Introduction to molecular biology

Prof. David Jones

d.t.jones@ucl.ac.uk

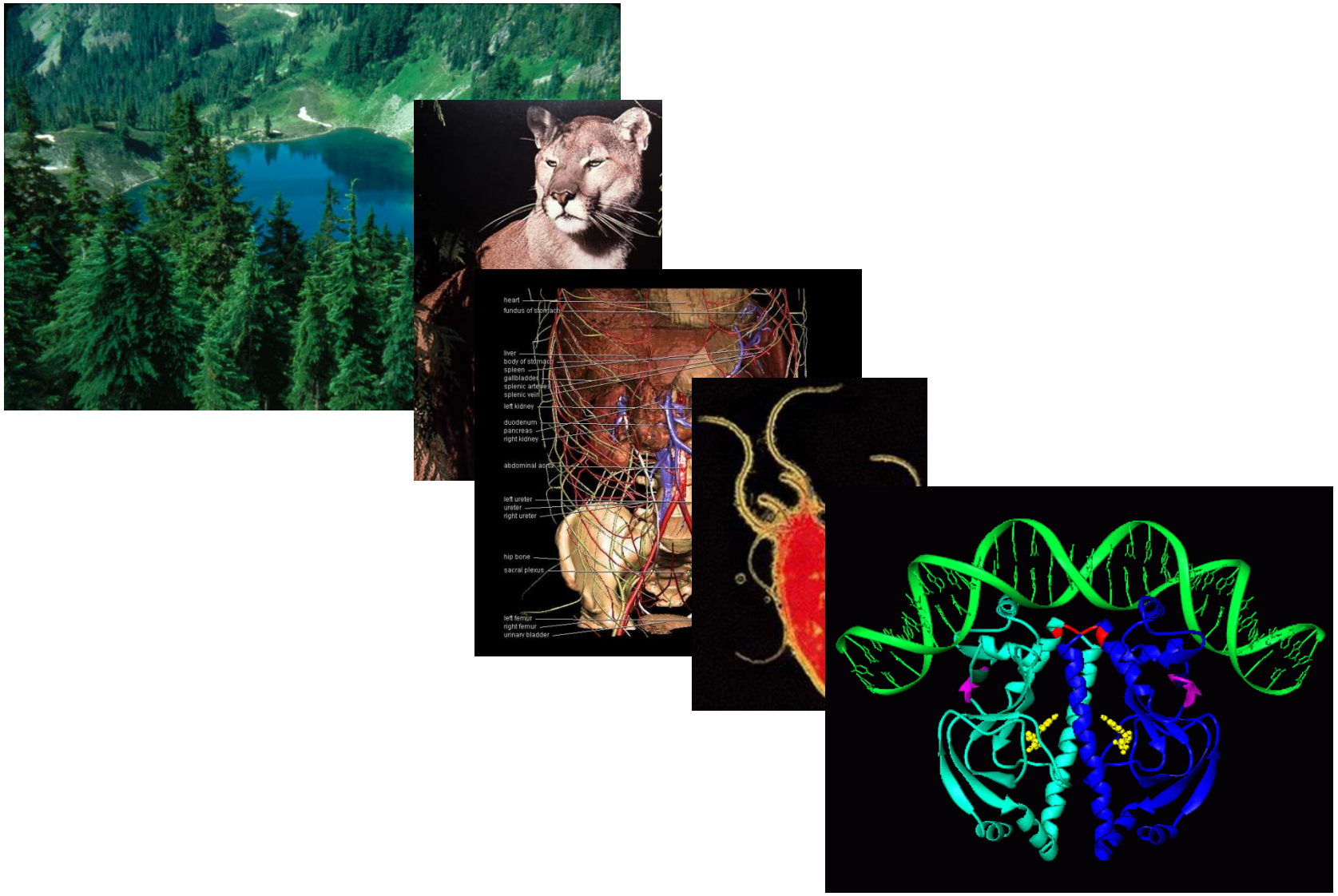
What is Biology?

- If it **moves** it's **biology**, if it **smells** it's **chemistry** and if it doesn't **work** it's **physics**
  - **Not a very useful definition!**
- **Wikipedia**
  - **Biology** (from Greek *Bìo* meaning life and *Λογος* meaning the study of) i.e. the study of life
  - It is concerned with the **characteristics**, **classification**, and **behaviors** of **organisms**, how species come into existence, and the interactions they have with each other and with the natural environment. Biology encompasses a broad spectrum of academic fields that are often viewed as independent disciplines. However, together they address phenomena related to living organisms (biological phenomena) over a **wide range of scales**, from biophysics to ecology.

What is Biology?

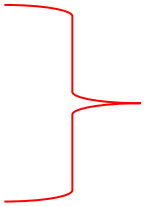
- **ChatGPT**
  - Biology is the study of living organisms and their interactions with each other and their environments. It is a broad field that encompasses many different subdisciplines, including genetics, evolution, ecology, and more. Biologists study a wide range of topics, from the smallest microorganisms to the largest mammals, and they use a variety of techniques and approaches to understand how living things work.

Biology at Different Scales



Biological Hierarchy

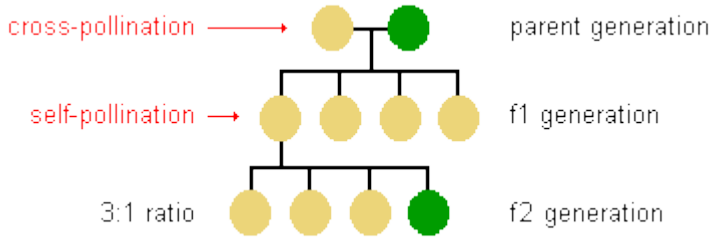
- molecules
- organelles
- cells
- tissues
- organs
- organisms
- populations
- communities
- ecosystems
- biosphere



This course

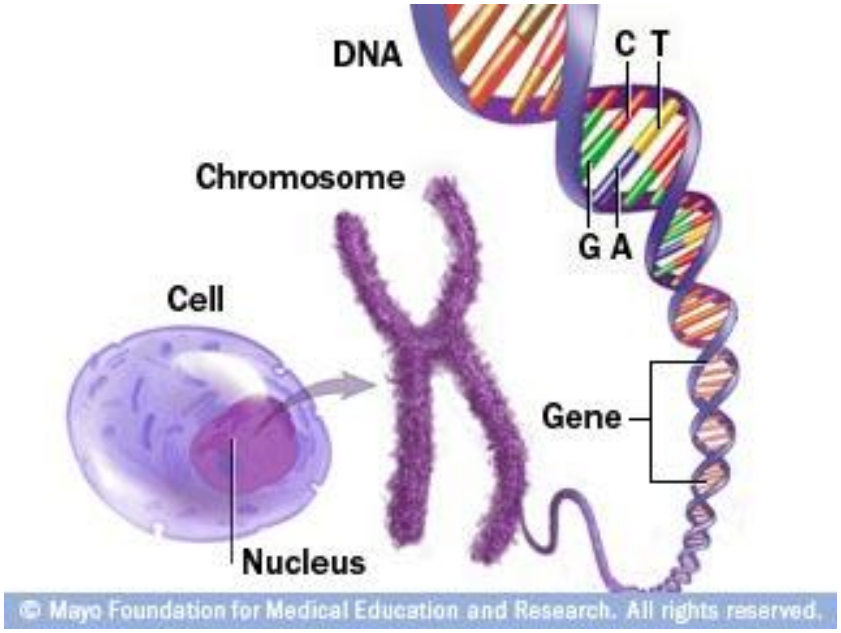
Genes, proteins and DNA

- A 19th-century monk called Gregor Mendel introduced the notion of **genes**: basic units responsible for possession and passing on of a single characteristic
- Initially it was thought that proteins carried genetic information
- Until mid 20th-century, when it was found that **DNA** did
- **Proteins** are the functional molecules in cells (i.e. they perform the majority of the reactions of life)



What is DNA?

- Deoxyribonucleic acid (DNA) is the molecule in the cell nucleus which holds the chemical information required to build proteins (Mitochondria in cells also contain some DNA (mtDNA))
- DNA is stored in the nucleus wrapped up in discrete units called chromosomes – humans have 23 pairs



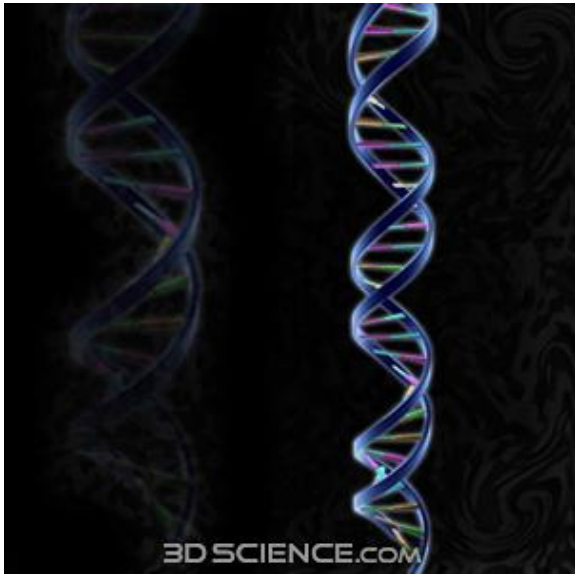
http://www.mayoclinic.org/tests-procedures/genetic-testing/multimedia/genetic-disorders/sls-20076216

- DNA is built up of a sugar-phosphate backbone and a sequence of **nucleotides**: **adenine (A), guanine (G), cytosine (C)** and **thymine (T)**

What is the structure of DNA?

