nGICE Phylogenomics

Hamid Ghanavi, Etka Yapar, Niklas Wahlberg

Introduction to phylogenetics

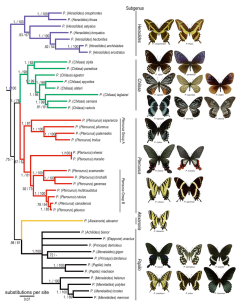
- Overview - why molecular phylogenetics?
- Some basic concepts e.g. phylogeny, monophyly, homology & analogy
- Exploring patterns in sequence data
- Alignment

The questions

- What is a phylogeny?
- Why are we interested in phylogenies?
- Why should we use molecular data (sequences) to infer phylogenies?

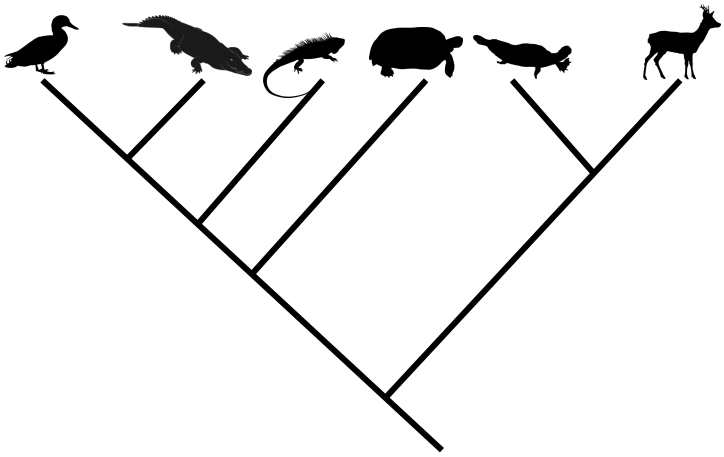
The very basic facts

- What we see today in nature is the outcome of what has happened in the past
- Ecology and evolution are inseparable
- "Species" or "genes" are not individual entities without any connections to other species or genes
  - phylogeny



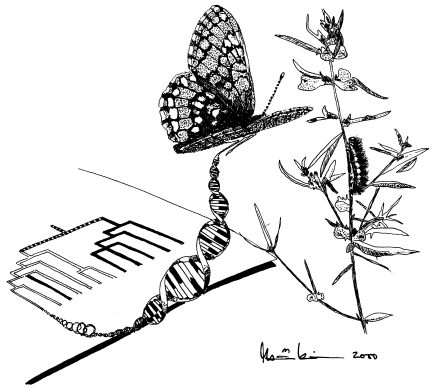
What is a phylogeny?

- A phylogeny is the historical genealogy of a group of species



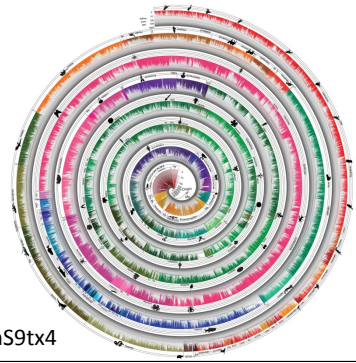
A phylogeny is an inference

- Envisioned as a dichotomously branching tree
- A phylogeny cannot be observed
- A phylogenetic hypothesis can be inferred from observed data



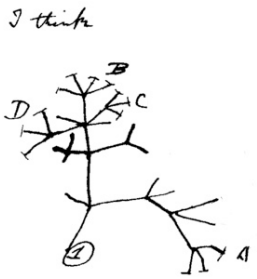
What we are after

- Phylogenies – the Tree of Life
- With phylogenies we are attempting to get a good working framework for Life
- Getting to the root of how evolution has worked



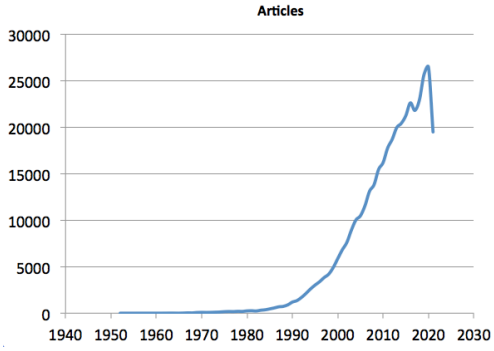
http://t.co/Q6RThS9tx4

- ”Nothing makes sense in biology except in the light of evolution”
  - Dobzhansky 1973
- ”Nothing in evolution makes sense except in the light of phylogeny”
  - Savage 1997



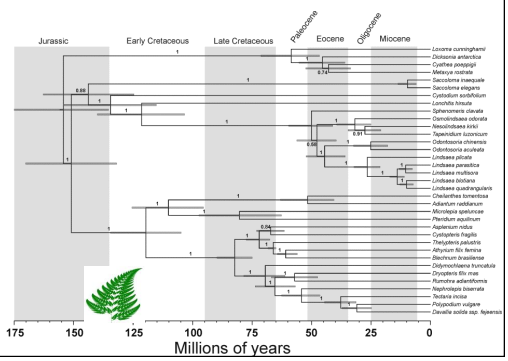
The rise of systematics

- Within the last 25 years the number of phylogenetic studies has skyrocketed
- Largely due to the advent of easy DNA sequencing methods
- Is helping us understand biodiversity and evolutionary processes better



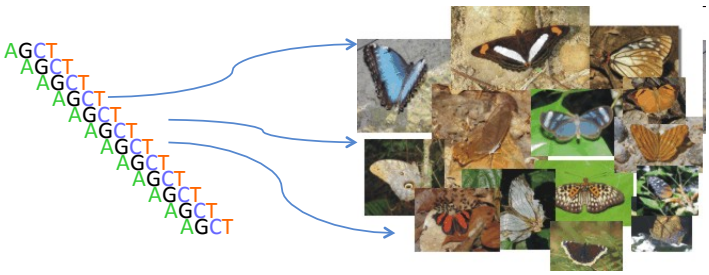
Systematics is…

- The study of the kinds and diversity of life
- The study of character evolution
- The study of historical biogeography
- The study of the temporal framework of evolution
- The study of molecular evolution



Why *molecular* systematics?

