

Molecular Evolution I: Protein Evolution

1. Protein Evolution

We begin our discussion of molecular evolution with protein evolution for two reasons: First, historically protein sequencing came before DNA sequencing. A method for determining the amino acid sequence of a given protein was developed in mid-1950s. Methods for DNA sequencing were not developed until the mid-1970s. So there was a period of about 20 years where the study of molecular evolution concentrated on protein sequences and this is the period when much of the early molecular evolutionary theory was developed. Today, advances in DNA sequencing have made it much faster and easier to sequence DNA than proteins. Thus the vast majority of protein sequences found in current databases, such as GenBank or SwissProt, were not determined by sequencing the amino acids of the proteins, but instead were inferred from DNA sequences using the universal genetic code.

Second, protein evolution is typically ‘simpler’ than DNA evolution. Although protein sequences are more complex than DNA (20 amino acids *versus* 4 nucleotides), proteins are generally more conserved throughout evolution and easier to align and compare between distantly related species. As we will see later, DNA is also more complicated because it may be coding or non-coding, and even within coding regions there are synonymous and non-synonymous nucleotide sites.

Here is an example of a partial amino acid (a.a.) sequence of a hypothetical protein:

Methionine	Proline	Valine	Serine	Threonine	Leucine	Glycine	Isoleucine	Lysine	Phenylalanine	Tryptophan	...
Met	Pro	Val	Ser	Thr	Leu	Gly	Ile	Lys	Phe	Trp	
M	P	V	S	T	L	G	I	K	F	W	

The 3-letter abbreviation is often useful when indicating coding regions of a DNA sequence, because a codon of 3 nucleotides designates 1 amino acid. The 1-letter abbreviation is most commonly used, particularly in sequence databases.

Now lets compare it with the a.a. sequence of the same protein from a different species:

M P A S T L G L K F W

All proteins begin with Met, so it is usually not considered for comparisons; thus we have a total of 10 a.a.’s to compare between species. We can calculate the simple statistic D , the proportion of differences, as: $D = 2/10 = 0.20$ or 20%. This method is good for sequences that are not too divergent, however, it ignores the possibility of 2 amino acid changes occurring at the same site. For example, what if there was an unobserved, intermediate sequence of:

M P G S T L G L K F W

Then we really had 3 amino acid replacements instead of 2. This is often called the “multiple hit” problem. Without the intermediate sequence, we do not know if this happened or not, but we can estimate the probability based on the observed divergence. The expected proportion of differences, K , between 2 sequences is given by the equation: $K = -\ln(1-D)$.

How do we get this?

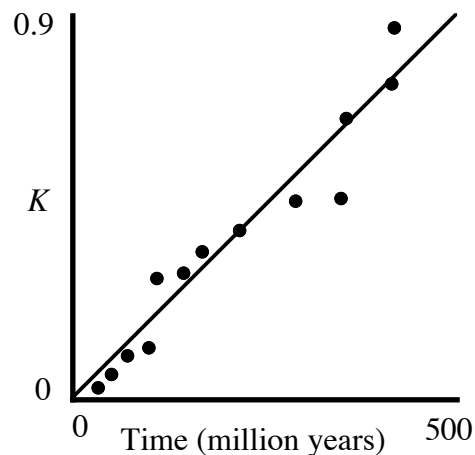
Let r = the rate of a.a. substitution (that is, the proportion of a.a.'s substituted per unit time, t). The expected proportion of substitutions, K , between 2 sequences is then $K = 2rt$. The 2 is there because there are 2 lineages on which changes can occur.

The proportion of a.a.'s that have NOT changed over consecutive units of time is $(1 - r)^{2t}$. The mathematical trick is that $(1 - r)^{2t} \approx e^{-2rt}$, where e is the base of the natural logarithm. The proportion of a.a.'s that are different is given by, $D = 1 - e^{-2rt} = 1 - e^{-K}$. Taking the natural log of both sides and re-arranging gives us $K = -\ln(1 - D)$.

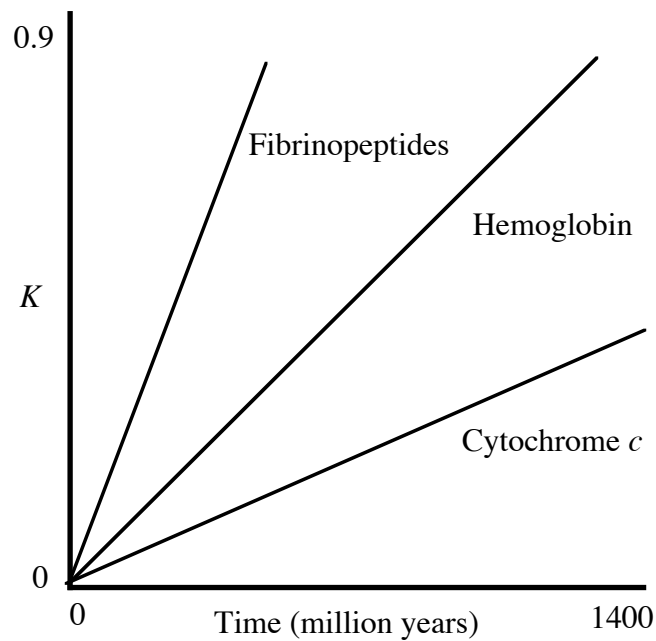
For our example, $K = -\ln(1 - 0.20) = 0.22$. Note that the difference between D and K becomes larger as D increases. For example, when $D = 0.50$, $K = 0.69$. Also, note that this method does not account for substitutions that make previously different amino acids identical, which are assumed to be rare.

