GENOME 453 J. Felsenstein
Evolutionary Genetics Autumn, 2015

Outline of lectures 16-19

Molecular Evolution and Phylogeny

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

2. Initially this work was very slow – early protein sequences had to be done with massive amounts of protein and direct chemical reactions (sperm whale myoglobin was the first globin sequences, and they used a ton of sperm whale meat bought from the whalers). The first DNA sequenced, by Walter Gilbert in the late 1970s, took 2 years for 20 base pairs–but the rate of accumulation of molecular data has increased exponentially.

3. When there are two sequences for the same locus in different species, we can simply align them (arrange them to match up the corresponding positions) and look at their similarities or differences (how many, where they are in the molecule, and so on). But actually 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. For more than two sequences from the same locus one needs to take the tree into account to understand the differences.

4. The trees from different loci are all expected to be the same. Within species, there are also treelike genealogies of copies of a gene, and a treelike pattern of descent of loci owing to gene duplication. We will talk about all three of these kinds of trees, separately.

5. DNA and protein sequences show strng signals of evolution. The more closely related species are, the more of their genomes will be alignable, and the more similar the alignable regions will be. Humans and chimps, for example, show 98.77% sequence similarity in alignable regions.

6. Is this because evolution has kept those two sequences similar? No, it is just that they have not had enough time to diverge more.

7. **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.

8. If we have (say) DNA sequences, we can look site by site. At each site, the different bases partition the species into groups. If the differences at a site arose by a single change in a single branch of the tree, the partitions correspond to a branch 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. There are a number of different methods: maximum parsimony, distance matrix methods, maximum likelihood methods, and Bayesian inference. *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 there are so vastly many (more than 10^{21}) that 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 there are 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. I will show an example with 5 species and 6 sites where we can calculate how many changes are needed (at a minimum) to explain the evolution of those sites on one particular tree (out of the 15 possible unrooted bifurcating trees). It turns out to need 9 changes. That is the parsimony score of that tree. Another tree turns out to need only 8 changes,, and it is the most parsimonious tree.

15. 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 accidentally 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.

16. 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.
- 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.
- Bayesian inference. Similar to Maximum Likelihood, and using the same models of evolution, it adds a distribution of prior belief among all trees. Then we can actually compute the probabilities of different trees given the data.

17. 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.

18. You can think of any branch as able to rotate, to flip the order of species beyond the branch. This a tree growing vertically with Gorilla, on the left and a group consisting of Human and Chimp on the right: (Gorilla,(Human,Chimp)) is really the same tree as one where the Gorilla lineage comes off on the right at that split: ((Human,Chimp),Gorilla), or where the order within the other group is Chimp-Human instead of Human-Chimp. Thus in this simple case, the order across the top of the tree could be any of four different orders and we would still be seeing the same tree. Beware of apparent "trends" that are really just a result of how the branches are flipped at the forks!

19. 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. Biologically speaking, rooted trees are much more useful: unfortunately most methods produce unrooted trees.

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

    - *The outgroup method.* If we have an unrooted tree, we can add to the analysis another species which we know is less related than any of the other species. This amounts to knowing in advance where it is. For example we may have an unrooted tree with 5 great apes and we add 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 (approximately) 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.

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

22. In many cases the molecular trees are very consistent with morphological trees that were over 100 years old.

23. An example is given in the lecture of the use of large-subunit ribosomal RNA sequences to investigate the relatives of the vertebrates. The outgroups were arthropods (a spider *Eurypelma* and the beetle *Tenebrio*) and molluscs (a scallop, *Placopecten* and a land snail, *Limicolaria*). The root of the tree is believed to be near the node that joins these two groups to the rest of the tree.

24. Within the deuterostomes (the vertebrates and their relatives), the acorn worm *Saccoglossus* is near the echinoderms (the sea urchin *Strongylocentrotus* and the brittle star *Ophiopholis*). There is some difference between different methods as to where the amphioxus *Branchiostoma* splits off, but there are hints it is more distant from us than are the tunicates (sea squirts) such as *Styela* and *Herdmania*).

25. Closer to us are the hagfish (*Myxine*) and the lamprey *Petromyzon*). These have traditionally been regarded as splitting off one after the other, with only the lampreys being true vertebrates. But there is increasing recognition that they are "sister groups" who are more closely related to each other than to us. So they are both vertebrates.

26. Within the vertebrates there is little resolution from these data, with only hints that the dogfish shark (*Squalus*) is least closely related to us, the rockfish (*Sebastoles*) next least close, and finally the remarkable no-longer-extinct coelacanth (*Latimeria*) closest to us. It is a lobe-finned fish with four "legs", more closely related to us than to ray-finned fishes such as the rockfish, and more closely related to us than to any other fish except for the even more-closely related lungfishes, who were not in this study. Who is "us"? For the purposes of this study we might as well be represented by our fellow tetrapod, our cousin the South African Clawed Frog (*Xenopus*).

27. In the lecture, I show an example of using two loci (essentially randomly picked from the alignments at the `Orthomam` web site, to infer the phylogeny of mammals. The first locus does badly, the second better, but when combined they do better yet.

28. Here are some major conclusions of four decades of work on molecular phylogeny:

    - Using immunological distances, Morris Goodman (1962 on) and later Wilson and Sarich (1966) showed that humans, gorilla, and chimps were a clade.
    - Wilson and Sarich (in that work, 1966) dated the divergence of humans from them to 5 million years.
    - Charles Sibley and Jon Ahlquist (1984) use DNA hybridization to argue for the clade humans-chimps.
    - Carl Woese (1978) used rRNA trees to (in effect) introduce evolution into microbiology, and to argue for the domain Archaea as separate from the Bacteria (the third major domain is the eukaryotes, Eukarya).

