

ESAT

Biology

DNA and protein synthesis

Free lesson notes

Covers	B5.1, B5.2, B5.3 and B5.4
Format	Lesson, worked examples, self-check
Price	Free

About this lesson

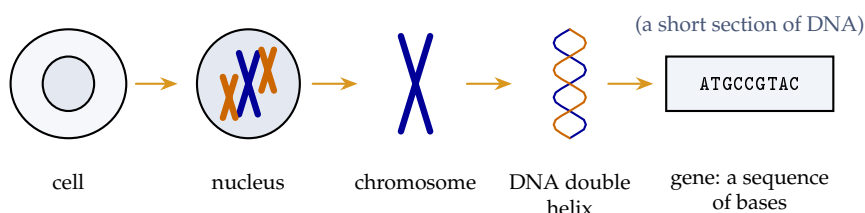
DNA is the molecule that carries the instructions for building every protein in your body, and so for building you. This lesson explains **what the genome is** (B5.1), **how DNA is built** from nucleotides and why its two strands pair up as they do (B5.2), **how the order of bases in a gene becomes a protein** (B5.3), and **what happens when the sequence changes**: gene mutations (B5.4). It rewards careful, mechanical working: reading triplets, counting bases, and tracking exactly what a change in one letter does to the protein.

Cover each worked example and try it first. There is no calculator in the real ESAT, so the numbers are friendly. Finish with the self-check, then the practice set.

Heurisco is an independent practice provider and is not affiliated with, endorsed by, or connected to UAT-UK, Pearson VUE, the UCAT Consortium, or any university. All questions are original.

1 The genome (B5.1)

Inside the nucleus of a cell, the DNA is organised in chromosomes. The **genome** of an organism is **the entire genetic material of that organism**: all of its DNA, every gene and all the DNA between the genes. The human genome has about 3 billion (3×10^9) base pairs, spread over the 23 pairs of chromosomes, and about 20 000 genes. Only a small part of it (a few per cent) actually codes for proteins; the rest includes regions that help to control which genes are used, and regions whose function is not fully understood. A **gene** is a section of DNA that codes for one protein (or polypeptide).



Scientists have now worked out ("sequenced") the whole human genome, in the Human Genome Project completed in the early 2000s. Knowing the genome allows scientists to find genes linked to inherited disorders and to risk of diseases, to design medicines suited to a person's genetic makeup, to trace the movement of human populations, and to compare species to understand their evolutionary relationships.

Worked example: genes and base pairs

A genome has 3×10^9 base pairs. Only 2% of it codes for proteins, and an average gene has 3000 coding base pairs. About how many genes are there?

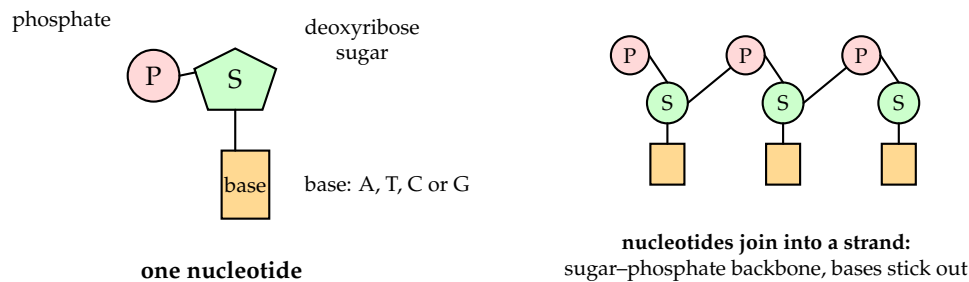
Coding DNA = $0.02 \times 3 \times 10^9 = 6 \times 10^7$ base pairs. Number of genes = $6 \times 10^7 \div 3000 = 2 \times 10^4 = 20\,000$.

Common mistakes

- Saying the genome is "all the genes". It is *all* the DNA, including DNA between genes.
- Confusing genome (the whole set of genetic material) with gene (one section) or with chromosome (one long DNA molecule).
- Believing most of the genome codes for proteins: only a small percentage does.

