

Molecular Evolution Lab

Designed for Biol 1B

By Weiwei Zhai

Fall 2005

Aims for this lab:

- 1) What is the molecular clock and its historical background
- 2) How to calculate the evolutionary rate for a protein assuming the molecular clock for pairs of sequences.
- 3) Rate variations among genes.
- 4) Demonstrate the molecular clock.
- 5) How to build an ultrametric “clock like” tree based on distances.
- 6) Use phylogenetic trees to date historical events.
- 7) Nonsynonymous/synonymous substitutions and the genetic code

1) Molecular clock and its historical background:

Molecular clock assumes: genes or genomes are evolving at a **relatively** constant rate over time. In other words, the number of substitutions at the DNA and amino acid sequence level that build up between pairs of sequences increases linearly with time since they diverged from each other.

This observation was first observed by Zuckerkandl and Pauling in 1962 based on protein sequence data. Ever since then, the idea of a Molecular Clock has been of great interest in Molecular Evolution and Population Genetics. In part based on the molecular clock observation, Kimura first proposed the evolutionary neutral theory in 1968. (Refer to page 506 of the 7th edition, page 503 in the 6th edition).

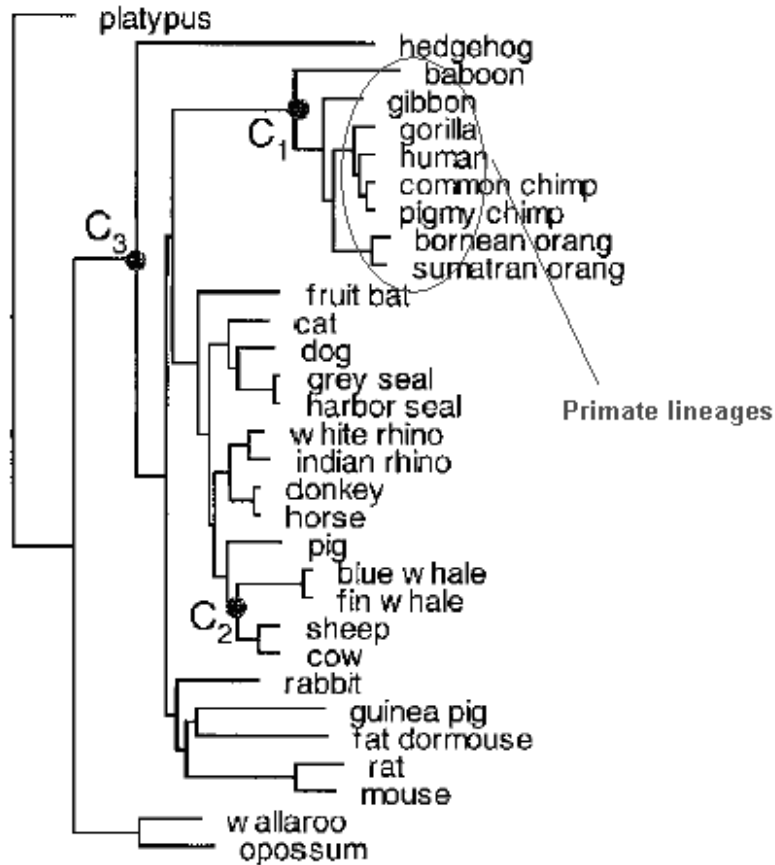
Phylogenetic trees drawn under the assumption of the molecular clock criteria are called “ultrametric trees”. For example, the following phylogenetic tree of the Placental Mammals is an ultrametric tree. (Concepts about ultrametric trees are available on page 500 7th edition for more details;* it is not available in the 6th edition).



<http://tolweb.org/tree?group=Eutheria&contgroup=Mammalia>

Note: the molecular clock assumption implies a relatively constant rate of evolution. In other words, if we pass the same amount of time, we will observe roughly the same amount of changes along the history of that protein or segment of DNA. (There will be random variation around this averaged evolutionary rate). This necessarily leads to the following: the root or any internal nodes are approximately equally far away from their current descendants respectively. This is illustrated in an ultrametric tree. Why is this so?

