Nucleic acids

International Baccalaureate · IB Diploma · Biology · HL

Scope and prerequisites

How can a molecule made from only four kinds of base carry so many different instructions? The answer lies in their sequence, the structure that keeps the sequence intact, and complementary pairing that allows information to be copied and used.

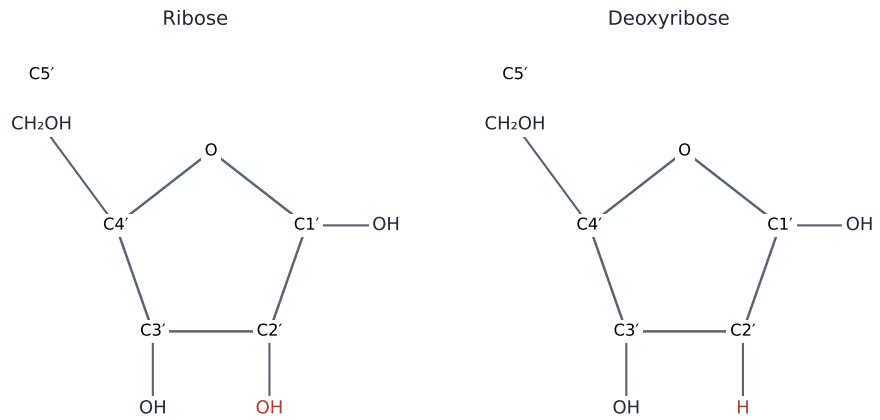
This handout teaches nucleotides, DNA and RNA structure, information storage, base pairing and the genetic code. Sections marked **AHL** are for Higher Level only. Prerequisites: covalent bonds, hydrogen bonds and condensation reactions. All sequence examples and numerical datasets below are original teaching examples, not real gene sequences or historical measurements.

Nucleotides: the units of nucleic acids

DNA is the genetic material of living organisms. RNA also has important cellular roles; some viruses have RNA genomes, but viruses are not classified as living organisms in this course. "DNA is genetic material" does not mean every mature cell retains a nucleus or that RNA never stores genetic information in a virus.

A **nucleotide** contains a phosphate group, a five-carbon pentose sugar and a nitrogen-containing base. Use a circle for phosphate, a pentagon for sugar and a rectangle for the base in a simplified diagram. Connect the base to the sugar and the phosphate to the sugar; the base is not inserted between sugar and phosphate in the backbone.

DNA contains deoxyribose; RNA contains ribose. In a sketch of the sugars, number the carbon positions and compare the 2′ position: ribose has an –OH group there, while deoxyribose has H in its place. "Deoxy" does not mean the sugar contains no oxygen. The prime mark distinguishes sugar-carbon numbering from other numbering in a nucleotide.



2′: OH in ribose; H in deoxyribose.
Other H and stereochemistry omitted.

The important comparison is the group at 2′; the diagram highlights connectivity and is not a full stereochemical formula.

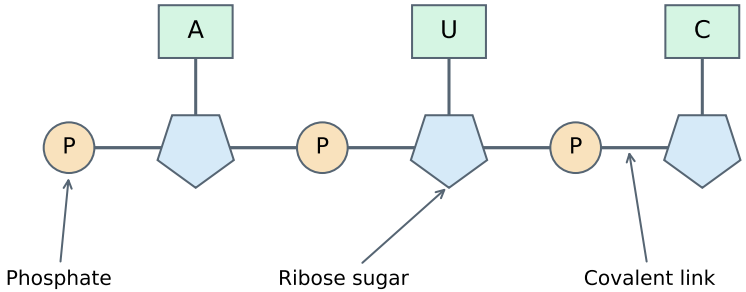
The DNA bases are adenine (A), thymine (T), cytosine (C) and guanine (G). RNA uses adenine, uracil (U), cytosine and guanine. A base name is not the name of a whole nucleotide: adenine alone has neither a sugar nor a phosphate.

Build a strand: a continuous covalent backbone

Nucleic acids are polymers of nucleotides. In a simplified condensation model, linking nucleotide units forms a sugar–phosphate connection with water released. The link is a covalent **phosphodiester bond**. Repeated sugar–phosphate links form the backbone, with bases projecting from the sugars. RNA can be represented as a single chain of these units.

