40-41. Substitution rates
[The updated version has several typographical errors corrected.]

The rate of substitution of neutral alleles is the mutation rate

Suppose there is a mutation at a site that changes the nucleotide present. A new single nucleotide polymorphism (SNP) is created at that site. The new nucleotide is a new allele, and its frequency immediately after it arises by mutation is

$$p = \frac{1}{2N}.$$

If it is neutral, the probability that it will ultimately be fixed because of genetic drift is $1/(2N)$.

The probability that a new nucleotide appears at a site is μ, the mutation rate per site. Because there are $2N$ chromosomes, the probability that there is a mutation on any chromosome at a particular site is $2N\mu$. If a mutation occurs, the probability that the new nucleotide is fixed because of genetic drift is $1/(2N)$. Therefore, the overall probability that a new nucleotide at a site is created by mutation and is ultimately fixed by genetic drift is

$$K = 2N\mu \times \frac{1}{2N} = \mu$$

independently of N.

Once fixation occurs, there has been a **substitution** of one nucleotide by another at that site. If you could compare sequences before and after the substitution, you would detect a different nucleotide at that site. By comparing sequences of different species, you can count the number of substitutions that have occurred. Furthermore, if you know when in the past those species had a common ancestor, you can estimate the **substitution rate** and hence the underlying mutation rate for individual nucleotides. The formula is

$$K = \frac{d}{2TL}$$

where d is the number of differences found when comparing L sites and T is the time separating two species. The 2 is needed because substitutions can occur in both species that are descended from the common ancestor, [Note: this formula is only approximate because it does not take account of the possibility that more than one substitution can occur per site.]

Be careful to **not confuse a mutation with a substitution**. A mutation appears in the offspring of a single individual. Most mutations will be lost and but a few will be fixed. When a mutation has become fixed, there has been a substitution.

One way to estimate the mutation rate is to compare third positions of codons that are **4-fold degenerate**, meaning that any of the 4 nucleotides present result in the same amino acid. The third positions of codons for Proline, Valine, Alanine and several other amino acids are 4-fold degenerate. Because a mutation of this type does not change the amino acid coded for, it is reasonable to assume that it is neutral. In the gene coding for β-globin, there are 78 4-fold degenerate codons (L=78). When sequences from mice and humans are compared, 30 of these 78 codons are found to differ at the third position (d=30). The fossil record indicates that the most recent common ancestor of humans and mice was present about 80 million years ago ($T=8\text{x}10^7$). Using the formula, you find $K=2.42\text{x}10^{-9}$ per year. That is an estimate of the **mutation per site per year**. That estimate is very close the to currently accepted average for mammals ($2.2\text{x}10^{-9}$), which is based on the analysis of a much large number of genes and species pairs.

Substitution rates of silent and replacement mutations differ

If the mutation of a nucleotide does not result in a change in the amino acid coded for, it is a **silent mutation**. A substitution of a silent mutation is called a **silent substitution**. The mutation rate μ is estimated from the rate of silent substitutions, which is called K_S.

If a mutation in a coding sequence results in a change in amino acid, it is a **replacement mutation**. Because of the genetic code, all mutations at the second-codon position are replacement mutations. When coding sequences of the same gene in different species are compared, the rate of replacement substitution, denoted by K_R, is almost always lower than the silent rate, sometimes much lower. If you found 13 differences at 102 second-codon positions of β-globin in humans and mice, you would estimate K_R to be $0.8\text{x}10^{-9}$, which is approximately the correct value for that gene. For insulin, K_R is $0.13\text{x}10^{-9}$. For two different histone genes, K_R is less than 10^{-13}. If $K_R<K_S$, you conclude that some replacement mutations are sufficiently deleterious that they cannot be fixed. In other words, there is **purifying selection**. You do not know how deleterious they are, however.