- There has been much progress on early radiation of angiosperms.
- The protostome-deuterostome tree of metazoans has been (more or less) replaced by deuterostome-lophotrichozoa-ecdysozoa tree. (What has really happened is that the protostomes have been split up into two groups, lophotrichozoa and ecdysozoa, with some groups such as nematodes and flatworms joining these two groups instead of splitting off before the deuterostomes did.
- Fungi are closer to animals than either is to plants.
- The symbiotic origin of mitochondria and of chloroplasts has been verified.
- The amphioxus diverged before split of tunicates from craniate chordates.
- Lots of horizontal gene transfer in prokaryotes, almost not a tree.

29. The inference of molecular phylogenies has become a major industry in biology. Most recently they are being used to detect regions of genomes that have low rates of evolutionary change, which indicates that these regions are functionally important and kept from changing by natural selection. For example, in the Human Genome Browser at UC Santa Cruz, multiple species comparisons using phylogenies are used to detect regions that have lower rates of change.

Rates and Causes of Molecular Evolution

1. Different parts of the genome are useful for different problems. fast evolving sequences are useful for recent events, but become saturated with overlying changes and unrecognizably different when comparing more distant relatives. Slow evolving sequences are useful around the base of the tree, but may not have any variability at all among close relatives.

2. 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.

3. 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.

4. 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). That is true because with neutrality, random genetic drift is equally likely to have any of the $2N$ copies in the population be the one that leaves all the descendants. So the rate of occurrence of new mutations that are destined to fix is $2N\mu_n \times \frac{1}{2N}$ The result is that the neutral mutation rate μ_n is also the predicted rate of neutral substitution at that locus.

5. The same result can be seen by 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.

6. If instead there were natural selection, we can use a 1929 result of J. B. S. Haldane that a mutant (in a large population) that has selective advantage s when present in one copy, in a 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 μ_a of advantageous mutations is $2N\mu_a \times 2s = 4N\mu_a s$.

7. 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 than expected if they were neutral.

8. Does this mean that most substitutions are advantageous? Not necessarily, because μ_a may be much, much smaller than μ_n.

9. In fact many 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.

10. Slightly deleterious mutations, ones that have selective disadvantage less then $1/(4N)$, can get fixed, so there will be some gradual deterioration of adaptation due to this, but it will be offset by occasional favorable mutations.

11. The Molecular Clock (which was mentioned previously) is the suggestion that genes change at a regular rate through time. It is not really suggested that the changes occur regularly, but rather that their rate is the same in different lineages. It is close to the assertion that two rock samples have the same radioactivity (and hence will cause the same average rate of clicks on a Geiger counter). The substitutions are random in time, like those clicks.

12. Different genes change at different rates – no one suggests that one clock applies to all loci.

13. At the same locus, closely related lineages change at very similar rates. As the biology of the species becomes different, the rates of change gradually become different.

14. Closely related species thus give a tree that is clocklike, in the sense that different tips of the tree are approximately equidistant from the root of the tree.

15. In the theory of molecular evolution it has been difficult to tell whether the substitutions we actually see are advantageous or not, but ...

16. Recently the relative rates of synonymous changes (that do not change the amino acid) and nonsynonymous changes (ones that do) have been used to detect evidence of positive selection. If a lineage has an excess of nonsynonymous changes, this indicates that it has been under positive selection for change.

17. This will not be true for all parts of the molecule, and it will also vary among lineages. This makes positive selection hard to detect.

18. Aside from that method we have few ways 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 naïvely hope. We can make predictions about what won't happen (substitutions that disrupt molecular function) more easily than we can predict where there will be improvements in molecules.

19. In addition to speciation events, there is another kind of branching event in molecular evolution: a gene duplication. If a locus duplicates in an individual, and thereafter there are two loci in different places in the genome (and that state fixes in the population) then both loci are carried on into both descendants at each speciation (fork) in the tree.

20. If we simply take all the copies of this "gene family" and make a tree, some splits will be the gene duplications. Beyond them the speciation events make identical splits in both subtrees. In a hypothetical example, a single locus is inherited by a frog, but a lineage ancestral to mammals then suffers a gene duplication, so there are now two loci, A and B, in all three species (human, monkey, and squirrel) descended from that point in the tree. The "species tree" (phylogeny) of these three species is supposed to be identical in the subtree of all A sequences and the subtree of all B sequences.

21. An example is given of a classic 1978 tree by Morris Goodman and his colleagues of the globin gene family. Myoglobin can be seen separating from the hemoglobins in a gene duplication after the lamprey lineage splits off, and somewhere in the fish α and β hemoglobins duplicate (there are also some more further on, such as γ (gamma) hemoglobin.

22. Sequences organized by gene families and aligned can be fetched from genomics sites such as Pfam (`http://www.pfam.org`) and a mammalian site, OrthoMAM (`http://www.orthomam.univ-montp2.fr/orthomam/html/`).

23. Thus there are trees within trees. Even within a single species such as humans, the loci (mostly) form a tree, descended from common ancestors by gene duplication events (and no speciation events). The similar trees in our relatives are not all independent, as those gene duplications are each shared, possibly by many species. Getting straight how these trees are interrelated is important, to avoid confusion.