When drawing an RNA polymer, preserve three relationships: a continuous alternating sugar–phosphate backbone, a base attached to each sugar, and no thymine in the RNA sequence. A chain is not held together mainly by hydrogen bonds between successive bases. Hydrolysis can break covalent links in a backbone; separating paired DNA strands is a different process.

For a simple linear condensation model, n separate mononucleotide units need $n - 1$ new links to make one chain. Thus six units need five links. This counting model is not a complete description of intracellular polymerization: cells use activated nucleotide substrates, and their reaction products must be considered separately. Do not assume one classroom condensation diagram describes every biochemical step.



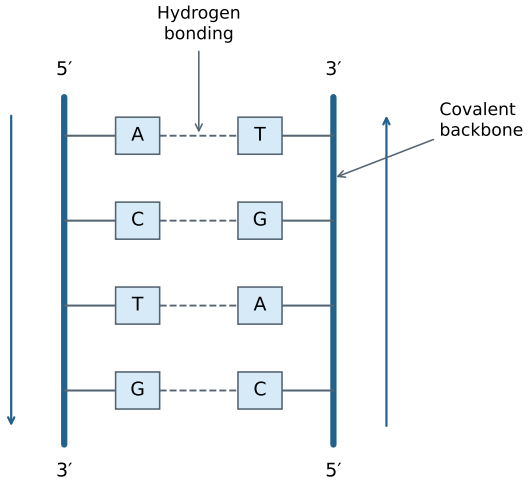
Continuous sugar–phosphate backbone.
Each base attaches to a sugar.

Bases attach to sugars; the strand is connected through covalent sugar–phosphate links.

DNA: complementary strands form a double helix

A DNA molecule normally contains two polynucleotide strands twisted into a double helix. Their sugar–phosphate backbones are on the outside and paired bases lie inside. The strands are **antiparallel**: they run in opposite directions. In a flat drawing, show opposing end labels or arrows rather than drawing both strands pointing the same way. A ladder sketch can represent the pairing without reproducing the helical twist.

Complementary base pairing means A pairs with T and C pairs with G in the standard DNA model. Hydrogen bonds form between the bases on different strands. Covalent bonds hold each backbone together. These two kinds of bond have different positions and roles; do not draw hydrogen bonds as the backbone links.



Reading across a pair gives its complement; reading along a strand gives a sequence.

For the invented top strand A–T–G–C–G–A, the paired bases in the aligned opposite strand are T–A–C–G–C–T. The matching is determined, but the sequence along a strand can vary. “Complementary” means matching by the pairing rule, not identical.

In ordinary double-stranded DNA, total A equals total T and total C equals total G because every pair contributes one of each matching base. This does **not** require A to equal G or every base to be 25%. It also does not require each individual strand to contain equal A and T. State whether a composition describes one strand or the entire double-stranded molecule.

■ Worked example: count both strands

Suppose an invented DNA molecule has 800 base pairs and 28% of its nucleotides contain adenine. It has $2 \times 800 = 1600$ nucleotides in total. A and T each make up 28%; together they are 56%. The remaining 44% is split equally between C and G, giving 22% each.

$$N_A = N_T = 0.28(1600) = 448,$$

$$N_C = N_G = 0.22(1600) = 352.$$

Check: $448 + 448 + 352 + 352 = 1600$. There are 448 A–T pairs and 352 C–G pairs, totaling 800 pairs. Mixing base pairs and individual nucleotides would give a factor-of-two error.

Compare DNA and RNA

Feature	DNA	RNA
Sugar	Deoxyribose	Ribose
Bases	A, T, C, G	A, U, C, G
Typical cellular arrangement	Two paired strands	One strand, often folded
Examples	Chromosomal DNA	mRNA, tRNA, rRNA

Messenger RNA carries sequence information used in protein synthesis. Transfer RNA helps match amino acids to codons, and ribosomal RNA is a structural and functional part of ribosomes. A single RNA strand can fold so that complementary regions pair; "single-stranded" does not mean it must remain a straight line. These are typical cellular comparisons, not a claim that every viral genome follows the same strand pattern.