- Ease of data generation for large numbers of taxa
- Ease of generating a large number of independent data sets for given taxa
- Molecular characters behind the morphological characters we see



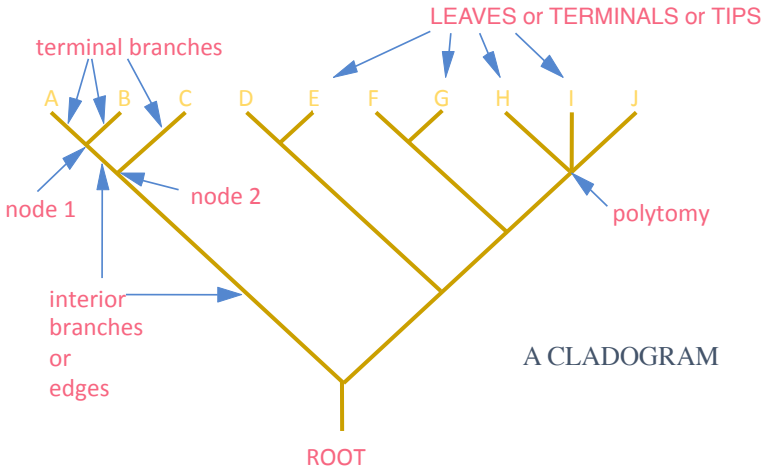
Molecular systematics as a part of understanding evolution

- Biochemistry — basic low-level processes (e.g., nucleotide substitution, amino acid interactions)
- Molecular genetics — fundamental genetic processes (e.g., DNA replication, recombination)
- Population genetics — micro-evolutionary processes
- Systematics — macro-evolutionary processes

Some basic concepts

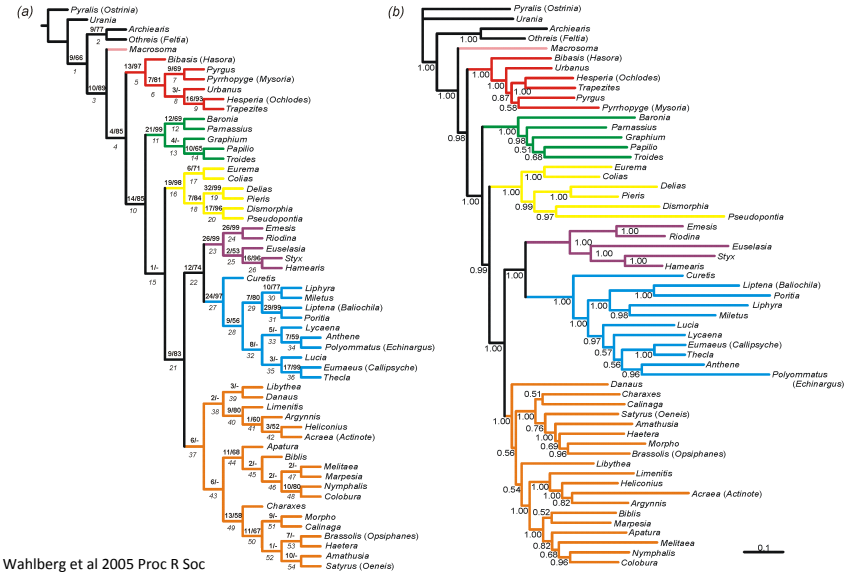
- Cladogram - a tree diagram which depicts a hypothesised evolutionary history
- Phylogram - a tree which indicates by branch length the degree of change believed to have occurred along each lineage
- Chronogram – a tree in which branch lengths are directly in proportion to time

Phylogenetic Trees



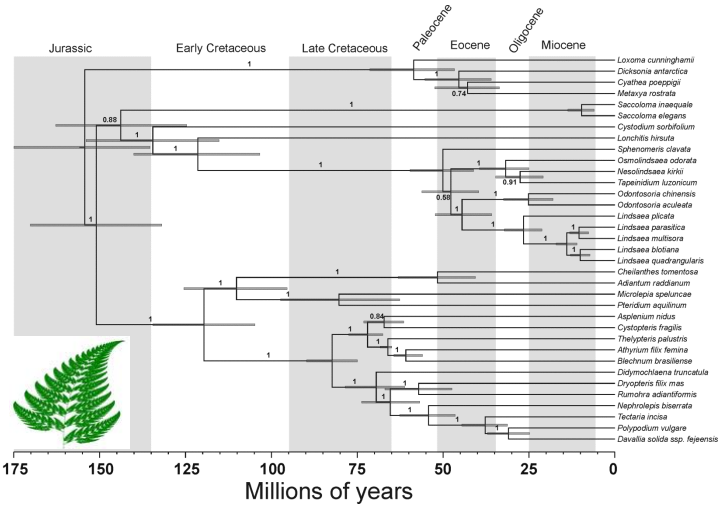
A CLADOGRAM

Cladograms and phylograms



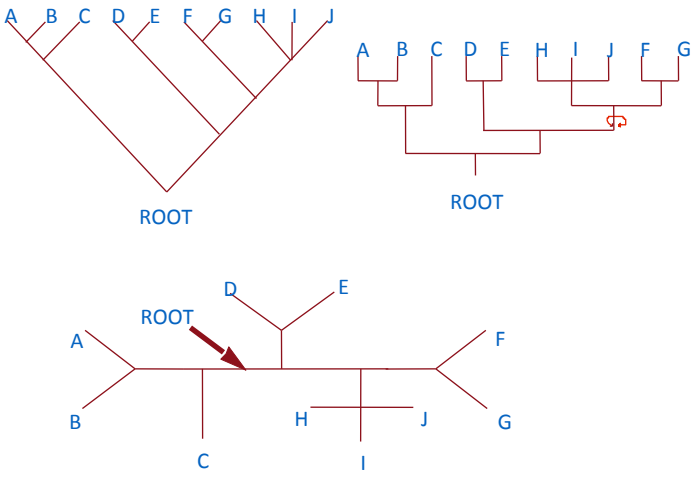
Wahlberg et al 2005 Proc R Soc

Chronogram



Lehtonen et al. 2012: Bot J Linn Soc 170:489.

