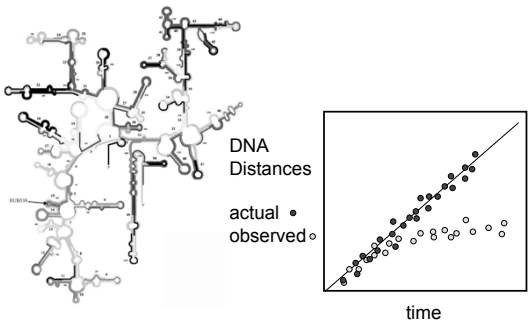
Systematics
Lecture 13 - Alignment, DNA data



Homoplasy - 3 types

1. **Convergence**: true analogy, e.g. wings of birds and insects; usually distantly related taxa

2. **Parallelism**: similar nonhomologous state in closely related taxa, often with same/similar development & genetic basis

3. **Reversal**: change to an earlier state, e.g. The aquatic lifestyle of whales is not homologous with that of fish

Note also distinguish between "real" homoplasy and artifactual (due to human error)

Outline

1. Four steps in Phylogenetic Inference

2. Molecular Data - Selection

3. Molecular Homology, alignment

4. Paralogs, Orthologs, Xenologs, gene trees

Four steps

1. **Character (data) selection** (not too fast, not too slow)

2. **Alignment of Data** (hypotheses of primary homology)

3. **Analysis selection** (choose the best model / method(s)) - data exploration

4. **Conduct analysis**

Four steps

Remember the following:

"The data are the things"

Much that is taught on phylogenetic inference deals with *methods* of analysis

Do not neglect the quality of the data
"Garbage in, garbage out"

"Black box or point-and-click phylogenetics"

1. **Data quality**: there are many considerations prior to analysis

2. **Analysis**: again, many considerations - issues to deal with…

Examples of poorly done phylogenetics are common - too many people* either (1) ignore the complexities or (2) are ignorant of them (* researchers, editors, reviewers, etc.)

"Black box or point-and-click phylogenetics"

Read for Wed: Grant, T., Faivovich, J., & Pol, D. (2003) The perils of 'point-and-click' systematics. Cladistics 19: 276-285.

- Critique of Hall's book "Phylogenetic trees made easy"
- Hall's book is, unfortunately, not just a "how-to" manual
- (Re-read Grant et al. at the end of the course when you understand more of what is discussed)

"Black box or point-and-click phylogenetics"

"Far from a step toward the elimination of 'point-and-click' systematics, the many misconceptions, inaccuracies, misrepresentations, and inconsistencies perpetuated throughout this book serve to exemplify **the perils of doing without knowing why**."

- this is the motivation behind this course: so you will be able to do & **know why**

Selection of Molecular characters

Character / discrete data: nucleotide or amino acid sequences (can be converted to distances)

"fast & slow" genes:
- there is variation in the rate of change among regions of the genome

e.g. rRNA (e.g. 18S) evolves slowly enough to hold information that is over 250 million years old

- whereas mtDNA (e.g. COII) evolves much faster and most information over 30-50 million yrs of age is probably gone (starts to go at 15-20 my)

Selection of Molecular characters

Higher-level phylogenetics: (families & above) use slower, conserved genes, nuclear genes
- evolve slowly due to **functional constraints**:
  e.g. some proteins "still work" with many potential amino acids
  others won't, e.g. histones are strongly conserved

- faster evolving regions, e.g. mtDNA,
  becomes **saturated with multiple hits**
  -information is overwritten
  -**back mutations**
  - yield nonsense phylogenies for deep splits

Selection of Molecular characters

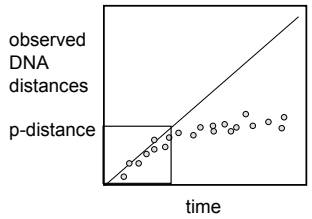
Lower-level phylogenetics: (subfamilies & below) use faster, less-conserved genes, mtDNA

- because slower genes would be identical across your species

- must select genes most appropriate for your study taxa

Selection of Molecular characters

Saturation graph
as time proceeds DNA distances also increase, to a point of saturation



Observable change increases in a linear fashion (x ~ y) for a while

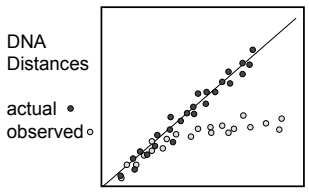
Only so much change is observable

Real change continues with time

Selection of Molecular characters

Saturation graph

as time proceeds DNA distances also increase, to a point of saturation



Observable change increases in a linear fashion (x ~ y) for a while

Only so much change is observable

Real change continues with time

Selection of Molecular characters

Why?

