Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 239 (2006) 226–235
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Comparisons of d*N*/d*S* are time dependent for closely related bacterial genomes

Eduardo P.C. Rocha[a,b], John Maynard Smith[c,†], Laurence D. Hurst[d], Matthew T.G. Holden[e], Jessica E. Cooper[d], Noel H. Smith[f], Edward J. Feil[d,*]

[a]*Atelier de BioInformatique, Université Paris VI, 75005 Paris, France*
[b]*Unité GGB, Institut Pasteur, 75015 Paris, France*
[c]*Haldane's Right Hand Side*
[d]*Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 1AJ, UK*
[e]*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
[f]*Veterinary Laboratories Agency Weybridge, Woodham Lane, New Haw, Addlestone, Surrey KT15 3NB, UK*

Received 14 February 2005; received in revised form 7 May 2005; accepted 15 May 2005
Available online 18 October 2005

Abstract

The ratio of non-synonymous (d*N*) to synonymous (d*S*) changes between taxa is frequently computed to assay the strength and direction of selection. Here we note that for comparisons between closely related strains and/or species a second parameter needs to be considered, namely the time since divergence of the two sequences under scrutiny. We demonstrate that a simple time lag model provides a general, parsimonious explanation of the extensive variation in the d*N*/d*S* ratio seen when comparing closely related bacterial genomes. We explore this model through simulation and comparative genomics, and suggest a role for hitch-hiking in the accumulation of non-synonymous mutations. We also note taxon-specific differences in the change of d*N*/d*S* over time, which may indicate variation in selection, or in population genetics parameters such as population size or the rate of recombination. The effect of comparing intra-species polymorphism and inter-species substitution, and the problems associated with these concepts for asexual prokaryotes, are also discussed. We conclude that, because of the critical effect of time since divergence, inter-taxa comparisons are only possible by comparing trajectories of d*N*/d*S* over time and it is not valid to compare taxa on the basis of single time points.

Keywords: Bacterial evolution; dN/dS ratio; Purifying selection

1. Introduction

Comparisons of the relative rates of change at synonymous (silent) and non-synonymous (replacement) sites have for many years been a central tenet of molecular evolution (Kimura, 1991). The relative frequencies of these changes are determined by a complex blend of stochastic and selective forces acting at many different levels from a single site, codon, genome or population (metagenome). Estimates for d*N* (the number of non-synonymous changes per non-synonymous site) and d*S* (the number of synonymous changes per synonymous site) are typically interpreted in terms of the selective consequences of these changes. Whilst this may be valid for comparisons between relatively diverged eukaryotic species, where most of the non-synonymous changes are fixed (i.e. they are substitutions), for comparisons within populations, or between closely related bacterial genomes where "species" are not easily defined, it is not valid to assume that the selective consequences of non-synonymous change are effectively instantaneous. In such cases, the possibility that the d*N*/d*S* ratio might change over time due to a lag in the removal of slightly deleterious mutations must be considered.

Synonymous substitutions are usually regarded as neutral, or at least as having a much smaller effect on

*Corresponding author. Tel.: +44 (0) 1225 340959; fax: +44 (0) 1225 386779.
E-mail address: e.feil@bath.ac.uk (E.J. Feil).
†Deceased.

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.08.037

fitness than non-synonymous substitutions. The dN/dS ratio resulting from the comparison between two orthologous genes therefore has both theoretical and practical implications as it can reveal the type of selection pressure acting on the genes. A low ratio (dN/d$S \ll 1$) indicates strong purifying (''stabilizing'') selection, whereas a high ratio (dN/d$S > 1$) indicates selection for diversification (''positive selection''). The calculation of dN/dS can therefore help to identify genes or domains under particular biochemical or ecological constraint, or conversely putative virulence factors or candidate vaccine targets subject to diversifying (frequency-dependent) selection from the host immune response (Smith et al., 1995) Although the average dN/dS ratio over a whole coding region is a fairly blunt tool for detecting positive selection and more sensitive approaches focusing on specific codons are becoming increasingly common, these approaches still focus on local variations of the dN/dS ratio (Suzuki et al., 2001; Nielsen and Yang, 1998; Yang and Bielawski, 2000a).