- Watson and Crick discovered the structure of DNA: a **double helix** – two sugar phosphate chains wrapping round each other, with the nucleotides sticking out – the nucleotide from strand 1 meets the nucleotide from strand 2 in the middle.

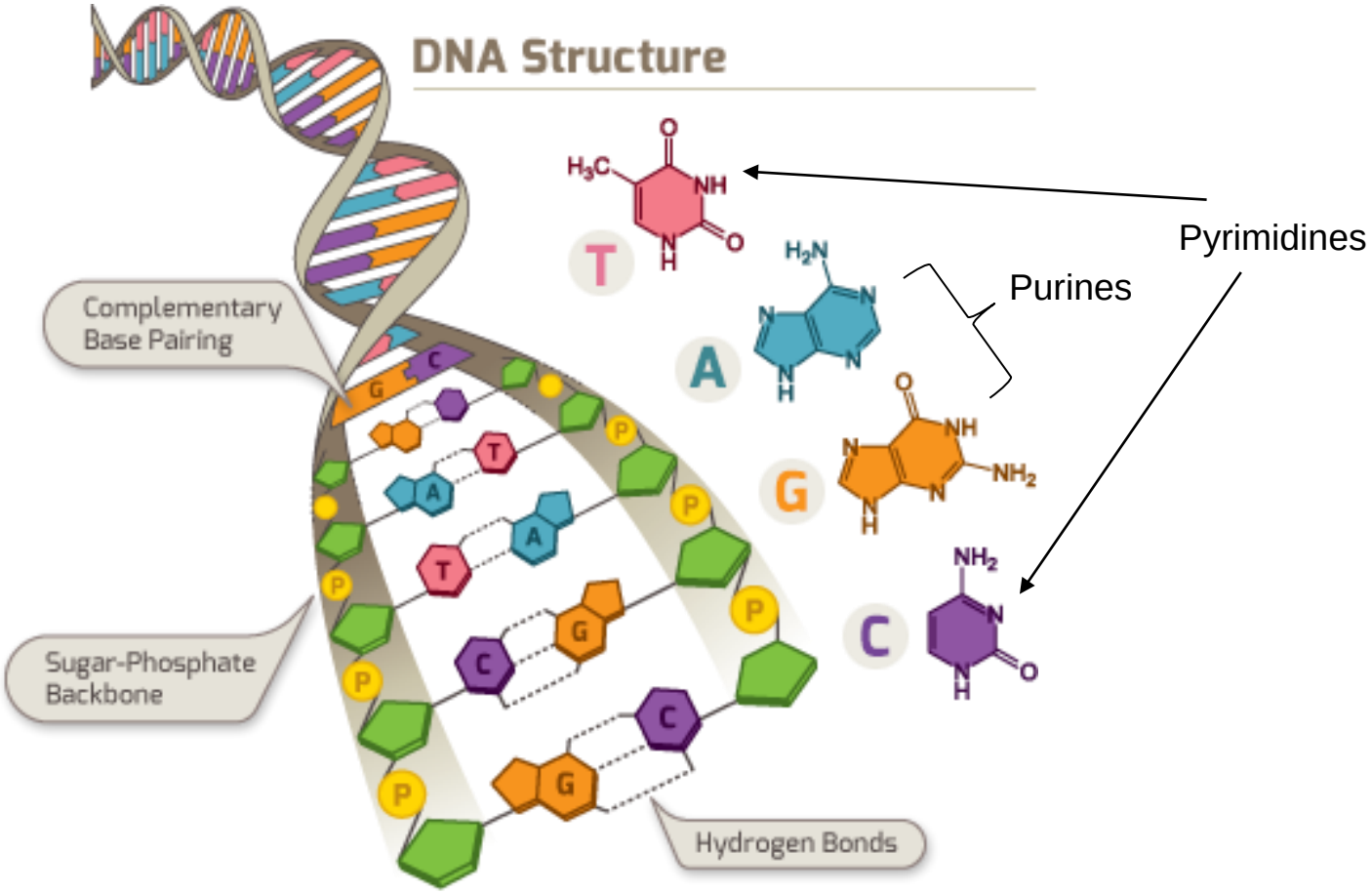
http://www.thehistoryblog.com/archives/25193



http://www.3dscience.com/3D_Models/Biology/DNA/DNA.php

- These pairs of nucleotides are complementary – where one strand has a C, the other has a G and vice versa; where one strand has an A the other has a T and vice versa.
- Human DNA consists of approximately 3 x 10^9 such **"base pairs".**

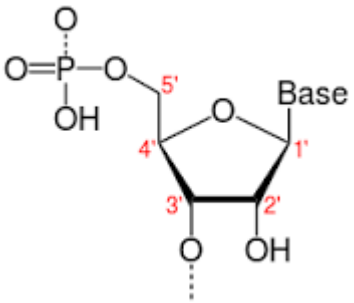
The DNA double helix



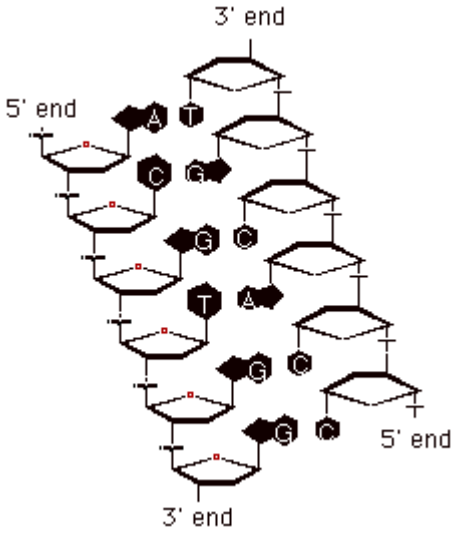
http://www.thednall.com/

DNA replication

- The DNA molecule is directional, because the sugars are asymmetrical – each sugar is connected to the strand **“upstream”** at its 5th carbon and **“downstream”** at its 3rd carbon. So you read the DNA sequence from the **“5 prime”** end to the **“3’ ”** end.



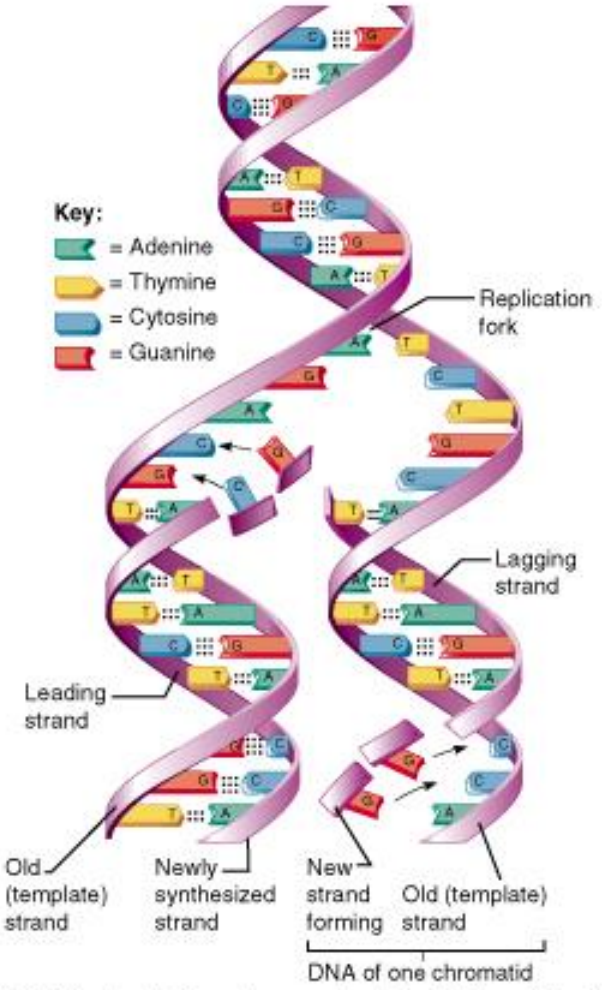
http://*en.wikipedia.org/wiki/Directionality_(molecular_biology)*



