Classification and cladistics

A3.2 · HL only

Which similarities reveal ancestry?

Dolphin and shark: similar shape, different ancestry

A useful classification connects observations, identification and evolutionary relationships.

A3.2 is HL only. We will read trees, compare sequence hypotheses, estimate dates and evaluate classification revisions.

Letter-labelled organisms and numerical data are invented. Named historical studies are linked in the handout.

Classification makes observations usable across studies

A named taxon connects information

Researchers can identify organisms, communicate consistently and compare observations from different places.

Grouping also suggests testable predictions about shared features of a newly identified member.

A scientific name reduces common-name ambiguity. A binomial contains both genus and specific epithet.

Traditional ranks organize nested groups

Domain lies above kingdom in the three-domain scheme

Kingdom → phylum → class → order → family → genus → species.

A genus can contain several species; one family can contain several genera.

A family rank is not a unit of age or genetic difference. Two families need not have diverged equally long ago.

A useful visible grouping can omit descendants

Test the boundaries against ancestry

A traditional reptile grouping that excludes birds omits descendants of its common ancestor.

An ancestry-based reptile clade includes birds. Birds did not descend from today's crocodiles.

Evolution branches at many levels; a fixed set of ranks does not naturally measure every split.

A clade includes an ancestor and all its descendants

Unranked clades avoid forcing every split into a rank

Clades nest within larger clades. Descendants can change, lose traits or develop distinctive features.

Shared ancestry can guide predictions about a feature, such as an enzyme pathway, in another member.

Predictions need testing: a shared history is not a promise that every descendant is identical.

Distinguish three explanations for similarity

Evidence	Interpretation
Shared ancestral state	May be too widespread to resolve relationships within the focal group
Shared derived state	Supports a common ancestor relative to the ancestral comparison
Convergent similarity	Similar function or form evolved independently; not evidence of a uniquely shared recent ancestor

An outgroup helps infer the ancestral state

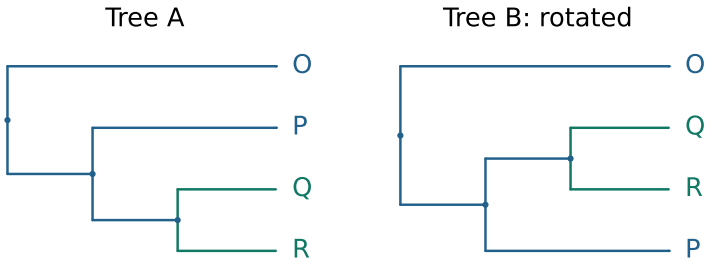
Choose the comparison using evidence

Record character states in a matrix before choosing the tree. A column is a character; its entries are states.

A suitable outgroup lies outside the focal group and helps infer which state is ancestral.

It is not assumed to be a surviving direct ancestor. Reversals and losses can complicate interpretation.

Rotate branches without changing the Q–R clade



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

Rotation preserves clades. Unscaled branch lengths give no dates.

Name the parts before reading a relationship

Part	Meaning
Root	The ancestral end of a rooted tree
Internal node	A hypothetical common ancestor at a split
Terminal branch	A lineage leading to a displayed taxon
Branch length	Time or change only if an appropriate scale is stated

Test three sets against their common ancestor

1. **Q and R.** Their common ancestor has just Q and R as displayed descendants: a complete clade.
2. **P and Q.** Their common ancestor also has descendant R: excluding R makes the set incomplete.
3. **P, Q and R.** All displayed descendants of their shared node are included: a complete clade.

Decision

P is a terminal lineage, not automatically an ancestor of Q. An unrooted diagram alone gives no ancestral direction.

Align homologous regions before comparing states

Base or amino acid sequences provide many characters

Use corresponding genes or proteins and align equivalent positions. Check sequence quality and gap handling.

Molecular evidence can reveal relationships hidden by visible form, especially among microorganisms.

Different DNA sequences can encode the same amino acid sequence. Neither data type is automatically error-free.

Read six invented aligned sites

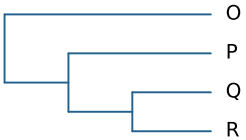
Taxon	Sites 1–2	Sites 3–4	Sites 5–6
O	AA	AA	AA
P	GG	AA	AA
Q	GG	CC	AA
R	GG	CC	TT

Identify which site patterns support a grouping

1. **Sites 1–2.** G is shared by P, Q and R relative to O's A: support the P–Q–R group.
2. **Sites 3–4.** C is shared by Q and R relative to A in O and P: support Q–R.
3. **Sites 5–6.** T is unique to R. It can mark R's terminal branch but does not alone select its sister group.