Although there have been a lot of examples supporting the molecular clock, there are also a lot of instances where the molecular clock assumption is violated. (Remember, we need the evolutionary rates to be **relatively** constant for all lineages in the evolutionary history. This is really a very strict condition.) For example, researchers have used the mitochondrial genome to build a phylogenetic tree for mammals. Much higher evolutionary rates are observed for lineages leading to primates.



Adopted from Yoder, A.D. and Z. Yang, 2000 with modification

Reasons for deviation from the molecular clock are complicated and we don't fully understand the reasons for differences in evolutionary rates in some groups. This also

depends on the scale of the time region you are looking at. Remember, we need evolutionary rates to be roughly the same for all the branches. The more species you compare, the more likely evolutionary rates in some lineages will differ from the others. Researchers have also developed methods to check the molecular clock assumption over the whole phylogenetic tree (global clock, for example the relative rate test) or part of the tree (local clock). These materials are currently beyond the scope of this course. In the literature, the phenomenon of a non-clock like tree is called rate variation among lineages. We will use a simple example to teach you how to draw an ultrametric tree if the molecular clock assumption is met.

For the rest of this lab, let's assume the molecular clock criterion is met. (Before you perform any analysis, always keep in mind the assumptions and possible reasons they may not hold).

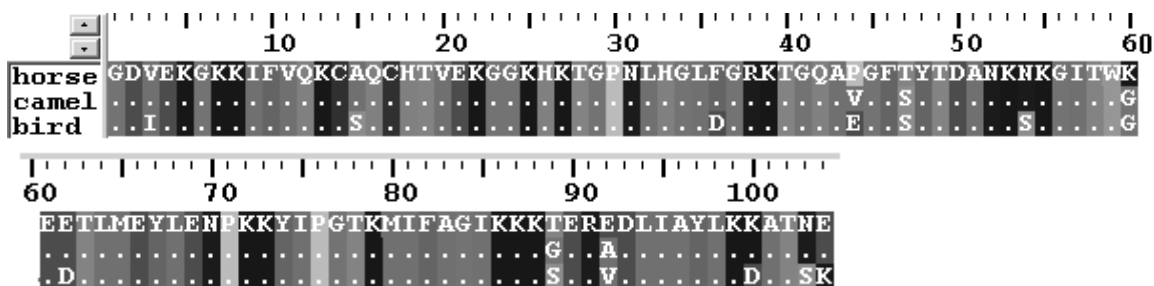
2) Calculate the evolutionary rate between pairs of sequences.

We have the amino acid sequences of the Cytochrome C gene from three species: horse, camel and bird. The sequence data is available for a large number of species, we are just using these three for illustrating some points. This gene is the terminal enzyme in the respiratory chain, located in the inner membrane of mitochondria and bacteria (Page 172 7th edition; page 168 6th edition).

Note: The amino acid sequences are abbreviated according to standard notation in the following table (page 79 7th edition; 72-73 6th edition).

Residue	Ala	Arg	Asn	Asp	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Abbr	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V

The dot in the sequences below means that in those amino acid position they are the same as the first sequence.



Step 1: Calculate p = the percentage of sites that are different.

1) Count how many amino acid sites are different between the three pairs of sequences and fill in the following table. (Keep in mind, the difference between A and itself is always 0 assuming small variation within species, also the difference between A and B is the same as the difference between B and A, so you only need to fill in either the upper or lower diagonal of the table.)

	Horse	Camel	Bird
Horse	0		
Camel		0	
Bird			0

Table 1: pair wise differences between three taxa

2) How many amino acids are there in the sequences?

3) What are the percentages of sites that are different between the three different pairs of sequences?

Horse vs. Camel:

Horse vs. Bird:

Camel vs. Bird:

Step 2: calculate the rate of evolution

(note: we are ignoring back and convergent mutations)

The relationship between evolutionary rate λ and the percentage difference p is approximately: $2\lambda t = p$. Here t is the divergence time between the two species we are comparing. (Why is there a factor of 2?)

