DNA, RNA, Proteins and Central Dogma

Manu Madhavan

Lecture 2

Outline

- More on DNA, RNA and Proteins
- Base pairing
- Central Dogma of Molecular Biology

The Genetic Material

- **DNA** (Deoxyribo Nucleic Acid) is the genetic material
- "What is the basis of inheritance?"
- "What allows living things to be different from nonliving things?"

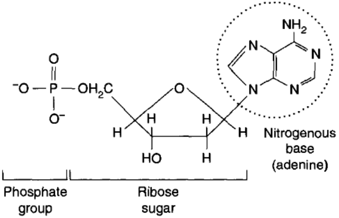
Genetics

It is the information stored in **DNA** that allows the organization of inanimate molecules into functioning, living cells and organisms that are able to regulate their internal chemical composition, growth, and reproduction.
As a direct result, it is also what allows us to inherit our mother's curly hair, our father's blue eyes, and even our uncle's too large nose.

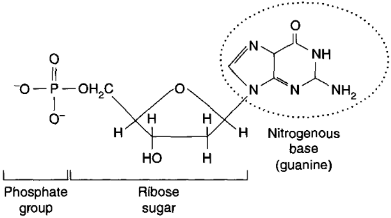
The Genetic Material

- Genes are **segments of DNA** that contain the code for a specific protein that functions in one or more types of cells in the body
- Genes themselves contain their information as a specific sequence of **nucleotides** that are found in DNA molecules
- Only four different bases are used in DNA molecules: guanine, adenine, thymine, and cytosine (G, A, T, and C).
- Each base is attached to a phosphate group and a deoxyribose sugar to form a nucleotide.
- The only thing that makes one nucleotide different from another is which nitrogenous base it contains.

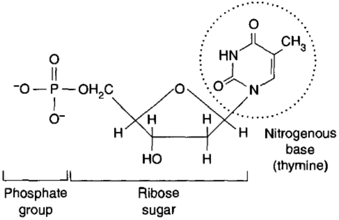
Chemical Structure of DNA

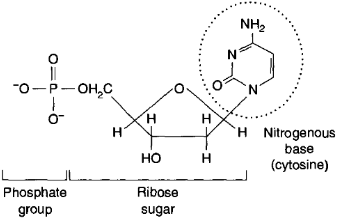


Adenosine-5'-monophosphate (AMP)



Guanosine-5'-monophosphate (GMP)





Chemical Structure of DNA

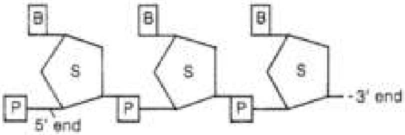
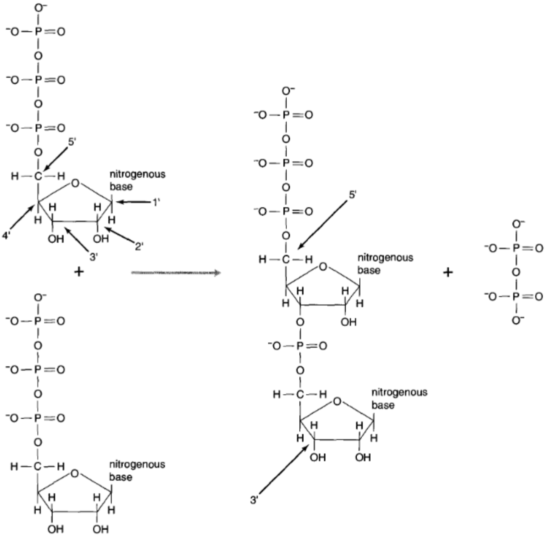


Fig. 257 **Polynucleotide chain.** General structure. P = phosphate, S = ribose or deoxyribose sugar, B = one of four bases.

- Strings of nucleotides can be attached to each other to make long **polynucleotide** chains
- The attachment between any two nucleotides is always made by way of a **phosphodiester** bond
- Connects the phosphate group of one nucleotide to the deoxyribose sugar of another

- Ester bonds are those that involve links made by oxygen atomsphosphodiester bonds have a total of two ester bonds, one on each side of a phosphorous atom.

Chemical Structure of DNA



Chemical Structure of DNA

- The phosphate group(s) of any single, unattached nucleotide are always found on its 5' carbon.
- Those phosphate groups are used to bridge the gap between the 5' carbon of an incoming deoxyribose sugar and the 3' carbon of a deoxyribose sugar at the end of a preexisting polynucleotide chain.
- As a result, one end of a string of nucleotides always has a 5' carbon that is not attached to another nucleotide, and the other end of the molecule always has an unattached 3' carbon.

Base Pairing-Complementary

- The information content in a DNA molecule comes from the specific sequence of its nucleotides
- The information in each strand of DNA molecule is redundant
- Not exact, **Complementary**
- For every G on one strand, a C is found on its complementary strand and vice versa (G == C)
- For every A on one strand, a T is found on its complementary strand and vice versa. (A == T)
- The chemical interaction (specifically, three hydrogen bonds that form between G's and C's and two hydrogen bonds that form between A's and T's) between the two different kinds of base pairs is actually so stable and energetically favorable that it alone is responsible for holding the two complementary strands together.