Trees - Rooted and Unrooted

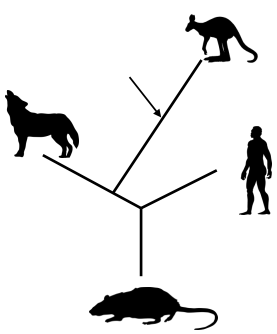


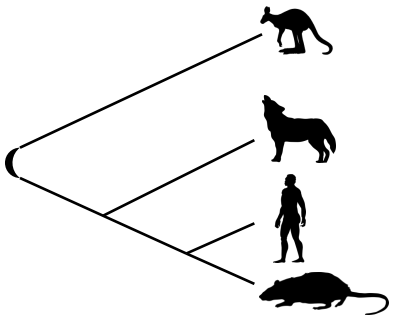
Rooting a tree

- Rooting a tree using outgroups
  - Place the root on the branch leading to the outgroup taxon
  - Use outgroup taxa in the analysis (rarely done)
- Other ways of rooting a tree
  - Assume a molecular clock
  - Midpoint rooting (root on the longest branch)

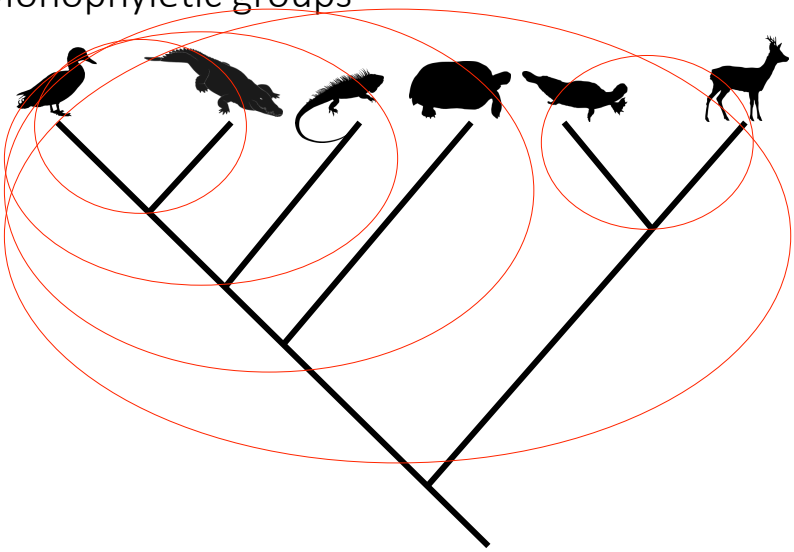
Outgroup rooting of unrooted trees

- Outgroup – related sequence that definitely diverged earlier (paleontological evidence)
- Not too distantly related (tree method becomes unreliable)

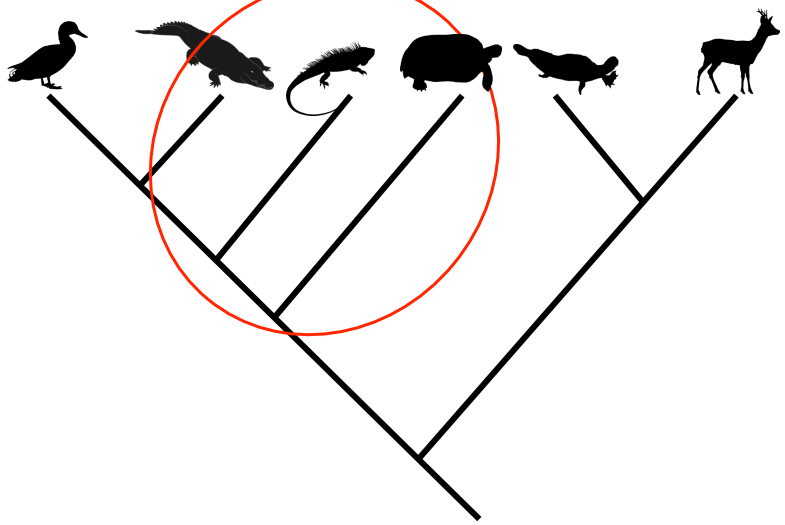




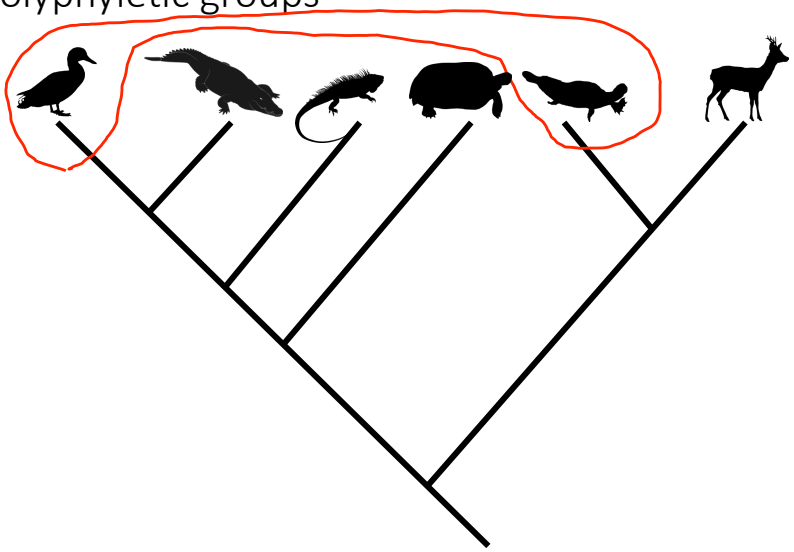
Monophyletic groups



Paraphyletic groups



Polyphyletic groups

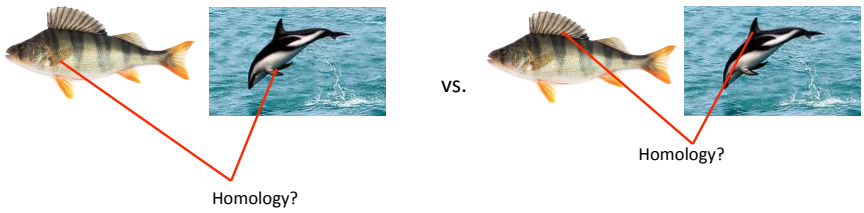


Some premises underlying phylogenetic inferences

- Phylogenetic inferences are premised on the inheritance of ancestral characters, and on the existence of an evolutionary history defined by changes in these characters
- Sees homology as evidence of common ancestry
- A tree-like model of evolution
  - paralogy and lateral transfer?

