Molecular Evolution II: DNA Evolution

1. DNA Evolution

DNA sequences can be compared between species much like protein sequences. Here is an example DNA sequence:

Adenine Guanine Cytosine Thymine …

A G C T C T A G G C

A and G nucleotides have a double-ring chemical structure and are known as purines; C and T have a single ring and are pyrimidines. Changes between two nucleotides of the same structural class (for example A -> G or C -> T) are known as transitions. Changes between two nucleotides of different structural classes (for example A -> C or T -> G) are known as transversions. Typically, transitions are more common than transversions with the transition/transversion ratio, κ (kappa) ≈ 2.

Here is the homologous DNA sequence from a different species:

A G T T C C A G A C

Again, we can calculate a simple statistic, d, which is the proportion of sites that differ. Here $d = 3/10 = 0.3$ or 30%. As with proteins, there can be a "multiple hit" problem. But now, since there are only four nucleotides, we must also correct for nucleotides that were previously different now being identical due to chance. The correction is:

$$k = -3/4 \ln (1 - 4d / 3)$$

For our example $k = -3/4 \ln (1 - 4 (0.3) / 3) = 0.38$. Note that d saturates at 0.75. That is, with only four nucleotides, two random DNA sequences are expected to be 25% identical just by chance.

There are different types of DNA sequences, such as: protein-coding sequences, introns, 5' and 3' untranslated regions (UTRs), 5' and 3' flanking regions.

Protein-coding sequences are made up of codons, each of which codes for an amino acid. Within codons, there are different types of sites: Nondegenerate, 2-fold degenerate, and 4-fold degenerate.

In coding regions, nucleotide changes that alter an amino acid are called "replacement" or "nonsynonymous" substitutions. Those that do not alter an amino acid are called "silent" or "synonymous" substitutions. In most cases (but not all!) first and second codon positions are nondegenerate, while third positions are 2-fold or 4-fold degenerate.

We can compare rates of evolution at different sites. In general, 4-fold degenerate sites evolve the fastest – at a rate similar to introns and pseudogenes. Nondegenerate sites evolve the slowest.

Pseudogenes are genes that typically arise through duplication and are no longer functional. Thus, they are expected to be under no selective constraint and accumulate changes according solely to the mutation rate of the organism. By comparison with pseudogenes, it appears that 4-fold degenerate sites are under little or no selective constraint. There is much stronger

constraint on non-degenerate sites, suggesting that changes at these sites are much more likely to have a negative effect on an organism's fitness. UTR's and flanking regions also evolve more slowly than pseudogenes or 4-fold degenerate sites, which suggests that they are under at least some selective constraint – probably for gene regulatory elements.

2. Synonymous/Nonsynonymous Substitutions

The ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site is often used as a measure of selective constraint on a protein. Sometimes this is designated as *Ka*/*Ks*, or d_N/d_S, or simply as ω ('omega').

There are three possibilities:

Ka/*Ks* < 1: This is the case for the vast majority of protein-coding genes. It indicates selective constraint against amino acid replacements (known as 'negative' or 'purifying' selection).

Ka/*Ks* = 1: This indicates that the amino acid sequence is under no selective constraint. This is expected for pseudogenes.

Ka/*Ks* > 1: Indicates positive (or 'diversifying') selection for amino acid replacement. This is sometimes observed in genes encoding antigenic proteins of pathogens, which are under strong selective pressure to change in order to avoid the immune response of the host - for example, the gene encoding HIV envelope protein or the hemagglutinin A gene of human influenza virus.

Searching for genes with *Ka*/*Ks* > 1 is a common method for identifying 'positively selected' genes from whole-genome comparisons between species.

Note that *Ka*/*Ks* > 1 is a very strict criterion for detecting positive selection. *Ka*/*Ks* < 1 does not mean that a gene has experienced no positive selection. For example, most amino acid sites in a protein may be under purifying selection and just a small number (or even just one) may have been subject to positive selection. In this case, *Ka*/*Ks* would be less than one. In other words, *Ka*/*Ks* > 1 is a very conservative test for positive selection. It typically only detects cases where there is strong selection to change many amino acid sites.

3. Codon Bias

An examination of many protein-coding sequences from a species usually indicates that all of the synonymous codons for a particular amino acid are not used with equal frequency as would be expected at random. This phenomenon is known as "codon bias". Certain codons are "preferred" and are used more frequently than "unpreferred" codons.

For example, Leucine is an amino acid that shows very high codon bias. Leu can be encoded by six different codons, CTG, CTA, CTC, CTT, TTG, TTA. At random, we would expect each codon to be used about 17% of the time. However, in highly expressed *E. coli* genes, CTG is used ≈ 90% of the time. In yeast, TTG is used ≈ 90% of the time. The preferred codons correspond to the most abundant tRNA in each species, suggesting that selection favors the use of codons that can be translated quickly and efficiently. In general, highly expressed genes show greater levels of codon bias than genes expressed at low levels.

An alternative explanation for codon bias is that there is a bias in mutation. If degenerate codon positions evolve completely neutrally, then we would expect their composition to reflect the underlying pattern of mutation. Since most preferred codons end in G or C, a bias towards G or C mutations could explain the observations. However, the third position GC content in highly biased genes is generally greater than the GC content of neighboring introns. This suggests that selection may be involved. Furthermore, most species tend to show a mutational bias towards A or T, which goes in the opposite direction of most preferred codons.

Selection acting on a particular synonymous codon site is expected to be very weak in most cases (on the order of $N_e s \approx 1$, where N_e is the effective population size and s is the selection coefficient). This means that it is typically very difficult to distinguish between selective and neutral explanations for codon bias. As expected, the clearest evidence for selection on synonymous codon usage comes from species with large N_e, such as bacteria, yeast, and *Drosophila*. It is not clear if selection influences codon usage in mammals (including human).

Note:
Synonymous sites may be under selection for reasons other than codon usage. For example, there is a growing list of human genetic disorders that are caused by synonymous mutations. These usually affect splicing, either by creating a new, incorrect 5' splice site or by preventing a correct 5' splice site from being used. Sequences within the exon may act as ESE, Exon Splicing Enhancers. Synonymous sites may also be involved in the formation of RNA structures.