http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/mbglossary/Image2.gif

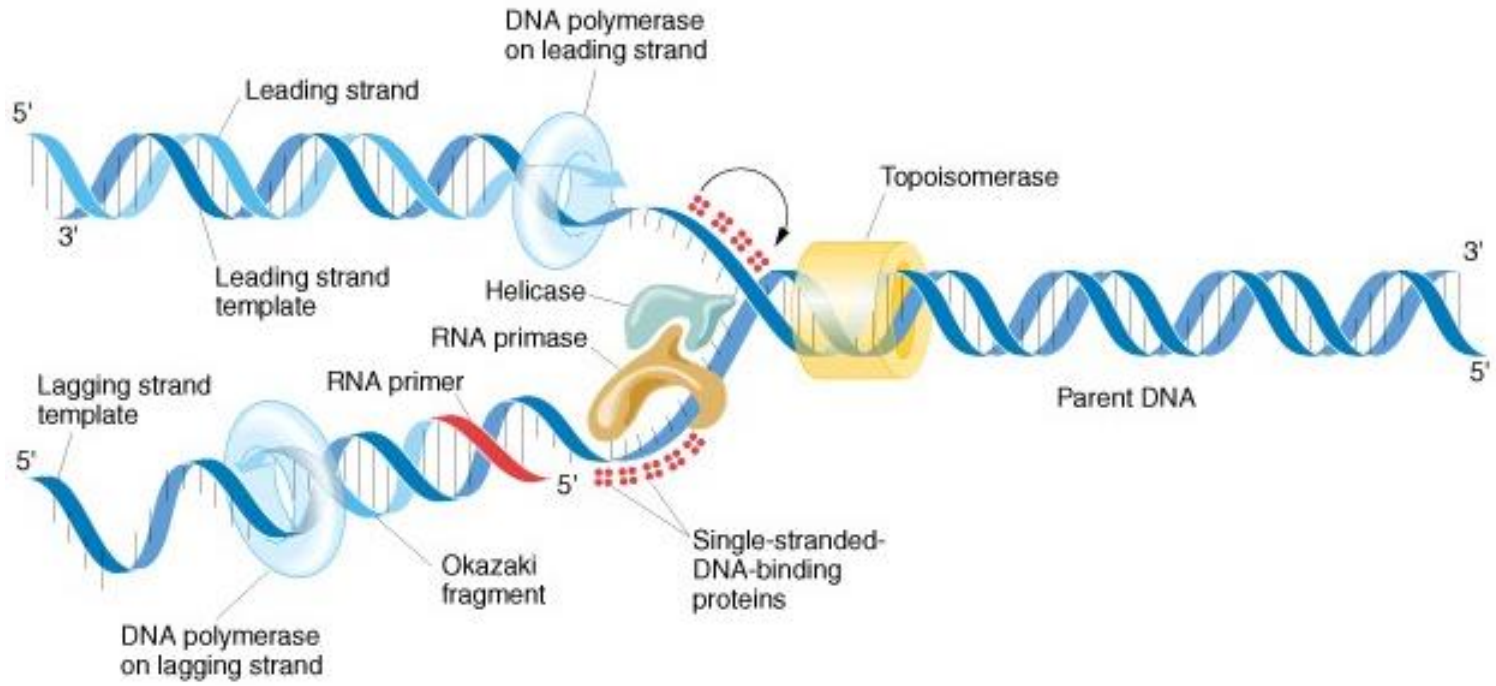
DNA replication

- In replication, the double helix becomes unzipped and free nucleotides bind to the their complementary pair nucleotides on the single strands. Thus each strand acts as a **template** for a new strand of DNA:



DNA replication in a bit more detail (FOR FURTHER READING)

- The reaction is catalysed by **DNA polymerase**- this causes the chain to elongate, but it can't start the formation of a new chain. For this a **primer** (short piece of DNA/ RNA) is required.

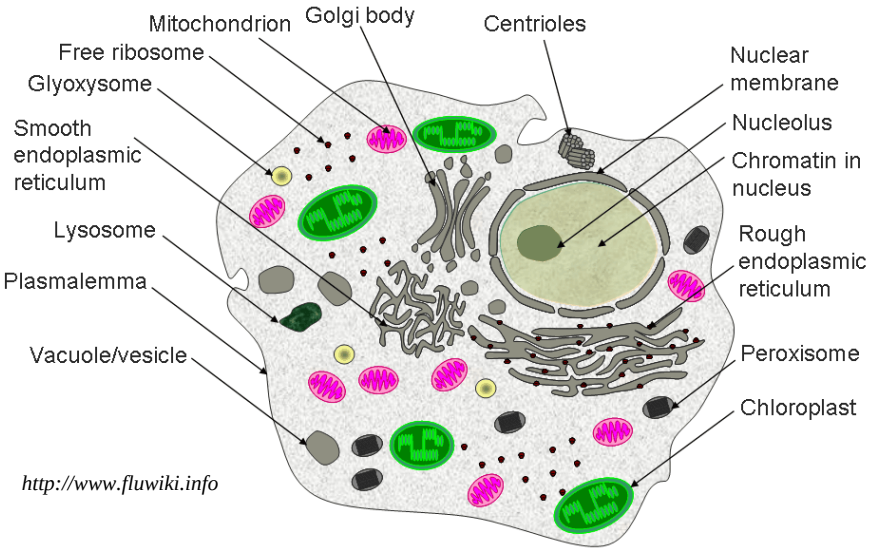


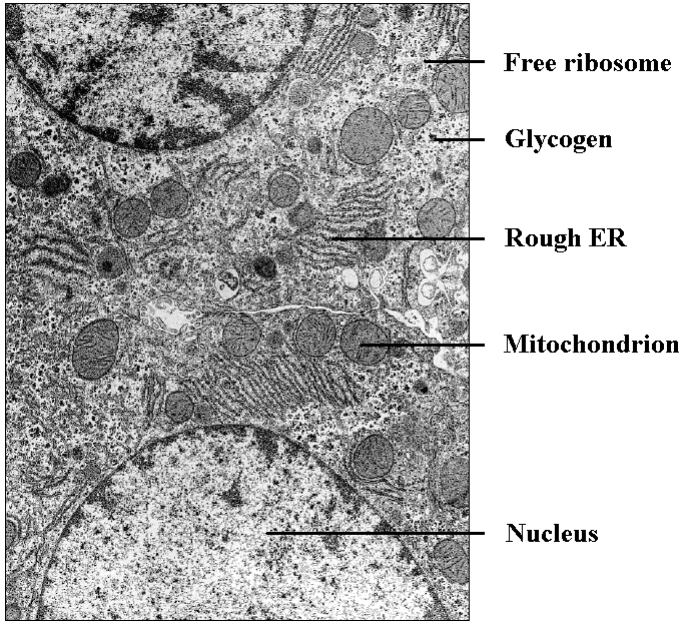
http://www.bio.miami.edu/dana/250/25008_4.html

What is a cell?

- Structural unit of most organisms
- Chemical factory enclosed by a semi-permeable membrane
- Different types of cell within an animal
- **Prokaryotic** cells are simple structurally
- Likely precursors of more complex eukaryotic cells
- E.g. Bacteria
- No compartments within the cell such as nucleus- just **cytosol** with **plasma membrane**
- **Eukaryotic** cells are compartmentalised- they contain **organelles** which perform specific functions – e.g. **nucleus, mitochondria**

Eukaryotic cell



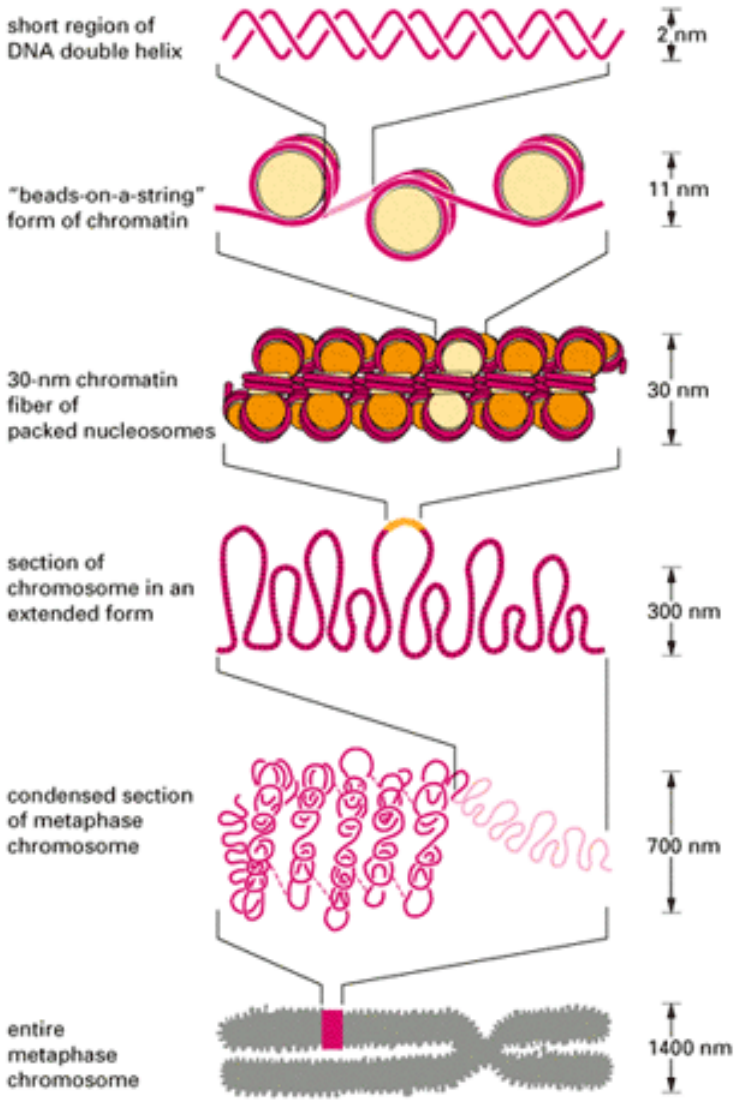


http://ibbiology.wikifoundry.com/page/Identify+the+structures+of+a+liver+cell+from+an+electron+micrograph

