Introduction to sequence biology

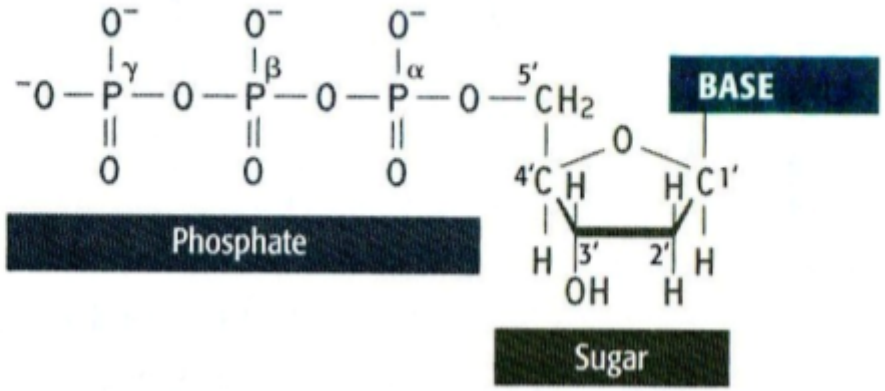
Deoxyribonucleic Acid (DNA): A non-branching polymer made up of four ‘nucleotides’. Each nucleotide consists of:

2' Deoxyribose (a sugar with 5 carbon atoms, called a *pentose*)

Nitrogeneous base (can be adenine, guanine, cytosine or thymine)

