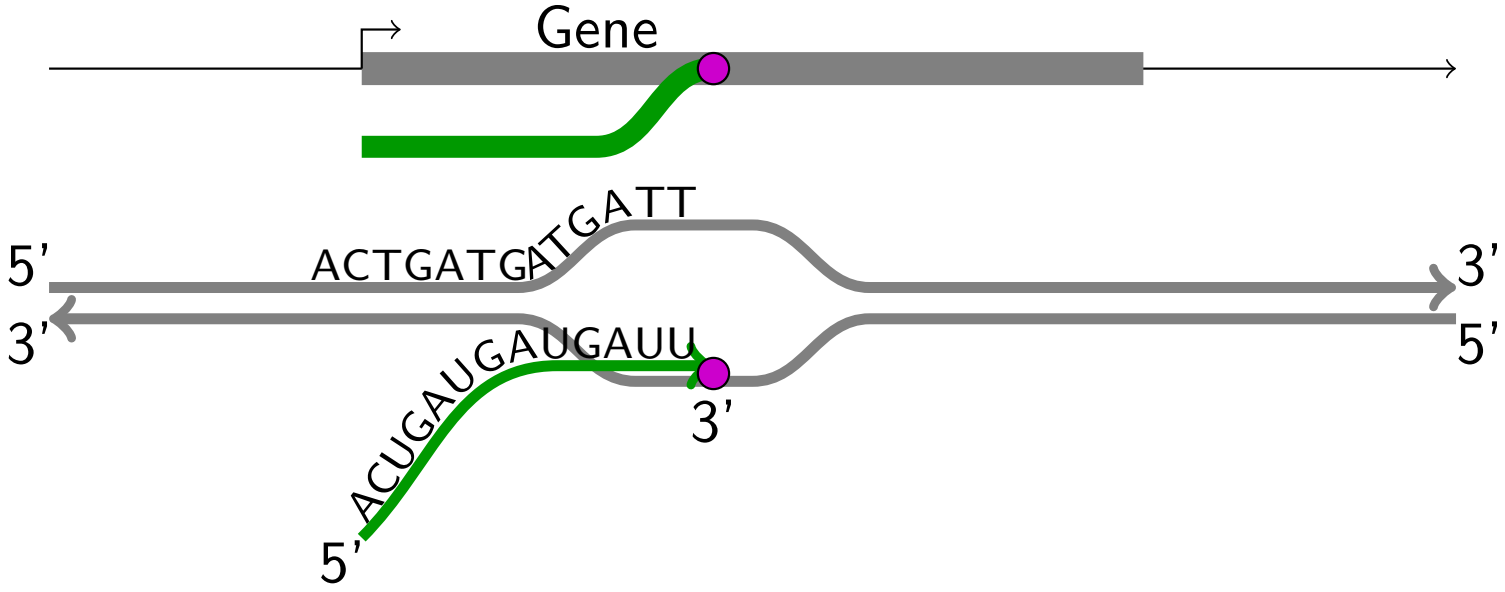
Strandedness!!



Strandedness!!



Strandedness!!



Encoding of Amino Acid sequence in DNA

Four nucleotides $\Rightarrow$ 20 Amino Acids

Word length?

- $1 \rightarrow 4^1 = 4$

  A, C, G, U

Encoding of Amino Acid sequence in DNA

Four nucleotides $\Rightarrow$ 20 Amino Acids

Word length?

- $1 \rightarrow 4^1 = 4$
  A, C, G, U
- $2 \rightarrow 4^2 = 16$
  AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, UU,

Encoding of Amino Acid sequence in DNA

Four nucleotides $\Rightarrow$ 20 Amino Acids
Word length?

- $1 \rightarrow 4^1 = 4$
  A, C, G, U
- $2 \rightarrow 4^2 = 16$
  AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, UU,
- $3 \rightarrow 4^3 = 64$
  AAA, AAC, AAG, AAU, ACA, ACC, ACG, ACU, AGA, AGC, AGG, AGU, ...

