

Handout: dN/dS ratios

A fundamental measure of the relative importance of selection and genetic drift in causing amino-acid substitutions is the dN/dS ratio. In this handout, you will learn how these ratios are calculated, as well as how they are interpreted.

Definitions

dN (alternatively designated K_a) is a measure of the degree to which two homologous coding sequences differ with respect to amino-acid content. Specifically, it indicates the degree to which two sequences differ at non-synonymous sites (nucleotide sites at which a substitution causes an amino-acid change). Formally, dN is the average number of nucleotide differences between the sequences *per non-synonymous site*.

dS (alternatively designated K_s) is a measure of the degree to which two homologous coding sequences differ with respect to silent nucleotide substitutions (substitutions that do not cause an amino-acid substitution). It indicates the degree to which two sequences differ at synonymous sites (sites at which a substitution does not cause an amino-acid substitution). Formally, dS is the average number of nucleotide differences between sequences *per synonymous site*.

Calculating dN and dS

Suppose we have the following two nucleotide sequences:

```

ACTCCGAACGGGGCGTTAGAGTTGAAACCCGTTAGA
 *   *   *   *   *   *   *   *   **
ACGCCGATCGGCGCGATAGGGTTCAAGCTCGTACGA
    
```

```

site # 12345678911111111112222222223333333
        012345678901234567890123456
    
```

```

syn  00100100 $\frac{1}{2}$ 001001 $\frac{1}{4}$ 0 $\frac{1}{2}$ 00 $\frac{1}{3}$  $\frac{1}{3}$ 0 $\frac{1}{3}$ 00 $\frac{1}{3}$ 001001 $\frac{1}{3}$ 0 $\frac{2}{3}$     sum = 7.5833
    
```

```

non  11011011 $\frac{1}{2}$ 110110 $\frac{3}{4}$ 1 $\frac{1}{2}$ 11 $\frac{2}{3}$  $\frac{2}{3}$ 1 $\frac{2}{3}$ 11 $\frac{2}{3}$ 110100 $\frac{2}{3}$ 1 $\frac{1}{3}$     sum = 28.4167
    
```

[For now ignore the two lines labeled "syn" and "non"]

The asterisks indicate nucleotide sites at which the two sequences differ.

The amino-acid translations of these sequences, using the standard genetic code, are

```

TPNGALELKPVR
 *   ***   *
TPIGAIGFKLVR
    
```

Handout: dN/dS ratios

From the protein sequence comparison, it is obvious that five of the nucleotide substitutions caused an amino acid change, so there were 5 non-synonymous substitutions. Since there are 10 nucleotide substitutions altogether, 5 must have been synonymous.

Next, we need to calculate the number of synonymous and non-synonymous nucleotide sites. Consider site 1, which is an A in both sequences. Given that the 3-base pair codon to which this A belongs is either ACT (sequence 1) or ACG (sequence 2), a change from A to any other nucleotide will cause an amino-acid substitution (e.g. CCT, GCT and TCT code for amino acids P, A, and S respectively). Site 1 is thus a non-synonymous site.

Consider site 2. The genetic code is such that no matter what codon one is dealing with, a change in the second amino acid of the codon always results in a new amino acid. Consequently, site 2 is also a non-synonymous site.

Next, consider site 3. ACT, ACG, ACC and ACA, which differ only at this 3rd site, all code for the same amino acid, T. Consequently any substitution at this site does not result in an amino acid substitution. All substitutions are synonymous. Consequently, site 3 is a synonymous site. It is also called 4-fold degenerate.

Moving on to site 4, any substitution of the first C in CCG will result in an amino acid change. This site is thus a non-synonymous site. Similarly, as the second position in a codon, site 5 is a non-synonymous site.

We keep moving on this way, designating each site as non-synonymous or synonymous. There are some complications, though. Consider site 9. In sequence 1, this is the third-position nucleotide in the codon AAC, which codes for amino acid N; and in sequence 2 it is the third-position nucleotide in codon ATC, which codes for amino acid I.

Consider what could happen if C in AAC mutates. The possible resulting codons are AAT, AAG, and AAA. These code for amino acids N, K and K respectively. The C in this case represents only a two-fold degenerate site and the site is considered 1/3 synonymous and 2/3 non-synonymous. Similarly, when the C in ATC mutates, it can result in ATT, ATA, or ATG, which code for I, I and M. This C is 3-fold degenerate, and the site is 2/3 synonymous and 1/3 non-synonymous. The overall average for this site, considering both sequences as the starting point for mutation, is thus

$$\frac{1}{2}[1/3 \text{ synonymous} + 2/3 \text{ non-synonymous}] + \frac{1}{2}[2/3 \text{ synonymous} + 1/3 \text{ non-synonymous}]$$

$$= 1/2 \text{ synonymous and } 1/2 \text{ non-synonymous.}$$

This site is thus a portion synonymous and a portion non-synonymous.