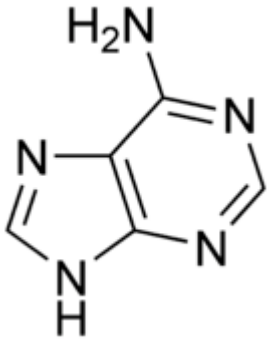
Phosphates

A nucleotide

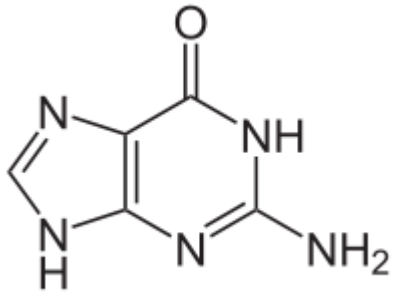


The four DNA nucleobases

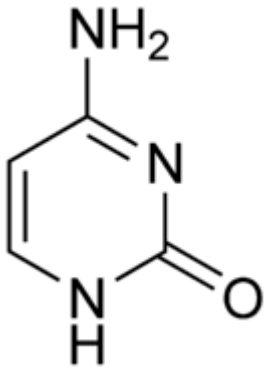
Adenine



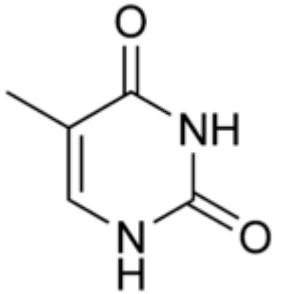
Guanine



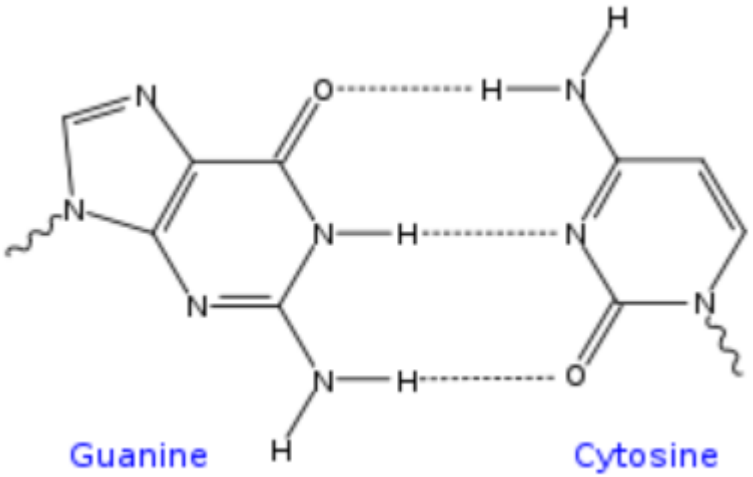
Cytosine



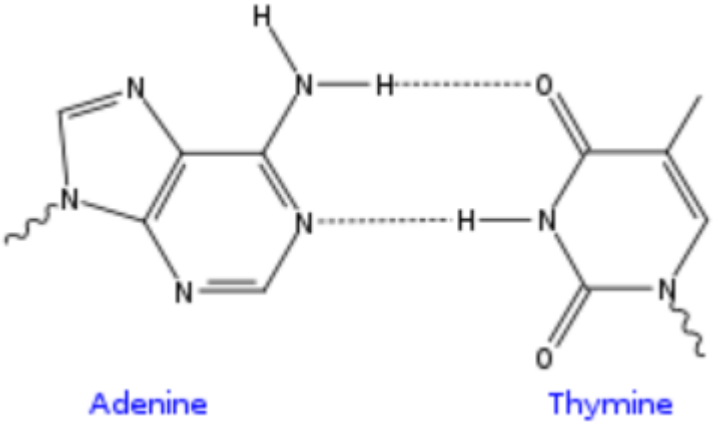
Thymine



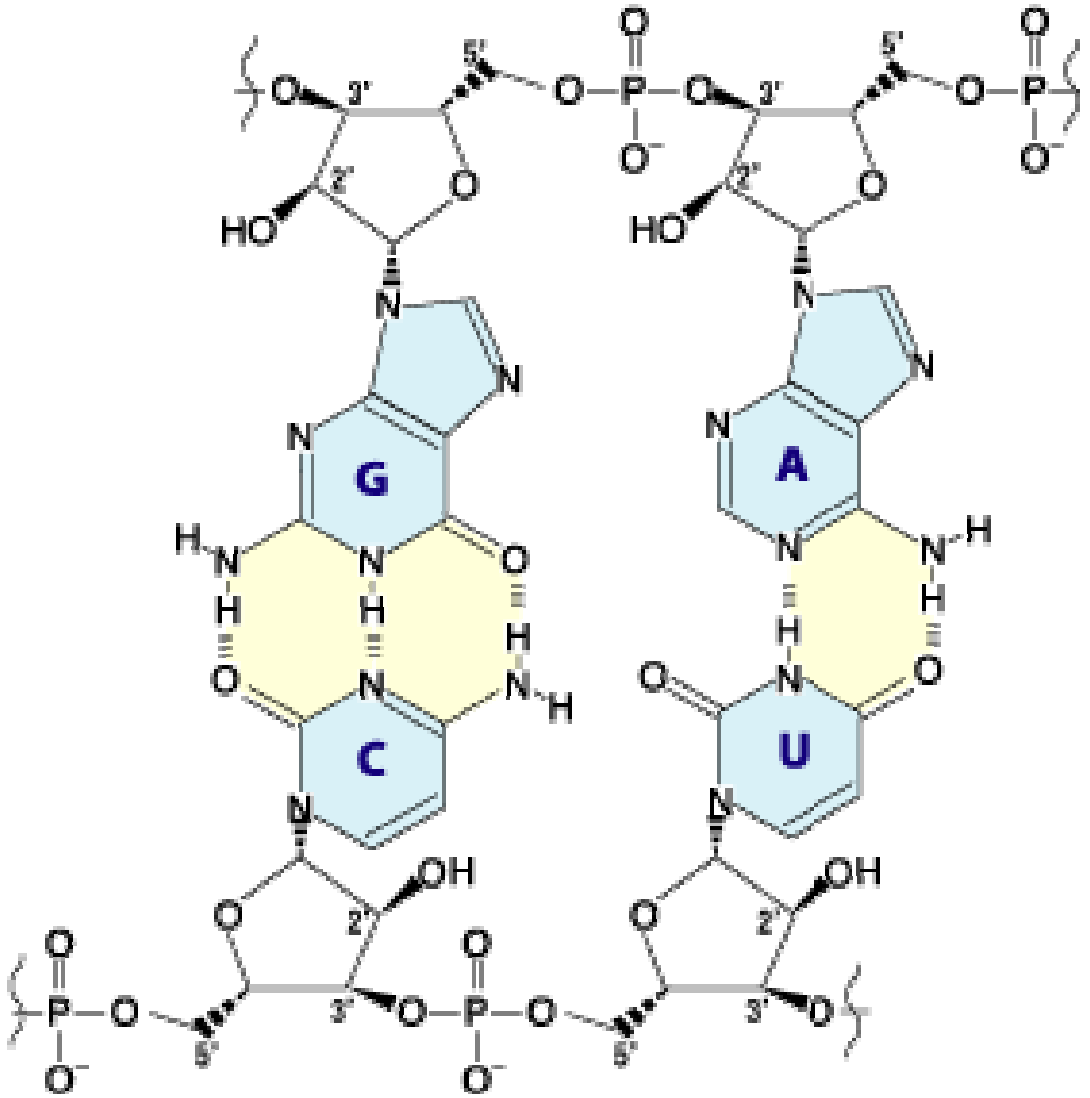
Base-pairing



3 hydrogen bonds

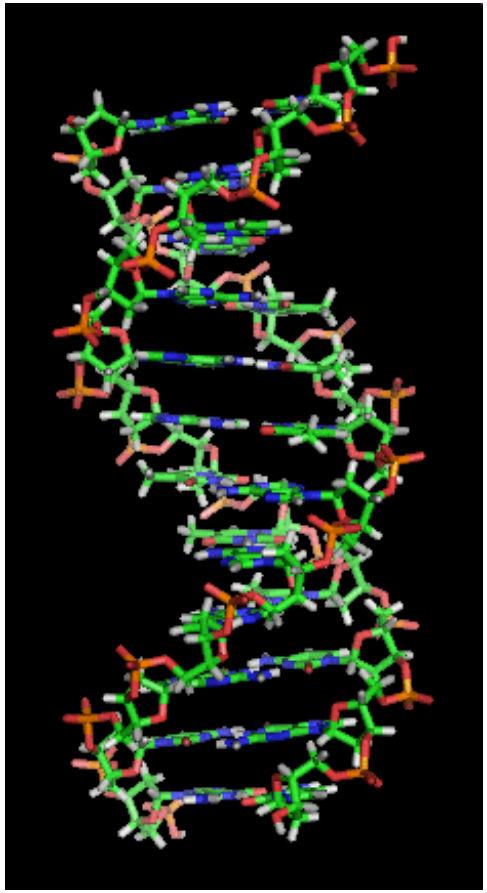


2 hydrogen bonds


G
A
C
U

Double Helix (Watson & Crick, 1953)

- Right-handed helix (although other confirmations exist)
- Bases paired up (A-T; G-C) in the interior of the helix
- One turn every 10 base-pairs