- constraints: for a given gene, some sites essentially do not change (preventing DNA distances from reaching 100%)
- even for regions that are variable, typically DNA distances can't go beyond 75% since 1/4 of the changes would be to the same nucleotide

-other sites do change: for a given comparison of 2 taxa a variable site might have changed:

once: (good)
two or more times: (bad) - "multiple hits" information lost…

Selection of Molecular characters

example

```
Species1     ATGCCTGGACTTATAA
Species2     ATGCCGGGAGATATAA
                   .   ..
```

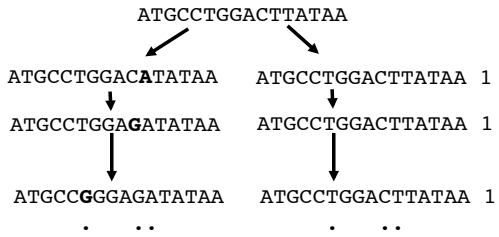
3 changes observed - this is the **minimum** and is only the ACTUAL number of changes if there have been no multiple hits (or back mutations)

i.e. each site changed only once since speciation / divergence

Selection of Molecular characters

Example - recent divergence, no saturation

Ancestral sequence

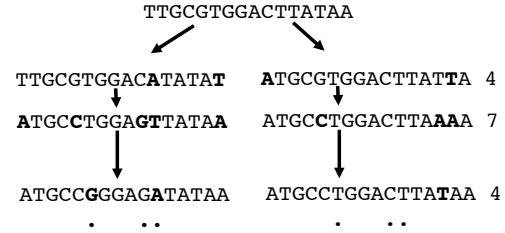


3 changes observed, 3 actual changes

Selection of Molecular characters

Example - ancient divergence, with saturation

Ancestral sequence



3 changes observed, 15 actual changes
(2 parallelisms (sites 1 & 5) = homoplasy)

Selection of Molecular characters

Implications are serious for:

1. Gene choice - select genes that are not saturated for your taxa (different genes depending on age of taxa)
2. Estimation of divergence dates / "molecular clock" estimation
3. Estimation of branch lengths (proportional to time and/or amount of change)
4. Estimation of homologies/synapomorphies and strength of support for a relationship
5. Use of distance methods with **uncorrected** data

Selection of Molecular characters

Three types of genes

tRNA - transfer RNA (short)
rRNA - ribosomal RNA (long, conserved)
mRNA - messenger RNA - protein coding (**exon)**

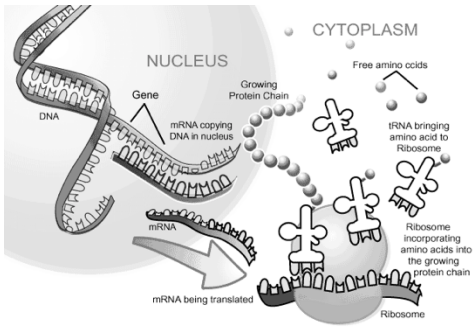
Also introns - non coding sequence sometimes inside a protein coding gene

Can be **Nuclear**

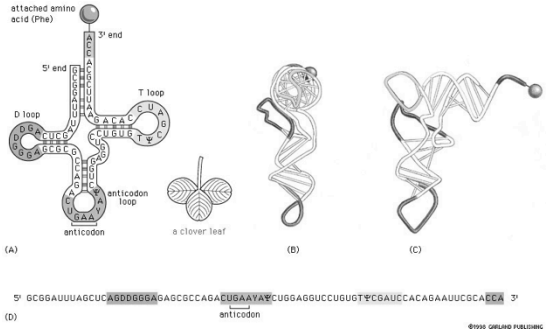
Typically slower evolving than mitochondrial
better for deeper (older) divergences

Can be **Mitochondrial**
Better for shallow (recent) divergences

Selection of Molecular characters

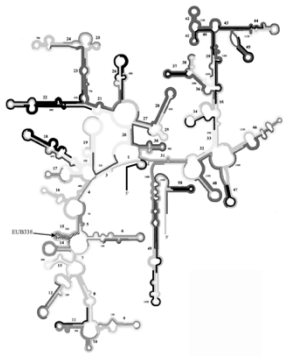


Selection of Molecular characters



tRNA Short, 70-150 base pairs, stems & loops, one for each amino acid, rarely targeted for sequencing - too few data