Organelles

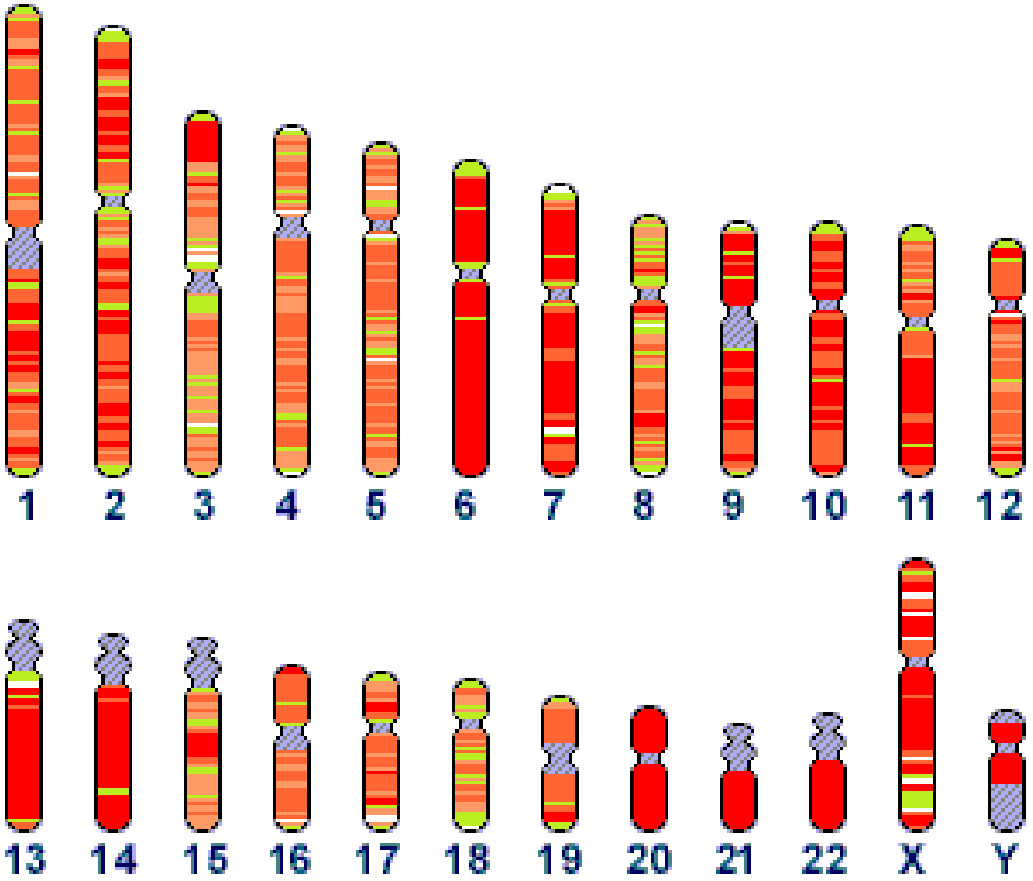
- The **nucleus** – houses the chromosomal DNA which is the genetic information store
- The **rough endoplasmic reticulum** – where most of the **ribosomes** reside, which are the sites for protein synthesis
- **Mitochondria** – are the powerhouses of the cell, producing energy for the reactions of life
- See http://www.biology.arizona.edu/cell_bio/tutorials/pev/page3.html for more details on organelles

DNA packaged into chromosomes



http://aqsaapbio0910.blogspot.co.uk/2009/12/dna-packaging.html

The chromosomes of the human genome



http://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/moreterms.html

The central dogma of Molecular Biology (FOR FURTHER READING)

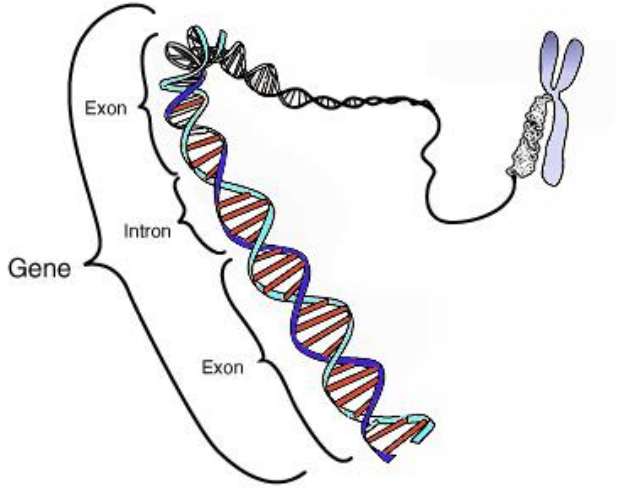
- 

- The expression of genetic information stored in DNA involves, first **transcription** into RNA and then **translation** into the functional protein molecules, in which the **amino acid** sequence is determined by the nucleotide sequence of the DNA.
- DNA replication is also often included as part of the dogma, but the core statement is summarised as “DNA makes RNA makes protein.”

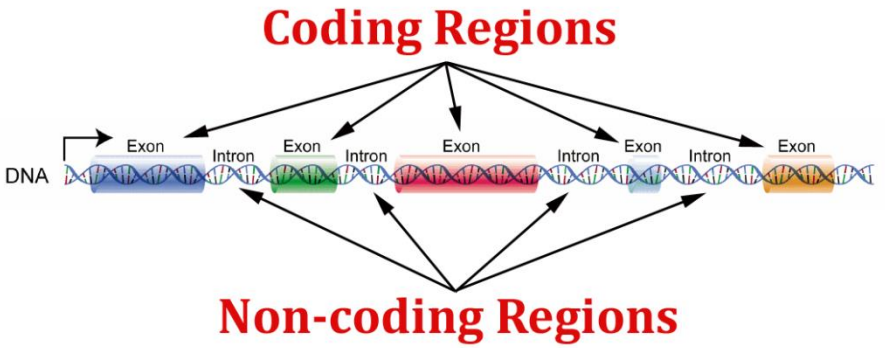
Genes

- A **gene** is a region of DNA that controls a discrete hereditary characteristic, usually corresponding to a single mRNA which will be translated into a protein.
- In eukaryotes, the genes have their coding sequences (**exons**) interrupted by non-coding sequences (**introns**).
- In humans, genes constitute only about 2-3% of DNA, the rest is **“junk”** DNA.

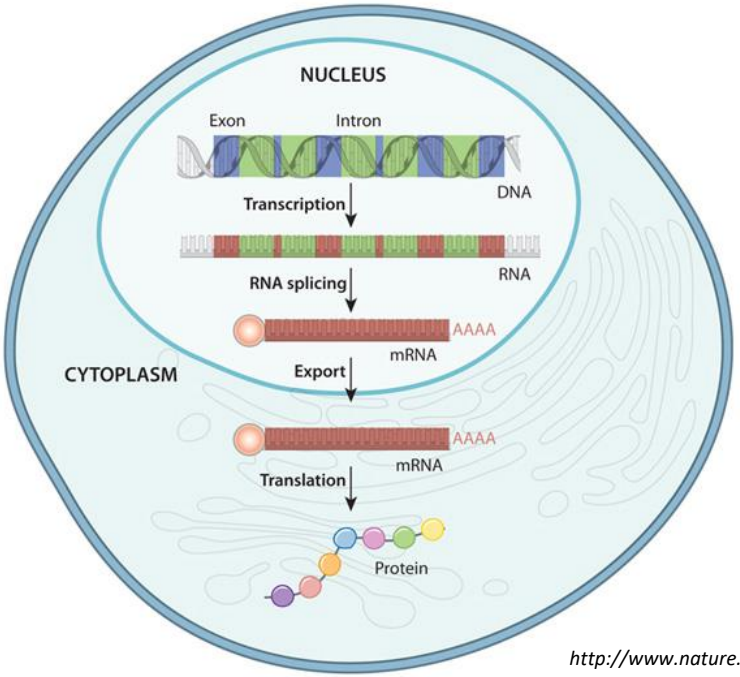
Introns and exons



http://upload.wikimedia.org/wikipedia/commons/0/07/Gene.png



http://imgarcade.com/1/intron-dna/



http://www.nature.com/scitable/topicpage/gene-expression-14121669

RNA

- RNA is like DNA but the sugar-phosphate backbone has a different sugar: ribose instead of deoxyribose.
- and where the DNA molecule has the nucleotide thymine (T), RNA has the nucleotide uracil (U).
- RNA is almost always a single stranded molecule whereas DNA always stored as a double helix in eukaryotes.
- RNA comes in different forms including:
  - Messenger RNA (mRNA) is transcribed from DNA and translated into protein.
  - Transfer RNA (tRNA) is a functional molecule used in the process of translation (see later).