If we know from the fossil record that Artiodactyls (including camel, even-toed ungulates) diverged from Perissodactyls (including horses, odd-toed ungulates) about 90 MY ago (MY = 10^6 years), also we know birds diverged from mammals about 300 MY ago. What are the values of λ in the unit of per amino acid site per 10^9 years for the three comparisons? (**be careful with the units)

Horse vs. Camel:

Horse vs. Bird:

Camel vs. Bird:

Take an average of the three values. It tells us that on average, for each 10^9 years, single amino acid positions change λ times.

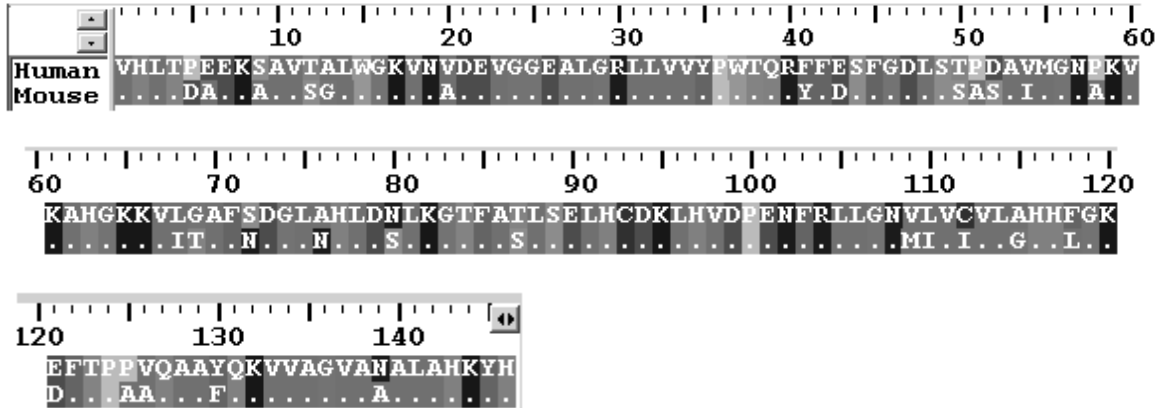
We need to wait for $1/\lambda * 10^9 =$ years on average for one change to happen in a single position.

3) Exercise/homework: Rate Variation among different genes

Although individual genes are evolving at an approximately constant rate, different genes show a tremendous amount of variation in evolutionary rate.

Here are two sequences of the hemoglobin beta from Human and Mouse which diverged from each other roughly 90 MY ago

Follow the previous example to calculate the evolutionary rate of this gene. (Is it different from Cytochrome C?)



A lot of effort has been devoted to calculating rate variation among groups of genes. Here is a table of protein evolutionary rate adapted from the literature.

Protein	Rate	Protein	Rate
Fibrinopeptides	9.0	Nerve growth factor	0.85
Growth hormone	3.7	Insulin	0.44
Ig lambda chain C region	2.7	Lactate dehydrogenase	0.34
Pancreatic hormone	1.7	Cytochrome C	0.22
Hemoglobin beta chain	1.2	Histone H4	0.01
Animal lysozyme	0.98	Ubiquitin	0.01

Rates of amino acid substitutions per amino acid site per 10^9 years in various proteins.

We notice from this table that the difference in the evolutionary rate between groups of genes can be several orders of magnitude from each other. Fibrinopeptides evolve 900 times faster than Histone H4. We also notice that the rate listed in this table for cytochrome C and hemoglobin are slightly different from the values we got from the examples. There are quite a few reasons for this. First of all, the results in the table are obtained from a much bigger dataset. It involves averaging over a lot of pairs of sequences from different species. Although the molecular clock assumption is not violated, it doesn't mean there is absolutely no variation in evolutionary rate among different parts of the evolutionary tree. (By this we mean they don't need to be for example 0.785 exactly everywhere, it can be 0.779 in this part of the tree and 0.812 somewhere else, but not a very different value such as 0.066. This is the reason we highlight the "relatively" in the first part when we introduce the definition of molecular clock. There are standard statistical methods we can employ to formally test the molecular clock hypothesis. For example, the relative rate test or likelihood ratio test. They are beyond the scope of this course). Furthermore, the method employed here is an approximation method. We have not taken into account the fact that there might be multiple, back and convergent mutations.