Copy and express information by base pairing

Complementary pairing makes a template useful. During DNA replication, the strands can separate and each can guide the formation of a complementary strand. The sequence of an existing strand constrains the new sequence. Accurate copying also depends on enzymes and checking processes; complementarity does not make mutations impossible.

Transcription produces RNA using a DNA template. RNA nucleotides pair with the template: DNA A directs RNA U, DNA T directs RNA A, and C pairs with G. The coding DNA strand has the same base sequence as the RNA transcript in the corresponding region, with T in DNA in place of U in RNA. Always state which DNA strand you were given.

Translation uses the codon sequence in mRNA to assemble a polypeptide. A **codon** is a group of three mRNA bases. Transfer RNA helps implement complementary matching during translation. A codon table is normally read using mRNA codons, not an unexplained DNA triplet. Stop codons signal termination; they do not specify an additional amino acid.

For the invented coding DNA segment ATG–GAA–TTC–TAA, the mRNA segment is AUG–GAA–UUC–UAA. Using the stated meanings AUG = methionine, GAA = glutamate, UUC = phenylalanine and UAA = stop, the model peptide has three amino acids. The four codons do not give four amino acids.

In a simplified complete coding segment with 90 mRNA bases, including one terminal stop codon, there are $90/3 = 30$ codons and $30-1 = 29$ amino acids. This assumes the segment begins with the initiating codon, contains no intron or untranslated region, and is translated as stated. A whole gene's DNA length cannot automatically be divided by three: genes can include other sequences.

The genetic code is **degenerate**: more than one codon can specify the same amino acid. For example, GAA and GAG both specify glutamate. A substitution between them can be silent with respect to the amino-acid sequence. Inserting or deleting a number of bases not divisible by three within a translated region can shift the reading frame. A mutation's consequence depends on position and context; not every sequence change alters the protein.

Sequence diversity and shared ancestry

With four possible bases at each of n specified positions, the number of possible sequences on one defined strand is 4^n. Six positions give $4^6 = 4096$ sequences; twenty give $4^{20} = 1,099,511,627,776$. The count grows rapidly even though the alphabet is small. The complementary strand is determined by the first; it is not a second independent choice of four bases at each position.

A particular molecule has a finite length and information capacity. The enormous general capacity of DNA comes from varying sequence and length, not from a single finite molecule holding literally infinite information. Not every mathematically possible sequence forms a functioning gene or an organism with high fitness.

Across living organisms, codons generally have the same meanings. This strong conservation of the genetic code supports common ancestry: related information-processing systems are inherited and modified. Small known code variations do not erase the broad pattern. Distinguish a shared code from identical DNA sequences; different organisms can use the same code to specify different proteins.

AHL: Directionality and molecular organization

- **The two ends are chemically different**

The 5′ and 3′ labels refer to sugar-carbon positions. The backbone connects the 3′ position of one sugar through phosphate to the 5′ position of the next. A conventional sequence is written 5′ to 3′ unless the question states otherwise.

DNA polymerases extend a strand by adding to its 3′ end, so new DNA is made 5′ to 3′ while its template is read in the opposite direction. Transcription similarly makes RNA 5′ to 3′. During translation, the ribosome reads mRNA codons in the 5′ to 3′ direction. These are distinct processes with a common need to track sequence direction.

If the top DNA strand is 5′–ATGCGA–3′, the aligned complement is 3′–TACGCT–5′. Written independently 5′ to 3′, that lower strand is 5′–TCGCAT–3′. Complementing without reversing when asked for an independently written 5′-to-3′ strand is a common error.

Pair widths and packaging

Adenine and guanine are double-ring **purines**; cytosine and thymine are single-ring **pyrimidines**. RNA's uracil is also a pyrimidine. Each standard DNA pair combines one purine with one pyrimidine, helping maintain a consistent helix width across different sequences. A purine–purine pair would be too wide and a pyrimidine–pyrimidine pair too narrow in the ordinary helix geometry.

In a eukaryotic nucleosome, DNA wraps around a core of eight histone proteins. Linker DNA joins neighbouring nucleosomes; an additional linker histone associates near the DNA entry/exit and linker region. That additional histone is not a ninth member of the eight-protein core. Packaging organizes a long DNA molecule and influences access to its sequence.

