Number of substitutions between two protein-coding genes

Computing the number of substitutions between two protein-coding sequences is generally more complicated than computing the number of substitutions between two noncoding sequences, because a distinction should be made between **synonymous** and **nonsynonymous substitutions**. Several methods have been proposed in the literature for estimating the numbers of synonymous and nonsynonymous substitution (Perler et al. 1980; Miyata and Yasunaga 1980; Li et al. 1985b; Nei and Gojobori 1986; Li 1993; Pamilo and Bianchi 1993; Ina 1995; Comeron 1995).

In studying protein-coding sequences, we usually exclude the initiation and the termination codons from analysis because these two codons seldom change with time.

If the number of nucleotide substitutions between two DNA sequences is small, such that the number of nucleotide differences equals the number of substitutions, and if we are only interested in the absolute number of substitutions, then the numbers of synonymous and nonsynonymous substitutions can be obtained by simply counting synonymous and amino acid-altering nucleotide differences. Consider the following comparison:

Ser	Thr	Glu	Met	Cys	Leu
TCA	ACT	GAG	ATG	TGT	TTA
↕	↕		↕		↕
TCG	ACA	GAG	ATA	TGT	CTA
Ser	Thr	Glu	Ile	Cys	Leu

There are four codon differences between the sequences. Three of the differences are synonymous (TCA ↔ TCG, ACT ↔ ACA, and TTA ↔ CTA), and one is nonsynonymous (ATG ↔ ATA). Since 18 nucleotide sites are compared, the total number of substitutions per site is estimated as $4/18 = 0.22$.

However, if we want to compute the number of substitutions per site for synonymous and nonsynonymous substitutions separately, we must determine the appropriate denominators, i.e., the numbers of **synonymous** and **nonsynonymous sites**, respectively. In general, these numbers cannot be easily determined, for two reasons. The first is that the classification of a site changes with time. For example, the third position of CGG (Arg) is synonymous. However, if the first position changes to T, then the third position of the resulting codon, TGG (Trp), becomes nonsynonymous. The second reason is that many sites are neither completely synonymous nor completely nonsynonymous. For example, a transition in the third position of codon GAT (Asp) will be synonymous, while a transversion to either GAG or GAA will alter the amino acid. Consequently, many sites must be counted as part synonymous and part nonsynonymous. Note also that transitions result in synonymous substitutions more frequently than transversions do; therefore, the substitution scheme used to determine the number of synonymous and nonsynonymous substitutions must take into account that transitions occur with different frequencies than transversions.

One way to deal with this problem was proposed by Miyata and Yasunaga (1980) and Nei and Gojobori (1986). In their method, nucleotide sites are classified as follows. Consider a particular position in a codon. Let i be the

number of possible synonymous changes at this site. Then this site is counted as $i/3$ synonymous and $(3 - i)/3$ nonsynonymous. For example, in the codon TTT (Phe), the first two positions are counted as nonsynonymous because no synonymous change can occur at these positions, and the third position is counted as one-third synonymous and two-thirds nonsynonymous because one of the three possible changes at this position is synonymous. As another example, the codon ACT (Thr) has two nonsynonymous sites (the first two positions) and one synonymous site (the third position), because all possible changes at the first two positions are nonsynonymous while all possible changes at the third position are synonymous. When comparing two sequences, one first counts the number of synonymous and the number of nonsynonymous sites in each sequence and then computes the averages between the two sequences. We denote the average number of synonymous sites by N_S and that of nonsynonymous sites by N_A.

Next, we classify nucleotide differences into synonymous and nonsynonymous differences. For two codons that differ by only one nucleotide, the difference is easily inferred. For example, the difference between the two codons GTC (Val) and GTT (Val) is synonymous, while the difference between the two codons GTC (Val) and GCC (Ala) is nonsynonymous. For two codons that differ by more than one nucleotide, the problem of estimating the numbers of synonymous and nonsynonymous substitutions becomes more complicated, because we need to determine the order in which the substitutions occurred.

Let us consider the case in which two codons differ from each other by two substitutions. For example, for the two codons CCC (Pro) and CAA (Gln), there are two possible pathways:

Pathway I: CCC (Pro) $\leftrightarrow$ CCA (Pro) $\leftrightarrow$ CAA (Gln)

Pathway II: CCC (Pro) $\leftrightarrow$ CAC (His) $\leftrightarrow$ CAA (Gln)

Pathway I requires one synonymous and one nonsynonymous change, whereas pathway II requires two nonsynonymous changes.

There are basically two approaches to deal with multiple substitutions at a codon. The first approach assumes that all pathways are equally probable (Nei and Gojobori 1986), so we average the numbers of the different types of substitutions for all the possible scenarios. This approach is called the **unweighted method**. For example, if we assume that the two pathways shown above are equally likely, then the number of nonsynonymous differences is $(1 + 2)/2 = 1.5$, and the number of synonymous differences is $(1 + 0)/2 = 0.5$.

The second approach, the **weighted method**, employs *a priori* criteria to decide which pathway is more probable. For instance, it is known that synonymous substitutions occur considerably more frequently than nonsynonymous substitutions (Chapter 4), and so, in the above example, pathway I is more probable that pathway 2. If we give a weight of 0 to pathway II, then the number of synonymous differences is 1 and the number of nonsynonymous differences is 1. Alternatively, we may give weights other than 0 and 1 to the