Transcription

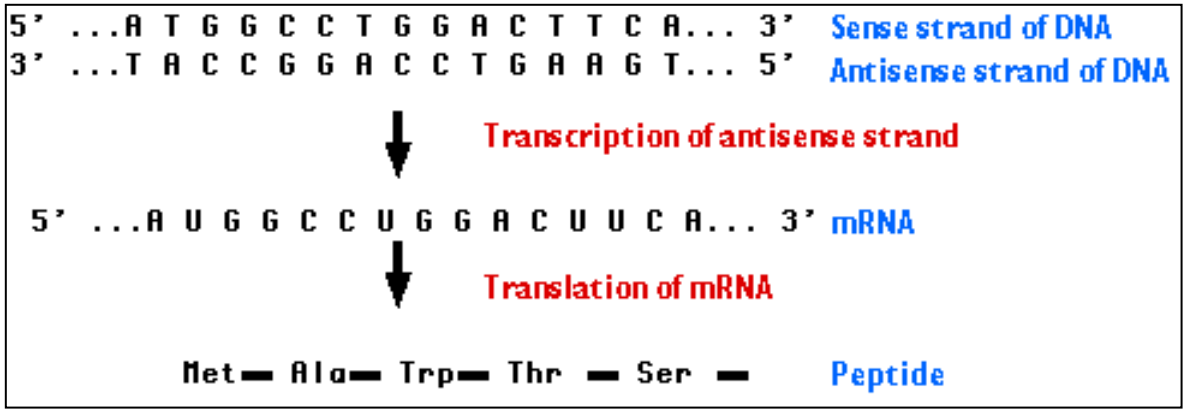
- The process of production of RNA from DNA is called **transcription**; it consists of three stages:

1) Initiation – the **RNA polymerase** enzyme binds to a **promoter site** on the DNA and unzips the double helix.

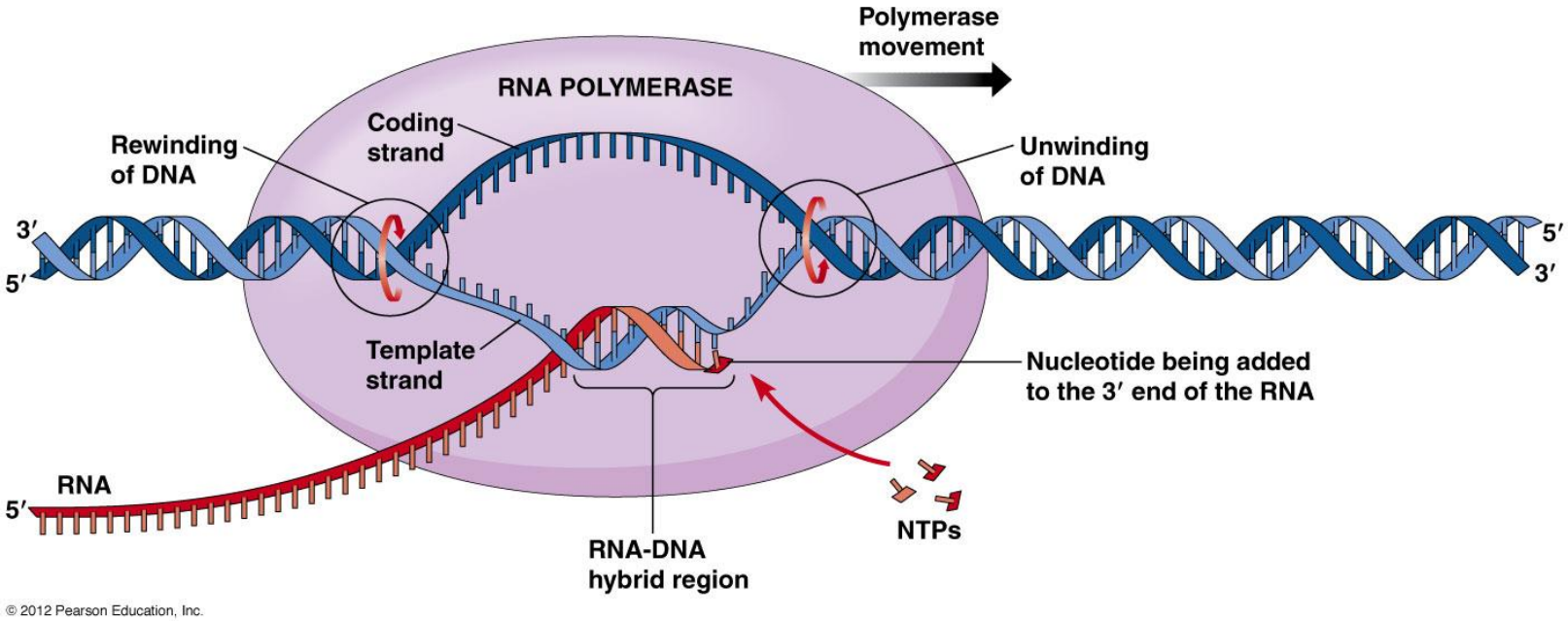
2) Elongation – free nucleotides bind to their complementary pairs on the **template strand** of the DNA elongating the RNA chain which is identical to the **informational strand** of DNA, except that the nucleotide thymine in DNA is replaced by **uracil** in RNA. The polymerase moves along the DNA in the 3' to 5' direction, extending the RNA 5' to 3'.

3) Termination – specific sequences in the DNA signal termination of transcription; when one of these is encountered by the polymerase, the **RNA transcript** is released from the DNA and the double helix can zip up again.

Transcription



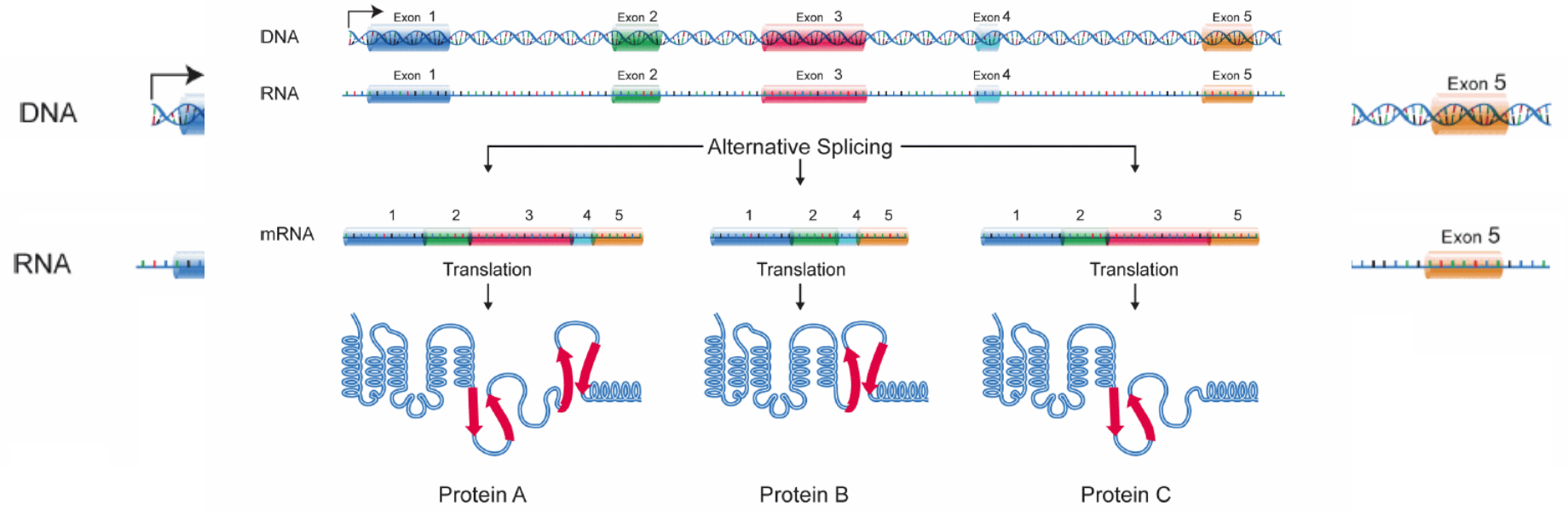
http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/A/AntisenseRNA.html



http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-21/CB21.html

Splicing

The original transcript from the DNA is called **heavy nuclear RNA (hnRNA).** It contains transcripts of both introns and exons. The introns are removed by a process called **splicing** to produce **messenger RNA (mRNA)** and the ends of the RNA molecule are processed.



https://*adapaproject.org/bbk/tiki-index.php?page=Leaf%3A+How+can+one+gene+be+transcribed+and+translated+to+produce+more+than+one+protein%3F*