There are a lot of explanations for rate variation among genes in evolutionary biology. An intuitive explanation would be: for functionally less important genes, they have more freedom to change. (For example, Fibrinopeptides are peptides cleaved out of fibrinogen by Thrombin during the blood clotting process. They don't seem to have any particular function after cleavage. Presumably a lot of mutations that occur in Fibrinopeptides will not perturb the function of the protein that much. (page 882 7th edition; page 885 6th edition).

The difference in the evolutionary rates for different proteins provide different yard sticks that can help us measure time at different scales if the molecular clock assumption is met. (For those of you who have taken physics before, this is the same as radioactive decay. Different atoms have different rates of radioactive decay. They are used to date geological time.) For estimating divergence times of distantly related species using the molecular clock, you need a slowly evolving protein, e.g. Histone, whereas for closely related species, faster evolving proteins are much more helpful. Why is this so?

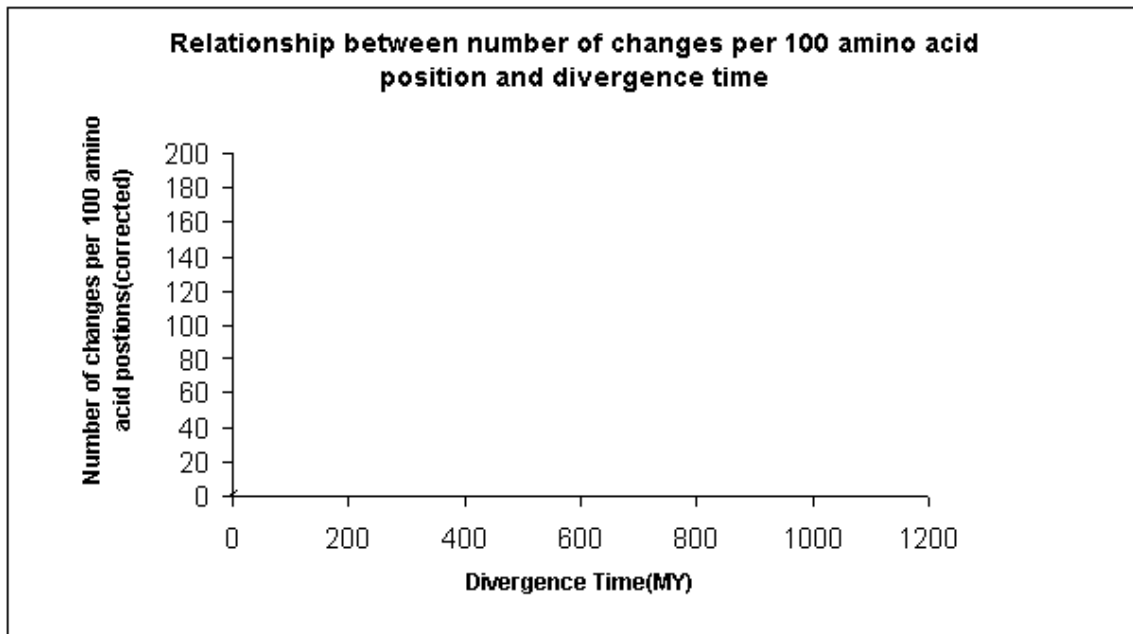
4) Molecular Clock

Imagine, you are a researcher 35 years ago like Dr. Dickerson, you obtained the following table of information for the hemoglobin gene family. (For details of the evolution of hemoglobins, please consult the textbook on page 379 7th edition; 359 6th edition)

Branch Point	Changes per 100 residues (corrected for multiple hits)	Age(MY)
Human, Gorilla β /other primates β	5.0	29
Primate δ /Primate β	7.0	40
Human γ /Mammalian β δ	31.6	180
Lamprey globin/Hemoglobins	138.6	780
Myoglobins/Hemoglobins	151.4	890
Lamprey Globin/Myoglobins	156.5	900
Hemoglobins/Insect globin	162.0	940
Myoglobins/Insect globin	163.5	970
Lamprey globin/Insect globin	174.2	1000

Modified from Dickerson (1971)

- 1) Plot the data points from the table on the following graph.
- 2) Fit a line **through the origin (why?)**. Is there a good linear relationship? Is this what is expected under the molecular clock?



