Genetics 453 J. Felsenstein
Evolutionary Genetics Winter, 2002

Outline of lectures 17-19

Molecular Evolution and Phylogeny

1. With the the development of protein sequencing methods at the end of the 1950's, people started comparing different sequences from different species. Linus Pauling and Emil Zuckerkandl pioneered the molecular study of evolution, along with others such as Margaret Dayhoff, Russell Doolittle, and Morris Goodman. This literature developed independently of the controversies over electrophoretic data in population genetics.

2. Initially this work was very slow–the first DNA sequenced, by Gilbert, took 2 years for 20 base pairs–but the rate of accumulation of molecular data has increased exponentially.

3. When there are two sequences we can simply align them and look at their similarities or differences (how many, where they are in the molecule, and so on). But when there are more than two sequences, they are related to each other as a result of a branching evolutionary history, and they are located at the tips of a tree of species. (This may not be true of sequences from within a species, or where horizontal transfer is occuring.)

4. Phylogenies (evolutionary trees) can be inferred from molecular sequences. They can be used to address a wide variety of questions about the course of evolution. They are needed not only to know the relationships of organisms, but to enable us to correctly interpret rates of evolution and the evolution of different parts of different molecules.

5. In examining a phylogeny diagram, it is important to look at the branching pattern and not be misled by the order of species names on the page, or the left-right order of branching of lineages.

6. Rooted trees indicate the direction of evolution, and the ancestor: unrooted trees only show patterns of relationship. Several rooted trees correspond to each unrooted tree, and biologically speaking rooted trees are much more useful: unfortunately most methods produce unrooted trees.

7. The root can be placed on a tree by two methods

   - The outgroup method. This amounts to knowing in advance where it is. For example we may have a tree with 5 great apes and one monkey, and know in advance that the first split in the tree separates the monkey from the apes.

- A molecular clock. If we assume that rates of change are about the same up all lineages, then the root will be a point which is equidistant from all the tips. It is not necessarily easy to see where this is. The *molecular clock* is the assumption that change occurs at a constant expected rate, so that all tips of the tree are equidistant from the root. This is better the more closely related the species are, but breaks down as their biology becomes more different. It is often a useful approximation.

8. If we have (say) DNA sequences, we can look site by site. At each site, the different bases partition the species into groups. When these partitions correspond to branches in the phylogeny, then that character supports that phylogeny, as it can evolve on it with no state having to arise more than once.

9. When there is conflict among characters as to what phylogenies they support, we need some way of choosing among them a best estimate. *Maximum parsimony* or simply *parsimony* chooses that phylogeny on which the characters can evolve with the fewest evolutionary events.

10. Characters that make a difference to a parsimony method are usually (and somewhat misleadingly) called *phylogenetically informative*. Many characters will not make a difference. For example, a site where all species have `G` except for one species which has an `A` can always evolve with only one change of state, no matter what the tree. Hence it does not affect the choice of tree by a parsimony method.

11. Such characters do, however, affect the outcome of other methods based on similarity or distance. Whether or not it is reasonable to use such characters is central to the debate between advocates of parsimony methods and other methods.

12. For small cases we can consider all possible trees: but above about 20 tips we must pick and choose which trees to evaluate, and we may miss the best tree. This is unavoidable because the number of possible trees is overwhelmingly large–there are more possible trees of 30 tips than atoms in the universe.

13. Parsimony actually infers an unrooted tree, because where the root is does not affect the number of changes of state (the number of DNA base substitutions, for example) that are needeed on the tree. It just alters our interpretation of which direction some of the changes were in.

14. Parsimony has one major weakness. It can be subject to a bias, where long branches that are near one another in the phylogeny, but not immediately adjacent "attract" each other by picking up parallel changes that make the two species appear to be related. Under unfavorable circumstances, this can lead to more evidence appearing in the data in favor of the wrong result than appears in favor of the trues groupings, and this is not cured even when vast amounts of data are collected.

15. Some other methods:

- Distance matrix methods. A table of pairs of "distances" between species is constructed. The phylogeny that does the best job of predicting these is the one that is preferred. The distances must be corrected for unobserved changes ("multiple hit correction"), and the prediction of them is made by adding up branch lengths along that path of the tree, bertween the two species.
- Invariants or "evolutionary parsimony". A less used method, too complex for discussion here.
- Maximum likelihood. The probability of the data is computed, given the tree and a probabilistic model of evolution. We choose that tree that gives the highest probability to the observed data, among all trees.