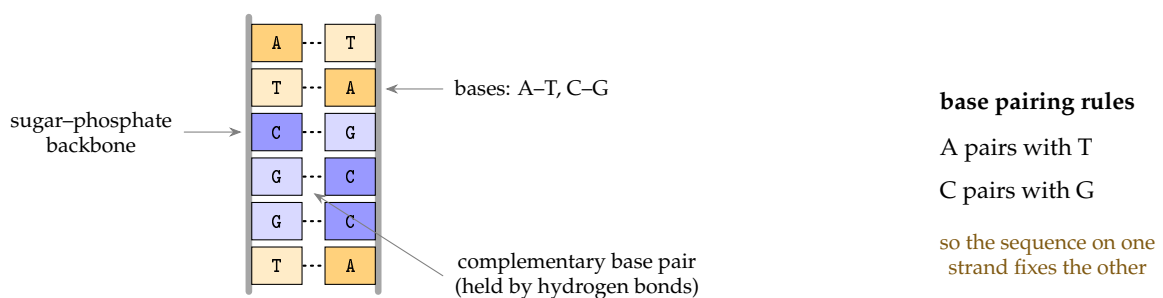
2 The structure of DNA (B5.2)

DNA (deoxyribonucleic acid) is a **polymer**, a very long molecule made of many repeating units called **nucleotides**. Each nucleotide has three parts: a **phosphate** group, a **sugar** (deoxyribose) and one of four **bases**: **A** (adenine), **T** (thymine), **C** (cytosine) and **G** (guanine).



Nucleotides join into a strand by linking the sugar of one to the phosphate of the next, forming a **sugar-phosphate backbone** with a base sticking out of each sugar. DNA is made of **two strands** wound around each other in a **double helix**. The strands are held together by weak bonds between the bases, and the bases pair up in a fixed way, called **complementary base pairing**:

A pairs with T C pairs with G.



Why only these pairs? Each pair has one larger base and one smaller base which fit exactly between the two backbones, and the bonds form between the matching shapes. The consequence is vital: **if you know the sequence of bases on one strand, you know the other**. A strand reading A-T-C-G-G-T is paired with a strand reading T-A-G-C-C-A.

Counting bases. In double-stranded DNA the number of A equals the number of T, and the number of C equals the number of G (so A + G is half of all the bases).

Worked example: base proportions

In a sample of double-stranded DNA, 28% of the bases are adenine. What percentages are thymine, cytosine and guanine?

T pairs with A, so T = 28%. A + T = 56%, leaving $100 - 56 = 44\%$ for C + G. C = G, so each is 22%.

Worked example: complementary strand

One strand of DNA has the base sequence TACGGCTA. Write the sequence of the complementary strand.

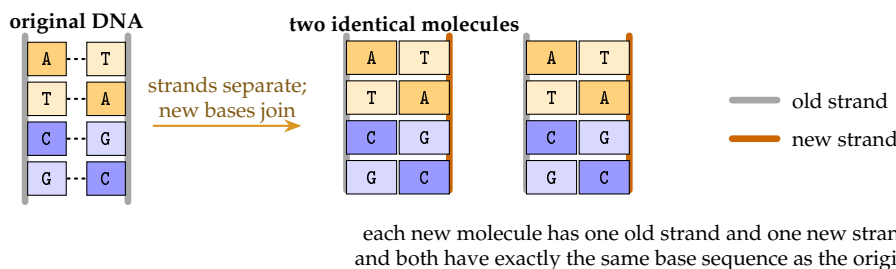
Swap each base for its partner (A ↔ T, C ↔ G): T → A, A → T, C → G, G → C, G → C, C → G, T → A, A → T, giving ATGCCGAT.

Common mistakes

- Pairing A with C or G with T, or pairing A with A. Learn the pairs: A-T, C-G.
- Thinking the strands are held together by the sugar-phosphate backbone. The *bases* are joined across the middle; the backbone runs along each strand.
- Using U in DNA. DNA has T; the base U (uracil) is found in RNA.

3 Copying DNA (B5.2)

Before a cell divides it must copy its DNA. The double helix **unwinds and the two strands separate**. Each separated strand is used as a pattern: free nucleotides line up against it by complementary base pairing (A with T, C with G) and are joined to make a new strand. The result is two DNA molecules, each of which has **one old strand and one new strand**, and both have *exactly the same sequence as the original*. This accurate copying is why cells made by mitosis have identical DNA.



