

A.3.2 · Classification and cladistics

International Baccalaureate · IB Diploma · Biology · HL

Scope and prerequisites

A dolphin and a shark have similar swimming shapes. A mushroom and a green plant both stay in one place. Which similarities help us identify an organism, and which reveal its ancestry? A useful classification must do more than collect things that look alike.

This is the HL-only A3.2 unit; there is no SL counterpart. It covers the need for classification, the limits of fixed ranks, clades, molecular evidence and clocks, constructing and reading cladograms, revising plant classification and the three domains. Prerequisites: species and genomes from A3.1 and base sequences from A1.2. All numerical datasets and letter-labelled organisms below are invented teaching examples, not records from real investigations.

AHL: Why classify such diverse organisms?

Classification arranges organisms into groups so we can identify, compare and study them. A named group is a taxon. A common system lets researchers connect observations made in different places, find information about a newly identified organism and ask whether a feature is shared across a group. Scientific names reduce the ambiguity of common names.

The traditional nested ranks are kingdom, phylum, class, order, family, genus and species. Domain is placed above kingdom in the three-domain classification. A species belongs within a genus, that genus within a family, and so on. A genus contains one or more species; species in one family need not belong to the same genus. The **binomial system** gives genus followed by specific epithet, for example *Homo sapiens*. The epithet alone does not name the species.

The ranks help organize information, but a rank is not a fixed amount of evolutionary time or genetic difference. Two families in different lineages need not be equally old or equally diverse. Evolution produces branching at many levels, without a natural rule that exactly seven named ranks must fit every branch. Assigning the same rank therefore does not establish that two groups diverged at the same time.

Traditional classifications sometimes emphasize a useful visible trait while excluding some descendants of the same ancestor. For example, a traditional grouping of reptiles that excludes birds does not include all descendants of the reptile group’s common ancestor. Birds belong within the reptile lineage in a classification based on ancestry. This does not mean birds descended from today’s crocodiles, or that birds lack their own distinctive features.

AHL: Clades connect classification to ancestry

Cladistics uses evidence of common ancestry to group organisms. A clade includes a common ancestor and **all** its descendants. Clades can be nested and can be named without assigning each one a traditional rank. This avoids forcing every evolutionary split into an arbitrary rank system.

A classification that follows evolutionary relationships supports predictions. If several members share a feature inherited from their common ancestor, we can investigate whether other members have it too. For example, a group sharing an enzyme pathway suggests a useful starting point for studying a newly identified member. Such predictions are testable: later loss, modification or incomplete knowledge can produce exceptions. A clade does not require every descendant to look identical.

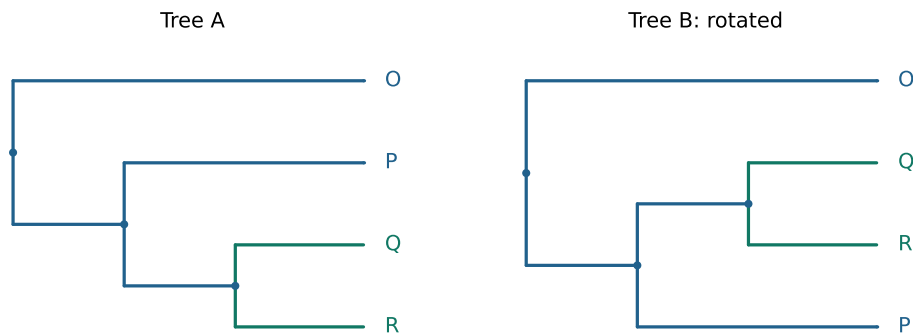
A **shared derived character** is a character state inherited from the common ancestor of a group and different from the ancestral state used for comparison. It supports a particular branching hypothesis. An ancestral state shared much more widely may tell us little about the closest relationships inside the group. A suitable outgroup, outside the group being studied, helps infer which state is ancestral; the choice must itself have evidence.