16. Parsimony has one major weakness. It can be subject to a bias, where long branches that are near one another in the phylogeny, but not immediately adjacent "attract" each other by picking up parallel changes that make the two species appear to be related. Under unfavorable circumstances, this can lead to more evidence appearing in the data in favor of the wrong result than appears in favor of the trues groupings, and this is not cured even when vast amounts of data are collected.

17. When trees are constructed for various molecules, it is found that we can use them to infer evolutionary history of those groups.

Rates and Causes of Molecular Evolution

18. Different parts of the genome are useful for different problems. fast evolving sequences are useful for recent events, but become saturated and unrecognizable when comparing more distant relatives. Slow evolving sequences are useful around the base of the tree, but don't have any variability at all among close relatives.

19. When we use multiple loci that may have duplicated from each other (such as the different hemoglobin loci) we see trees that show not just speciation events, but also gene duplication events. Once a gene duplicates, the two trees (one for each locus) can have the same species in them – hopefully one finds the same relationships between those species using the hemoglobin α locus as using the hemoglobin β locus.

20. Different regions of molecules evolve at different rates. A summary of the patterns is:

    - DNA distant from genes evolves very quickly (at about one substitution per 10^8 years),
    - Flanking regions upstream and downstream from a gene evolve less quickly than that,
    - Introns evolve less quickly than those, though not much less,
    - Third positions of codons evolve less quickly than introns,
    - First and second positions of codons evolve less quickly than that,
    - Within a protein,

  - active sites evolve very slowly,
  - sites that bind heme, or interact with other proteins evolve a bit faster but also very slowly,
  - interior sites evolve less quickly than exterior sites,
  - substitutions that involve less radical changes of the amino acid (i.e. that change to a rather similar amino acid) happen more readily.
- Of base changes, transitions (`A` $\leftrightarrow$ `G` or `C` $\leftrightarrow$ `T`) happen several times more readily than transversions (all other changes).
- Between protein-coding loci, some (fibrinopeptide, for example) evolve rapidly, some less so (hemoglobins, cytochromes), and some (histones, for example) change very slowly.

21. *(this wasn't covered in lecture).* With species, polymorphism is found in the DNA just as it was already known to be in protein coding loci. If we compare two sequences in the same population, about one nucleotide every 1,000 shows a difference between the two sequences. Within protein coding regions, studies such as Marty Kreitman's pioneering work on Alcohol Dehydrogenase (ADH) show that much of this variation is in 3rd codon positions and little or none in 1st and 2nd positions.

22. Motoo Kimura's work on the neutral theory of molecular evolution brought together within- and between-population studies. The neutral theory not only predicts the level of within-species heterozygosity, it also predicts the rate of molecular evolution. With neutral mutation rate μ_n at a locus, if there are N individuals we expect $2N\mu_n$ neutral mutations per generation. A fraction $1/(2N)$ of these will fix (simply because a neutral gene's probability of fixation is its initial gene frequency). The result is that the neutral mutation rate μ_n is also the predicted rate of neutral substitution at that locus.

23. The same result can be seen y considering the lineage of genes that stretch between one species and another – the number of differences will be the number of mutations expected to occur on that lineage.

24. If instead there were selection, we can use a 1929 result of J. B. S. Haldane that a mutant (in a large population) that has selective advantage s in heterozygote will have approximately probability $2s$ of escaping the random genetic drift that could cause it to get lost, and instead becoming established in enough copies that they spread into the population. Then the rate of substitution expected if there is a rate mu_a of advantageous mutations is $2N\mu_a \times 2s = 4N\mu_a s$.

25. Even though most favorable mutants thus actually are lost (if $s = 0.01$, 50 get lost for every one that becomes established), selection thus has a major effect. If $\mu_a = \mu_n$, for example, the advantageous mutations would substitute at a rate $4Ns$ higher.

26. Does this mean that most substitutions are advantageous? Not necessarily, because μ_a may be much (much) less than μ_n.

27. In fact most mutations in protein coding regions will probably be deleterious. Both neutral and non-neutral theories allow for this. Most of the inequalities of rate that we have just talked about are due to deleterious mutations being screened out. In fact a deleterious mutation that has selective disadvantage s will be kept from fixing if $s > 1/(4N)$, which can be quite a small number. So with $N = 1,000,000$, a deleterious mutation must reduce fitness by less than 1/4,000,000 before it will act as if neutral.

28. In the end, the theory of molecular evolution is frustratingly unable to tell us whether the substitutions we actually see are advantageous or not. We have no easy way to find out at what rate evolution is improving the function of molecules. This means that the theory and empirical study of molecular evolution does not deliver as much insight into the rate of progress in evolution as one might hope.