Use a molecular viewer, such as the RCSB view of structure 4QLC, with an experimentally determined nucleosome structure to distinguish DNA from protein, rotate the structure and identify the wrapped DNA path. In 4QLC the linker-associated component is a globular H5 domain, not a complete additional core histone. A cross-section may hide parts of the molecule, so inspect more than one view. Record the structure identifier and representation settings. A diagram alone does not replace this three-dimensional inspection.

AHL: Evidence for DNA as genetic material

Hershey and Chase: track different components

Hershey and Chase investigated bacteriophage T2 infection of bacteria. In separate preparations, radioactive phosphorus-32 labelled DNA, while sulfur-35 labelled sulfur-containing phage proteins. DNA contains phosphorus in its backbone but no sulfur; proteins containing cysteine or methionine can incorporate sulfur. Keep the two labels and preparations distinct.

After phages attached and infection began, blending detached much of the external phage coat. Centrifugation separated a bacterial pellet from the supernatant containing detached material. Much of the DNA-associated label remained with the bacteria, whereas much of the protein-associated label was outside them. DNA-associated material also contributed to progeny phage. This supported DNA, rather than the bulk of the external protein coat, as the material directing phage reproduction.

The experiment did not mean a phage contains no protein, that the protein coat has no function in attachment, or that every labelled molecule crossed the bacterial membrane. Radioactive tracing allowed components to be followed through a process; it was evidence interpreted with controls and the experimental separation, not a photograph of a gene controlling a cell. This is a historical explanation, not a school procedure involving radioisotopes.

Chargaff: test a prediction with composition data

A rigid repeating tetranucleotide model containing equal numbers of A, T, C and G predicts about 25% of each base regardless of the organism. Chargaff's compositional findings showed that base proportions differed across biological sources, while matching relationships between A/T and G/C were evident. These observations contradicted the equal-four-base prediction and helped constrain later structural explanations.

For an invented comparison, sample X has A = T = 30%, C = G = 20%; sample Y has A = T = 21%, C = G = 29%. Both satisfy complementary-pair totals but neither has four equal proportions. Even a single reliable counterexample can falsify a universal equal-proportion claim. A sample with four 25% values would fit that prediction, but would not prove that every DNA molecule consists of a repeating four-base unit. Composition also does not reveal the order of the bases.

Practice

All sequences and numerical data are invented. Use the stated simplified models.

1. Draw a nucleotide using the standard three shapes. Name each component and show where the base attaches.
2. Sketch ribose and deoxyribose to show their key difference. Explain why deoxyribose does not mean "no oxygen".
3. Draw three RNA nucleotide units as one chain. Identify the covalent backbone, attached bases and the number of new links in the simple condensation model.
4. Compare DNA and RNA in sugar, bases and usual cellular strand arrangement. Name two functional types of RNA and explain why folded RNA can still be one strand.
5. For the aligned DNA strand A–T–G–C–G–A, write the six paired bases. Distinguish complementary pairing from an identical sequence.
6. Double-stranded DNA has 800 base pairs and 28% adenine among all nucleotides. Calculate the total number of each base and check the sum.
7. For coding DNA ATG–GAA–TTC–TAA, give the mRNA sequence and peptide using the codon meanings supplied above. Why is the coding strand not the template strand?
8. A complete mRNA coding segment has 90 bases including one terminal stop codon. Find peptide length. Repeat for 63 bases, and state the assumptions that make these calculations valid.
9. GAA and GAG both encode glutamate. Explain degeneracy and a possible silent substitution. Contrast this with inserting one base inside the translated region.
10. Calculate the number of possible six-base sequences on one defined strand. Explain why knowing its complementary strand does not double the number of independent positions.

11. Explain how conserved codon meanings support common ancestry. Distinguish this from a claim that all organisms have the same base sequence.
12. **AHL.** Write the complement of 5′–ATGCGA–3′ first aligned with it, then independently 5′ to 3′. State synthesis and mRNA-reading directions, and explain how purine–pyrimidine pairing helps helix geometry.
13. **AHL.** Describe a nucleosome's eight-protein core, wrapped DNA and linker region. What would you inspect in a molecular viewer to distinguish a ninth core protein from a linker-associated protein?
14. **AHL.** Explain what the Hershey–Chase label separation supported. Then use invented samples X and Y above to test an equal-four-base hypothesis. State one limitation of each type of evidence.