The mutation rate is almost certainly the same for silent and replacement mutations. Therefore, the ratio K_R/K_S estimates the fraction of replacement mutations that are neutral. If $K_S=2.2\text{x}10^{-9}$, then roughly 0.8/2.2=0.36 of the replacement mutations in β-globin are neutral while only 0.13/2.2=0.06 of the replacement mutations in insulin are neutral.

Furthermore, detecting protein subunits for which rates are very low can help identify the most important parts of proteins. For example, $K_R=0.24\text{x}10^{-9}$ for the part of β-globin the codes for the heme pocket and $K_R=1.35\text{x}10^{-9}$ for the surface amino acids.

Substitution rates in other parts of the genome can also be compared with K_S. The rate of substitution in pseudogenes is roughly the same as K_S. In 5′ untranslated and flanking regions, rates are roughly $K_S/2$ in mammals.

There appears to be strong purifying selection in some parts of the genome not coding for proteins. Numerous (481) segments of 200 bases or more in noncoding regions of the

human genome have been found to be perfectly conserved between humans, mice and rats. There were no substitutions detected. These regions have been called **ultraconserved**. Ultraconserved regions have almost certainly had an important function in mammals but we do not know what that function is or was.

There have been some functional studies of ultraconserved elements. Ahituv et al. (2007) created transgenic mice in which one of four different ultraconserved elements was deleted. None of the four deletions resulted in an obvious loss of viability and fertility under laboratory conditions.

Molecular clocks

Rates of substitution of genes from pairs of species that diverged at different times in the past are roughly the same. Silent rates are similar for different genes but replacement rates for the same genes in different species pairs are roughly the same. The molecular clock makes it possible to estimate the divergence times for pairs of species for which no estimate can be based on the fossil record.

In some cases, deviations from a molecular clock can indicate **positive selection** acting on a protein in a species. One example is the protein lysozyme, which is a bacteriolytic enzyme found in all animals. It is expressed in saliva and other bodily fluids and plays a role in defense against bacteria. In cows and other ruminants with foregut fermentation, lysozyme is also expressed in the stomach and plays a role in the extraction of nutrients from bacteria that aid in the fermentation of plants that are eaten. The langur, an Asian leaf-eating monkey also has foregut fermentation and also expresses lysozyme in the digestive system. The amino acid sequences of lyzozymes from humans and baboons differ at 14 positions, humans and langurs at 18 positions and baboons and langurs at 14 positions (Stewart et al. 1987). These data imply that the rate of replacement substitution on the branch leading to the langur is much higher than on the branch leading to the baboon, which indicates that selection favored additional amino acid changes in the langur. There is support for the idea that positive selection caused some of these changes because 5 of 11 changes on the langur branch are the same as those found in lysozyme in cows.

Bejerano et al. (2004) Ultraconserved elements in the human genome. *Science* 304: 1321-1325 (http://www.sciencemag.org/cgi/content/abstract/304/5675/1321)

Gross L (2007) Are Ultraconserved Genetic Elements Really Indispensable? PLoS Biology 5:e253
http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0050253

Ahituv N et al. (2007) Deletion of Ultraconserved Elements Yields Viable Mice. PLoS Biology 5:e234
http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0050234

Stewart CB, Schilling JW, Wilson AC (1987) Adaptive Evolution in the Stomach Lysozymes of Foregut Fermenters. Nature 330:401-404.
http://www.nature.com/nature/journal/v330/n6146/abs/330401a0.html

Problems.

40.1 If $\mu=2x10^{-9}$ per site per year, how many substitutions would you expect to see if you compared 200 neutral sites in humans and mice ($T=8x10^7$ years)?

Answer: For neutral sites, $K=\mu$. Therefore, $d=2TL\mu=2x8x10^7x200x2x10^{-9}=64$.

40.2 What fraction of the replacement mutations affecting amino acid positions in the heme pocket of β-globin are neutral?

Answer: 0.24/2.2=0.11

40.3 Are all of the amino acid substitutions on the lineage leading to the langur caused by positive selection?

Answer: Probably not. Amino acid substitutions also occurred on the baboon lineage which does not have foregut fermentation.