Encoding of Amino Acid sequence in DNA

Four nucleotides $\Rightarrow$ 20 Amino Acids
Word length?

- $1 \rightarrow 4^1 = 4$
  A, C, G, U
- $2 \rightarrow 4^2 = 16$
  AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, UU,
- $3 \rightarrow 4^3 = 64$
  AAA, AAC, AAG, AAU, ACA, ACC, ACG, ACU, AGA, AGC, AGG, AGU, ...

$\rightarrow$ triplet code used.

Encoding of Amino Acid sequence in DNA (2)

Summary

- Amino acid encoded by codons that contain 3 nucleotides each.
- 61 codons $\rightarrow$ 20 amino acids
- 3 codons $\rightarrow$ STOP
- 1 codon $\rightarrow$ START (AUG encodes Met)

Slightly simplified, but generally good enough.

The *standard* code

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F)	UCU	(Ser/S)	UAU	(Tyr/Y)	UGU	(Cys/C)	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L)	UCA		UUA	Stop	UGA	(Stop)	A
	UUG		UCG		UAG	Stop	UGG	(Trp/W)	G
C	CUU		CCU	(Pro/P)	CAU	(His/H)	CGU	(Arg/R)	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln/Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I)	ACU	(Thr/T)	AAU	(Asn/N)	AGU	(Ser/S)	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K)	AGA	(Arg/R)	A
	AUG	(Met/M)	ACG		AAG		AGG		G
G	GUU	(Val/V)	GCU	(Ala/A)	GAU	(Asp/D)	GGU	(Gly/G)	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E)	GGA		A
	GUG		GCG		GAG		GGG		G

Non-standard codes

1 The Standard Code
2 The Vertebrate Mitochondrial Code
3 The Yeast Mitochondrial Code
4 The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code
5 The Invertebrate Mitochondrial Code
6 The Ciliate, Dasycladacean and Hexamita Nuclear Code
9 The Echinoderm and Flatworm Mitochondrial Code
10 The Euplotid Nuclear Code
11 The Bacterial, Archaeal and Plant Plastid Code
12 The Alternative Yeast Nuclear Code
13 The Ascidian Mitochondrial Code
14 The Alternative Flatworm Mitochondrial Code
16 Chlorophycean Mitochondrial Code
21 Trematode Mitochondrial Code
22 Scenedesmus obliquus Mitochondrial Code
23 Thraustochytrium Mitochondrial Code
24 Rhabdopleuridae Mitochondrial Code
25 Candidate Division SR1 and Gracilibacteria Code
26 Pachysolen tannophilus Nuclear Code
27 Karyorelict Nuclear Code
28 Condylostoma Nuclear Code
29 Mesodinium Nuclear Code
30 Peritrich Nuclear Code
31 Blastocrithidia Nuclear Code
33 Cephalodiscidae Mitochondrial UAA-Tyr Code

Example codes

- The standard code

```
  AAs  = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M------**--*----M---------------M----------------------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

- The vertebrate mitochondrial code

```
  AAs  = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG
Starts = ----------**--------------------MMMM----------**---M------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

- The yeast mitochondrial code

```
  AAs  = FFLLSSSSYY**CCWWTTTTPPPPHHQQRRRRIIMMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ----------**----------------------MM--------------M------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

[1] https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Reading frames

```
5’ ATCAGATAGATATTACCGATAGACAG 3’
```

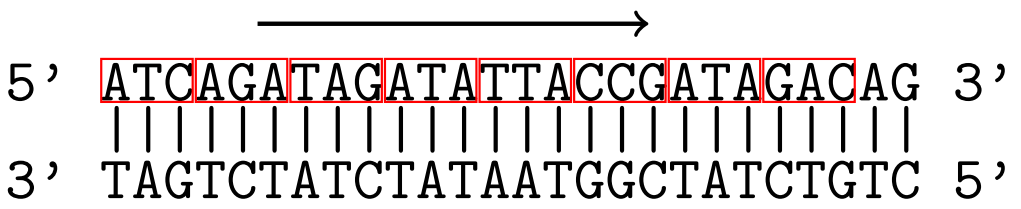
This can encode amino acids in 6 different ways:

Reading frames