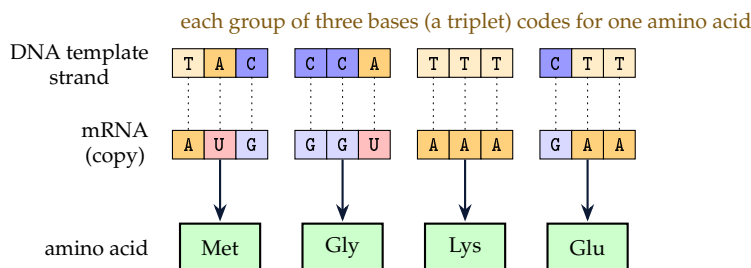
Common mistakes

- Saying the DNA is copied during mitosis. It is copied earlier, in interphase, before the cell divides.
- Thinking the new molecules are different from the original. Complementary base pairing makes them identical.

4 The genetic code: from bases to amino acids (B5.3)

Proteins are the molecules that do most of the work in the body: enzymes, many hormones, antibodies, structural proteins such as collagen. Every protein is a long chain of **amino acids** joined together, called a **polypeptide**. There are twenty different amino acids, and a protein is defined by *which* amino acids it contains and *in what order*.

A gene stores that order. The code works in groups of **three bases**, called **triplets**: each triplet codes for **one amino acid**. (Why three? There are only four bases, so a code of one base could specify only 4 amino acids, and two bases only $4^2 = 16$. Three bases give $4^3 = 64$ combinations, enough for all twenty amino acids.) Some amino acids are coded by more than one triplet, and one triplet, called the **stop codon**, tells the cell where the polypeptide ends.



In the picture, the cell reads the DNA template strand and makes a copy called **messenger RNA (mRNA)**, with the bases paired as A→U, T→A, C→G, G→C. (RNA has U, uracil, in place of T.) Each group of three mRNA bases, a **codon**, specifies one amino acid. A short extract from the genetic code (the full code has 64 codons):

AUG = Met (start)	GGU, GGC = Gly	AAA, AAG = Lys	GAA = Glu
GAU = Asp	CGG = Arg	GUA = Val	UGG = Trp
UAA, UGA = STOP	CCU = Pro	UUU = Phe	

Counting. A polypeptide of n amino acids needs $3n$ bases in its coding region, plus 3 more for the stop codon.

Worked example: how many bases?

A polypeptide has 150 amino acids. How many bases are needed in the gene to code for the amino acids? How many are there including the stop codon?

$150 \times 3 = 450$ bases for the amino acids; including the stop codon $450 + 3 = 453$.

Worked example: reading mRNA

Translate the mRNA AUG GGU AAA GAA using the table.

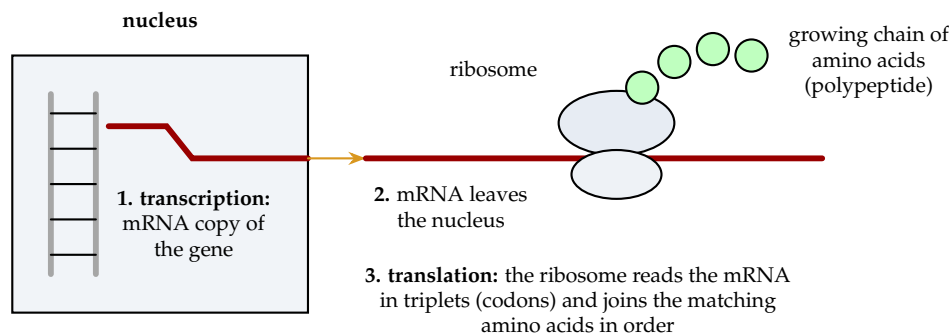
AUG = Met, GGU = Gly, AAA = Lys, GAA = Glu: Met–Gly–Lys–Glu.

Common mistakes

- Saying one base codes for one amino acid. It is a group of three.
- Starting the reading in the wrong place: the codons must be read from the start codon in unbroken groups of three.
- Forgetting to convert DNA to mRNA when asked for mRNA (T is replaced by A, A by U).