5) How to build an ultrametric tree-a distance based method

With DNA/Protein sequence datasets of different organisms, how can we build a phylogenetic tree from this information? This has been an important topic in the area of Phylogenetics for several decades.

There are three major methodologies in building a phylogenetic tree based on different rationales respectively. We briefly summarize all three here. Students are not required to understand all of these except for the general principle of distance-based methods. But exposure to this information may stimulate some students in pursuing deeper interests in phylogenetics (page 501 in 7th edition, less in 499 in 6th edition).

The first type of method is based on a principle called **Parsimony**. Parsimony assumes that evolution takes the pathway that will minimize the number of required changes along the evolutionary history. Accordingly phylogenetic trees that require fewer changes are “better” in the parsimony sense than all possible alternative trees. Researchers have also developed algorithms to search over all possible trees (we call this the tree space) for the tree that will require the least changes. Parsimony has produced a lot of debate among researchers with respect to its philosophical and scientific merits. Trees built in the parsimony principle are called maximum parsimony trees.

The second type of method is based on the statistical rational called **Maximum Likelihood**. The likelihood is the “probability” of observing the real data given the parameters. In the first step, the evolution of sequences or morphological characters is modeled with specific probability models. These probability models assume specific parameters in the modeling process. Students can think of the likelihood as the probability of observing the real dataset under the statistical framework we assume. The likelihood methods are standard methodologies inherited from Statistics. For the situation here, we would simply search over all the tree spaces for the tree that will give the maximum likelihood. In other words, it is the tree that has the maximum probability of “producing” the dataset we observe in real life.

The last type of methods is called **distance based methods** (or sometimes, clustering methods). The method we introduce below is one of the distance methods. The process starts with measuring distances between all pairs of sequences resulting in a matrix of pair-wise distances: the distance matrix. The methodology involves grouping or clustering “closest” pairs of sequences (or taxa) progressively until we resolve into a phylogenetic tree. Besides the method we introduced here, there are also other distance based methods, e.g the Neighbor-Joining (NJ) method has the same philosophy but is based on a different definition of the closest pair of sequences. The distances matrix can be derived from nucleotide/amino acid sequences as we have done in the previous section. (Table 1) or from other factors, e.g, morphological data, immunological binding essays (so called immunological distances.)

We will introduce here a method called **UPGMA-(Unweighted Pair Group Method with Arithmetic Mean)** developed in the 1950’s (Sokal and Michener 1958). It is a clustering method assuming the molecular clock assumption (You will understand why

this is true once you learn the algorithm itself). The version we introduce here is just a simplified version of the general algorithm (please search “UPGMA” on the internet for those students who are really interested in developing methodologies.)

Since it is a distance-based method, it definitely starts with the distance matrix. Let’s imagine we obtained a distance matrix for three taxa as follows:

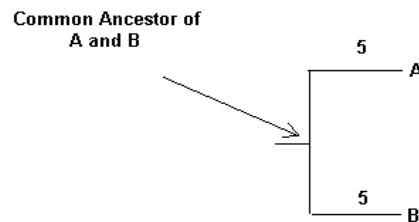
- 1) find the pair that is the closest.

	A	B	C
A	0	10	17
B		0	15
C			0

We find that distance between A and B is 10 and is the smallest distance among the three distances.

- 2) Connect the closest pair to their common ancestor each with half the distance between them.

$10/2=5$. So we build a tiny tree like this:

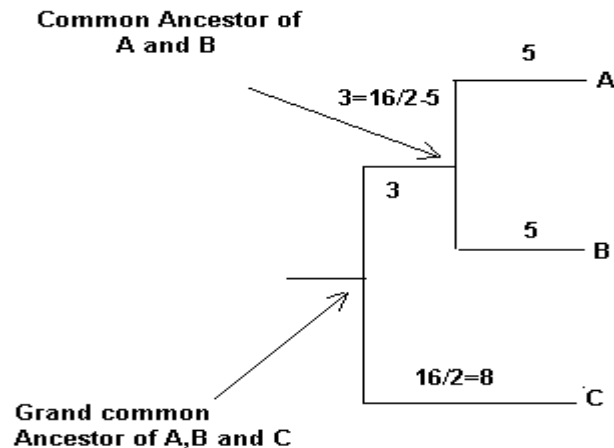


- 3) Average the distances between the taxa that are left (C) and the two taxa we already joined.