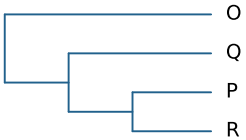
Compare the minimum changes on two hypotheses

Tree I: Q + R



Sites 1–2	2
Sites 3–4	2
Sites 5–6	2
Total changes	**6**

Tree II: P + R



Sites 1–2	2
Sites 3–4	4
Sites 5–6	2
Total changes	**8**

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

With equal substitution costs, tree I requires six changes; tree II requires eight.

exercises.lol exercises.com.cn

Explain why tree II costs two extra changes

1. **Shared G sites.** Sites 1–2 each need one change: two in total on either tree.
2. **Shared C sites.** Tree I: one per site. Tree II separates Q and R around P: two per site, by independent gains or a gain plus reversal.
3. **Unique T sites.** Sites 5–6 each need one terminal change. Totals: I $= 2 + 2 + 2 = 6$; II $= 2 + 4 + 2 = 8$.

Decision

Parsimony prefers I under this model. The score is a minimum explanation, not a count of directly observed historical events.

exercises.lol exercises.com.cn

A tied distance does not uniquely choose the tree

1. **P–Q.** Two differing sites out of six: $2/6 \times 100 \approx 33.3\%$.
2. **Q–R.** Also two out of six: 33.3%. P–R differs at four sites: about 66.7%.
3. **Interpretation.** The equal shortest pairwise distances do not resolve the tie. Shared derived states and the outgroup support Q–R.

A short parsimonious tree is still a hypothesis

Test with independent data

Repeated substitutions, reversals and convergence can hide historical changes. Short samples can give weak or conflicting support.

A gene may have a different history from other genomic regions. Additional genes and morphological characters can test agreement.

Reusing the same sites is not independent confirmation. Different inference models can select different hypotheses.

Difference is an observed quantity; time needs a model

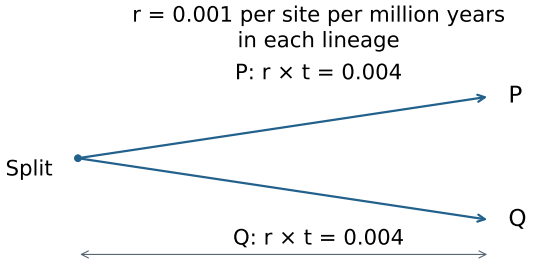
Use a calibrated rate and compatible units

Observed difference d = differing aligned sites / sites compared. It is a fraction, not a date.

Present states can hide repeated substitutions or reversals. Raw differences can underestimate historical changes.

Check alignment, region, rate definition and calibration. Unscaled branch length supplies none of these automatically.

Add the two lineages' accumulated differences



Invented model: equal constant rates; negligible repeated changes.

Pairwise difference $\approx 0.004 + 0.004 = 0.008$.

Two lineages contribute: pairwise divergence $\approx 2rt$.

Estimate time with a per-lineage rate

1. **Invented data.** Eight differences among 1,000 sites give $d = 8/1000 = 0.008$.
2. **Model.** Each lineage has constant $r = 0.001$ per site per million years; repeated changes are negligible. Then $d \approx 2rt$.
3. **Calculate.** $t = 0.008/[2(0.001)] = 4$ million years.

Decision

0.8% is 0.008. Equal constant rates and negligible repeated changes are assumptions, not conclusions from the eight differences.

A pairwise rate already includes both branches

1. **Rate definition.** For the same invented alignment, d = 0.008. Pairwise divergence rate = 0.002 per site per million years.
2. **Calculate.** $t = d/0.002 = 0.008/0.002 = 4$ million years.
3. **Check.** Do not multiply a pairwise rate by two again. Read what the rate measures before using a formula.

Turn rate uncertainty into a time range

1. **Slower rate.** For invented data with $d = 0.008$, each lineage has $r = 0.0008$ per site per million years. Then $t = 0.008/0.0016 = 5.0$ million years.
2. **Faster rate.** At $r = 0.0012$ per site per million years in each lineage, $t = 0.008/0.0024 \approx 3.3$ million years.
3. **Interpret.** About 3.3–5.0 million years under the simple model. Faster change requires less time for the same difference.

Molecular clocks estimate rather than certify dates

Biology and measurement both contribute uncertainty

Rates can differ with generation time, population processes, selective pressures and sequence function.