Similarity can also arise independently through convergent evolution. Similar swimming outlines in dolphins and sharks are useful in water, but do not place them together as the closest pair among mammals and fish. Study multiple characters and distinguish a shared inherited structure from a similar function. Reversals and losses can also obscure history. One striking trait should not override all other evidence automatically.

AHL: Read the branching, not the positions on the page

A **cladogram** is a hypothesis about branching relationships. Trace connections rather than counting the gaps between tips.

Part	What it represents	What it does not establish by itself
Root	The ancestral end of a rooted tree	The identity of a directly observed first organism
Internal node	A hypothetical common ancestor at a split	A named modern species at a different tip
Terminal branch	A lineage leading to a displayed terminal taxon	A ladder of improvement from lower to higher life
Branch length	Time or amount of change only if an appropriate scale is given	An age measured from the drawing’s unscaled length



Invented taxa. Same nodes, same clades.
No branch-length scale is given.

Both trees have the same relationships. The coloured Q–R branches identify one complete clade; their position on the page changes.

In the illustrated tree, Q and R share a more recent common ancestor with each other than either shares with P. The clade beginning at their shared node contains Q and R. The next deeper clade contains P, Q and R. Grouping P and Q while excluding R omits a descendant of their common ancestor, so that set is not a complete clade on this tree.

Rotating branches around a node preserves the connections. It may move two tips next to each other without making them sister taxa. P is not an ancestor of Q merely because P's branch leaves a deeper node; P represents its own descendant lineage. Modern terminal taxa have all continued evolving since their shared ancestors.

An unscaled cladogram does not say whether a long-looking branch took longer or accumulated more mutations. A dated tree or a phylogram can supply such information, but the key must say what the scale measures. Identify the root before using “more recent”; an unrooted diagram alone does not specify the direction of ancestry.

AHL: Build a tree from sequence evidence

Base sequences of homologous genes, or amino acid sequences of homologous proteins, allow many comparable characters to be studied. Molecular evidence can reveal relationships that visible form hides, especially among microorganisms. The comparison still needs care: use corresponding regions, align equivalent positions, check sequence quality and avoid mixing unrelated genes.

The following six aligned DNA sites are invented. O is the outgroup, and for this simple exercise its A states are taken as ancestral. A column is a character; its bases are states. A gap in a real alignment must be handled by a stated method, not silently counted as whichever base is convenient.

Taxon	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6
O	A	A	A	A	A	A
P	G	G	A	A	A	A
Q	G	G	C	C	A	A
R	G	G	C	C	T	T

Sites 1 and 2 support grouping P, Q and R relative to O. Sites 3 and 4 support Q with R. Sites 5 and 6 distinguish R, but a state unique to one terminal taxon does not by itself choose its sister group. The same reasoning works with amino acid states, although different DNA sequences can encode the same amino acid sequence.

Worked example: choose by parsimony

Parsimony compares the minimum number of changes needed on candidate trees. Here assume equal cost per base substitution and count each column independently.

- Tree I has the rooted branching O, then P, then the Q–R pair. Each of sites 1–2 needs one A-to-G change on the P–Q–R stem. Each of sites 3–4 needs one A-to-C change on the Q–R stem. Each of sites 5–6 needs one A-to-T change on R. Total: $2 + 2 + 2 = 6$ changes.
- Tree II has O, then Q, then the P–R pair. Sites 1–2 still need two changes in total. For each of sites 3–4, Q and R share C but are separated by P's A state: the minimum is two changes per site, by independent gains or a gain plus reversal. Sites 5–6 still need one change each. Total: $2 + 4 + 2 = 8$ changes.

Tree I: Q + R		Tree II: P + R	
Sites 1–2	2	Sites 1–2	2
Sites 3–4	2	Sites 3–4	4
Sites 5–6	2	Sites 5–6	2
Total changes	6	Total changes	8

Invented alignment; O states taken as ancestral.
Minimum substitutions; equal cost per site.

The shorter explanation is selected under the stated equal-cost model. Six changes do not mean six observed mutation events in a real history.