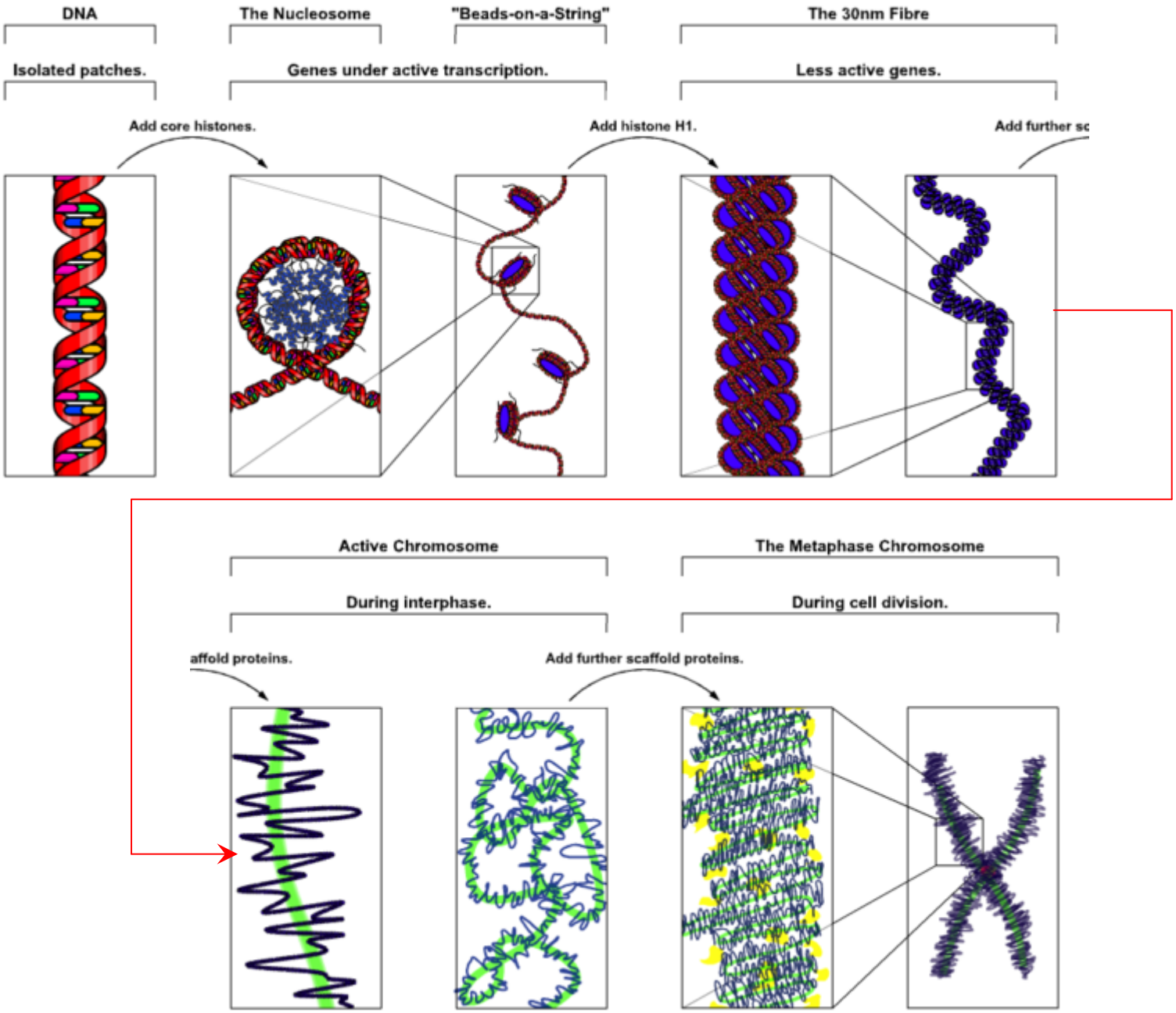
DNA orientation

A strand of DNA has an orientation - 5' end and 3' end

Nomenclature is based on the numbering of the carbon atoms in the sugars of the backbone