Homology

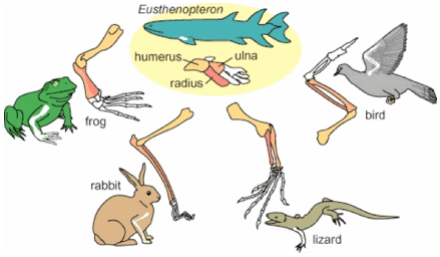
- The most fundamental concept in inferring phylogeny is homology
- We need to be sure the characters we are studying are homologous, ie ”the same” character in different organisms
- Otherwise our analyses will be misled

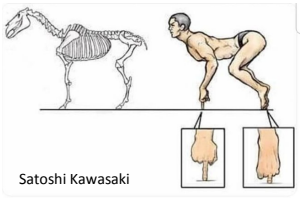


Owen's definition of homology

Homologue: the same organ under every variety of form and function (true or essential correspondence)

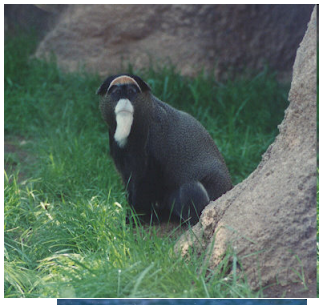
Richard Owen 1843

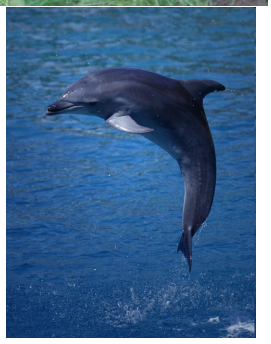




Homologies can be:

- Apomorphic
  - Shared derived
- Plesiomorphic
  - Shared ancestral





Character evolution

- Heritable changes (in morphology, gene sequences, etc.) produce different character states
- Similarities and differences in character states provide the basis for inferring phylogeny (i.e. provide evidence of relationships)
- The utility of this evidence depends on how often the evolutionary changes that produce the different character states occur independently

Unique and unreversed characters

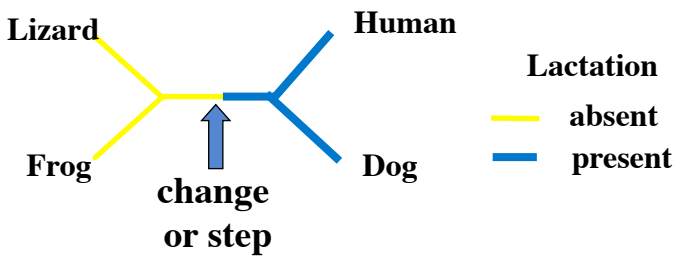
- Given a heritable evolutionary change that is unique and unreversed (e.g. the origin of lactation) in an ancestral species, the presence of the novel character state in any taxon must be due to inheritance from the ancestor
- Similarly, absence in any taxon must be because the taxa are not descendants of that ancestor

Unique and unreversed characters

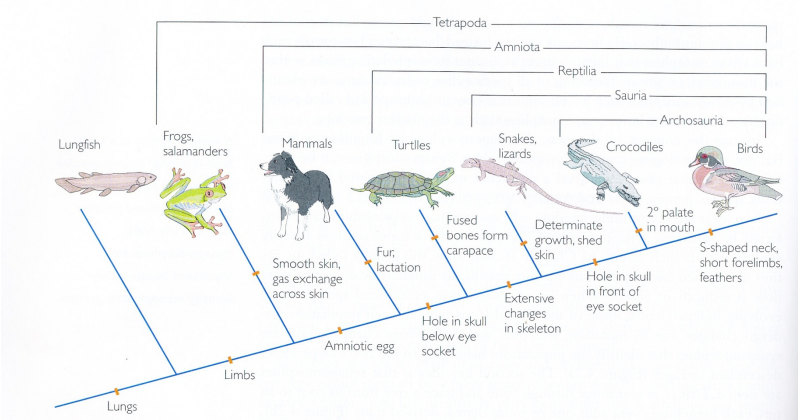
- The novelty is a *homology* acting as badge or marker for the descendants of the ancestor
- The taxa with the novelty are a clade (e.g. Mammalia)

Unique and unreversed characters

- Because lactation evolved only once and is unreversed (not subsequently lost) it is *homologous* and provides unambiguous evidence of relationships



In the ideal world



The converse of homology

- Analogy: superficial or misleading similarity
  - Homoplasy

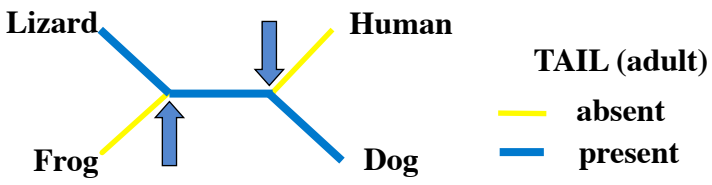
Falcon	Bat	Pterodactyl	Butterfly
↑	↑	↑	↑
Ancestral bird	Ancestral mammal	Ancestral reptile	Ancestral hexapod

Homoplasy - Independent evolution

- Homoplasy is similarity that is not homologous (not due to common ancestry)
- It is the result of independent evolution (convergence, parallelism, reversal)
- Homoplasy can provide misleading evidence of phylogenetic relationships (if mistakenly interpreted as homology)

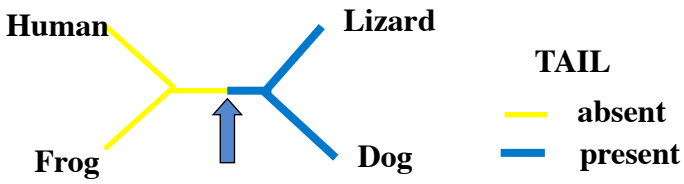
Homoplasy - Independent evolution

- Loss of tails evolved independently in humans and frogs - there are two changes on the true tree



