Trees of Life (BIOL30004)
Lecture 7
Gene trees and species trees

Mark Beaumont

Spring 2017

Motivation for Lecture 7

A gene tree based on samples taken within a population has a characteristic timescale, and depends on the effective population size and generation time.

Phylogenetic trees are gene trees that cover a much longer timescale.

Typically in phylogenetic analysis we have one random sequence from each species that is being examined.

Naïvely we might assume that the coalescent has little relevance on long time scales. However this is not necessarily the case, and this lecture will cover situations where it does matter.

Summary of Lecture 7

1. Inconsistent phylogenies of humans and apes.
2. Incomplete Lineage Sorting
3. Estimation of ancestral population sizes and speciation times.
4. Examples: humans, apes, and finches.
5. Implications for phylogenetics

Phylogeny of humans and apes I

Satta *et al.* (*Mol Phylog Evol*, 2000) looked at 45 loci (47,000 bp) in humans, chimps, gorillas.

They reconstructed phylogenies for these genes and found that:

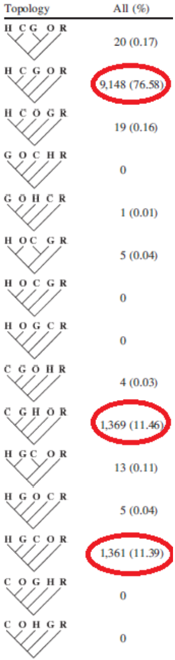
- 60% of loci favour ((human,chimp),gorilla)
- 20% of loci favour ((human, gorilla),chimp)
- 20% of loci favour ((gorilla,chimp), human)

Phylogeny of humans and apes II

This problem was revisited again by Ebersberger *et al* (2007) who looked at 23,210 DNA alignments (sequences that can be compared) for human, chimp, gorilla, and orangutan, with rhesus as the outgroup, and found the following:

You can see that it follows the Satta *et al* result, but with more sequences showing the ((human,chimp),gorilla) pattern. (Only sequences that gave high support for a particular topology are used.)

If we include the orangutan you can see that a variety of phylogenies are supported by different sequences.



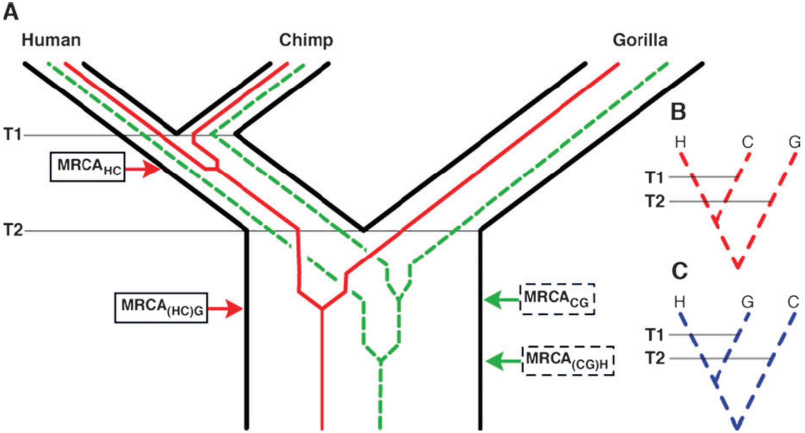
Topology	All (%)
H C G O R	20 (0.17)
H C G O R	9,148 (76.58)
H C O G R	19 (0.16)
G O C H R	0
G O H C R	1 (0.01)
H O C G R	5 (0.04)
H O C G R	0
H O G C R	0
C G O H R	4 (0.03)
C G H O R	1,369 (11.46)
H G C O R	13 (0.11)
H G O C R	5 (0.04)
H G C O R	1,361 (11.39)
C O G H R	0
C O H G R	0

Incomplete Lineage Sorting

This pattern is known as incomplete lineage sorting and is commonly found in closely related species (but note that we are going back 16 million years for the common ancestor of humans and orangutans, so they don't have to be very closely related. . .).

What explains this pattern?