Calibration, alignment and sampling add uncertainty; repeated changes can be hidden by matching present states.

A larger observed difference does not alone prove an earlier split when rates or regions differ.

exercises.lol exercises.com.cn

An old family is a testable ancestry claim

Find its common ancestor and all the descendants

A well-supported tree may scatter the old family among separate lineages or place an excluded group inside its clade.

Check organisms sampled, genes, quality and independent evidence before changing the boundaries.

Revising a family changes the classification. It does not cause organisms to evolve instantly or reject common descent.

The broad figwort family did not form one complete clade

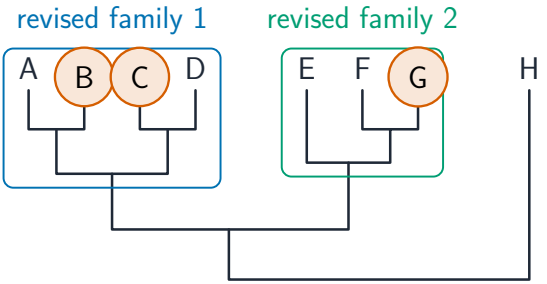
A real study; our letter-labelled trees are invented

A 2001 study of broad Scrophulariaceae compared three plastid genes and identified at least five distinct monophyletic groups.

Traditional family boundaries therefore needed revision. Similar-looking flowers had not reliably identified a single complete ancestry group.

Explain the evidence and revision; memorizing every transferred genus is unnecessary. The study is linked in the handout.

Revising a family



old family by look:
B, C and G
each revised family:
one ancestor and
all its descendants

The revised groups each enclose one complete clade

A shared absence of a nucleus hides molecular differences

rRNA permits comparison across cellular life

Bacteria and archaea are prokaryotic in organization, but their rRNA sequences reveal major distinctions.

All cellular organisms have ribosomes. Conserved rRNA regions help alignment; variable positions supply relationship evidence.

rRNA is a ribosomal component, not messenger RNA for one protein. Compare many homologous positions.

exercises.lol exercises.com.cn

Place organisms within the three-domain scheme

Domain	**Interpretation**
Bacteria	One major domain of cellular life
Archaea	Molecularly distinct from Bacteria; not restricted to extreme habitats
Eukarya / Eucarya	Includes animals, plants and fungi; those kingdoms are not separate domains

Separate a discovery from the later naming proposal

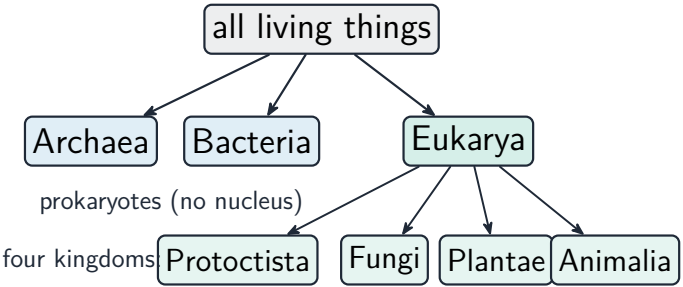
1977 lineage evidence; 1990 formal domain proposal

Woese and Fox's 1977 rRNA work identified three major lineages. The formal domain rank and names followed in 1990 with Kandler and Wheelis.

Viruses have no ribosomes of their own and are outside these three domains of cellular life.

Molecular evidence transformed classification; the named scheme does not settle every detail of life's rooted history.

Three domains



Ribosomal RNA evidence revealed the domains

Try a clock problem and an ancestry decision

State the model before calculating or grouping

Two lineages differ at 12 of 600 sites. Each has a constant rate of 0.0025 per site per million years. Estimate time under the simple model.

Would using a pairwise rate of 0.005 change the answer?

On the rooted O, then P, then Q–R tree, explain why P–Q alone is not a complete clade.

Reveal: rate definition and complete descendants matter

1. **Fraction.** $d = 12/600 = 0.020$.
2. **Time.** $t = 0.020/[2(0.0025)] = 4$ million years. A pairwise rate of 0.005 gives the same result without another factor of two.
3. **Clade.** The common ancestor of P and Q also has descendant R. Excluding R makes the group incomplete.

Check yourself

Trace nodes, count sites, define rates and test revisions

A clade includes a common ancestor and every descendant. Tip order and unscaled lengths do not measure relationships or dates.

Compare minimum site changes under stated assumptions, and test the hypothesis using independent data.

Convert percentages to fractions and distinguish per-lineage from pairwise rates. Explain revisions using evidence, not an unchanging list of names.