Selection of Molecular characters



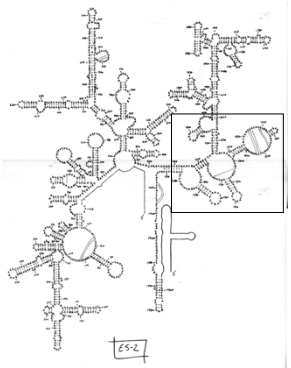
rRNA
e.g. 16S rRNA small subunit

Long: > 1000 sites

Many stems & loops
Complicated 2ndary structure

Forms part of the ribosome that assists with protein synthesis

Selection of Molecular characters



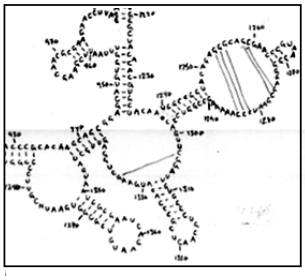
rRNA

Of variable length among taxa

Difficult to align / determine homologous sites

Stems **tend** to evolve more slowly than loops

Selection of Molecular characters



rRNA

Conserved and universal

Used to estimate deep divergences (families +)

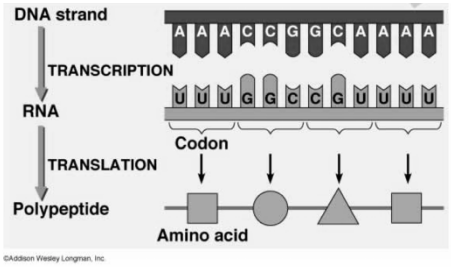
Conserved regions virtually 100% identical among species in a genus

(See Doublet Model – Kjer 2004)

Stem sites **not** independent

Selection of Molecular characters

mRNA (protein coding DNA) - codons



Selection of Molecular characters

Genetic code

1st position (5' end)	2nd position U	2nd position C	2nd position A	2nd position G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

64 codons
20 amino acids

Degenerate (redundant) code

AUG (ATG) = start codon

Selection of Molecular characters

Genetic code - not universal: different for mitochondria of various taxa

Table 14-4 Some Differences Between the "Universal" Code and Mitochondrial Genetic Codes*

Codon	"Universal" Code	Mitochondrial Codes: Mammals	*Drosophila*	Yeasts	Plants
UGA	STOP	*Trp*	*Trp*	*Trp*	STOP
AUA	Ile	*Met*	*Met*	*Met*	Ile
CUA	Leu	Leu	Leu	*Thr*	Leu
AGA, AGG	Arg	*STOP*	*Ser*	Arg	Arg

*Italics and color shading indicate that the code differs from the "universal" code.

20 Amino Acids & stop codon

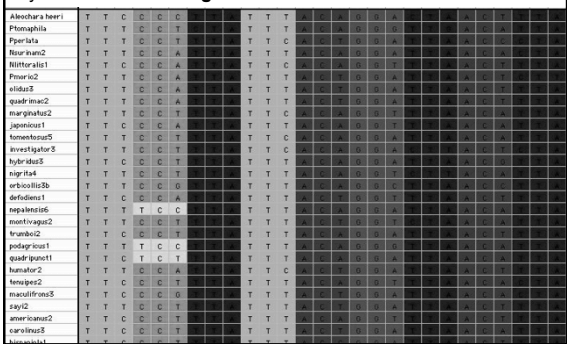
3 Letter Code	1 Letter Code	Full name	mRNA nucleotide triplets (codons)
Ala	A	Alanine	GCA, GCC, GCG, GCU
Arg	R	Arginine	AGA, AGG, CGA, CGC, CGG, CGU
Asn	N	Asparagine	AAC, AAU
Asp	D	Aspartic acid	GAC, GAU
Cys	C	Cysteine	UGC, UGU
Glu	E	Glutamic acid	GAA, GAG
Gln	Q	Glutamine	CAA, CAG
Gly	G	Glycine	GGA, GGC, GGG, GGU
His	H	Histidine	CAC, CAU
Ile	I	Isoleucine	AUA, AUC, AUU
Leu	L	Leucine	UUA, UUG, CUA, CUC, CUG, CUU
Lys	K	Lysine	AAA, AAG
Met	M	Methiodine	AUG
Phe	F	Phenylalanine	UUC, UUU
Pro	P	Proline	CCA, CCC, CCG, CCU
Ser	S	Serine	AGC, AGU, UCA, UCC, UCG, UCU
Thr	T	Threonine	ACA, ACC, ACG, ACU
Trp	W	Tryptophan	UGG
Tyr	Y	Tyrosine	UAC, UAU
Val	V	Valine	GUA, GUC, GUG, GUU
STOP			UAA, UAG, UGA