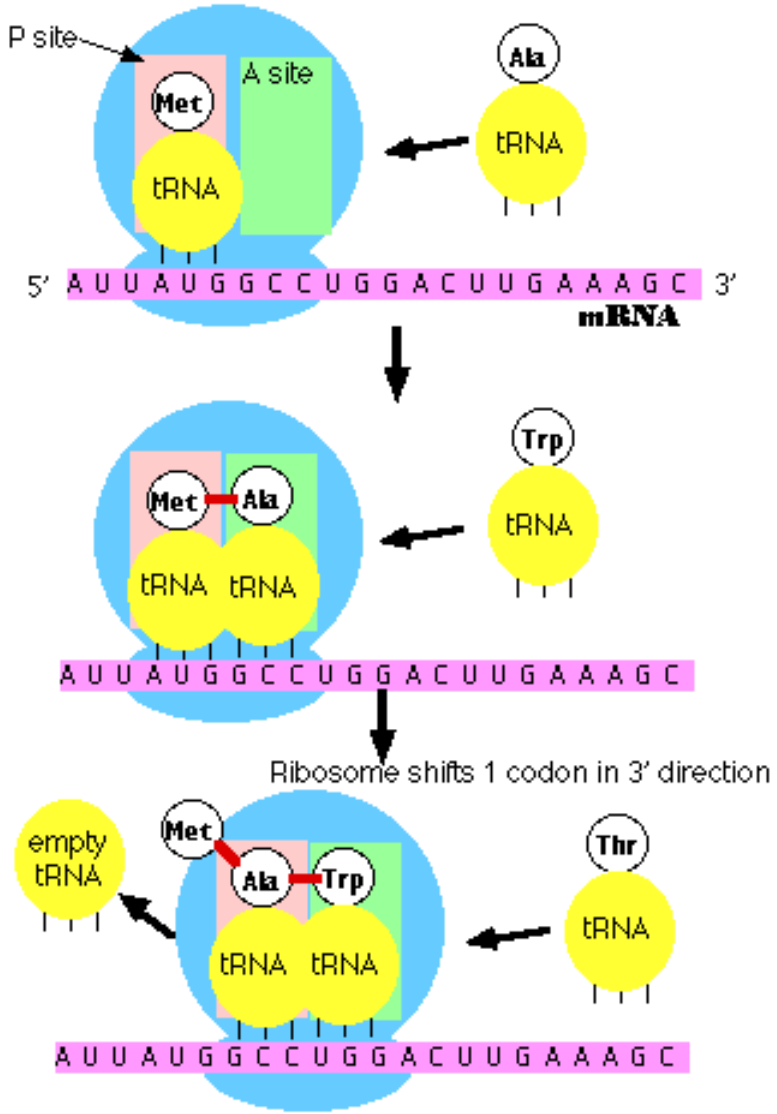
NOTE: One gene can be spliced in multiple ways (different exon combinations) to produce multiple gene products – this is called *alternative splicing*. This means that one gene can code for many proteins.

Translation
(FOR FURTHER READING)

- Special molecules called **transfer RNAs (tRNAs)** recognise both an **amino acid** and a triplet of nucleotides (a **codon**). The tRNA molecule has an **anticodon** on one end which binds to a codon on the mRNA and to a specific amino acid on the other end. It thus enforces the **genetic code** in which a codon codes for a specific amino acid.

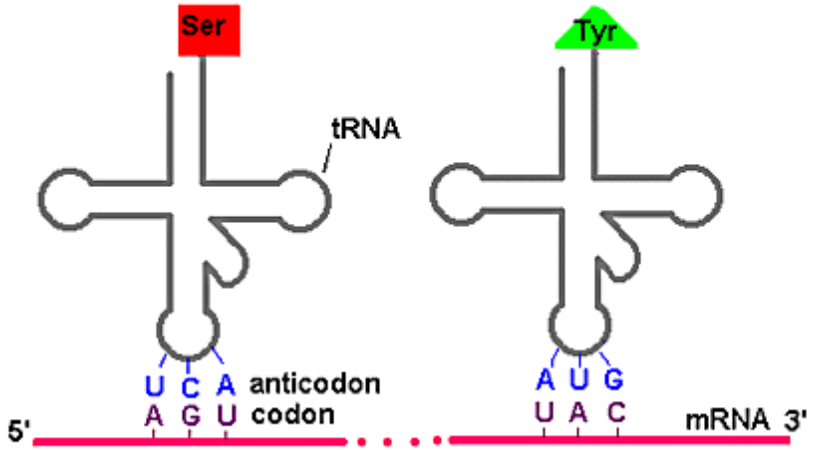
- Protein synthesis takes place on the ribosomes. The tRNAs position themselves for reading the genetic message in the mRNA. The first tRNA binds to a **start codon** (**AUG**) on the mRNA and then each tRNA adds an amino acid to a growing **polypeptide (protein) chain**.

Translation



http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/T/Translation.html

The genetic code



2nd base in codon

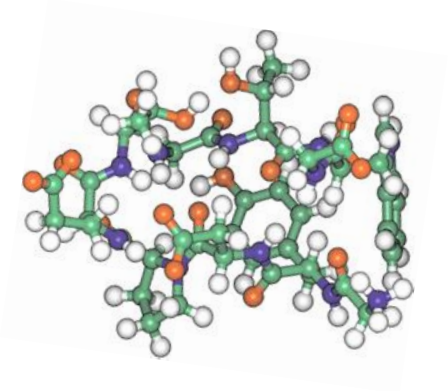
1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

The Genetic Code

http://www.accessexcellence.org/RC/VL/GG/genetic.php

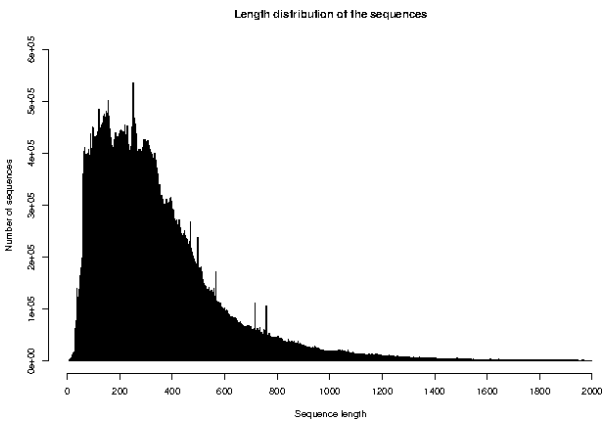
What is a protein?

- A protein is a linear polymer of amino acids linked together by **peptide bonds**.
- The median length of a protein is c. 200 amino acids, but some can contain thousands of amino acids.
- Proteins are the main functional chemicals in the cell, carrying out many functions, for example catalysis of the reactions involved in **metabolism**.
- Proteins have a complex structure which can be thought of as having four structural levels.



Designed 10 aa peptide/protein "chignolin" (PDB code 1UAO)

View it here:
https://www.ncbi.nlm.nih.gov/Structure/icn3d/?mmdbid=1UAO&bu=0



Protein structure
(FOR FURTHER READING)

- **Primary structure** – the sequence of amino acids in the protein chain
- **Secondary structure** – the local spatial arrangement of the protein; short stretches of the chain fold up to form structures such as **alpha-helices** and **beta-sheets**
- **Tertiary structure** – the long range 3D structure of the chain – how the beta-sheets, etc. relate to each other in space (they pack into **domains**)
- **Quaternary structure** – a protein may consist of more than one linear chain molecule; the quaternary structure determines how these chains fold around one another

The amino acids

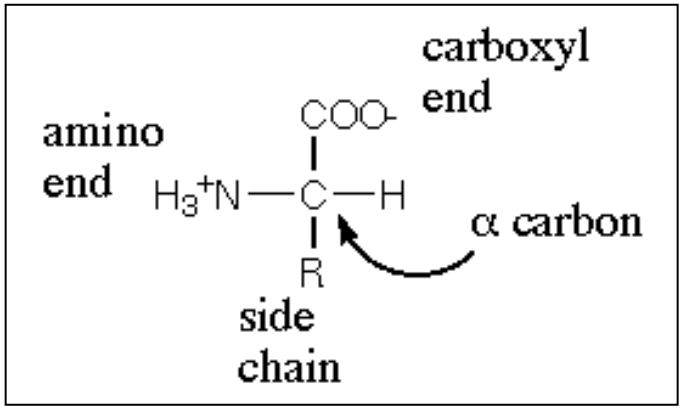
Proteins are polymers of the 20 naturally occurring amino acids. Each amino acid has a three-letter code and a single letter code:

Start Learning these NOW!