Tree I is more parsimonious for these data. This is a reasoned choice among hypotheses,

not proof that evolution always takes the fewest possible steps. Multiple substitutions at one site, convergence and reversals can conceal changes. Different genes may have different histories, and a short alignment may supply weak or conflicting support. Add independent genes or characters and ask whether the grouping persists; do not repeat the same six sites and call that independent confirmation.

A percentage difference is useful but is not itself a complete tree-building method. P–Q and Q–R each differ at two of six sites here, yet only Q–R share the derived C states at sites 3–4. A tied pairwise distance cannot select between these relationships without further reasoning. Real analyses may use models beyond equal-cost parsimony.

AHL: A molecular clock estimates a date

Over generations, sequence changes can accumulate in diverging lineages. A molecular clock uses an estimated rate to relate divergence to elapsed time. The observed fraction of different sites is

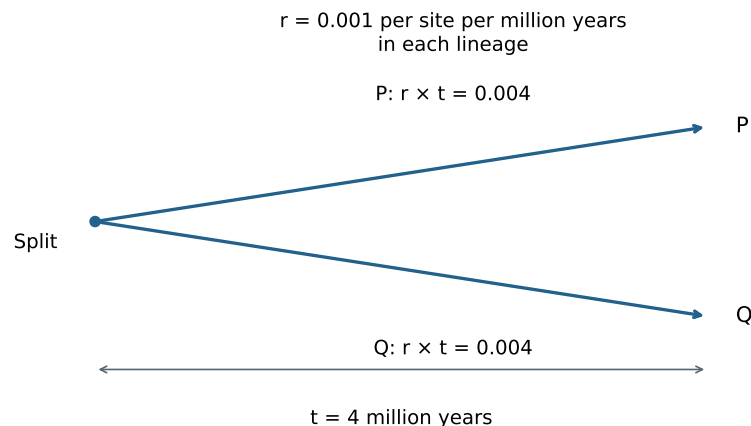
$$d = \frac{\text{number of differing aligned sites}}{\text{number of sites compared}}.$$

Observed difference is not automatically the total number of historical substitutions. A site can change more than once or change back. For distant lineages, simply counting present differences may underestimate the changes that occurred. Select a suitable region and model, and calibrate rates using independent evidence where possible.

■ Worked example: keep the rate definition explicit

In an invented comparison, two lineages differ at 8 of 1,000 aligned sites. Assume a constant substitution rate $r = 0.001$ substitutions per site per million years **in each lineage**, equal in both lineages, with negligible repeated changes. Both lineages accumulate differences after their split, so

$$d = \frac{8}{1000} = 0.008, \quad d \approx 2rt, \quad t = \frac{0.008}{2(0.001)} = 4 \text{ million years.}$$



Invented model: equal constant rates; negligible repeated changes.
Pairwise difference $\approx 0.004 + 0.004 = 0.008$.

Add the two lineage contributions. A rate defined for pairwise divergence already includes both; do not double it again.

If the stated rate were instead a **pairwise** divergence rate of 0.002 per site per million years, use $t = d/0.002 = 4$ million years. Check the definition before substituting. A percentage and a fraction must also be consistent: 0.8% is 0.008, not 0.8.

A clock gives an estimate, not an exact historical timestamp. Rates can vary with generation time, population processes, selective pressures and sequence function. Errors in alignment, sampling and calibration add uncertainty. If the per-lineage rate in this example could be 0.0008–0.0012, the same model gives about 5.0–3.3 million years. A faster rate implies less time to produce the same difference. Without a rate or calibration, a difference percentage alone supplies no numerical date.

AHL: Classification changes when evidence changes

A cladistic analysis can test whether an existing named group contains one ancestor and all its descendants. A traditional family may prove to contain several separate lineages, or a group excluded because it looks different may belong inside the same clade. Revision changes the classification to better fit the evidence; it does not make the organisms themselves suddenly evolve into different organisms.

The figwort family, Scrophulariaceae in its older broad sense, provides a real example. A published 2001 study compared sequences from three plastid genes and found that the broad grouping included at least five distinct monophyletic groups. Several traditional family boundaries therefore needed revision. You do not need to memorize all the transferred genera to explain the scientific reasoning. Similar-looking flowers alone had not reliably identified one complete ancestry group. The study's abstract and bibliographic record identify the actual analysis; the invented trees in this handout

are not its data.