Molecular Alignments

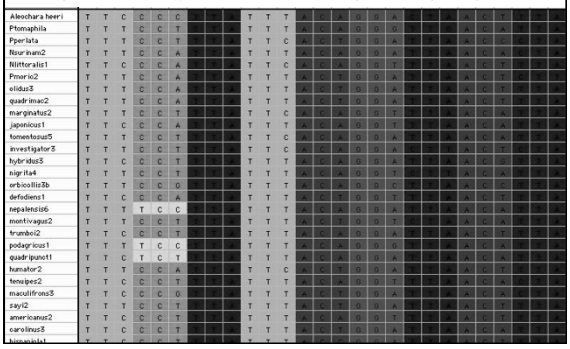
Protein coding genes - alignment usually trivial due to conserved codon structure (if no introns)
- often done by "eye" with reference to known amino acid sequence [CLUSTAL]
- homologous sites are known with certainty

Non-protein coding - more challenging due to variation in length of sequence among taxa / OTUs (like Morphology!)
[**OTU =** operational taxonomic unit]
- can be done by "eye" with reference to secondary structure (e.g. Kjer 2004)
- typically aligned by computer software

Codon structure of protein-coding genes makes alignment easy, "trivial," IF you know the Amino Acid sequence - i.e. if you know the **reading frame** / codon structure

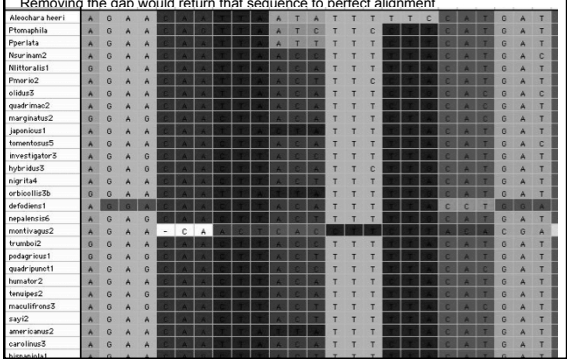


Note: 3rd codon position most variable, (**synonymous substitution**), AA doesn't change - if 1st or 2nd position changes this usually causes an AA change (**nonsynonymous**)



A single gap can cause a 'frame shift' that is obviously misaligned - this gap would need to be removed, it is an artifact

Removing the gap would return that sequence to perfect alignment



Molecular Alignments

Protein coding genes

- A joy to work with because one huge problem, that of hypotheses of primary homology, is greatly reduced

- We KNOW which data belong to which characters (sites)

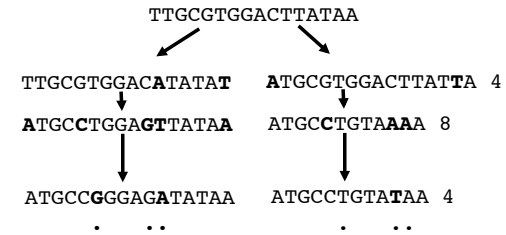
- **There still may be plenty of homoplasy but it won't be an artifact of human error!**

(unlike morphology & non-protein coding DNA which can have plenty of homoplasy due to human error)

Alignment of non-protein coding DNA

Example - deletion event

Ancestral sequence



3 changes observed, 16 actual changes (one deletion event)
(2 parallelisms (sites 1 & 5) = homoplasy)

