Chapter 3  The molecular clock

Objective  One of the exciting aspects of population biology is that it entails not only a set of biological facts and principles, but that those observations can be compared to theoretical models that arise from pure mathematics. One of the most compelling results from this kind of theorizing is that neutral mutations ought to accumulate in lineages in a way that allows one to relate the number of differences between a pair of sequences to the age of common ancestry. This theory implies that molecules are turning over in a fairly constant or "clock-like" manner. In this chapter we will compare this theoretical prediction to real data, and see where it fits the data and where it does not.

Early observations

Zuckerkandl and Pauling (1965) plotted the numbers of amino acid differences between the cytochrome c molecules of a number of organisms against the estimated time back to the common ancestor of those organisms, and got a picture that looked like this:

They coined the term molecular clock to describe the remarkable constancy of the rate of amino acid substitutions that they saw. They noted that one could turn this process around, and if one knew the calibration for the clock for a particular protein, one could estimate the time of common ancestry of an organism for which there is no good fossil dating.
Theoretical reasons to expect a clock

In chapter 2 we outlined the reasoning of Motoo Kimura, which first led to the expectation that molecules should exhibit neutral variation that changes in a clock-like fashion. One key issue is whether the differences are neutral. For now we will tolerate this assumption -- that changes in the gene result in no difference in function--but of course we know that there will be exceptions. Kimura's reasoning was that there should be $2Nu$ new neutral mutations in a population with $2N$ genes ($N$ diploid individuals) where $u$ is the neutral mutation rate. Each new mutation has an initial frequency of $1/(2N)$, and we saw before that the chance of fixation of each new mutation is therefore $1/(2N)$. If we multiply the number of new mutations per generation times the probability of going to fixation, we get the number of new mutations that go to fixation each generation. This is $2Nu \cdot 1/(2N) = u$. Thus $u$, the neutral mutation rate, is the number of newly fixed substitutions that occur per generation. $u$ is typically a small number, on the order of $10^{-9}$ changes per nucleotide site per year.

Note that the rate of the molecular clock does not depend on the size of the population. You might think that genes would evolve faster in a large population because large populations face a larger number of mutations. On the other hand, a large population has more "momentum" in the sense that any new mutation is a smaller minority in a large population, so the chance of a new mutation is less in a large population compared to a small population. These two things precisely cancel, so that the molecular clock rate is independent of population size.

So what about changes that kill the function of a protein?

Mutations that cause premature termination of a protein, or otherwise destroy the function of the protein will be selected against, and will not be expected to ever go to fixation. The way the neutral theory deals with this is to admit that such mutations occur and to basically not count them when considering the mutation rate. Note that we defined $u$ as the neutral mutation rate. That is, $u$ is the rate of mutations to new alleles that have no deleterious (or advantageous) effects. Other mutations may occur, but we will just ignore them for now. In terms of the rate of turnover of the molecular clock, this idea works just fine.

Things that might make a molecular clock run unevenly

Before we turn to more real data, consider what determines the molecular clock rate and what sort of factors might influence its rate. One thing that could vary is the mutation rate. The rate of mutation might vary from one gene to another (it does!), and it might vary from one organism to another (it does!). These two sources of variation lead us to expect that the molecular clock should run at different rates in different genes, and that different organisms also might have different clock rates. Another factor that is a bit more subtle is the generation time. In developing the model, we referred to the rate of mutation in terms of
mutations per generation. If two organisms have the same mutation rate per generation, but one has a 20 year generation time and another has a 1 year generation time, then we expect that the organism with the one year generation time ought to have a molecular clock that runs 20 times faster. Let's look at some examples from real data and see just how good the idea of a molecular clock is.

The molecular clock runs at different rates for different proteins

We already saw that cytochrome c shows a very good fit to a clock-like pattern of substitution. The same figure above also showed what happens when the same organisms are compared in their sequences of beta globin and of fibrinopeptide. Note that the globin clock is running more than twice as fast as the cytochrome c clock, and that fibrinopeptide is racing along at around 20 times the speed. We mentioned that a difference in neutral mutation rate might be the reason, so let's think about what this means. It is possible that the actual probability of a variant DNA occurring is the same for all 3 genes, but the cytochrome c has a much lower neutral mutation rate because a large fraction of the mutations are actually not neutral. That is, the rate of the molecular clock depends on the degree to which the molecular sequence is constrained by natural selection. If cytochrome c were very fussy, so that most amino acids are important to the function of the protein, and if changed give a deleterious result, then the neutral mutation rate is much lower than the total mutation rate. On the other hand, if fibrinopeptide has subunits whose function is unconstrained, then fibrinopeptide can tolerate most changes in these regions, and its neutral mutation rate may be close to the total mutation rate.