The two lines above labeled "syn" and "non" give the proportions of each site that are synonymous and non-synonymous respectively. Summing these values over sites yields 7.583 synonymous sites and 28.417 non-synonymous sites in the 36-site sequence.

We can now calculate dN and dS:

Handout: dN/dS ratios

$$dN = \frac{\text{No. non-synonymous substitutions}}{\text{No. non-synonymous sites}} = \frac{5}{28.417} = 0.176$$

$$dS = \frac{\text{No. synonymous substitutions}}{\text{No. synonymous sites}} = \frac{5}{7.583} = 0.659$$

The ratio is then

$$\frac{dN}{dS} = \frac{0.176}{0.659} = 0.269$$

Interpreting dN/dS ratiosCase 1: All non-synonymous mutations are neutral.

Consider a single synonymous site. Mutations at this site are all neutral. The number of mutations fixed per generation is, from neutral theory

$$dS = \text{mutations fixed/generation} = (\text{mutations arising per generation}) \times (\text{prob. of fixation})$$

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

Next, consider a single non-synonymous site. Mutations at this site are also neutral, by assumption. Consequently the number of mutations fixed per generation is

$$dN = \text{mutations fixed/generation} = (\text{mutations arising per generation}) \times (\text{prob. of fixation})$$

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

In this case, then, $dN/dS = \mu/\mu = 1$.

Conclusion: When non-synonymous mutations are all neutral, $dN/dS = 1$.

Case 2: A fraction f of all non-synonymous mutations are neutral, the rest are deleterious.

Synonymous mutations are still neutral, so $dS = \mu$ again.

Consider non-synonymous mutations. A fraction f are synonymous, and these fix at a rate of μ per generation. A fraction $1 - f$ are deleterious, and none of these fix. Thus the overall number of non-synonymous mutations fixed per generation is

$$dN = f\mu + (1 - f) 0 = f\mu$$

Handout: dN/dS ratios

so that

$$dN/dS = \frac{f\mu}{\mu} = f.$$

Note that by definition, $f < 1$, so $dN/dS < 1$.

Conclusion: When some non-synonymous mutations are deleterious and the rest neutral, $dN/dS < 1$. A value of $dN/dS < 1$ thus indicates the operation of purifying selection.

Case 3: A fraction f of all mutations are non-deleterious and a fraction $1 - f$ are deleterious. Of the non-deleterious mutations, a fraction θ are advantageous, while a fraction $(1 - \theta)$ are neutral.

Synonymous mutations are still neutral, so $dS = \mu$ again.

Of the non-synonymous mutations, a fraction $1 - f$ are deleterious and do not fix. A fraction $f(1 - \theta)$ are neutral, and fix at a rate of μ per generation. Finally, a fraction $f\theta$ are advantageous. These arise at a rate of $2N\mu$ per generation and fix with probability s , where s is the selection coefficient associated with the advantageous mutation. Overall, then, the number of non-synonymous mutations fixed per generation is

$$dN = (1 - f) 0 + f(1 - \theta) \mu + f\theta 2N\mu s$$

so that

$$dN/dS = \frac{f(1-\theta)\mu + f\theta 2N\mu s}{\mu} = f(1 - \theta) + f\theta 2Ns .$$

Notice that in this case, dN/dS may be > 1 . This will happen if θ is large enough; in particular, if

$$\theta > \frac{1-f}{f} \frac{1}{(2N-s)} .$$

Conclusion: A value of $dN/dS > 1$ indicates the operation of repeated positive selection in causing some amino-acid substitutions. Note that a value of $dN/dS < 1$ does not necessarily mean positive selection has not acted, only that it can't be detected with this approach.

Summary

1. If $dN/dS = 1$, amino-acid substitutions may be largely neutral. However, there is also the possibility that positive selection just cancels purifying selection, so that some amino-acid substitutions were driven by natural selection. This situation is thus ambiguous.
2. If $dN/dS < 1$, purifying selection (selection against deleterious non-synonymous substitutions) has definitely operated. Some amino-acid substitutions *may* have been caused by selection, just not enough to overcome the effects of purifying selection.
3. If $dN/dS > 1$, selection has caused some amino-acid substitutions. Some substitutions *may* also have been caused by genetic drift. Purifying selection also likely operates, but is not strong enough to overcome the effects of positive selection.