Alignment of non-protein coding DNA

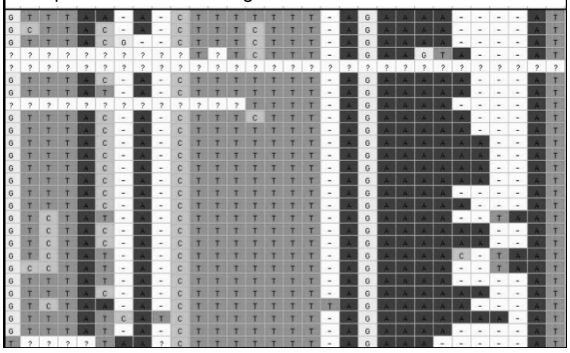
Indels

- insertions / deletions

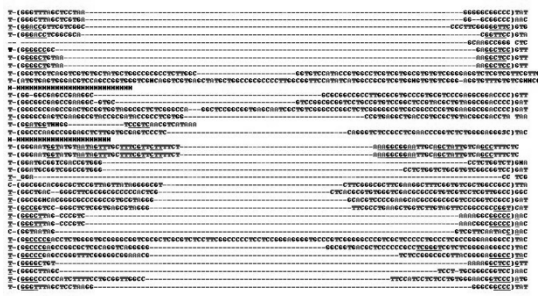
```
Species1    AA-TCGG
Species2    AAATCGG
```

- don't know if species1 lost (deletion) an A
- or if species2 gained (insertion) an A
- issue of polarity (next lecture)
- large indels sometimes coded as an extra character (see also lecture on inapplicable characters)

tRNA & rRNA (& introns) have **no codon structure** - making alignment much more challenging - different taxa can have sequences of different length

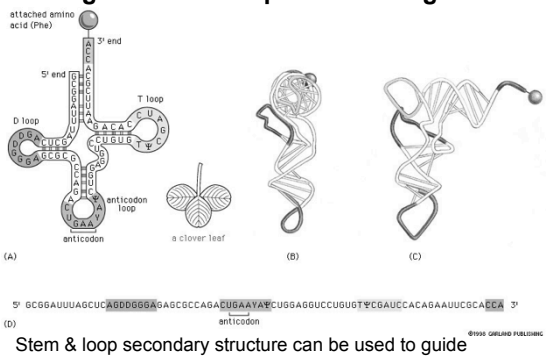


Alignment of non-protein coding DNA



Some regions are unalignable - these are often excluded from analysis (prefer a lack of data to misleading data)

Alignment of non-protein coding DNA



Stem & loop secondary structure can be used to guide alignment - **sites in stems must pair across the stem**

Alignment of non-protein coding DNA

Non-protein coding DNA alignment has issues similar to alignment (homology assessment) of morphological data

Biological criteria, prior to analysis, can help establish hypotheses of homology

- e.g. Remane's 3 criteria (morphology)

- e.g. 2ndary structure for tRNA & rRNA data

(- e.g. codon structure for mRNA data)

Alignment of non-protein coding DNA

However, use of 2ndary structure is difficult, tedious and has been criticized and rejected by those who prefer computerized alignments

- Critics suggest that such 2ndary structural methods generate irreproducible alignments (different workers would generate different alignments)

- This is somewhat true, but is no more a problem than for morphological character coding, and if done carefully the alignments will be highly congruent with each other & hopefully with "reality"

Alignment of non-protein coding DNA

Computer alignments with software like CLUSTAL or MALIGN requires user to select (subjectively) a **gap cost penalty**

- This specifies the "cost" for the software to insert gaps to align the data: high = fewer gaps inserted, low = more gaps inserted

- Allows others to replicate the alignment using the same gap cost penalties & software - thus reproducible, but still subjective

Alignment of non-protein coding DNA

BUT the alignment is not done with reference to the 2ndary structure - thus it may select "impossible" alignments - 100% reproducible but **wrong**

- See the Kjer (2004) reading: computerized alignments of 18S yield phylogenies that disagree with known groups / other data

- Alignment using 2ndary structure yields phylogeny in greater agreement with known groups / other data

Alignment of non-protein coding DNA

Why use secondary structure ?