Humans have a slower clock than monkeys

The dating of the time of common ancestry of humans and other primates has long been of interest to molecular evolutionists, and these date estimates depend on assumptions about molecular clocks. How can the constancy of a clock be determined? Consider three species with relationship as indicated to the right:
In this case we know that species A and B are more closely related than either is to C. This is the kind of situation where we can use sequences from these three species to test the clock idea. In particular, if the molecular clock is constant across this tree, then the number of substitutions from the common ancestor down to species A is expected to be the same as the number of substitutions from the common ancestor down to species B. This is known as a relative rate test, because even if we do not know when the common ancestor lived, we can determine if the relative rates of evolution along the branches that end up at species A and at species B are the same.

To apply the test, we need sequence that we believe will be relatively unconstrained. Intron sequences are probably not entirely unconstrained, but they are also clearly less constrained than coding sequences. Sequences were collected for humans (species A on the figure), Old World monkeys (species B on the figure), and New World monkeys (species C). In a test like this, the New World monkeys are used as a sort of standard against which the other two species are compared. For this analysis, such a species is called an outgroup. When the relative rate test is applied to these data, it was found that the lineage leading to humans had a rate that was 76% that of the rate on the lineage to Old World monkeys. Statistical tests were done to verify that this slow-down was significant.

<table>
<thead>
<tr>
<th>Species pair</th>
<th>Percent divergence</th>
<th>Time (MY)</th>
<th>Rate x 10^{-9}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human vs Old World monkeys</td>
<td>7.1</td>
<td>25</td>
<td>1.5</td>
</tr>
<tr>
<td>Human vs New World monkeys</td>
<td>14.7</td>
<td>35</td>
<td>2.1</td>
</tr>
<tr>
<td>Mouse vs Rat</td>
<td>14.4</td>
<td>20</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Rodents have a fast clock

In order to apply the relative rate test to rodents vs. humans, it is necessary to get an outgroup species that diverged before the common ancestor of humans and rodents. Gu and Li used chickens as an outgroup, because the common ancestor of mammals and chickens is thought to have lived 300 million years ago, well before the time of common ancestry of humans and rodents. Gu and Li compared 54 proteins that were sequenced in the three species and counted up the number of amino acid changes in the following way.
They let \( N_R \) represent the number of amino acids that were the same in humans and chickens but different in mice. Looking at the tree, these are amino acids that have probably changed along the rodent-specific lineage. They also tallied the number of amino acids that were the same in chicken and mouse, but were different in humans (\( N_H \)). These changed along the lineage that ends in humans. There were 600 rodent-specific differences and 416 human-specific differences. If the rate of evolution were the same in both lineages, one would expect \( 1016/2 = 508 \) changes along each branch. The chi-square is \( (600-508)^2/508 + (416-508)^2/508 = 33.32 \). This is much larger than one would expect by chance, and in fact the probability of getting a value this large or larger is less that 0.001. We conclude that rodents have a faster rate of molecular evolution than do humans.

**Virus clocks run a million times faster than those of higher organisms**

Influenza virus has a life cycle that includes pigs, ducks, and people. One of the means by which influenza virus maintains a high level of diversity is through recombination among viruses in the various hosts. It also has an exceptionally rapid rate of sequence evolution. When humans are infected with a new influenza virus, it may take our immune system a while to make antibodies with the appropriate specificity, but after a few days we generally are successful, and the virus is subsequently eliminated. The only way an influenza virus can invade again is if it changes it exterior protein coat enough that the same antibodies no longer recognize it. This results in a sort of race between the virus changing its coat and our immune system trying to detect and eliminate the virus.
The reason influenza virus is still a source of sickness in people is that it is able to evolve at an astounding rate. Note that the axis on this graph is not in millions of years, but is in decades! The molecular clock of viruses runs at about $1 \times 10^{-3}$ substitutions per site per year, or one million times faster than the molecular clock of higher organisms.

The human immunodeficiency virus (HIV) also has a molecular clock rate that is about the same as influenza. In the case of HIV the result is that the virus is actually undergoing evolution within a host during the progression of a single infection. As the host immune system starts to recognize parts of the HIV surface, new forms of virus are already being made that can evade the host's immune system.

**Summary**

1. The rate of the molecular clock depends only on the neutral mutation rate and does not depend on the size of the population.

2. In a population that is in balance between the addition of variation by new mutations and the fixation/loss of variation by random genetic drift, the rate of turnover is simply the neutral mutation rate.

3. Molecules that are highly constrained will have a lower neutral mutation rate, because most of the mutations are deleterious and are rapidly eliminated, leaving a small fraction of the total mutations that are neutral.

4. Cytochrome c appears to be more constrained than beta globin, and that beta globin is probably more constrained than is fibrinopeptide. We conclude this because the order of rates of molecular evolution is fibrinopeptide $>$ globin $>$ cytochrome c.

5. Different lineages of organisms have different molecular clock rates. Primates appear to have an unusually slow clock, and rodents have an unusually fast clock. Viruses are in
part so successful evolutionarily because their molecular clock rate is one million times faster than ours.