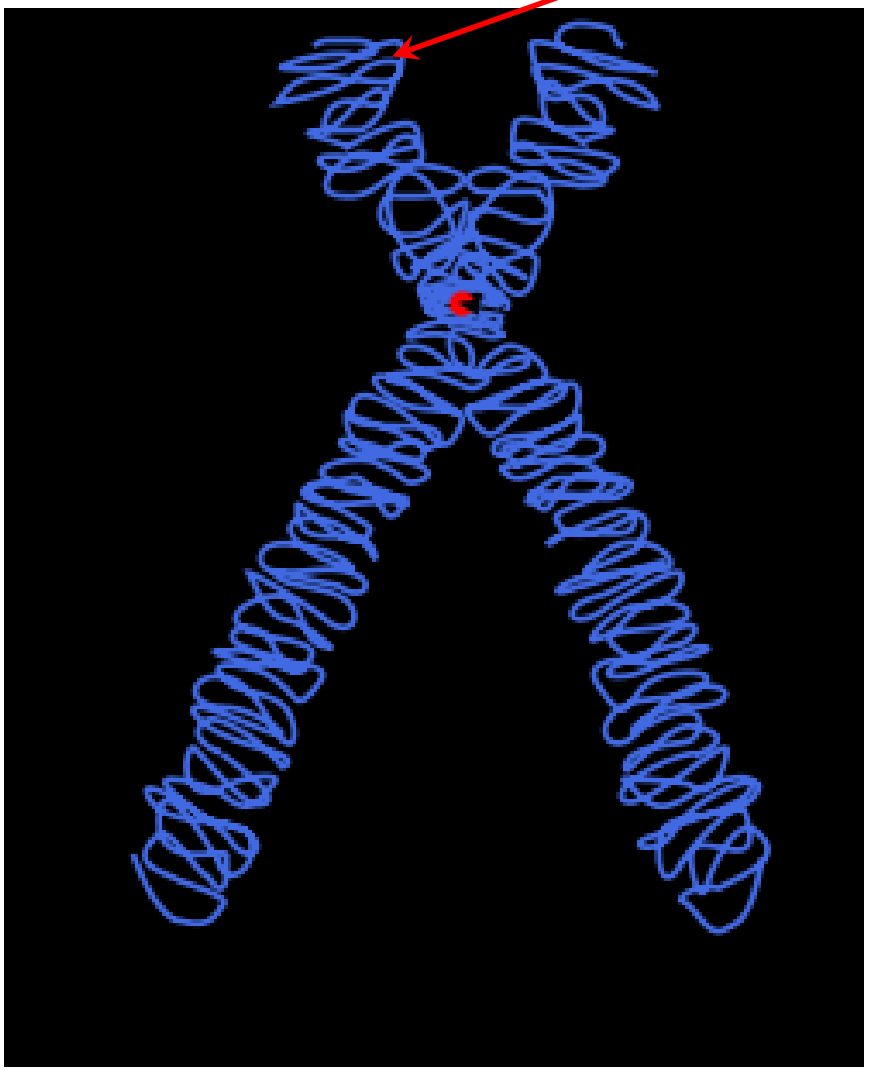
By convention DNA sequences are written in the 5' to 3' direction

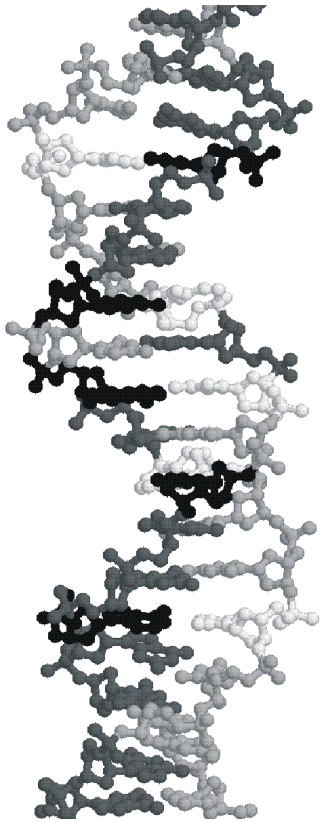
Strands of the double helix run in opposite directions – one strand is often referred to as ‘forward’ and the other ‘reverse’


DNA
The Nucleosome
"Beads-on-a-String"
The 30nm Fibre
Isolated patches.
Genes under active transcription.
Less active genes.
Add core histones.
Add histone H1.
Add further sc
Active Chromosome
The Metaphase Chromosome
During interphase.
During cell division.
affold proteins.
Add further scaffold proteins.

Chromosome (condensed form during cell division)

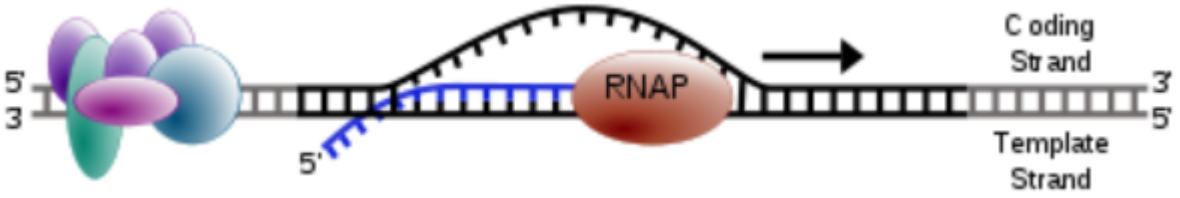
Sister chromatids (double stranded DNA in each)