Base Pairing-Antiparallel

- The 5’ end of one strand corresponding to the 3’ end of its complementary strand and vice versa
- Consequently, if one strand’s nucleotide sequence is 5’-GTATCC-3’, the other strand’s sequence will be 3’-CATAGG-5’
- Strictly speaking, the two strands of a double-stranded DNA molecule are reverse complements of each other
- Sequence features that are 5’ to a particular reference point are commonly described as being **upstream** while those that are 3’ are described as being **downstream**

Try yourself

- Upstream: 5'-ACCCGTGGTAAGG-3'
  Downstream: ?
- Downstream: 3'-CAAGTTTGA-5'
  Upstream: ?

```
In [1]: from Bio.Seq import Seq

In [3]: my_seq = Seq("AGTACACTGGT")
        my_seq

Out[3]: Seq('AGTACACTGGT')

In [4]: my_seq.complement()

Out[4]: Seq('TCATGTGACCA')

In [5]: my_seq.reverse_complement()

Out[5]: Seq('ACCAGTGTACT')
```

Central Dogma of Molecular Biology

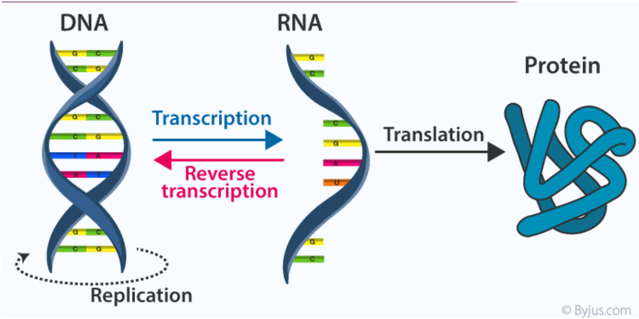
- An explanation of the flow of genetic information within a biological system
- DNA $\rightarrow$ RNA $\rightarrow$ Protein

Central Dogma: Francis Crick, 1970

According to central dogma, the genomic region of DNA is copied into RNA (known as messenger RNA (mRNA) by a process known as **transcription**. The mRNA travels to the protein production site and converted into protein by a process known as **translation**. Protein is never translated back to RNA or DNA. Further, DNA is never produced from RNA, except for retroviruses.

Central Dogma of Molecular Biology

- An explanation of the flow of genetic information within a biological system
- DNA $\rightarrow$ RNA $\rightarrow$ Protein



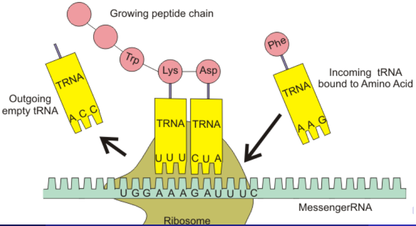
Transcription

- DNA, which is a double stranded molecule, unzips down the center through the action of an enzyme called RNA polymerase.
- An RNA template is laid down on top of one of the DNA strands, making a near mirror image of the DNA code, as shown in the table below.
- This template is laid down in the 5' to 3' direction.
- The RNA strand is now called messenger RNA, or mRNA, and leaves the nucleus of the cell.
- Enzymes facilitating the process include **RNA polymerase and transcription factors**

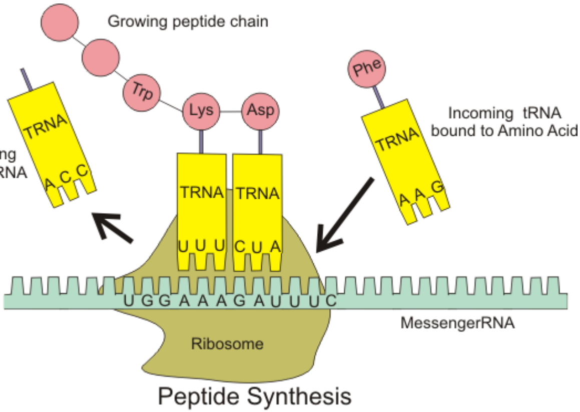
Whenever the DNA had this nucleotide base...	the RNA will have this nucleotide base
Cytosine (C)	Guanine (G)
Guanine (G)	Cytosine (C)
Adenine (A)	*Uracil* (U)
Thymine (T)	Adenine (A)

Translation

- The mRNA arrives at a **ribosome**. Each three letter snippet of RNA is called a **codon**
- Each codon calls for the synthesis of a particular amino acid
- As the ribosome moves across the mRNA strand, reading the 3-letter codons, a special transport molecule called **transfer RNA (tRNA)** is dipatched to pick up the amino acid specified by the codon.
- The tRNA delivers the correct amino acid back to the ribosome, lands on the mRNA codon (the tRNA molecule has a corresponding **anti-codon**), and releases its amino acid cargo, where it is attached to a growing peptide chain.



Translation



Watch: `https://www.youtube.com/watch?v=gG7uCskUOrA`

Promotor Sequence

- **Gene expression**, the process of using the information stored in DNA to make an RNA molecule and then a corresponding protein
- Sequences express for particular proteins
- Must be able to determine which genes code for proteins that are needed at any particular time.
- Particular combinations of nucleotides are not as likely to occur by chance, and the greater the number of nucleotides involved, the smaller a chance occurrence becomes
- Promoter sequences are DNA sequences that define where transcription of a gene by RNA polymerase begins.
- Promoter sequences are typically located directly upstream or at the 5' end of the transcription initiation site.
- RNA polymerase and the necessary transcription factors bind to the promoter sequence and initiate transcription.
- Promoter sequences define the direction of transcription and indicate which DNA strand will be transcribed; this strand is known as the sense strand.

Genetic code

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

Genetic code

- Different amino acids are coded for by more than one codon.
- This feature of the genetic code is called **degeneracy**.
- It is therefore possible for mistakes to occur during DNA replication or transcription that have no effect on the amino acid sequence of a protein. This is especially true of mutations (heritable changes in the genetic material) that occur in the third (last) position of a codon.

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

Next

- Few more terminologies: ORF, Exons, introns, ...
- Bioinformatics Tools and Databases