Worked answers

1. A phosphate circle links to a pentose pentagon; a base rectangle attaches to the sugar. One complete three-component unit is a nucleotide, whereas the base alone is only one component.
2. At 2′, ribose has –OH and deoxyribose has H instead. Both still contain oxygen elsewhere, including in the ring. Label the compared carbon rather than replacing every oxygen in the sugar.
3. Draw an alternating sugar–phosphate chain with a base attached to each of three sugars. Its backbone links are covalent phosphodiester bonds. Three separate units need $3 - 1 = 2$ new links in this stated linear condensation model.
4. DNA: deoxyribose, A/T/C/G, typically two paired strands. RNA: ribose, A/U/C/G, typically one strand. mRNA carries coding information; tRNA helps match amino acids to codons, or rRNA forms part of ribosomes. Intramolecular pairing can fold one continuous RNA chain without adding a second chain.
5. T–A–C–G–C–T, aligned opposite the given sequence. Each base follows A–T or C–G pairing; complementarity is a specific matching relationship rather than copying the same letters across the pair.
6. $N = 2(800) = 1600$ nucleotides. $N_A = N_T = 0.28(1600) = 448$. C and G each make up $(100 - 56)/2 = 22\%$, so $N_C = N_G = 352$. The sum is 1600; there are 800 pairs.
7. AUG–GAA–UUC–UAA gives methionine–glutamate–phenylalanine, then termination: three amino acids. The template pairs with the RNA; the coding strand matches the RNA's corresponding sequence except T replaces U.
8. $L = 90/3 - 1 = 29$ amino acids; for 63 bases, $L = 63/3 - 1 = 20$. Assume one complete translated coding region beginning at initiation, one terminal stop and no untranslated region or intron included in the count.
9. Degeneracy means multiple codons can specify the same amino acid. GAA → GAG preserves glutamate and can leave the amino-acid sequence unchanged. A one-base insertion within the translated sequence can shift subsequent codon grouping, though the exact effect depends on position and sequence.
10. $N = 4^6 = 4096$. There are six independent positions on the defined strand; each paired base is then constrained. A finite molecule has finite capacity even though varying sequence and length allow an enormous range of information.
11. Shared codon meanings are a conserved feature of the information-processing system, consistent with descent from shared ancestors. Sequences can differ while the same codon-to-amino-acid mapping is retained. The inference comes from a broad pattern, not a requirement that every code variant be identical.
12. **AHL.** Aligned: 3′–TACGCT–5′; independently: 5′–TCGCAT–3′. DNA/RNA synthesis extends 5′ to 3′ by adding to the 3′ end; a ribosome reads mRNA 5′ to 3′. One double-ring purine with one single-ring pyrimidine helps maintain standard pair width.
13. **AHL.** DNA wraps around a core of eight histones; linker DNA connects cores, with an additional histone associated near entry/exit and the linker. Rotate a viewer model and distinguish the core's protein chains from the separate linker-associated chain; the latter is not a ninth core member.
14. **AHL.** DNA-associated phosphorus label remained largely with infected bacteria and contributed to progeny; sulfur-labelled coat material was largely outside, supporting DNA's reproductive-information role in this system. Separation was not perfect and protein still had roles. X and Y depart from four equal 25% proportions while preserving A/T and C/G matches, contradicting the rigid equal-proportion prediction. Composition alone gives no base order, and one matching sample could not establish a universal repeating sequence.

Exam tips

- Identify sugar, phosphate and base separately. Distinguish backbone covalent bonds from hydrogen bonds across a pair.
- Count base pairs and total nucleotides carefully; state the strand or whole-molecule conditions behind a percentage calculation.
- Identify coding DNA, template DNA and mRNA before using codons. A stop codon adds no amino acid; a whole gene is not automatically all coding sequence.
- **AHL.** Keep 5′/3′ ends explicit, and connect an experiment's observation to the precise hypothesis it supports or contradicts.

References: IB A1.2; Hershey and Chase, original research; Chargaff, original research.