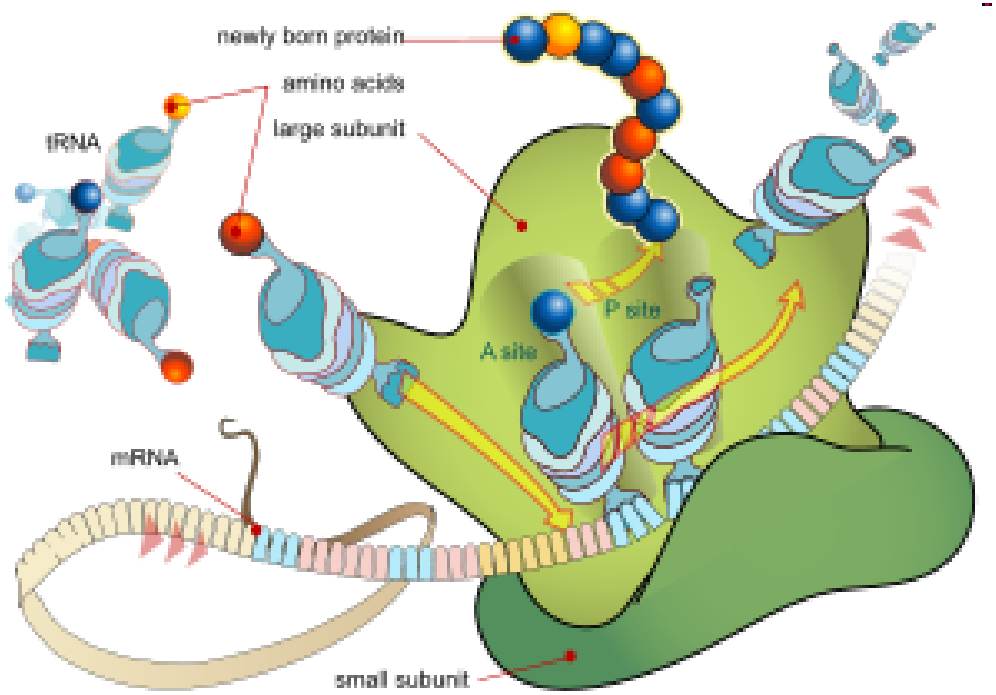


A C T G T A C T T...

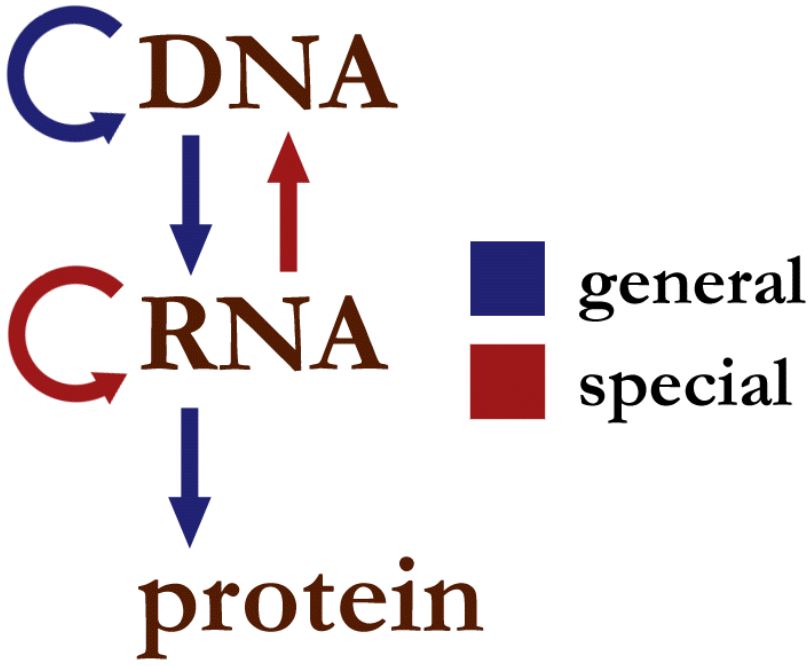
Transcription (DNA => RNA)



Translation (RNA => Protein)



Central dogma of molecular biology (Francis Crick, 1958)



The ‘Universal’ genetic code: 61 codons encode 20 amino acids

Second base in codon

First base in codon	U	C	A	G	Third base in codon
U	Phe	Ser	Tyr	Cys	**U**
	Phe	Ser	Tyr	Cys	**C**
	Leu	Ser	STOP	STOP	**A**
	Leu	Ser	STOP	Trp	**G**
C	Leu	Pro	His	Arg	**U**
	Leu	Pro	His	Arg	**C**
	Leu	Pro	Gln	Arg	**A**
	Leu	Pro	Gln	Arg	**G**
A	Ile	Thr	Asn	Ser	**U**
	Ile	Thr	Asn	Ser	**C**
	Ile	Thr	Lys	Arg	**A**
	Met	Thr	Lys	Arg	**G**
G	Val	Ala	Asp	Gly	**U**
	Val	Ala	Asp	Gly	**C**
	Val	Ala	Glu	Gly	**A**
	Val	Ala	Glu	Gly	**G**

UUC - UUU : synonymous
UUC - UUG: non-synonymous

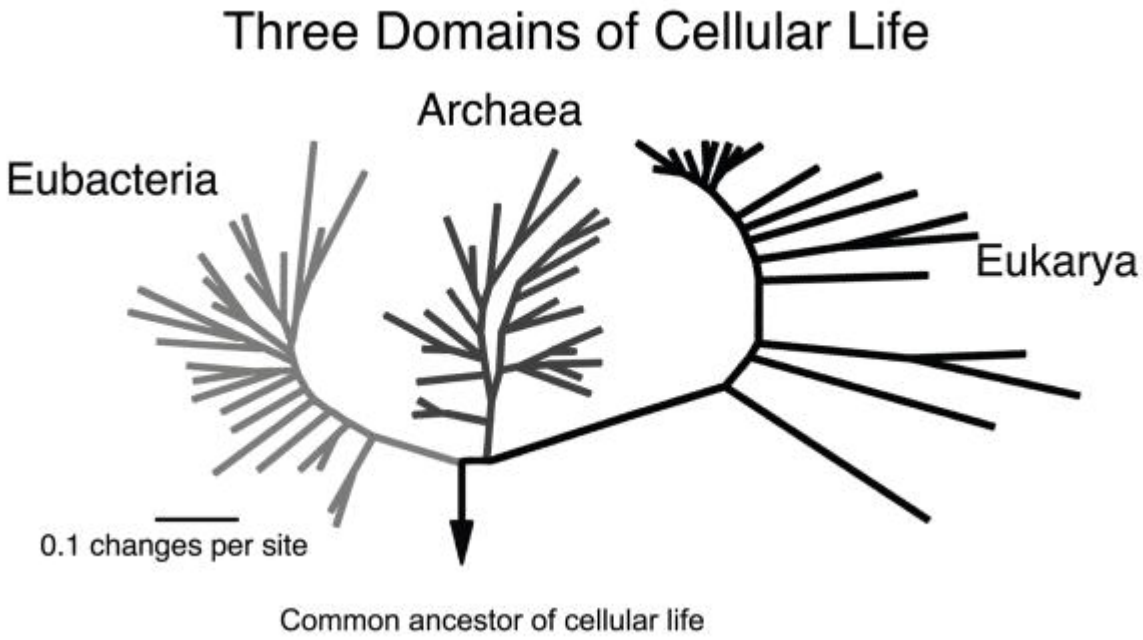
Cellular life can be divided into:

Prokaryotes – Archaea & Bacteria

- single celled (almost always)
- no nucleus (or membrane-bound organelles)
- usually simple gene structure

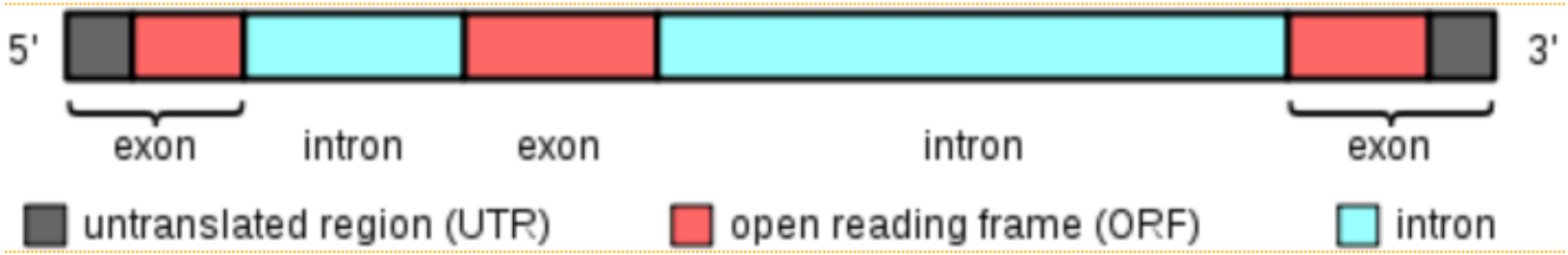
Eukaryotes (Greek: *karyon = nut or kernel*)