$(17+15)/2=16$. That’s the average distance between A and C and B and C.

- 4) Connect C and the group AB to their grand common ancestor, each with half of the average distance in step 3.

We resolve the final tree like this:



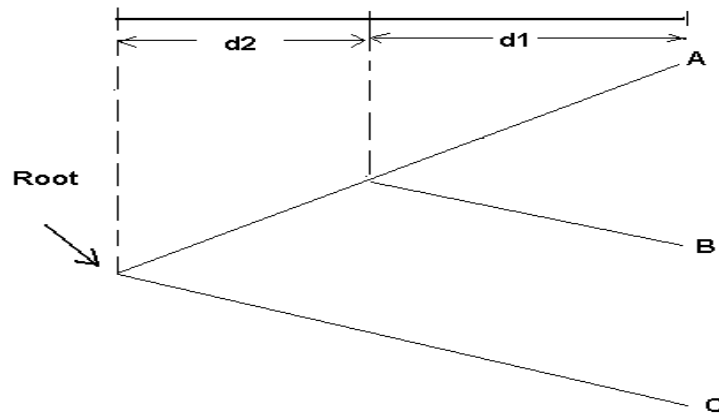
Numbers marked on top of the horizontal lines are the branch lengths. If you watch carefully enough along the process, you will find that it is an ultrametric tree. However, we also notice that the original distances between pairs of sequences are not always preserved. For example the distance in the distance matrix between A and C is 17 while in the tree, the distance became A and C became 16, which is the average distance we

calculated in step 3. However, the distance between A and B is preserved. (Think intuitively, we need to average over the distances to preserve the global proportionality=ultrametric property. This can also lead to the phenomena that branch lengths can be fractional numbers instead of real integers as we will see in the later example).

Exercise/Homework: use the distance matrix you developed in table 1 and draw a three taxa tree for the species.

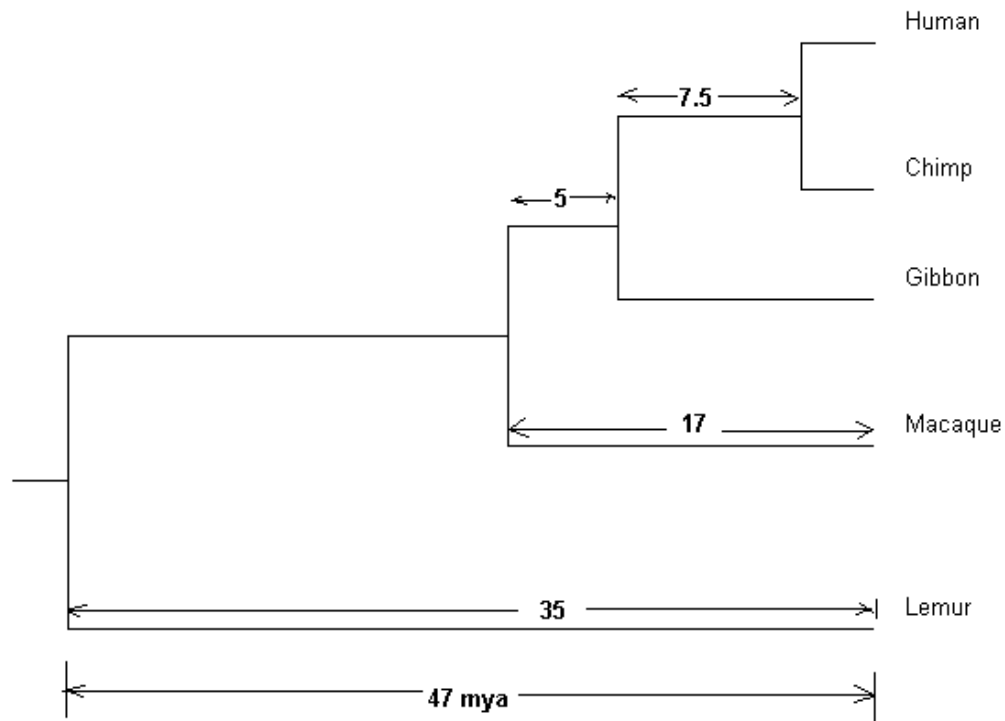
6) Dating Historical Events:

Since trees met with the molecular clock assumption have this ultrametric property, we can calibrate the tree by using known information. For example, given the three taxa ultrametric tree listed below,



Suppose we know from fossil record that species A and species B diverged from each other t_1 time ago. From molecular sequence data we also have the ultrametric tree shown above. Then we can calculate the time t_2 when species C diverged from the common ancestor with A and B, simply by solving this linear relationship: $d_1/(d_1+d_2)=t_1/t_2$. Molecular clocks provide an important way to scale time in evolutionary history.

Suppose we use the UPGMA method and build the phylogenetic tree for a group of primates. We also know that Lemur lineage diverged from the rest of the primates around 47 mya ago. What can we say about human/chimp divergence time?



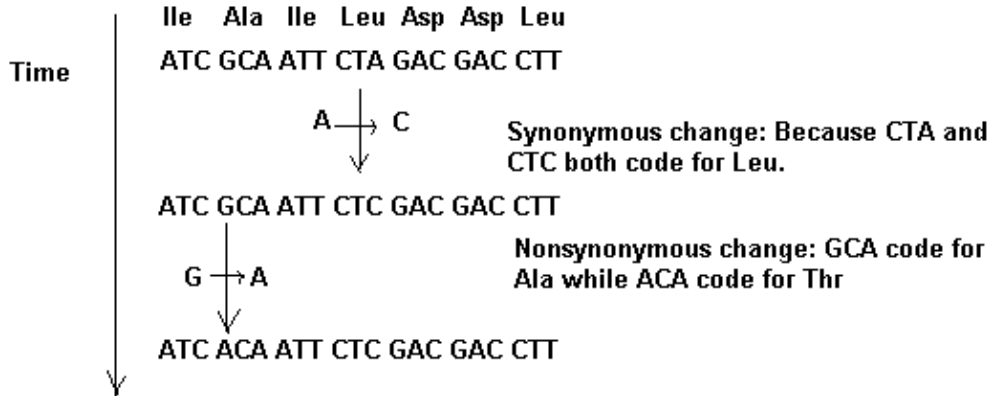
(this page is intentionally left blank for exercise)

7) Homework/Exercise: Synonymous/nonsynonymous changes

Students, who have studied genetics or molecular evolution before may have come across terms like synonymous/nonsynonymous changes. When we are talking about synonymous/ nonsynonymous changes, we are always referring to changes happening in nucleotide sequences and whether or not these changes at the DNA level result in a different amino acid at the protein level.

A synonymous substitution (also called a silent substitution) is a substitution of one base for another in a gene coding for a protein, such that the protein sequence produced is not modified. Nonsynonymous substitutions (aka replacement changes) are the opposite of synonymous changes. They are base changes that will change the amino acid sequences.

Imagine a small segment of nucleotide sequence like that listed below, the corresponding amino acid sequence is listed just above. It undergoes two changes in nucleotide sequence. The first change from A->C causes no amino acid change (synonymous change), while the second change (G->A) causes the corresponding amino acid to change from Ala to Thr.



Listed below is the table of genetic code.

Table of Genetic Code

	T	C	A	G
T	TTT Phe (F) TTC Phe(F) TTA Leu (L) TTG Leu(L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC Tyr(Y) TAA Stop TAG Stop	TGT Cys (C) TGC Cys (C) TGA Stop TGG Trp (W)
C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)
A	ATT Ile (I) ATC Ile (I) ATA Ile (I) ATG Met (M)	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)
G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)

If we observe carefully enough, we will find that: changes in the third codon position are usually synonymous (We usually say that the third positions are very redundant, remember, there are $4 \times 4 \times 4 = 64$, we only have 20 amino acids, there definitely should be a lot of redundancy) and most of the changes in the first positions are nonsynonymous (95%). All the changes that happen in the second position are nonsynonymous.

