Calculating Ka/Ks

We can analyze changes in protein sequences over evolutionary time to determine what kind of selection (if any) a particular gene is under. Most often, changes to a protein's sequence interfere with the function of the protein, reducing an organism's fitness. For this reason, most changes to protein sequences are selected against (subject to negative selection), so that they are eliminated from the population and the protein's sequence remains unchanged. Occasionally, however, changes to a protein's sequence are selected *for* (subject to positive selection), because they give the protein a new function that is advantageous for the organism. This is one way that species can adapt, becoming better able to survive and reproduce in a changing environment.

These adaptations can be detected when you compare the sequences of the protein-coding portion of the gene in two or more species. By comparing the DNA sequences of the genes, researchers can determine the frequency with which nucleotide differences in the sequences (base substitutions) change the amino acid encoded. Because the genetic code is degenerate, only some substitutions in the DNA sequence will cause changes in the amino acid sequence. For proteins whose functions are so important that changes are not tolerated at all, over time you would only see the accumulation of substitutions that are silent (do not change the protein sequence) – all of the substitutions that change amino acids would be purged from the population by negative selection.

In a Ka/Ks calculation, the frequency with which synonymous and nonsynonymus substitutions are observed in orthologous genes (genes that are descended from the same gene in a shared ancestor) is analyzed. The Ka/Ks ratio is used to compare the rates of nonsynonymous and synonymous changes found in the protein-coding sequences of two orthologous genes, and it can be used to determine what type of selection that gene is under. If the Ka/Ks ratio is greater than 1, it means that nonsynonymous changes are relatively more common than synonymous changes, and the gene is under positive selection. If the Ka/Ks ratio is equal to 1, it means that nonsynonymous and synonymous changes occur with approximately equal frequency, and the gene is under neutral selection. If the Ka/Ks ratio is less than one, it means that nonsynonymous changes are relatively less common than synonymous changes, and the gene is under purifying (or negative) selection. This ratio uses the relative frequency of each type of change because in a particular sequence, there are more options for making nonsynonymous changes than synonymous ones, and you must take this into account to perform this analysis. In other words, using the ratio provides an internal control for the rate of DNA sequence changes overall.

This tutorial will walk you through the calculation of a Ka/Ks ratio between two protein-coding sequences from two different species. Though often performed by computers, these calculations can be done on a small scale by hand, and learning the procedure will give you a good sense of what is involved, and provide you with insights into what the results tell you about the evolutionary pressures affecting a particular gene.

To calculate the Ka/Ks ratio for a particular gene, you need two sequences, one from each species that you want to compare. Let's say that you have the following two sequences:

Species 1	`CAC ACC CCG GGA`
Species 2	`CAG ACA CCG GGG`

First, you designate one of those sequences as the starting sequence, and the other as the "derived" sequence. We will tell you which sequence to use as the starting sequence – for this example, we'll say that species 1 is the starting sequence.

Using the genetic code, translate the sequence into amino acids, and assign a value (S, N, or M) to each site in the starting sequence – either a change in the nucleotide at each site will be synonymous ("S"), nonsynonymous ("N"), or mixed ("M"). If changing the base at a site DEFINITELY changes the amino acid encoded, it counts as nonsynonymous. If changing the base at a site DEFINITELY DOESN'T change the amino acid encoded, it counts as synonymous (you ONLY find these at the 3rd codon position). If changing the base SOMETIMES changes the amino acid but sometimes doesn't, it counts as mixed. Most mixed sites are at the 3rd codon position, but codons for leucine and arginine can have mixed sites at the 1st position as well.

	His Thr Pro Gly
Species 1	CAC ACC CCG GGA
	NNM NNS NNS NNS
Species 2	CAG ACA CCG GGG

Calculate two values for each "mixed" site – the first value is the frequency with which a change in the "mixed" site would be synonymous ("S"), and the second value is the frequency with which a change in the "mixed" site would be nonsynonymous ("N"). For the sequence above, the third base in the CAC codon is the only one that's mixed – the third "C" could change either to an A, T, or G. A change to "T" would be synonymous, while a change to "A" or "G" would be nonsynonymous. Thus this site is considered 1/3 "S" and 2/3 "N."

> * A Note *
> Most codons with a "mixed" third position occur in a 2/2 split between two amino acids, which is why we most often assign "mixed" sites a value of 2/3 N and 1/3 S. However, the codons that code for isoleucine and methionine are distributed differently – only ATG codes for methionine, while ATT, ATC, and ATA all code for isoleucine. For this reason, if the ancestral codon was for methionine, it is counted as "N," while if the ancestral codon was for isoleucine, it is counted as "M" and divided 2/3 S and 1/3 N. Calculations involving the codons for tryptophan and cysteine must be similarly adjusted, with the additional note that they share their first two bases with a stop codon as well.

Add up the total number of synonymous and nonsynonymous sites, adding 1 to each total for an N or S, and 1/3 or 2/3 to each total for a mixed site. For the example above, we would have

Nonsynonymous = 8 N's + 1 M (2/3N) = 8.67 nonsynonymous sites
Synonymous = 3 S's + 1 M (1/3 S) = 3.33 synonymous sites

These are all the POSSIBLE differences that you could see. Now look at the differences that ACTUALLY APPEAR – they have been underlined in the sequence below. Synonymous differences are bold and green, while nonsynonymous differences are bold and purple. You can tell if they are synonymous or nonsynonymous based on the translation of the sequence, which is why we don't worry about whether or not things were "mixed" at this point – we're not looking at the possible differences, but the ones that are found in the sequence.

	His Thr Pro Gly
Species 1	CA**C** AC**C** CCG GG**A**
Species 2	CA**G** AC**A** CCG GG**G**
	Gln Thr Pro Gly

You can see from this that there are 2 synonymous differences and 1 nonsynonymous difference.

Calculate the Ka and Ks values, and determine the ratio between the two.

The Ka value, which indicates what fraction of nonsynonymous differences are observed relative to the total number of nonsynonymous differences possible, can be calculated:
Observed nonsynonymous changes/possible nonsynonymous changes = 1/8.67 = 0.1153

The Ks value, which indicates what fraction of synonymous differences are observed relative to the total number of synonymous differences possible, can also be calculated:
Observed synonymous changes/possible synonymous changes = 2/3.33 = 0.6006

The Ka/Ks value is the ratio between these two numbers:
0.1153/0.6006 = 0.1920

Once you have a Ka/Ks value, you need to evaluate it to determine what type of selection is occurring. If Ka/Ks is greater than one, then change is favored, and the gene is undergoing positive selection. If Ka/Ks is much lower than one, change is selected *against*, and the gene is undergoing negative selection. If Ka/Ks is close to 1, then the sequence is evolving neutrally.

In our example, the Ka/Ks ratio was 0.1920 – much less than one. This means that, relatively speaking, in this sequence fewer nonsynonymous changes are seen than synonymous changes – there were many more potential nonsynonymous changes, but you only see one, while of the three possible synonymous differences, two are actually observed. This suggests that changes that alter amino acids are selected against, and the protein is under negative selection.

A final note is that the Ka/Ks ratio is generally limited to use in comparing the sequences of a gene in two different species, not between individuals within a species. This calculation requires that there be sufficient substitutions in the DNA sequence to determine an accurate rate of synonymous and nonsynonymous changes, which is generally not possible within a species since DNA sequence changes overall are quite rare within species. There are other measures of molecular adaptations, such as Fst, that are better suited to the study of changes within species.