To evaluate a proposed revision, ask which organisms and genes were sampled, how the tree was inferred, and whether additional evidence agrees. One unusual sequence might reflect an error or a particular gene's history. Agreement across independent evidence is stronger. A well-supported new tree can falsify a specific claim that an old family is one complete clade. It does not falsify the broader conclusion that organisms evolve. Classification is a scientific framework open to improvement, rather than a list that must remain unchanged.

AHL: Three domains from ribosomal RNA evidence

Bacteria and archaea are both prokaryotic in cell organization: neither has a membrane-bound nucleus. This shared absence does not establish that they form one uniform evolutionary group. Molecular comparisons of ribosomal RNA revealed major differences hidden by that simple structural description.

Ribosomal RNA is useful because cellular organisms share ribosomes, and corresponding rRNA regions can be compared across very diverse groups. Conserved regions help alignment; variable positions supply information about divergence. Researchers compare many positions, rather than deciding from the presence of one base. rRNA is a structural and functional RNA component of ribosomes, not the messenger RNA encoding one protein.

The three-domain classification distinguishes Bacteria, Archaea and Eukarya (also written Eucarya). Domain lies above kingdom; animals, plants and fungi belong within Eukarya, not in three separate domains. Archaea are not simply bacteria living in extreme habitats: their molecular distinctions matter, and not all archaea are extremophiles. Viruses have no ribosomes of their own and are not assigned to these three domains of cellular life.

Historically, Woese and Fox's 1977 rRNA work identified three major lineages; Woese, Kandler and Wheelis formally proposed the domain rank and names in 1990. Distinguish the underlying discovery from the later naming proposal. The three-domain scheme demonstrates how molecular evidence can transform a familiar classification. Deep relationships continue to be investigated; three named domains should not be read as a claim that every detail of life's rooted history is settled.

Practice

Use the invented six-site alignment and trees above where specified. State assumptions in calculations.

1. Give three reasons for classification. Write the traditional ranks from kingdom to species and explain why two families need not be equally old.

2. Define a clade. On the tree O, then P, then Q–R, decide whether each set is a complete clade: Q and R; P and Q; P, Q and R. Explain using their common ancestors.
3. Explain root, internal node and terminal branch. Does rotating Q–R around their shared node alter their relationship? Does the longest unscaled branch identify the oldest taxon?
4. Distinguish an ancestral character, a shared derived character and a convergent similarity. Explain how an appropriate outgroup helps a character matrix.
5. For the six-site alignment, identify the sites supporting P–Q–R, those supporting Q–R, and those unique to R. Why do the last sites not identify R's sister taxon alone?
6. Recalculate the minimum equal-cost scores for trees I and II. Explain why two changes per site are required for sites 3–4 on tree II. Which tree is preferred?
7. Calculate P–Q, Q–R and P–R percentage differences across the six sites. Why is choosing one tied shortest distance insufficient to settle the tree?
8. Two invented alignments have 8 differences in 50 sites and 12 in 60 sites. Calculate both percentages. Can these alone establish which pair split earlier? Explain.
9. Two lineages differ at 12 of 600 sites. Each has a constant rate of 0.0025 substitutions per site per million years. Estimate their divergence time, stating the simple model. What changes if 0.005 is instead given as the pairwise rate?
10. For $d = 0.008$, calculate the time range when each lineage's rate is between 0.0008 and 0.0012 per site per million years. Explain the direction of the relationship and give two limitations of clock estimates.
11. An old plant family appears in three separated parts of a well-supported new tree. Explain how to test the old grouping and revise the classification. Distinguish revising a family from rejecting evolution.
12. Name the three domains, locate animals and fungi, and explain why “both lack a nucleus” is insufficient to combine archaea and bacteria. Why is rRNA suitable evidence, and why are viruses excluded from this scheme?
13. Evaluate: “Molecular data always give the true tree; a shorter tree proves the exact history.” Propose two genuinely independent checks, and distinguish the 1977 rRNA finding from the 1990 domain proposal.