Coalescences in closely related species I

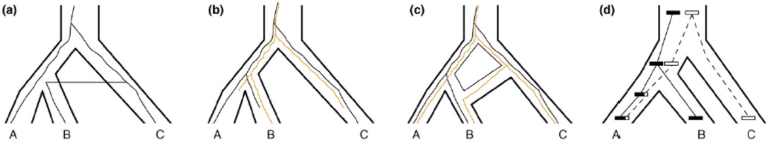


Coalescences in closely related species II

- If we take a sequence in a chimp and one in a human, they necessarily have to wait until they are back into the common ancestor of both before they can coalesce.
- However if they don't coalesce before the time of splitting with the gorilla then the human, chimp and gorilla sequences are just random samples from the ancestor of all three. So there is an equal chance of all three topologies.
- That is why the proportion of sequences that support ((human, gorilla),chimp) is very close to the proportion that support ((gorilla,chimp), human), because they had an equal chance of occurring in the ancestral population.
- The ((human,chimp),gorilla) topology is commonest because most of the time the sequences coalesce in the human/chimp ancestor before getting back to the ancestor with gorillas.

Multispecies Coalescent

- The genealogy of gene sequences from many species is often referred to as the multispecies coalescent.
- A nice general review is by Degnan and Rosenberg (TREE, 2009).
- Degnan and Rosenberg point out that a number of evolutionary factors can give rise to discrepancies between loci in the phylogenetic trees:
  - a Horizontal gene transfer
  - b Gene duplication and loss
  - c Hybridisation
  - d Recombination



(From Degnan and Rosenberg)

Ancestral inference with phylogenetic samples I

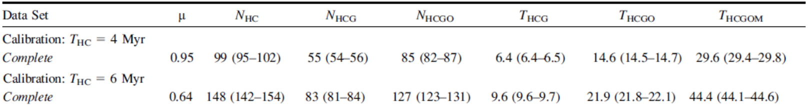
- It is possible to infer the ancestral population sizes and splitting times from just single sequence copies taken from each species, provided multiple loci are used.
- Intuition: for sequences from a pair of species (provided there is no gene flow. . .) coalescence can only occur in the common ancestor.
- So for each locus the coalescence time is determined by speciation time + coalescence time in the ancestor.
- The speciation time is a fixed quantity that is the same for all loci.
- On the other hand, the coalescence time in the ancestor is random and varies between loci and depends on the effective population size in the ancestor.
- So if the effective size of the ancestor is large compared to the speciation time, then we expect a lot of variability between loci in pairwise divergence of sequences taken from two different species because the branch lengths vary a lot.
- Otherwise, if it is small, then the only variability should come from randomness in the number of mutations, but the branch length will be very similar for each locus.

Ancestral inference with phylogenetic samples II

- The arguments in the previous slide were first noted by Takahata (1986), and since then there have been many studies that have attempted to estimate ancestral population sizes and divergence times, particularly for humans and apes.
- The MCMC coalescent-based method of Rannala and Yang (2003) is often used to obtain Bayesian estimates of these parameters. This is implemented in the program MCMCcoal.
- For example Burgess and Yang (2008) used this approach to estimate effective sizes and divergence times in humans and apes.

Estimation of effective sizes and divergence times in apes and humans

- Burgess and Yang use the chimpanzee divergence to calibrate the mutation rate (which is why the human-chimp divergence time is not shown in the table).
- The estimates below are for two different opinions on the human-chimp divergence time.
- The mutation rate is in units of 10^{-9} per base position per year. Population sizes are in 1000s. Times are in millions of years.



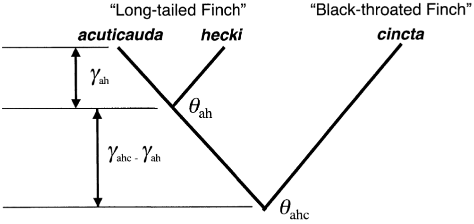
Data Set	μ	N_{HC}	N_{HCG}	N_{HCGO}	T_{HCG}	T_{HCGO}	T_{HCGOM}
Calibration: $T_{\mathrm{HC}} = 4$ Myr							
Complete	0.95	99 (95–102)	55 (54–56)	85 (82–87)	6.4 (6.4–6.5)	14.6 (14.5–14.7)	29.6 (29.4–29.8)
Calibration: $T_{\mathrm{HC}} = 6$ Myr							
Complete	0.64	148 (142–154)	83 (81–84)	127 (123–131)	9.6 (9.6–9.7)	21.9 (21.8–22.1)	44.4 (44.1–44.6)

Example: Australian Grass Finches

- Jennings and Edwards (2007) used Sanger sequencing to genotype 30 anonymous sets of sequences (‘loci’; average size around 500bp) in one representative individual from each of 3 closely related species of Australian grass finch (*Poephila acuticauda*, *P. hecki*, and *P. cincta*.
- They appear to have chosen one of the two homologous sequences per individual per locus (remember each bird will have two copies; the paper is unclear how they chose this).
- So the data are similar to human/ape examples discussed earlier.