5 Protein synthesis step by step (B5.3)

DNA never leaves the nucleus, but proteins are made on ribosomes in the cytoplasm. The message is carried by mRNA:



1. **Transcription** (in the nucleus): the DNA unwinds at the gene, and a single-stranded mRNA copy is made using complementary base pairing.
2. The mRNA leaves the nucleus and goes to a ribosome in the cytoplasm.
3. **Translation**: the ribosome reads the mRNA codon by codon; amino acids (brought by small carrier molecules called tRNA) are joined in the order the codons specify, building the polypeptide chain.

	DNA	mRNA
strands	two (double helix)	one (single strand)
bases	A, T, C, G	A, U, C, G (U in place of T)
where	stays in the nucleus	made in the nucleus, moves to the ribosome
job	stores the genetic code permanently	carries a copy of one gene to the ribosome

Worked example: from gene to polypeptide

The DNA template strand has the sequence TAC CCA TTT CTT. Find the mRNA and the amino acids. mRNA (A → U, T → A, C → G, G → C): AUG GGU AAA GAA. Amino acids: Met–Gly–Lys–Glu.

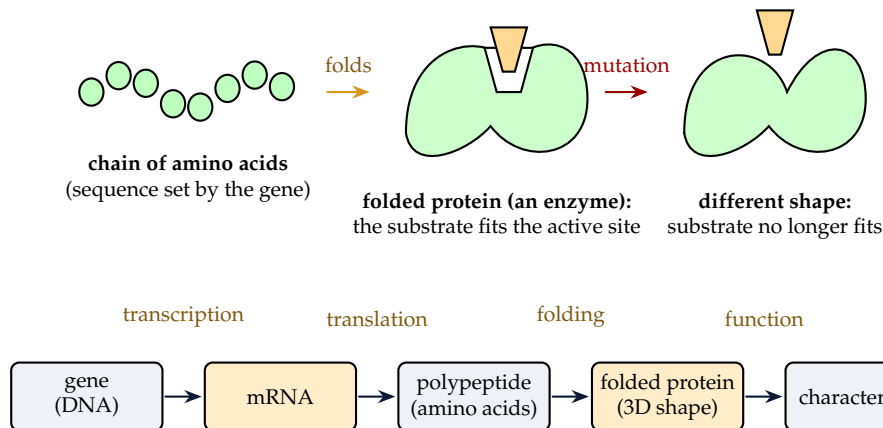
Common mistakes

- Saying proteins are made in the nucleus. They are made at ribosomes in the cytoplasm.
- Mixing up transcription (DNA → mRNA, in the nucleus) with translation (mRNA → polypeptide, at the ribosome).
- Thinking mRNA is double-stranded. It is a single strand.

6 Protein shape and function (B5.3)

A newly made polypeptide folds up into a precise **three-dimensional shape**. The shape is determined by the **sequence of amino acids**: different amino acids attract or repel each other, so the chain folds in a particular way. The shape matters because it decides what the protein can do. An **enzyme**, for example, has an **active site** whose shape matches its substrate,

like a lock and key; an antibody has a shape that fits a particular antigen; collagen forms strong fibres. Change the sequence and the shape (and so the function) can change.