```
5' ATCAGATAGATATTACCGATAGACAG 3'
   ||||||||||||||||||||||||||
3' TAGTCTATCTATAATGGCTATCTGTC 5'
```

This can encode amino acids in 6 different ways:

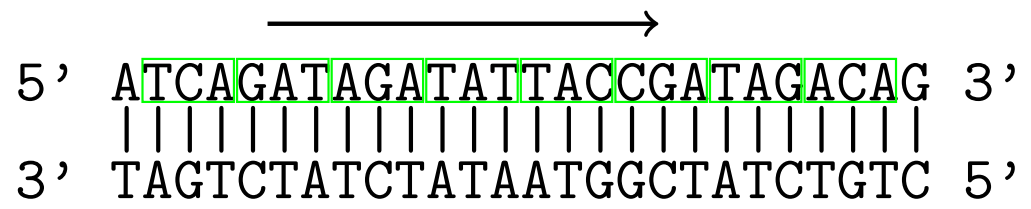
Reading frames



This can encode amino acids in 6 different ways:

```
frame 1 .. ATC AGA TAG ATA TTA CCG ATA GAC AG
           Ile Arg     Ile Leu Pro Ile Asp
```

Reading frames

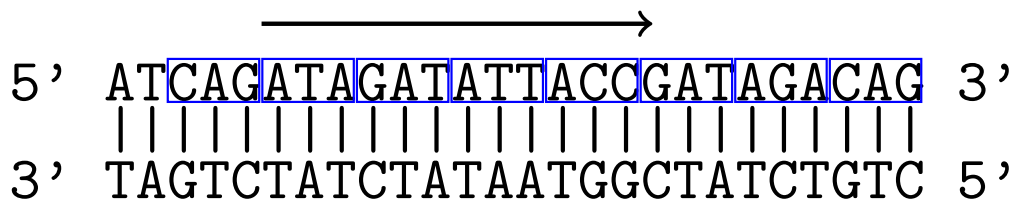


This can encode amino acids in 6 different ways:

frame 1 `.. ATC AGA TAG ATA TTA CCG ATA GAC AG`
`Ile Arg Ile Leu Pro Ile Asp`

frame 2 `.A TCA GAT AGA TAT TAC CGA TAG ACA G`
`Ser Asp Arg Tyr Tyr Arg Thr`

Reading frames



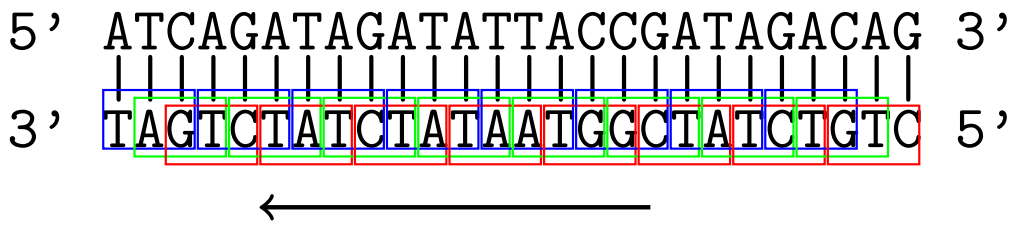
This can encode amino acids in 6 different ways:

frame 1 `.. ATC AGA TAG ATA TTA CCG ATA GAC AG`
`Ile Arg Ile Leu Pro Ile Asp`

frame 2 `.A TCA GAT AGA TAT TAC CGA TAG ACA G`
`Ser Asp Arg Tyr Tyr Arg Thr`

frame 3 `AT CAG ATA GAT ATT ACC GAT AGA CAG`
`Gln Ile Asp Ile Thr Asp Arg His`

Reading frames



This can encode amino acids in 6 different ways:

frame 1
```
.. ATC AGA TAG ATA TTA CCG ATA GAC AG
   Ile Arg     Ile Leu Pro Ile Asp
```

frame 2
```
.A TCA GAT AGA TAT TAC CGA TAG ACA G
   Ser Asp Arg Tyr Tyr Arg     Thr