1. Stems are more conserved than the actual nucleotides - data changes but stems remain across divergent taxa - seek **conserved motifs**
2. rRNA function is largely determined by its structure
3. Computerized alignments - gap cost penalty should vary among different parts of the molecule:
   - perfectly conserved regions should have a penalty of infinity
   - hypervariable regions should have a very low penalty

Alignment of non-protein coding DNA

Computerized alignments

- can save time for protein coding data & typically produce +/- same alignment as "by eye"
- final alignment must be checked visually sometimes nonsensical alignments are produced (Fig. 3.5 text)
- some programs perform "direct optimization" which doesn't produce an alignment - it aligns & searches for trees simultaneously & chooses alignment that produces optimal tree but the alignment is never seen / can't be checked - [e.g. POY - popular with Cladists]

Alignment of non-protein coding DNA

Computerized alignments

Those that select the alignment which produces the optimal (shortest) tree might be removing "real" homoplasy

- example:

```
species1    CTATTGCATTT
species2    ATATTGCATTT
species3    ACGCCGCATTT
```

Say there was a parallelism with site1 (A) - one extra step on the tree = homoplasy

Alignment of non-protein coding DNA

Computerized alignments

- example:

```
species1    C-TATTGCATTT
species2    A-TATTGCATTT
species3    A-CGCCGCATTT
species4    TACGCCGCATTT
```

Another species is added which requires a gap be inserted for species 1-3

- here, the homoplasy remains

Alignment of non-protein coding DNA

Computerized alignments

- example:

```
species1    C-TATTGCATTT
species2    A-TATTGCATTT
species3    -ACGCCGCATTT
species4    TACGCCGCATTT
```

A computerized alignment using parsimony can eliminate the homoplasy (which yields a more parsimonious tree) - but a "real" parallelism has been removed from the data

Alignment of non-protein coding DNA

	Head	Wing color	Legs	Tail
species1	narrow	?	hairy	with spines
species2	narrow	?	smooth	no spines
species3	wide	black	hairy	with spines

Alignment of non-protein coding DNA

	Head	Wing color	Legs	Tail
species1	narrow	?	hairy	with spines
species2	?	narrow	smooth	no spines
species3	wide	black	hairy	with spines

Relevant paper - used MALIGN with and without secondary structure:
Titus,T. A., & Frost, D. R. 1996. Molecular Homology Assessment and Phylogeny in the Lizard Family Opluridae (Squamata: Iguania). Mol. Phyl. Evol. 1:49-62.

Alignment of non-protein coding DNA

Summary of approaches to alignment

1. Some methods base hypotheses of homology on biological information (codon structure, secondary structure)
2. Other methods ignore this information and use a computer calculated score, e.g. parsimony (shortest tree)
3. Can be combined - computerized methods using biological information, e.g. 2ndary structure

Alignment of non-protein coding DNA

Summary of importance of alignment

1. Different alignments of the same data can yield different estimates of phylogeny
2. A good alignment is critical to the analysis
3. A good alignment minimizes homoplasy due to human error (artifactual homoplasy) - but watch out about elimination of *real* homoplasy
4. Important to state how one did their alignment (of course in a paper, but also in talks)

Alignment of non-protein coding DNA

Some good references to cite regarding the value of secondary structure to guide rRNA alignment

- Hickson et al, 1997. Mol. Biol. Evol. 13:150
- Hickson et al., 2000. Mol. Biol. Evol. 17:530
- Kjer 1995. Mol Phylogenet. Evol. 4:314
- Morrison and Ellis, 1997, Mol. Biol. Evol. 14:428
- Titus and Frost, 1996. Mol Phylogenet Evol 6:49
- Buckley et al. 2000 Insect Molecular Biology 9(6), 565–580
- Page, R.D.M. 2000. Nucleic Acids Research 28(20):3839-3845

Gene Trees vs Species Trees

With genetic data we are actually inferring **gene trees**

- We hope the gene tree (splitting events of genes) will mirror the species tree (splitting events of populations)
- But it may not…
- Another potential source of Phylogenetic Error
- More of a problem for recent divergences

Gene Trees 1: Serial homology of genes

- Just like you wouldn't want to compare data taken from the mid-legs of species1 to those of the hind-legs of species2 (serially homologous structures)

- you also wouldn't want to compare DNA data taken from serially homologous genes (**paralogs**)

Paralogy is serial homology due to gene duplication
- some genes exist simultaneously as multiple, different copies (with their own unique histories) within the same organism

Gene Trees 1: Serial homology of genes

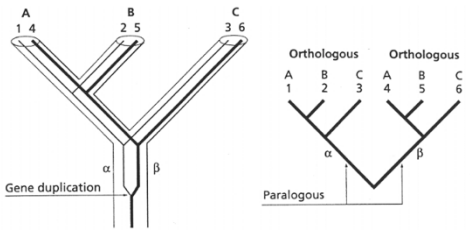
When you compare DNA data among OTUs

- you want to compare **orthologs** to each other
- like comparing the hind legs among OTUs