- have a nucleus
- frequently muticellular
- often more complex gene structure

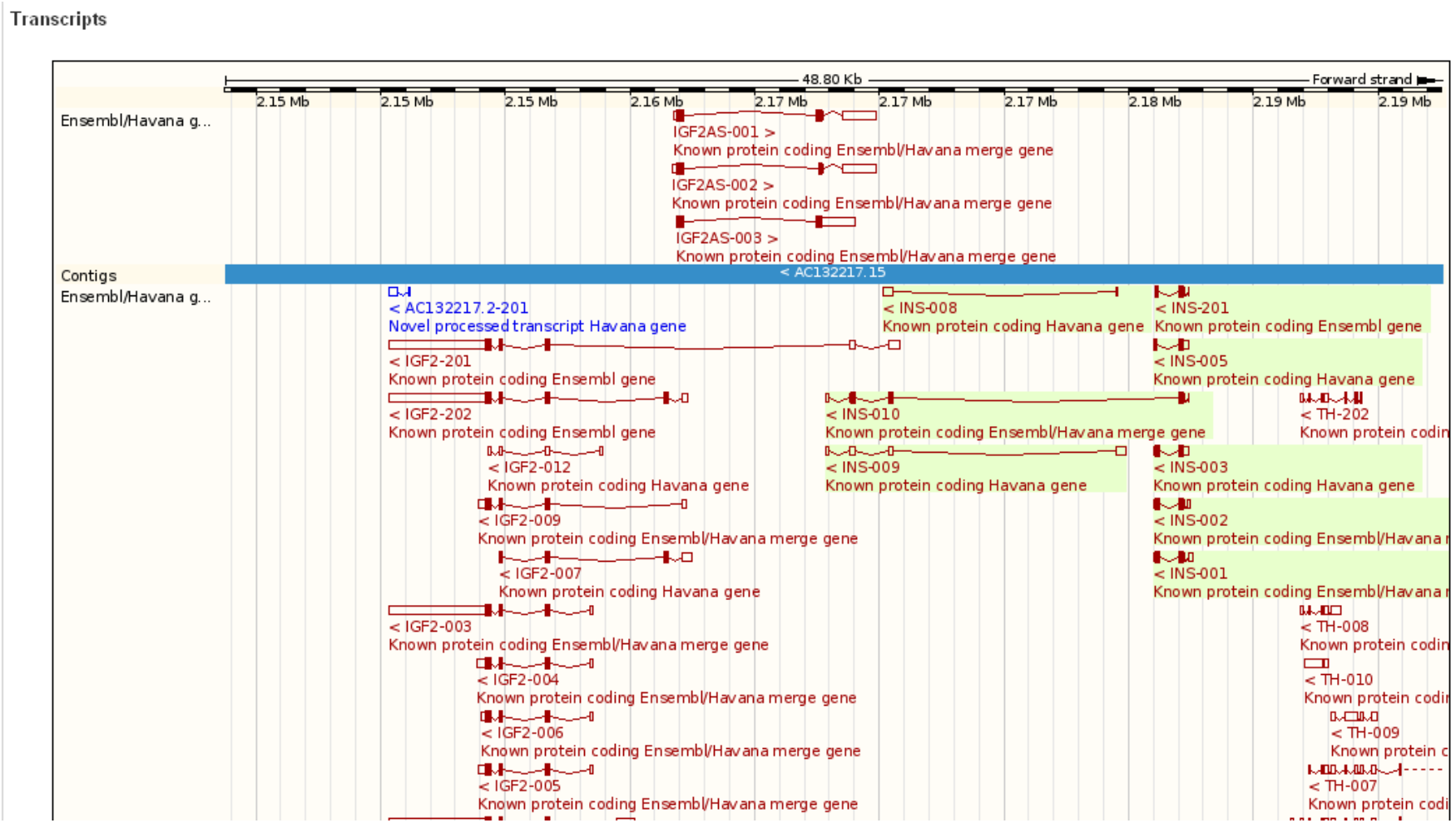


From Varki *et al.*

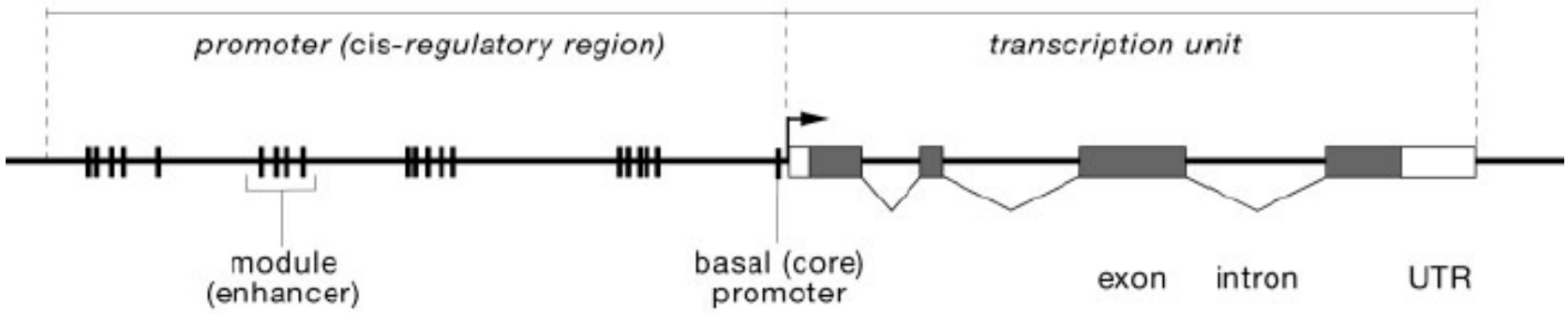
Structure of a eukaryotic gene



A sample region of the human genome

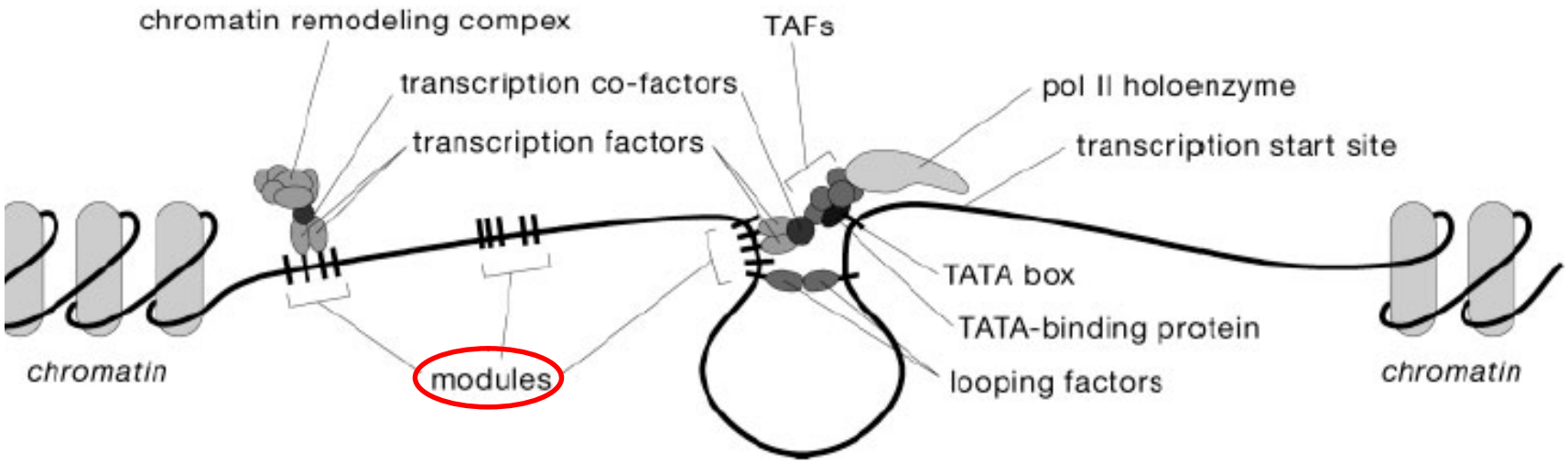


Organization of a gene and its promoter



Wray et al. MBE 2003

Organization of an idealized promoter



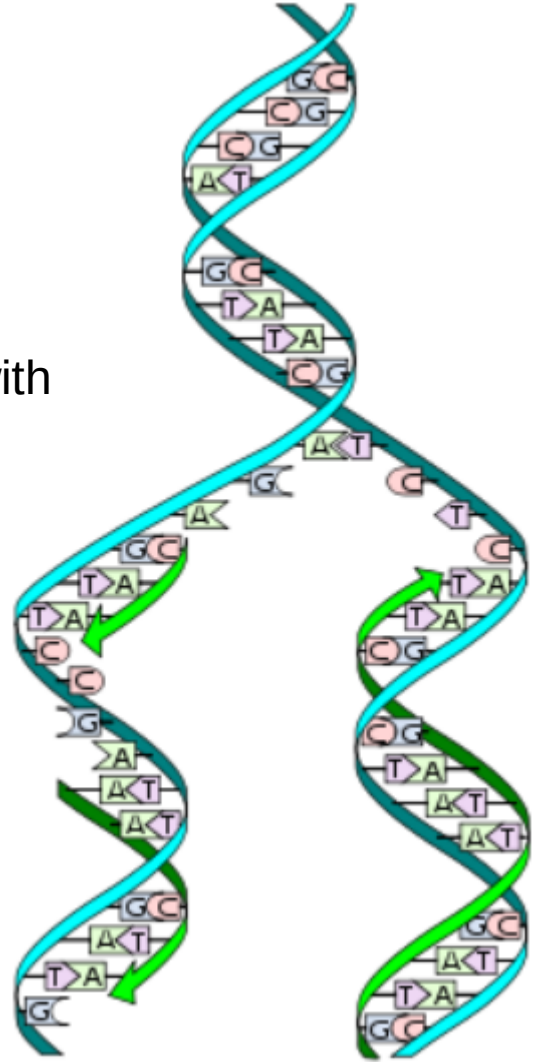
Wray et al. MBE 2003

DNA replication

DNA strands unwind

Formation of replication fork

Copy of DNA produced by base paring with template strand



Wikimedia

Mutation

Error in DNA replication

Such errors are significantly reduced in frequency through the proofreading activity of the polymerase enzymes

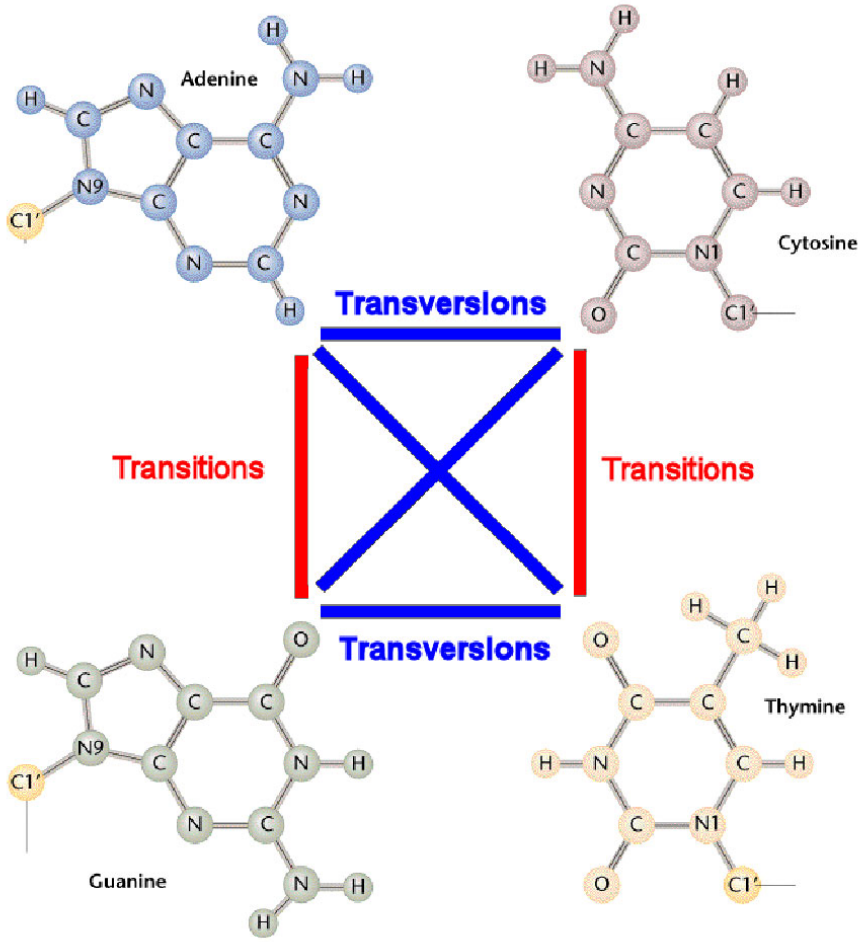
Errors in nucleotide replication lead to base-pair mismatching

Mismatch repair mechanisms further reduce the error rate in many organisms

Mutation of a single nucleotide most commonn type (‘point mutation’)

Recent results indicate that simultaneous mutations of multiple nucleotides occur

Mutations can also result in insertions or deletions of several nucleotides (often associated with ‘runs’ of repeated DNA bases)



purines pyrimidines

CpG Islands

C & G occur at frequency of 0.42 in human

CpG expected frequency is 0.044 (0.21 X 0.21) but observed frequency is ~1%

Most CpGs are methylated

Methylated CpGs are subject to rapid mutation (to TpG or CpA)

CpG islands occurs mostly in islands around ‘promoters’ of ‘ubiquitously expressed genes’