■ Worked answers

1. Classification supports identification and communication, organizes comparisons, and helps predict or investigate shared features. Kingdom → phylum → class → order → family → genus → species. Ranks are conventions, not units of time or divergence; evolutionary splitting does not occur at just seven fixed levels.
2. A clade includes an ancestor and all its descendants. Q–R is complete; P–Q is incomplete because their common ancestor also has descendant R; P–Q–R is complete. Include the ancestor concept even when only terminal labels are listed.

3. Root: ancestral end of a rooted tree; internal node: hypothetical common ancestor at a split; terminal branch: lineage to a terminal taxon. Rotation leaves the connections unchanged. Unscaled branch length gives no date, and one modern tip is not automatically another's ancestor.
4. An ancestral state predates the group of interest; a shared derived state originated in its common ancestor relative to the comparison. Convergence produces similar states independently. An evidence-based outgroup helps infer ancestral states, but is not assumed to be a surviving direct ancestor.
5. Sites 1–2 support P–Q–R; sites 3–4 support Q–R; sites 5–6 uniquely distinguish R. A state present only in R can be placed on its terminal branch on either candidate tree, so it does not select R's sister group.
6. Tree I: $2 + 2 + 2 = 6$; tree II: $2 + 4 + 2 = 8$. On II the C-bearing Q and R do not form a clade excluding A-bearing P. At least independent C gains, or one gain and one reversal, are needed per affected site. Prefer I under equal-cost parsimony, without treating its score as an observed event count.
7. P–Q: $2/6 \times 100 \approx 33.3\%$; Q–R: also 33.3% ; P–R: $4/6 \times 100 \approx 66.7\%$. The equal shortest distances do not uniquely choose a topology. The distribution of derived states, including the outgroup comparison, supports Q–R in this dataset.
8. $8/50 \times 100 = 16\%$ and $12/60 \times 100 = 20\%$. The larger percentage does not alone prove an earlier split: regions, rates, selection and repeated substitutions can differ. A date needs calibration or a rate model.
9. $d = 12/600 = 0.020$; assuming equal constant per-lineage rates and negligible repeated changes, $t = d/(2r) = 0.020/0.005 = 4$ million years. A pairwise rate of 0.005 already includes both contributions: $t = 0.020/0.005 = 4$ million years, with no extra factor of two.
10. At 0.0008, $t = 0.008/0.0016 = 5.0$ million years; at 0.0012, $t = 0.008/0.0024 \approx 3.3$ million years. The range is about 3.3–5.0 million years. A faster rate requires less time. Variable rates and repeated substitutions are two limitations; alignment, sampling and calibration uncertainty also matter.
11. Locate the group's common ancestor and test whether the named family includes all and only that ancestor's descendant lineages. Check sampling, sequence quality and independent genes or characters. Revise family boundaries to follow supported clades, which may require moving species. This rejects a particular grouping, not common descent itself.
12. Bacteria, Archaea and Eukarya/Eucarya; animals and fungi are in Eukarya. A shared absence of a nucleus hides major molecular differences. rRNA has homologous conserved and variable regions and occurs in all cellular life. Viruses lack their own ribosomes, so this cellular classification does not place them in the three domains.
13. Sequence quality, alignment, short samples, convergence, reversals and gene-specific histories limit inference. Parsimony chooses fewer required changes under a model, not a proven history. Test different independent genes and a separately

observed character dataset, rather than reuse the original sites. The 1977 work identified major lineages using rRNA; the formal domain proposal and names followed in 1990.

Exam tips

Name the common ancestor and all its descendants when identifying a clade. Trace nodes rather than tip spacing. Count changes site by site under the stated model. Write a difference as a fraction before using a rate, and check whether that rate is per lineage or pairwise. Explain classification revisions through evidence and a specific rejected grouping.

References: IB Biology guide, A3.2; 1977 rRNA study; 1990 domain proposal. The figwort study is linked in its teaching section.