Australian Grass Finches: data analysis

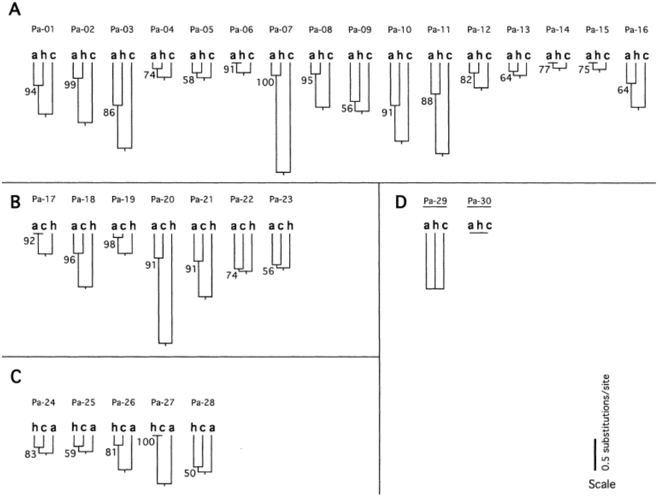
Jennings and Edwards (2007) use a MCMC program (MCMCcoal; Rannala and Yang, 2003) to fit the parameters of the following model:



- They have already decided on the tree topology.
- They want to estimate the population size multiplied by mutation rate for the ancestor of long-tailed finches and the ancestor of all three finches.
- Estimate the time interval (multiplied by mutation rate) of the two speciation events in the history of these three species.
- They will use an estimate of mutation rate to then get the ancestral N_es and times.

Australian Grass Finches: gene trees

These trees have been estimated by maximum likelihood. Note the different branch lengths and topologies.


A
Pa-01 Pa-02 Pa-03 Pa-04 Pa-05 Pa-06 Pa-07 Pa-08 Pa-09 Pa-10 Pa-11 Pa-12 Pa-13 Pa-14 Pa-15 Pa-16
a h c
94
99
86
74
58
91
100
95
56
91
88
82
64
77
75
64
B
Pa-17 Pa-18 Pa-19 Pa-20 Pa-21 Pa-22 Pa-23
a c h
92
96
98
91
91
74
56
C
Pa-24 Pa-25 Pa-26 Pa-27 Pa-28
h c a
83
59
81
100
50
D
Pa-29 Pa-30
a h c
0.5 substitutions/site
Scale

Australian Grass Finches: parameter estimates

Using an estimated mutation rate of $\mu = 3{\cdot}6 \times 10^{-9}$ substitutions/site/ year, and a generation time of one year. They combine the information from all the loci to obtain the following estimates:

- Interval between ancestral speciation events: 0.10 my (0.01 my to 0.34 my)
- Interval up to the *acuticauda-hecki* speciation: 0.61 my (0.35 my to 0.86 my)
- Effective size of the *acuticauda-hecki* ancestor: 98,889 (12,222 to 275,278)
- Effective size of the common ancestor of all three species: 373,056 (240,625 to 538,611)

So you can see that there is quite a lot of error in these estimates (the Bayesian credible intervals are in parentheses), and more loci are needed, but in principle the information is there to be able to make these estimates.

Implications for phylogenies I

- The phylogeny you get may depend on the gene you choose.
- Phylogenetic divergence times necessarily include the ancestral coalescence time, which depends on its effective population size.
- Degnan and Rosenberg (PLOS Genetics, 2006) have proved that in species trees with 4 or more species there are always sets of parameters where the most probable gene tree is not the species tree.