Alanine	Ala A
Cysteine	Cys C
Aspartic AciD	Asp D
Glutamic Acid	Glu E
Phenylalanine	Phe F
Glycine	Gly G
Histidine	His H
Isoleucine	Ile I
Lysine	Lys K
Leucine	Leu L
Methionine	Met M
AsparagiNe	Asn N
Proline	Pro P
Glutamine	Gln Q
ARginine	Arg R
Serine	Ser S
Threonine	Thr T
Valine	Val V
Tryptophan	Trp W
Tyrosine	Tyr Y

The amino acids

- Consist of a central carbon atom (the alpha-carbon) connected to an amino group, a carboxyl group, a hydrogen atom and a side chain. The side chain differs between the different amino acids but the rest is the same:

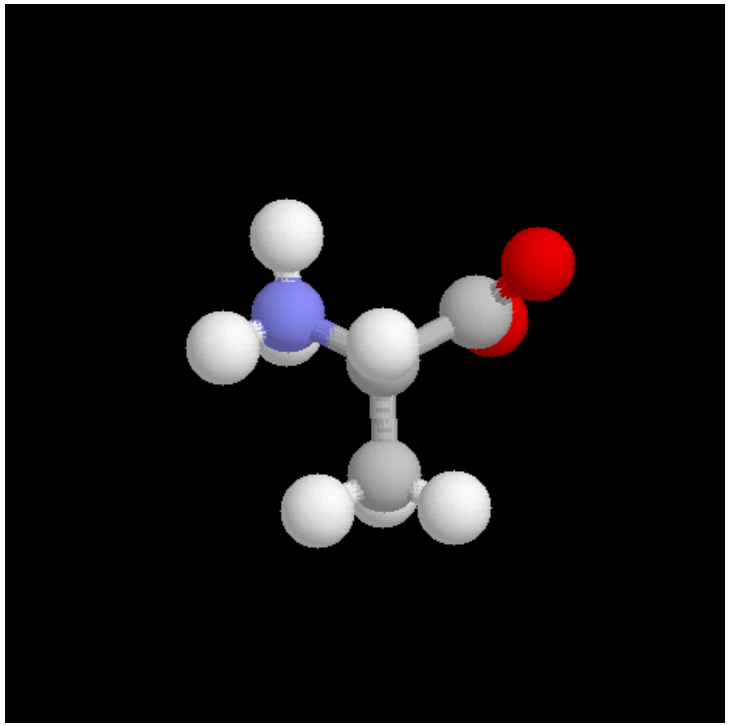


http://web.mit.edu/esgbio/www/lm/proteins/aa/aminoacids.html

- Have different properties because their side chains have different shapes and chemical groups.
- Hydrophobic/hydrophilic, acidic/basic/neutral, aliphaptic/aromatic, conformationally important

Example amino acids

Grey=carbon, white=hydrogen, red=oxygen, blue=nitrogen

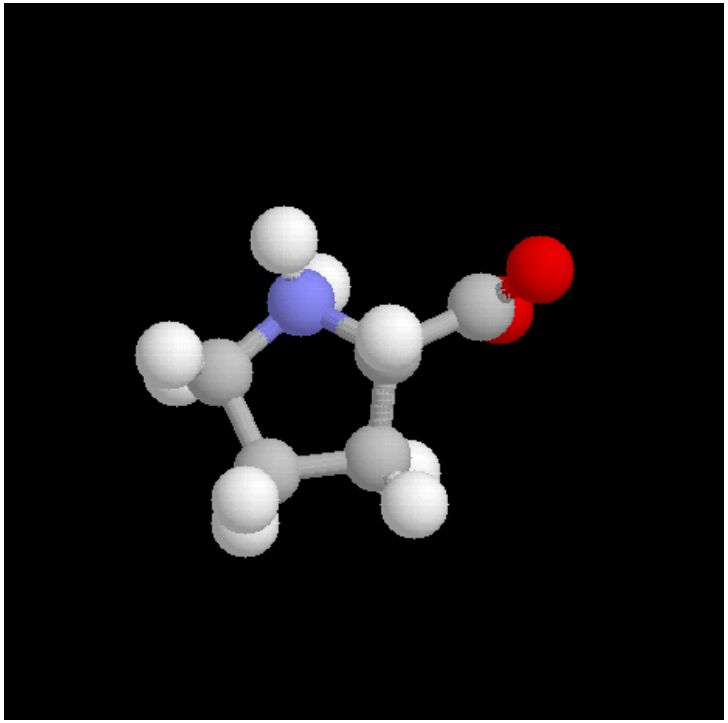


Alanine: $C_3H_7NO_2$

Side chain: CH_3

Proline: $C_5H_9NO_2$

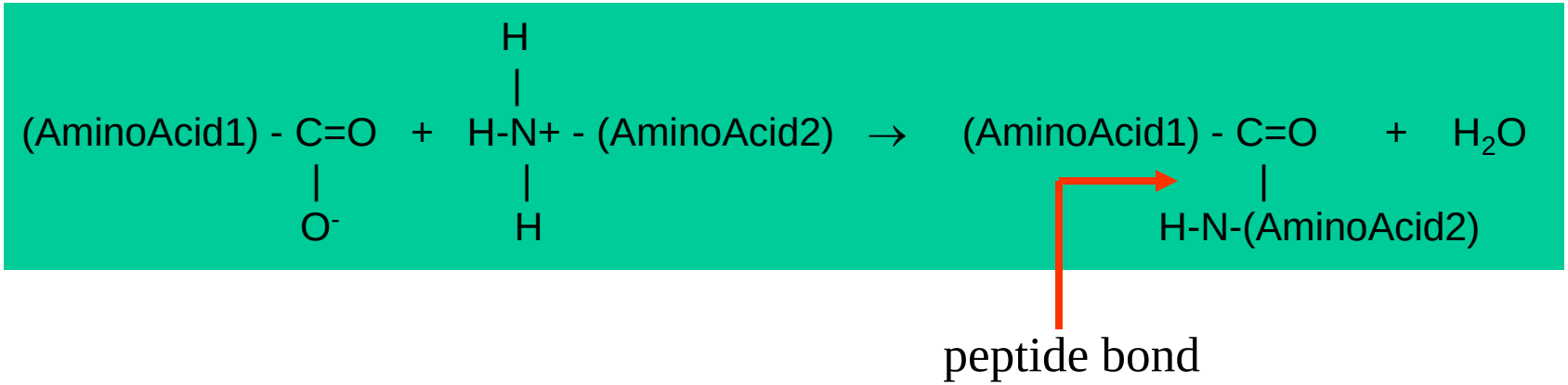
Side chain: C_3H_6, links to N in amino group



http://www.cryst.bbk.ac.uk/PPS2/course/section2/index.html, Sami Raza

The peptide bond

AminoAcid1 + AminoAcid2 → Dipeptide + Water, i.e.



- Polypeptides are just long chains of amino acids linked by peptide bonds. Proteins are made up of one or more polypeptide chains (cf. quaternary structure).
- Name comes from peptide group –CONH-.

Primary structure

- Primary structure of a protein is simply the linear sequence of amino acid in its polypeptide chain(s) (NB proteins are written in order from the amino-terminal end to the carboxy-terminal end, so Ala-Cys-Phe is different from Phe-Cys-Ala)

- E.g. Pancreatic trypsin inhibitor protein:

  MKQSTIALALLPLLFTPVTKARPDFCLEPPTGPCKARII
  RYFYNAKAGLCQTFLYGGCRAKRNNFKSAEDCMRTCGGA

  This sequence contains all the information required to determine the higher levels of structure. The linear polypeptide chain folds in a particular arrangement, giving a three-dimensional structure, but the information on how to fold is contained in the sequence.

Alpha helix

- An alpha-helix is a tight rod-like helix formed out of the polypeptide chain.
- The polypeptide main chain makes up the central structure, and the side chains extend out and away from the helix.
- The CO group of one amino acid (n) is hydrogen bonded to the NH group of the amino acid four residues away (n+4):

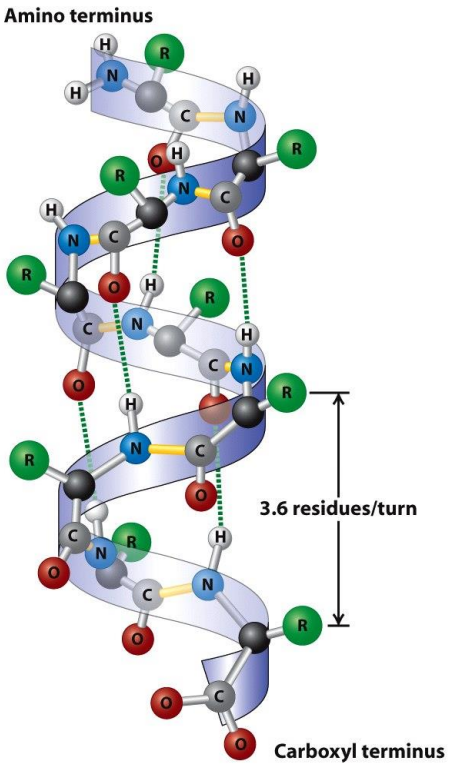


Figure 3-4
Molecular Cell Biology, Sixth Edition

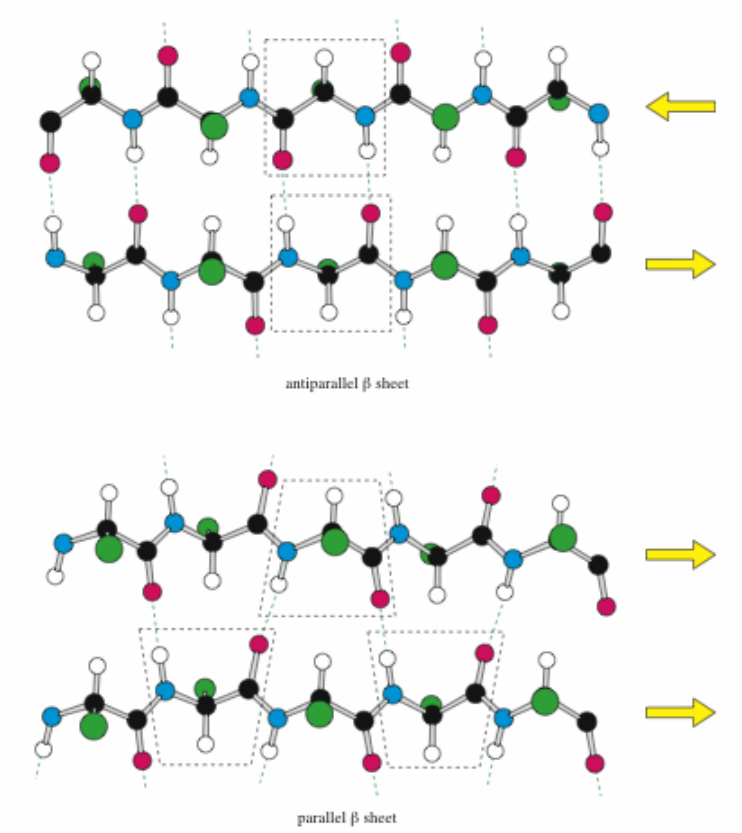
http://www.bio.miami.edu/tom/courses/bil255/bil255goods/03_proteins.html

Alpha helix

- Alpha helices are most commonly made up of hydrophobic amino acids, because hydrogen bonds are generally the strongest attraction possible between such amino acids.
- Between one amino acid and the next is a rise of 1.5Å and a turn of 100°.
- Alpha helices are found in almost all proteins to various extents , e.g. haemoglobin -75% (see Stryer).
- Proteins have right-handed helices.