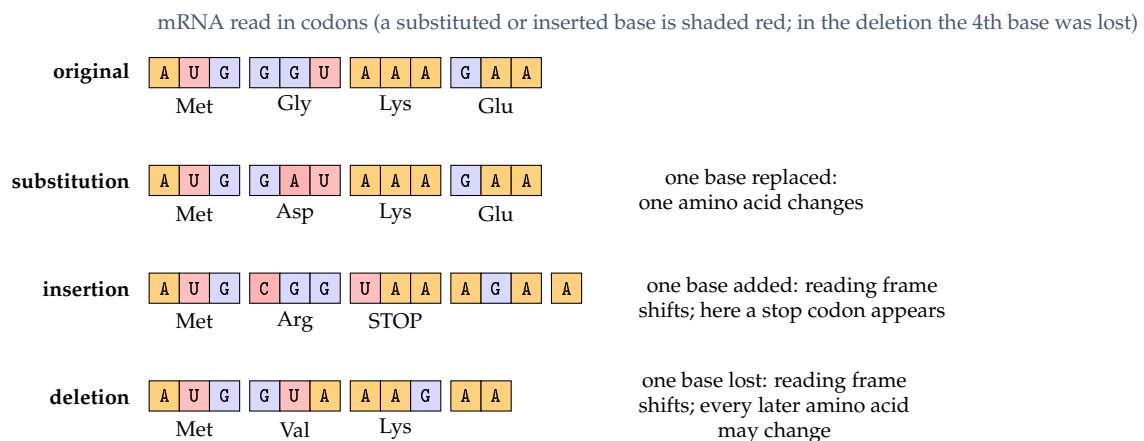
This is the chain of reasoning to remember: **gene (sequence of bases) → sequence of amino acids → folded 3D shape → function → characteristic**. If a protein cannot fold properly, it will not work.

Common mistakes

- Saying the DNA sequence directly determines the shape. It determines the amino acid sequence, which in turn determines the shape.
- Saying a protein's shape is the same whatever amino acids it contains.
- Saying a denatured or altered enzyme "uses up" its substrate faster. It usually can no longer bind the substrate.

7 Gene mutations (B5.4)

A **gene mutation** is a change in the base sequence of DNA. Mutations occur spontaneously (random errors when DNA is copied), and their rate is increased by **mutagens** such as ionising radiation (X-rays, gamma rays), ultraviolet light and certain chemicals (including some in tobacco smoke). There are three basic types, shown here as changes in the mRNA codons:



- **Substitution:** one base is replaced by another. Only one codon is changed, so *at most one amino acid* changes. Because several codons can code for the same amino acid (GGU and GGC both code for Gly), a substitution may change nothing at all (a *silent mutation*).
- **Insertion and deletion:** one or more bases are added or removed. Unless the number of bases added or removed is a multiple of three, all the codons after the change are read in a different grouping (a **frameshift**), so *many* amino acids may be changed, or a stop codon may appear early, giving a short, non-functional protein. If three bases are lost, the code is still read in the same frame afterwards and only one amino acid is missing.

Effects. Most mutations have no effect on the protein (silent, or in DNA that does not code for a protein) or little effect. Some are harmful: if an amino acid change alters the shape of an enzyme's active site, the substrate no longer fits and the enzyme does not work; for example, sickle cell disease results from a single base substitution that replaces one amino acid in haemoglobin, and cystic fibrosis is commonly caused by the deletion of three bases that removes one amino acid. Very rarely a mutation gives a protein with a new, useful property, and this is a source of the variation on which natural selection acts. A mutation in a gamete can be inherited; a mutation in a body cell cannot (but can sometimes cause cancer).

Worked example: effect of a substitution

A gene region is read as the mRNA AUG GGU AAA GAA (Met–Gly–Lys–Glu). Find the effect of changing the third base of the third codon from A to G (AAA→AAG).
AAG also codes for Lys, so the amino acid sequence is unchanged: a silent mutation.

Worked example: effect of a deletion

The 4th base (a G) of the same mRNA is deleted. Find the first three complete codons that remain.
The sequence becomes AUG GUA AAG AA... giving Met–Val–Lys, and the rest (previously Gly–Lys–Glu) is read in a different frame. All the amino acids after the first are changed, a frameshift.

Common mistakes

- Saying every mutation is harmful. Many have no effect on the protein.
- Saying a substitution always changes the amino acid: the code has more than one codon for some amino acids.
- Forgetting that deleting one base shifts every later codon, while deleting three (a whole codon) removes a single amino acid.
- Saying a mutation in a body cell is passed to children. Only mutations in gametes (or the cells that make them) can be inherited.

8 Three full walkthroughs (B5.1 to B5.4)

Each walkthrough combines several ideas from this lesson.