Researchers have shown that: if we assume that all codons are equally frequent in the genome and the probability of substitution is the same for all pairs of nucleotides, the proportion of nonsynonymous mutations is about 71%, excluding nonsense mutations (mutations that mutate to stop codons). In other words, about 29% of mutations at the nucleotide level cannot be detected by analyzing sequences at the protein level. Although these assumptions are not true for the real genome, these numbers are roughly correct and you can get a feeling of magnitude of the ratio for nonsynonymous and synonymous changes.

One of the important reasons we want to have this division is: nonsynonymous mutations are often disruptive to the functional protein. A lot of human diseases are associated with nonsynonymous mutations. The most classical example is sickle-cell disease. (page 84 and 328 7th edition; page 75 and 323 6th edition). A single non-synonymous mutation (from T->A) in the hemoglobin molecule leads to an amino substitution at position 6 from Glu(CTT) to Val(CAT). This single point change causes the molecules to polymerize within themselves and distort erythrocytes into sickled shape. The deformed and rigid erythrocytes can interfere with normal blood flow in microcirculation and thereby induce ischemia in tissues distal to the vascular blockage, the basis for many sickle cell disease complications. It is a relatively common disease especially among individuals with African descent (1 out of every 400) (page 267 7th edition; page 262 6th edition).

On the other hand, synonymous mutations, mutations that do not change the functional entity-protein, are “usually” selectively neutral (neither good nor bad). We put quotes around usually because there is evidence indicating that selection can work on codon usage (preference for different codon of the same amino acid due to abundance in tRNA) so that synonymous mutations are not always selectively neutral. However, it is generally true that synonymous mutations are selectively neutral. This provides an important yardstick for measuring rate of evolution.

Miscellaneous:

- 1) In this lab, we “adopted” quite a few “old” examples from the seventies. Don’t feel that they are too old to mean anything today. We use these simply because they are so classical and played very important roles in the history of science. (Everybody likes to cite them. ☺)
- 2) The examples we use/simulate are really very ideal and they are introduced here to illustrate the conceptual definitions. Molecular clock assumptions are not always true. Students should potentially ask themselves questions like: what can we do if we don’t have the molecular clock? How do different methods of tree

construction differ and relate to each other and in practice, how can I choose from them? What are the reasons or explanations for the molecular clock? All these questions could/will be answered in part during more advanced courses.

3) The methods/datasets we introduce here are really the tip of the iceberg. There are a lot more areas you can explore in this field. (Imagine, if you have 20 species, the number of rooted, bifurcating tree is 8,200,794,532,637,891,559,275. How can you possibly search over them to find the most parsimonies tree?) I hope this section has helped you to get a taste of molecular evolution, hopefully stimulating your interest in Phylogenetics/Population Genetics/Evolution, at least for some of the students. Remember: “nothing makes sense in biology except in the light of evolution”—by Theodosius Dobzhansky.

References:

Zuckerkandl, E and L. Pauling 1962. Molecular Disease, evolution, and genetic heterogeneity. In *Horizons in Biochemistry*. Ed. Kasha, M and B. Pullman, pp189-225. Academic Press, New York.

Kimura, M 1968 Evolutionary rate at molecular level. *Nature* 217: 624-626

Dayhoff, M.O 1978 Survey of new data and computer methods of analysis in M.O. Dayhoff, ed., *Atlas of Protein Sequence and Structure*, vol 5, supp. 3, pp2-8. Silver Springs, Md.: Biomed. Res. Found

Dickerson, R.C. 1971. The structure of cytochrome C and the rates of molecular evolution. *J. Mol. Evol.* 1:26-45

Sokal, R.R and C.D. Michener 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38:1409-1438

Yoder, A.D. and Z. Yang. 2000 Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, 17(7):1081-1090