Orthology is homology due to speciation (common ancestry)

- another source of Phylogenetic Error is mistakenly comparing non-orthologous genes (paralogs)
- best to use genes that are not known to have copies (paralogs)

Gene Trees 1: Serial homology of genes



Gene duplications yield families of genes - widespread & important evolutionary phenomenon

Gene Trees 2: Ancestral polymorphism

Even restricting analysis to orthologous genes cannot, in principle, guarantee that gene tree = species tree because of ancestral polymorphism and differential survival of alleles (lineage sorting)

At speciation, lineage A was polymorphic, with one allele more closely related to lineage B's allele than to the other lineage A one.

If the polymorphism persists until a subsequent speciation event, gene tree will support ((A$_2$,B),A). (Fig 1)

Lineage sorting may eliminate the alternative allele and the gene tree will match the species tree (Fig 2)

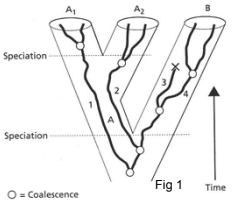


Fig 1

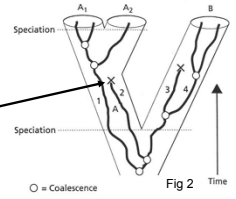
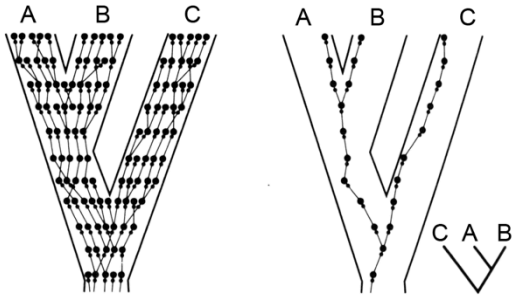


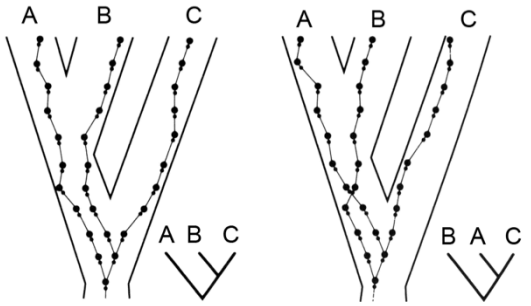
Fig 2

Gene Trees vs Species Trees



Species tree (A&B are sister spp.) gene tree that agrees

Gene Trees vs Species Trees



Two gene trees that do not agree with the species tree - due to **ancestral polymorphism**

Xenology

- gene was obtained by organism through horizontal transfer

  e.g. transposable elements

- also a potential source of confusion and Phylogenetic error, if mistaken for orthologous genes

- fortunately, rarely a source of error
- related to issues of **hybridization** & **introgression**

Summary

1. Alignments critical to reducing artifactual homoplasy (due to incorrect alignment) - want an **unambiguous alignment**

2. Protein coding genes can usually be aligned without worry of artifactual homoplasy. Difficult to do this for morphology, rRNA, & tRNA and introns

3. Phylogenetic error can result from using non-orthologous genes or ancestral polymorphisms - the latter problem is most common for recently divergent taxa

Terms - from lecture & readings

"point-and-click" phylogenetics
"fast & slow" genes
Higher & lower level phylogenetics
Saturation
Multiple hits
Back mutations
tRNA, rRNA, mRNA
Nuclear & mitochondrial
Stems & loops
Codons
Codon structure
Reading frame
Frame shift
Two kinds of homoplasy:
- artifactual homoplasy
- "real" homoplasy

Indels
Introns / exons
Synonymous / nonsynonymous substitutions
CLUSTAL
MALIGN
POY
OTUs
Gap cost penalty
Orthology
Paralogy
Xenology, introgression
Gene tree vs species tree
Ancestral polymorphism

Study questions

Why do we need to select gene(s) of the appropriate evolutionary rate? What problems might arise if we didn't? (for both higher and lower level investigations)

Why does saturation happen? Implications of saturation?

Which of the codon positions evolves the fastest (is most variable)?

Why are stems typically slower to evolve than loops? Why might one want to use secondary structure to align rRNA data?

Alignment of which type(s) of the 3 kinds of genes is most like primary homology assessment using morphology? And why?