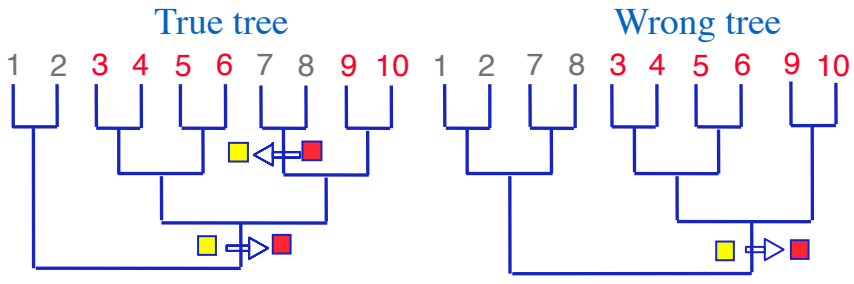
Homoplasy - misleading evidence of phylogeny

- If misinterpreted as homology, the absence of tails would be evidence for a wrong tree: grouping humans with frogs and lizards with dogs



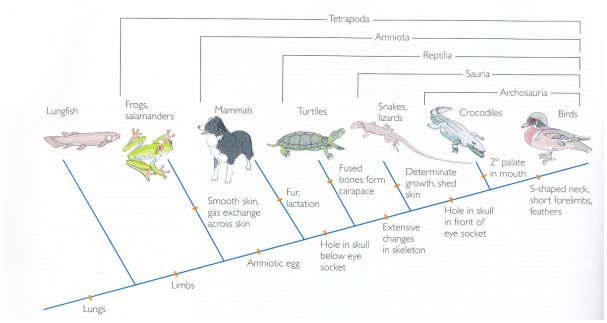
Homoplasy - reversal

- Reversals are evolutionary changes back to an ancestral condition
- As with any homoplasy, reversals can provide misleading evidence of relationships



Homoplasy - a fundamental problem of phylogenetic inference

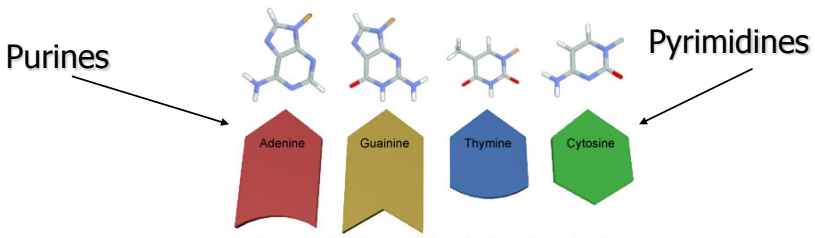
- If there were no homoplastic similarities inferring phylogeny would be easy - all the pieces of the jig-saw would fit together neatly
- Distinguishing the misleading evidence of homoplasy from the reliable evidence of homology is a fundamental problem of phylogenetic inference



Homoplasy in molecular data

- Incongruence and therefore homoplasy is common in molecular sequence data
  - There are a limited number of alternative character states (e.g. only A, G, C, T and "gap" in DNA)

Figure B-3: The Four Nitrogenous Bases

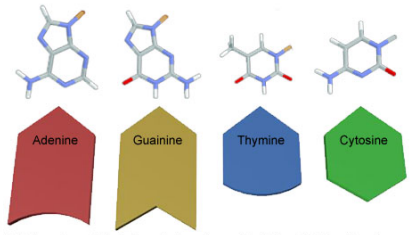


Each base has a distinct shape that can be used to distinguish it form the others. 3D representations of the four bases are shown, with the corresponding chemical structures drawn above.

Homoplasy in molecular data

- Character states are chemically identical
- Homology and homoplasy are equally similar
- Cannot be distinguished by detailed study of similarity and differences

Figure B-3: The Four Nitrogenous Bases



Each base has a distinct shape that can be used to distinguish it form the others. 3D representations of the four bases are shown, with the corresponding chemical structures drawn above.

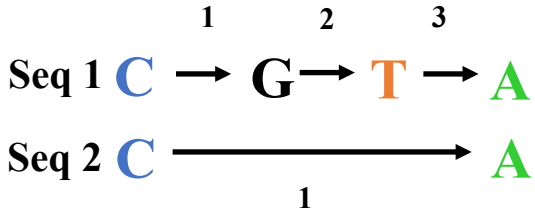
Saturation in sequence data

- Saturation is due to multiple changes at the same site subsequent to lineage splitting
- Models of evolution attempt to infer the missing information through correcting for "multiple hits"
- Most data will contain some fast evolving sites which are potentially saturated (e.g. in protein genes often position 3)
- In severe cases the data becomes essentially random and all information about relationships can be lost

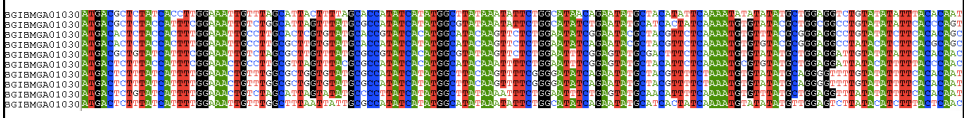
Multiple changes at a single site - hidden changes

```
Ancest GGCGCG
Seq 1  AGCGAG
Seq 2  GCGGAC
```