2. The Molecular Clock

Interestingly, for many proteins the rate of amino acid substitution appears to be constant over long periods of time (where time is estimated from the fossil record). This observation led to the hypothesis of a "molecular clock". For example, when comparing sequences of the blood protein α -globin from various vertebrate species (modified from Hartl and Clark):



The molecular clock does not always tick at the same rate. Although the rate of substitution for a given protein appears to be constant, different proteins may have different rates, for example (modified from Hartl and Clark):



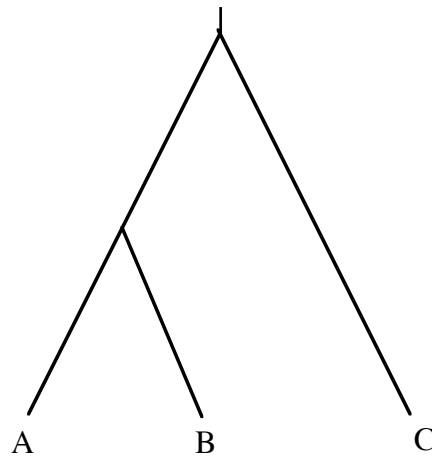
One possible explanation for these patterns is that all of the observed changes in amino acid sequence are neutral. That is, they are neither favored nor disfavored by natural selection. Different proteins have different *neutral* mutation rates, and thus accumulate substitutions at different rates. This hypothesis is the basis of the neutral theory of molecular evolution.

An important implication of the molecular clock is that once we know the rate of substitution for a given protein, we can use it to determine the time of divergence of two species for which we have amino acid sequences of that protein. Similarly, if the molecular clock is constant, we can use amino acid sequence divergence to correctly infer phylogenetic relationships among species.

It is important to note that the molecular clock is constant with time – not with generations. This is somewhat surprising, given that generation times vary greatly among species and mutations causing changes in protein sequences can occur in each generation. For example, if rodents have a generation time of 1 year and primates have a generation time of 15 years, then we might expect proteins to evolve 15 times faster in rodent lineages. This does not appear to be the case in general (although smaller generation-time effects have been reported). A possible explanation for this is that many amino acid changes may be very slightly deleterious (‘nearly neutral’) and that there is a strong negative correlation between generation time and population size. Thus, more slightly deleterious amino acid changes become fixed due to drift in species with low effective population size (N_e) and long generation times. This may compensate for the reduced fixation rate of neutral mutations due to the smaller number of generations per unit of time.

3. Relative Rate Test

Is the molecular clock really constant? A simple and commonly-used way to test this is the relative rate test. For this you need protein sequences from two relatively closely related species and one more distantly related species to be used as an outgroup. Consider the example:



Species A and B should be equally divergent from the outgroup species C. If the molecular clock holds, K_{AC} should equal K_{BC} , and thus $K_{AC} - K_{BC} = 0$.

Two examples (A = mouse, B = rat, C = human)

Lactate Dehydrogenase: $K_{AC} = 0.804$, $K_{BC} = 0.803$, $K_{AC} - K_{BC} = 0.001$

Thyroglobulin β : $K_{AC} = 0.774$, $K_{BC} = 0.927$, $K_{AC} - K_{BC} = -0.153$

Here Lactate Dehydrogenase conforms well to the molecular clock, while Thyroglobulin β does not. The significance of the deviation from 0 may be tested computationally using maximum likelihood methods or, alternatively, the number of substitutions along the two branches may be used for a contingency table test, such as a chi-squared test. In general, many proteins appear to evolve in a clock-like manner, although many exceptions have been found in certain proteins and in certain lineages. Thus, there remains much debate about the existence and/or the reliability of the molecular clock.

4. Clock Dispersion

If the clock-like evolution observed for many proteins is due only to the random accumulation of neutral amino acid substitutions, then we expect to see a particular statistical distribution of substitutions over time. Specifically, we expect a Poisson distribution, which is a distribution describing the occurrence of rare events. An important feature of the Poisson distribution is that the mean (μ) and variance (σ^2) are equal. This suggests a test of the molecular clock by testing the ratio of the variance to the mean:

$$R = \sigma^2/\mu$$

Under neutrality, $R = 1$.

The value of R has been observed to vary greatly from protein to protein, but in general there is an excess of variance, that is $R > 1$. For example, in a comparison of 9 mammalian proteins R ranged from 0.16 to 35.6. Some proteins actually appear to follow the clock more regularly than expected ($R < 1$), while many others show greater dispersion than expected under a neutral model of molecular evolution ($R > 1$).