Implications for phylogenies II

- Phylogeneticists often use a 'majority-rule' consensus method for combining phylogenies. In this case Degnan and Rosenberg's result says that you will definitely converge on the wrong phylogeny as you add more loci.
- Another commonly-used method in phylogenetics is to concatenate sequences together and construct a phylogeny on that.
- Kubatko and Degnan (2007) show that the range of parameter values that lead to the 'majority-rule' failing will also lead to similar problems of estimation for the concatenation method, with the most strongly supported tree being the wrong one.
- **So population genetics is important!**

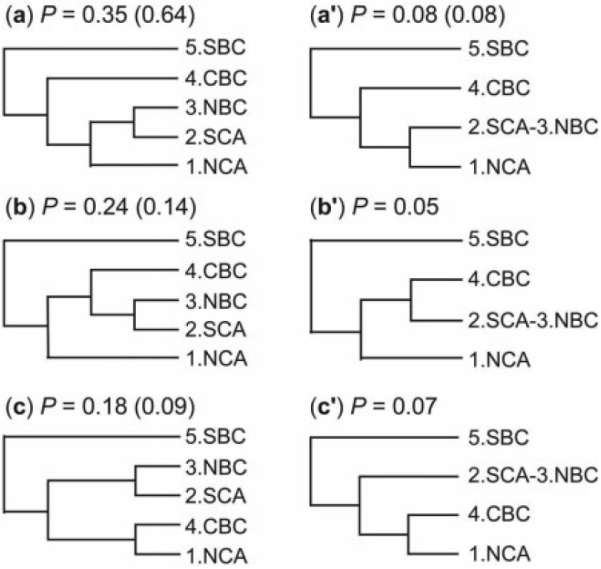
Bayesian Phylogenetics and Phylogeography (BPP)

- These results suggest that we should aim to carry out phylogenetic analysis using a coalescent framework.
- Unfortunately this is difficult to do because the space of genealogies is very large.
- There is a great deal of research in this area. Many contributions have come from the collaboration of Ziheng Yang (UCL) and Bruce Rannala (UC Davis), who have developed increasingly powerful MCMC methods, implemented in their BPP package.
- The BPP package finds natural groupings in the data, without the need for prior species-designation, using the multi-species coalescent model.

Example: Analysis of Coast Horned Lizard

- This example comes from Rannala and Yang (MBE 2014).
- Californian lizard; samples originally split by mitochondrial clades into 5 groups.
- Question: how many species do we actually have and what are their relationships?
- Data: 2 nuclear loci (529bp and 1,100 bp); sample size around 130 sequence copies of each.
- Conclude that a 5-species model fits better than a 4-species model.
- The inferred phylogenies are shown in the next slide (the P-values are Bayesian posterior probabilities)

Example: Analysis of Coast Horned Lizard



Conclusions

- Taking a population genetics approach to phylogenetics is very challenging.
- There is going to be difficulty extending MCMC methods to many species, using many loci.
- Current models are simple splitting models, and ignore migration, and many other aspects of demographic history.
- Perhaps other, more approximate, methods will work better?

Further Reading I

- Burgess, R., & Yang, Z. (2008).
  Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors.
  Molecular biology and evolution, 25(9), 1979–1994.

- Degnan, J. H., & Rosenberg, N. A. (2006).
  Discordance of species trees with their most likely gene trees.
  PLoS Genet, 2(5), e68.

- Degnan, J. H., & Rosenberg, N. A. (2009).
  Gene tree discordance, phylogenetic inference and the multispecies coalescent.
  Trends in ecology & evolution, 24(6), 332–340.

Further Reading II

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M., & Von Haeseler, A. (2007).
Mapping human genetic ancestry.
Molecular Biology and Evolution, 24(10), 2266–2276.

Jennings, W. B., & Edwards, S. V. (2005).
Speciational History of Australian Grass Finches (Poephila) Inferred from Thirty Gene Trees.
Evolution, 2033–2047.

Kubatko, L. S., & Degnan, J. H. (2007).
Inconsistency of phylogenetic estimates from concatenated data under coalescence.
Systematic Biology, 56(1), 17–24.

Further Reading III

Rannala, B., & Yang, Z. (2003).
Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci.
Genetics, 164(4), 1645–1656.

Satta, Y., Klein, J., & Takahata, N. (2000).
DNA archives and our nearest relative: the trichotomy problem revisited.
Molecular phylogenetics and evolution, 14(2), 259–275.

Takahata, N. (1986).
An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced.
Genetical research, 48(03), 187-190.

Further Reading IV

Yang, Z., & Rannala, B. (2014).
Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci.
Molecular Biology and Evolution, 31(12), 3125–3135.