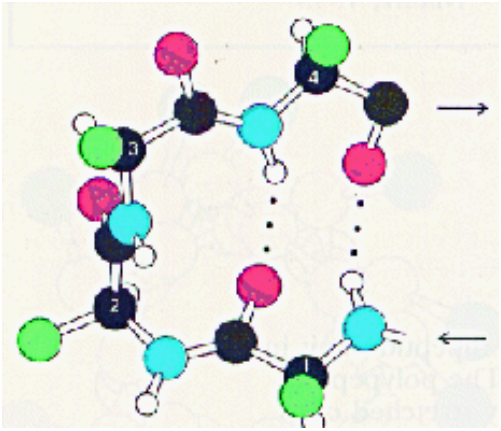
Beta Sheet

- A beta pleated sheet is another type of secondary structure. Sheets can either be parallel or anti-parallel.



http://biologicalphysics.iop.org/cws/article/lectures/46259

Anti-parallel sheets have hairpin turns like this (hydrogen bonds between amino acid n and n+4):



http://dwb4.unl.edu/Chem/CHEM869K/CHEM869KLinks/esg-www.mit.edu/esgbio/lm/proteins/structure/structure.html

Beta sheet

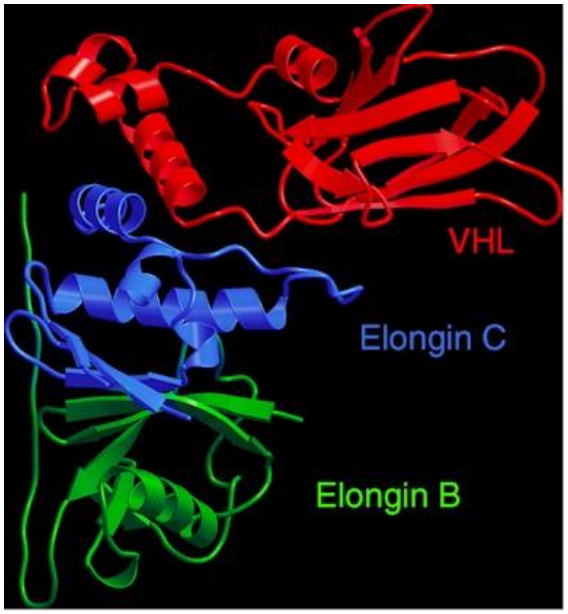
- The individual lines of amino acids within the protein are called strands.
- Typically a sheet will have 2-5 parallel or anti-parallel strands.
- The axial distance between two amino acids is 3.5Å.

Loops

- Between the alpha helices and beta strands in the protein are **loop** regions.
- These are less regular, although may still have some structure.
- Loops tend to end up on the outside of the proteins when the protein folds up to form its full 3D structure (tertiary structure), so they are exposed to water- the loop regions thus tend to be rich in hydrophilic **residues** (amino acid side chains) - cf. prediction.
- Loops often ~ binding sites for other molecules.

Schematic diagrams of secondary structure

- Secondary structure elements are often visualized in ribbon diagrams like this:



VHL protein

Stebbins et al, *Science*, 284:455.

View it here:

https://www.ncbi.nlm.nih.gov/Structure/icn3d/?mmdbid=1VCB&bu=0

Coiled ribbons = alpha helix, arrow ribbons = beta strand, thin strings = loops, etc.

A protein domain

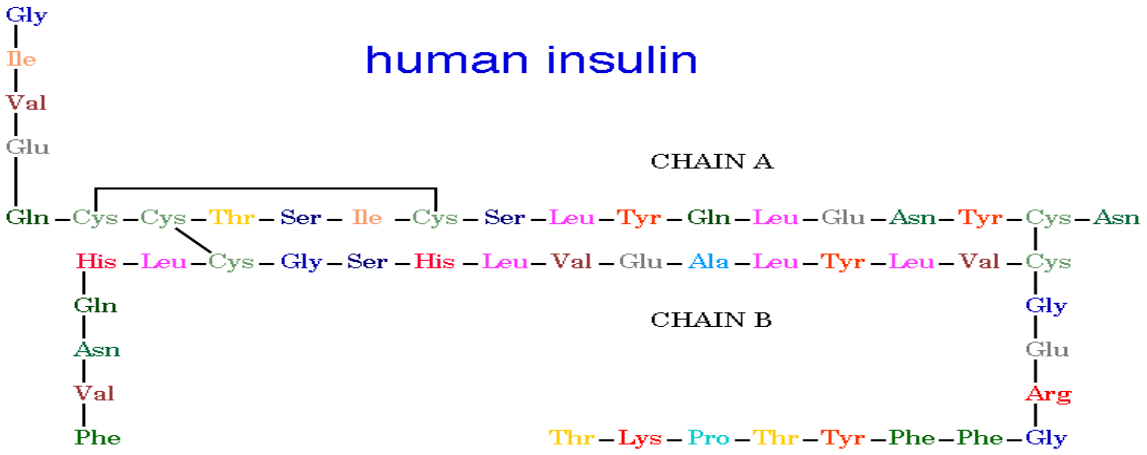
- The tertiary structure is the way in which the secondary structure elements (e.g. alpha helices) fit together in the full 3D structure.

- The protein folds up so that amino acids which are far apart in the linear sequence may be close together in space, forming a domain:

http://www.cryst.bbk.ac.uk/PPS2/course/section10/1pfk.gif

Disulphide bridges

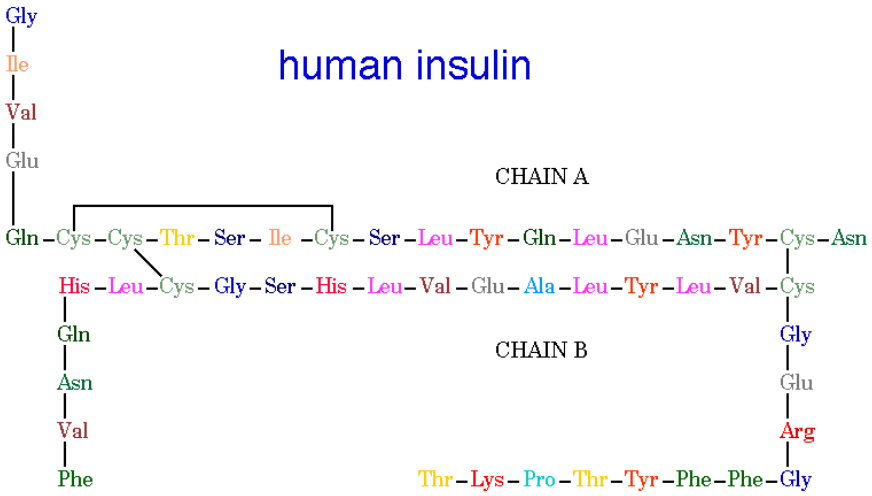
- The tertiary structure also describes the pattern of disulphide bridges, which form between the 2 cysteine (C) amino acids. Cysteine (like methionine) contains a sulphur atom and when two cysteines are spatially close they can form covalent –S-S- bonds:



http://www.chem.uwec.edu/Chem406/Webpages/Ying/overview.htm

A protein with more than one chain

- Insulin is also a good example of a protein with more than one polypeptide chain:



http://www.chem.uwec.edu/Chem406/Webpages/Ying/overview.htm

In the quaternary structure, the B chain is wrapped around the A chain, which forms a compact central unit.

- In fact the structure of insulin is more complicated: it forms hexamers with six insulin molecules (A and B chain) around two zinc ions and 6 water molecules.

Conclusions
(ALL FOR FURTHER READING)

- This has been a general introduction to molecular biology, introducing the key molecules of life:
  - **DNA** (the store of genetic information)
  - **RNA**
  - & **protein** (the function molecules of the cell)
- **Central Dogma**: (replicated) DNA is **transcribed** to form RNA which is **translated** to form protein
- Key processes: DNA replication, transcription, splicing, translation
- We discussed DNA and protein structures (which are important for their functions).