```

frame 3
```
AT CAG ATA GAT ATT ACC GAT AGA CAG
   Gln Ile Asp Ile Thr Asp Arg His
```

frame -1
```
.. CTG TCT ATC GGT AAT ATC TAT CTG AT
```

frame -2
```
.C TGT CTA TCG GTA ATA TCT ATC TGA T
```

frame -3
```
CT GTC TAT CGG TAA TAT CTA TCT GAT ..
```

Reading frames (2)

DNA is[4] double stranded and has 6 reading frames.

RNA is single stranded[5]; it has only 3 reading frames!

[4] Well in general anyway. It can be single stranded as well, but then you'll usually be informed.

[5] Well, it can be double-stranded as well, but...

Open Reading Frame (ORF)

- Refers to a stretch of codons (nucleotide triplets) in the same frame that do not contain any stop codons (UUA, UAG, UGA).
- Sometimes required to begin with a start codon (AUG), but this depends on circumstance.
- Long ORFs indicate the presence of protein coding genes.

Frames and mutations

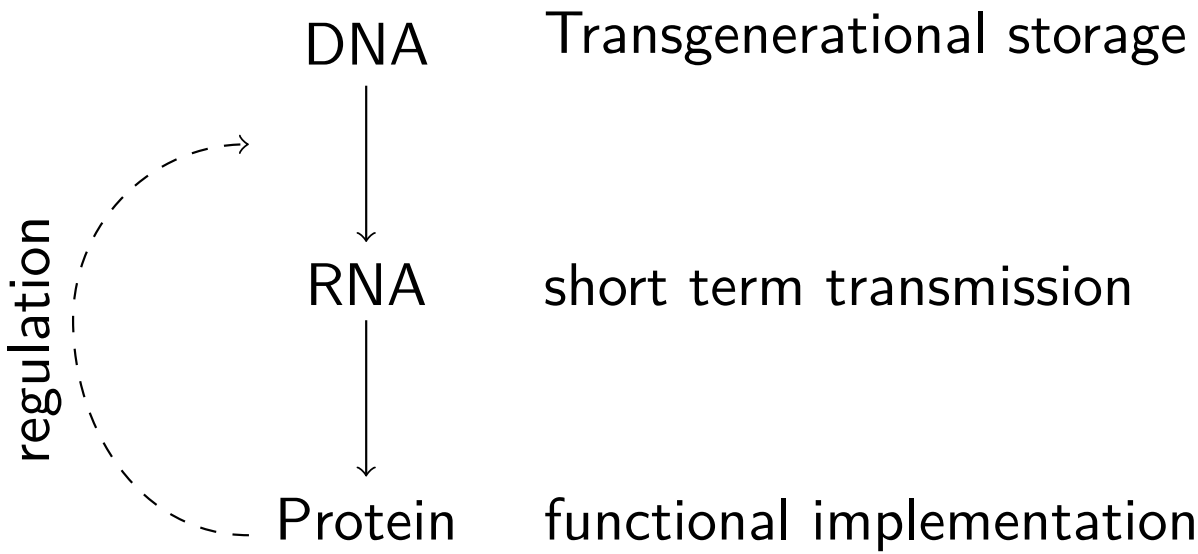
4 types of mutations

- substitution
- insertion
- deletion
- recombination events

ORF effects

- amino acid change (?)
- frame shift
- frame shift
- complex change

Summary



Regulated by DNA-RNA-protein interactions

Set of proteins, RNA and active promoters determine cell phenotype