Worked example: from DNA to polypeptide and back

The template strand of a gene is TACGGTTTCCTT. (a) Write the complementary DNA strand. (b) Write the mRNA. (c) Use the table (AUG = Met, CCA = Pro, AAA = Lys, GAA = Glu) to write the polypeptide. (d) How many bases are needed to code for a polypeptide of 90 amino acids?
(a) T→A, A→T, C→G, G→C, G→C, T→A, T→A, T→A, C→G, C→G, T→A, T→A: ATGCCAAAGGAA. (b) The mRNA is a copy of the template made by pairing A→U, T→A, C→G, G→C: AUGCCAAAGGAA. (c) AUG CCA AAG GAA: Met–Pro–Lys–Glu (AAG codes for Lys as well as AAA). (d) $90 \times 3 = 270$ bases.

Worked example: what does the mutation do?

Using the mRNA AUG CCA AAG GAA (Met–Pro–Lys–Glu), decide the effect of (i) a substitution of the fifth base C by U, (ii) the insertion of a G after the sixth base, (iii) deletion of the last three bases.
(i) CCA becomes CUA, which changes the second amino acid (Pro is replaced by Leu; CUA codes for leucine), so one amino acid changes and the rest is the same. (ii) The mRNA becomes AUG CCA GAA GGA A...: the third codon is now GAA (Glu) instead of AAG (Lys), and every later codon is different (a frameshift). (iii) The last codon GAA is lost, so the polypeptide is one amino acid shorter (Met–Pro–Lys); the earlier amino acids are unchanged.

Worked example: percentages and numbers together

A bacterial plasmid is a double-stranded DNA ring of 6000 base pairs, 30% of whose bases are guanine. How many thymine bases does it contain?
Total bases = $2 \times 6000 = 12000$. G = 30%, so C = 30%, and C + G = 60%. A + T = 40%, so T = 20% of 12000 = 2400.

Common mistakes

- Not checking whether a question gives you the template strand or the coding strand: the mRNA is a copy of the template.
- Giving the number of bases in a *base pair* count as if it were the number of bases (one base pair is two bases).

- Forgetting that more than one codon can code for the same amino acid when predicting a substitution.

9 Putting it together: exam technique

1. For base-pairing questions, write the partner under every base before answering. A with T (or U in RNA), C with G.
2. For percentages: $A = T$ and $C = G$; subtract from 100%, then halve.
3. To read a code, split into groups of three from the start, convert to mRNA if needed, and look up each codon in the table that is given.
4. Counting: $3 \times \text{amino acids} = \text{bases}$ (add 3 for the stop codon).
5. For a mutation, find the position of the change, work out which codon(s) are altered, and say whether the amino acid changes; for insertions and deletions, check whether the number of bases is a multiple of three.
6. Link the chain: base sequence $\rightarrow$ amino acids $\rightarrow$ 3D shape $\rightarrow$ function.

Spec coverage (B5.1 to B5.4). The genome; DNA structure (nucleotides, complementary base pairing); protein synthesis (triplets, polypeptides, 3D shape from the amino acid sequence); gene mutations.

Check yourself

Try these without looking back, then compare with the answers on the next page.

1. Define the genome.
2. A DNA strand reads GGATC. Write the complementary strand.
3. In a DNA molecule 35% of the bases are cytosine. What percentage are adenine?
4. How many bases code for a polypeptide of 200 amino acids (not counting the stop codon)?
5. Using the table, translate the mRNA CCG UGG GGC.
6. Where are proteins made in a cell, and what carries the instructions there?
7. Why can a substitution mutation have no effect on the protein?
8. Why is a deletion of one base usually more serious than a substitution?

Answers to the self-check

1. The entire genetic material (all the DNA) of an organism.
2. CCTAG.
3. $G = 35\%$, so $C + G = 70\%$; $A + T = 30\%$, so $A = 15\%$.
4. $3 \times 200 = 600$.
5. Arg–Trp–Gly.
6. At ribosomes in the cytoplasm; messenger RNA (mRNA).
7. Several codons code for the same amino acid (the code is degenerate), so the amino acid may not change.
8. It shifts the reading frame, changing every codon after it, whereas a substitution changes at most one codon.

Next step

Heurisco's *ESAT Biology: DNA and Protein Synthesis Practice Set* has twenty original exam-style questions on this lesson, with diagrams and a codon table, in order of difficulty, with a full worked solution and a note on the trap in each.