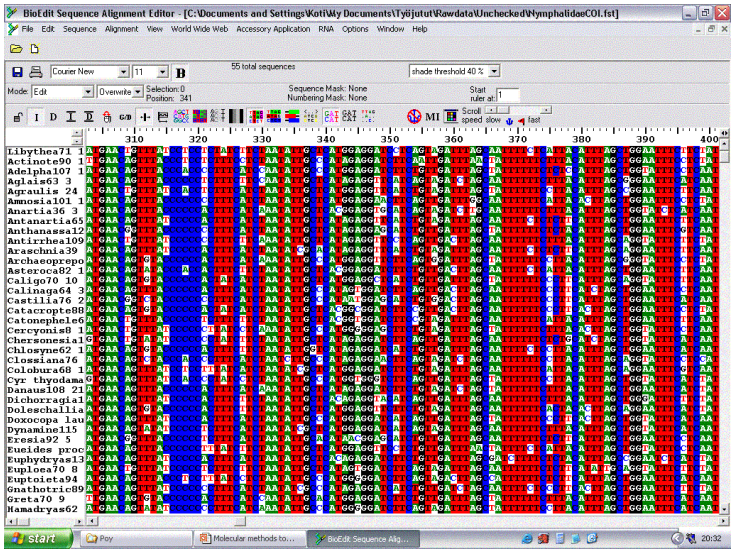
Number of changes



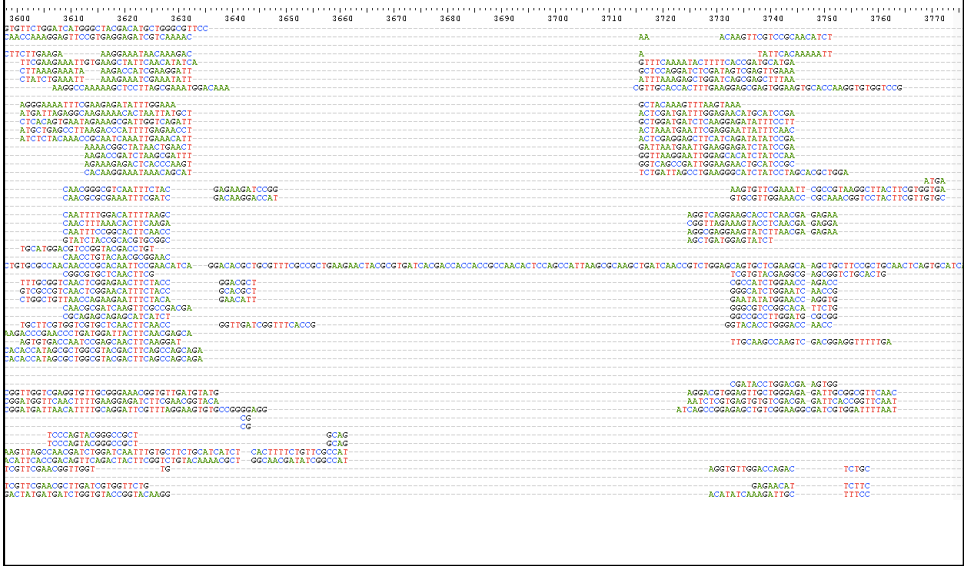
Multiple Sequence Alignment



Alignment can be easy…

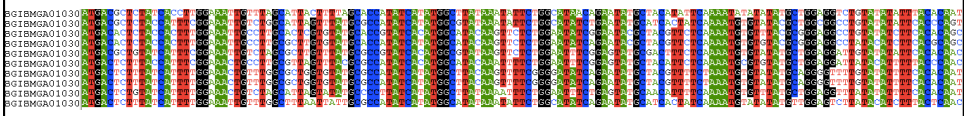


…or difficult



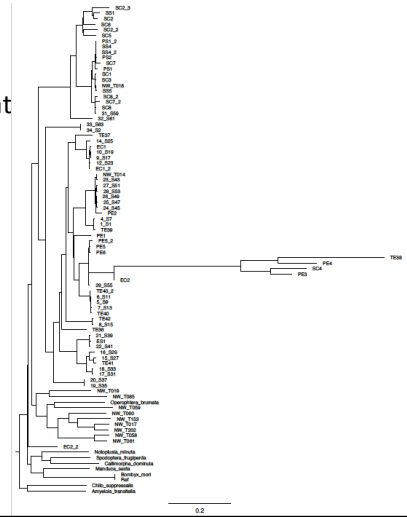
Homology: Definition

- Homology: similarity that is the result of inheritance from a common ancestor - identification and analysis of homologies is central to phylogenetic systematics
- An alignment is a hypothesis of positional homology between bases/amino acids



Multiple sequence alignment- goals

- To generate a concise, information-rich summary of sequence data
- Alignments can be treated as models that can be used to test hypotheses
- Does this model of events accurately reflect known biological evidence?



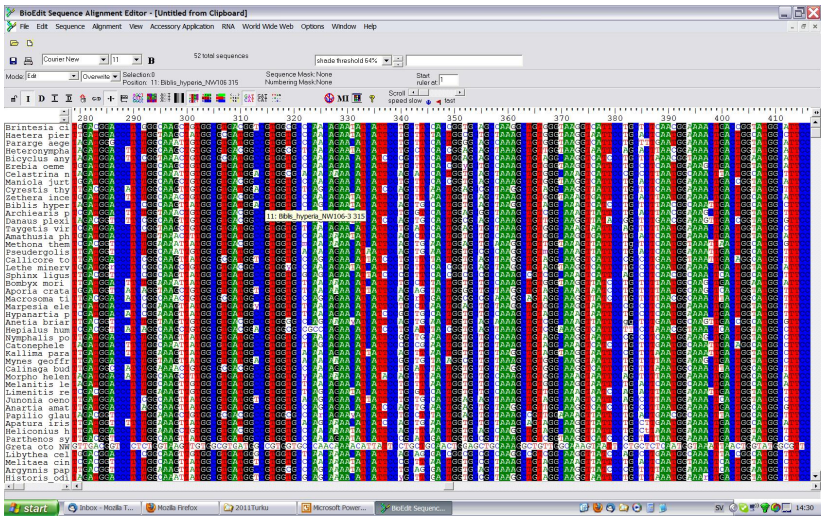
Multiple sequence alignment

- Manual
- Dynamic programming
- Heuristic methods
  - Progressive alignment
  - Consistency-based scoring
  - Iterative refinement methods

Manual alignment - reasons

- Might be carried out because:
- Alignment is easy
- There is some extraneous information (structural)
- Automated alignment methods have encountered a local minimum problem
- An automated alignment method can be “improved”

Protein-coding genes can often be manually aligned



How to align these sequences:

AGGGCTTTAA
AGGCTA
AATGGCTCTAA
GGAGCCCTAA

How to align these sequences:

```
A-GGGCTTTAA
A--GGCT--A-
AATGGCTCTAA
GGAG-CCCTAA
```

How to align these sequences:

```
-AGGGCTTTAA
-A-GGC--TA-
AATGGCTCTAA
-GGAGCCCTAA
```

Multiple sequence alignment

- Is not easy! How to be objective?
- Dynamic programming
- Heuristic methods
  - Progressive alignment
  - Consistency-based scoring
  - Iterative refinement methods

Dynamic programming

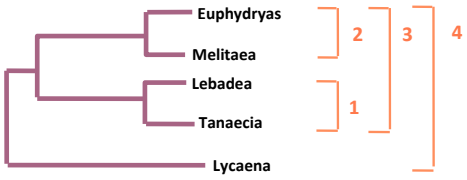
- For two sequences, the best alignment can be found by scoring all possible pairs of aligned nucleotides and penalizing gaps
- An optimality criterion
- Time and computer memory needed grows exponentially with number of sequences
- Becomes impossible to align more than 4 sequences of modest length
- Fails to fully exploit phylogeny and does not incorporate an evolutionary model

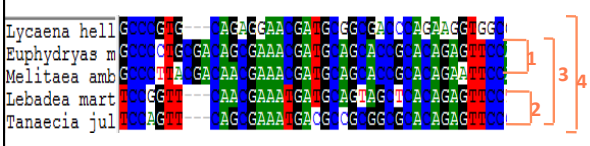
Heuristics: Progressive alignment

- Devised by Feng and Doolittle in 1987
- A heuristic method and as such is not guaranteed to find the 'optimal' alignment
- Requires *n-1+n-2+n-3...n-n+1* pairwise alignments as a starting point
- Most successful implementation is Clustal
  - ClustalX
  - Clustal Omega

Overview of Clustal procedure

Euphydryas 1 -
Melitaea 2 .17 -
Lebadea 3 .59 .60 -
Tanaecia 4 .59 .59 .13 -
Lycaena 5 .77 .77 .75 .75 -





Quick pairwise alignment: calculate distance matrix

↓

Neighbour-joining tree (guide tree)

↓

Progressive alignment following guide tree

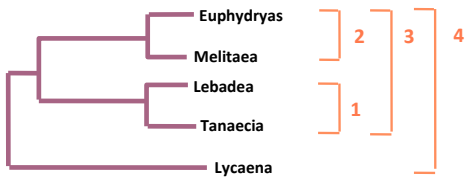
Clustal - pairwise alignments

- First perform all possible pairwise alignments between each pair of sequences
- Calculate the 'distance' between each pair of sequences based on these isolated pairwise alignments
- Generate a distance matrix

Euphydryas	**1**	-			
Melitaea	**2**	**.17**	-		
Lebadea	**3**	**.59**	**.60**	-	
Tanaecia	**4**	**.59**	**.59**	**.13**	-
Lycaena	**5**	**.77**	**.77**	**.75**	**.75**	-

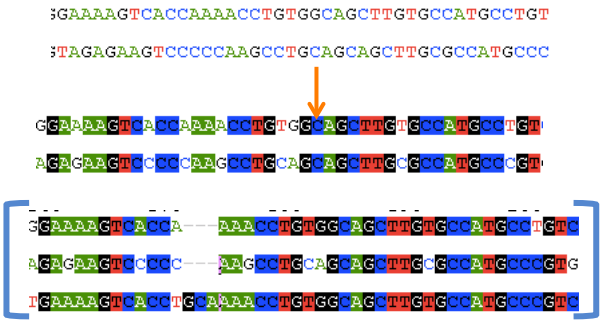
Clustal - guide tree

- Generate a Neighbour-Joining 'guide tree' from these pairwise distances
- This guide tree gives the order in which the progressive alignment will be carried out



Multiple alignment- first pair

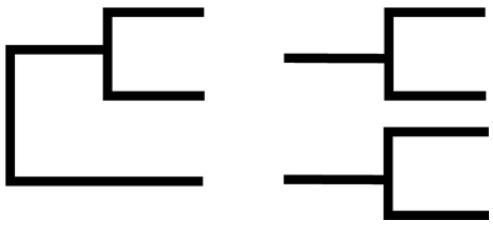
- Align the two most closely-related sequences first
- This alignment is then 'fixed' and will never change
- If a gap is to be introduced subsequently, then it will be introduced in the same place in both sequences, but their relative alignment remains unchanged



Clustal - decision time

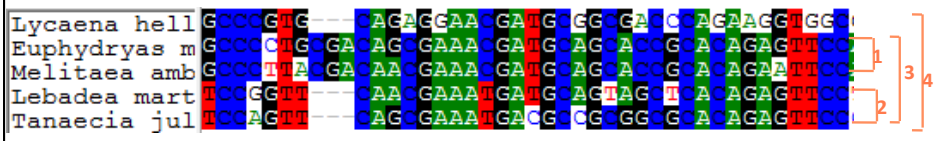
- Consult the guide tree to see what alignment is performed next.
  - Align a third sequence to the first two

  Or
  - Align two entirely different sequences to each other.



Clustal - progression

- The alignment is progressively built up in this way, with each step being treated as a pairwise alignment, sometimes with each member of a 'pair' having more than one sequence

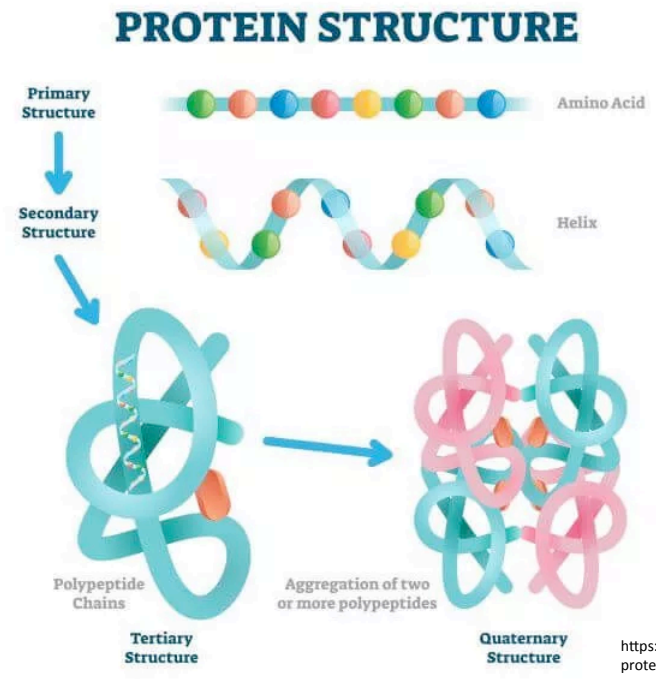


Clustal - good points/bad points

- Advantages:
  - Speed
- Disadvantages:
  - Hierarchic structure introduced that is not necessarily phylogenetic
  - No way of quantifying whether or not the alignment is good
  - No way of knowing if the alignment is 'correct'
  - Local minimum problem. If an error is introduced early in the alignment process, it is impossible to correct this later in the procedure
  - Arbitrary alignment

Increasing the sophistication of the alignment process

- Should we treat all the sequences in the same way?
  - some sequences are closely related and some sequences are distant relatives.
- Should we treat all positions in the sequences as though they were the same?
  - they might have different functions and different locations in the 3-dimensional structure.
  - codon structure – how to retain this?



https://biologydictionary.net/protein-structure/

Consistency-based scoring

- One way to avoid the problems of getting stuck in local minima or fixed gaps
- Based on optimizing a multiple alignment using information from all pairwise alignments
- Identifies those nucleotides that are aligned most consistently across the different alignments
- Used in e.g. T-Coffee

Iterative refinement methods

- Initial alignments split into two groups randomly
- Within groups the alignment is kept fixed
- Dynamic programming used to align the two groups to each other
- This is repeated until score converges
- Used in e.g. Muscle and MAFFT

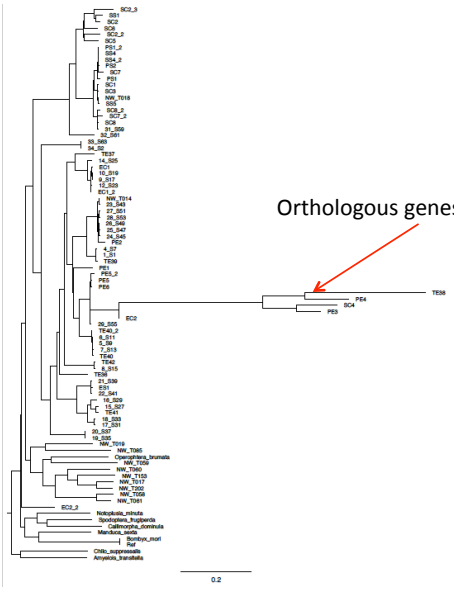
Using models in alignment

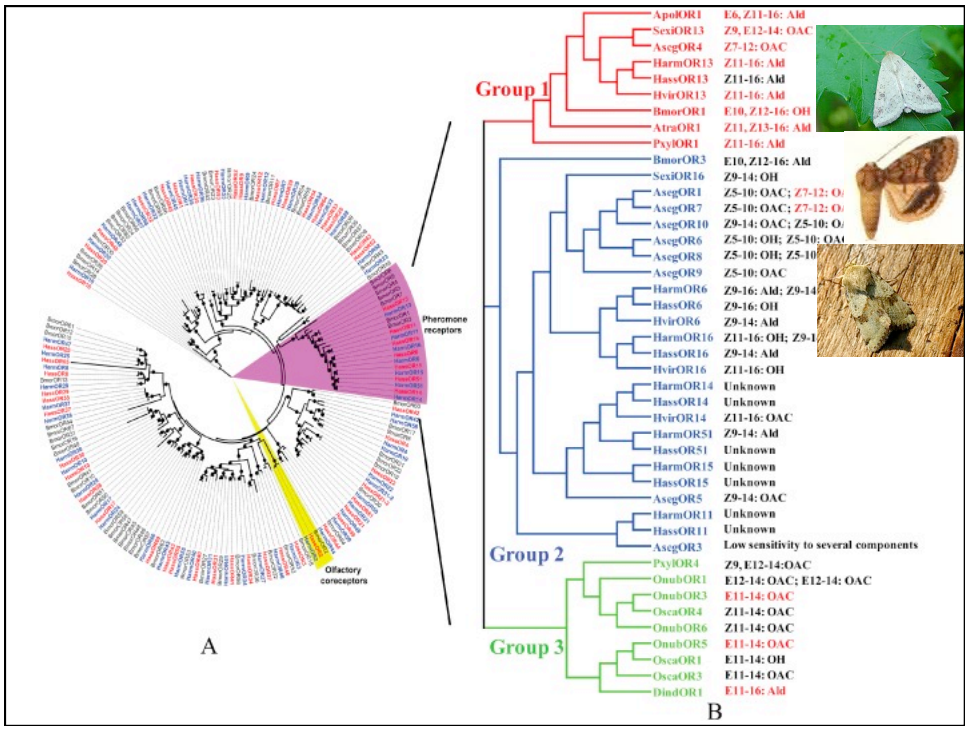
- New methods are being developed all the time
- Latest methods include using a Bayesian statistic framework, DNA evolutionary models and alignment concomitantly with estimation of phylogentic relationships
- Still not feasible with a moderately sized dataset

Genomics: orthology and alignment!

- We need to know that the genes we are studying are the same (homology)
- Genomics relies on bioinformatic methods to determine orthology and alignment
- For phylogenomics single copy, protein coding, orthologous genes are often preferred
  - Benchmarking Universal Single-Copy Orthologs (BUSCO)
  - Taxon specific sets
- For functional genomics, gene family dynamics are of interest
  - Orthology is crucial!
  - Alignment is not straightforward!

Contamination – a problem to be aware of





Bottom line

- Alignments are extremely important in phylogenetics
- A bad alignment means many wrong statements of homology, which means pure rubbish as output
- A good alignment can be hard to attain