Most studies on the relative rates of change at synonymous and non-synonymous sites have focused on sequences which shared a common ancestor millions of years ago. For highly diverged sequences the problem of multiple substitutions occurring at a single site complicates the calculation of dN and dS and a great deal of effort has been spent in perfecting methods which correct for this problem (Li et al., 1985; Nei and Gojobori, 1986; Yang and Nielsen, 2000b). Far less attention has been paid to much more closely related sequences belonging to the same named bacterial species and differing by, say, $<2\%$ of nucleotide sites. Although such sequences are far less likely to have experienced multiple substitutions, these comparisons are not drawn without difficulties. The paucity of nucleotide changes necessitates the use of large amounts of sequence data in order to achieve statistically meaningful results, and the estimates are extremely sensitive to sequencing error. Fortunately, complete genome sequence data are now available for a number of bacterial taxa. These data alleviate the statistical problems and allow closer examination of the relative rates of synonymous and non-synonymous change at a fine evolutionary scale; that is between sequences which shared a common ancestor from perhaps decades to hundreds of thousands of years ago.

Sequencing error, however, remains a critical issue for comparisons between closely related genomes, as each error becomes proportionately more important when the true number of changes is small. Importantly, too, one would expect sequencing errors to tend to make an observed dN/dS ratio approach unity as there need be no bias with respect to codon position. To eliminate this problem as far as possible, we have been highly selective as regards the genomes to employ and have rechecked putative changes by re-sequencing and by re-analysis of the trace files.

There are reasons to be confident that sequencing errors will not generally compromise this analysis. In their comparison of the two genomes of *Mycobacterium tuberculosis*. Gutacker et al. (2002) checked ~300 synonymous changes and reported that 91% of these changes were accurate. Qualitatively similar results were obtained for the analysis of close genome sequences of *Bacillus anthracis* (Read et al., 2002) and *M. tuberculosis* (Fleischmann et al., 2002). We have even more confidence in the accuracy of other genome sequences. We checked 285 unique single nucleotide changes within the MSSA476 genome of *Staphylococcus aureus* and found them to be 100% accurate, and a sample of 30 other unique single nucleotide changes in other *S. aureus* genomes were re-sequenced and also found to be 100% correct. Furthermore, statistical analysis of the quality of reads in assembled genome sequence projects at the Sanger Centre, UK, also point towards low error rates. In the case of the *S. aureus* MSSA476 genome sequence, the predicted sequence error rate as calculated from the consensus confidence of all bases in the assembled sequence is 1 in 7807752 i.e. (0.37 bases per genome) (MTGH unpublished data).

The dN/dS ratio of bacterial genes differing at 1–2% of nucleotides, and assumed to be under stabilizing selection, generally fall within the range of 0.04–0.2 (Feil et al., 2003; Dingle et al., 2001; Jolley et al., 2000; Jones et al., 2003; Meats et al., 2003). However, in cases where sequences are even more closely related, a relative preponderance of non-synonymous change is noted. Among the previously cited works, Gutacker et al. (2002) identified only ~900 single base changes within coding regions when they compared two genomes of the very uniform species *M. tuberculosis*, corresponding to ~2.5 changes per 10,000 sites. The authors noted that 65% of these single base changes were non-synonymous, corresponding to a dN/dS ratio of ~0.6. Holden et al. presented an even more extreme example in their comparison of two genomes of *S. aureus* (Holden et al., 2004). These genomes correspond to the same Sequence Type (ST) by Multilocus Sequence Typing (MLST) (Maiden et al., 1998; Holden et al., 2004) and only differ by 285 single base changes within coding regions, corresponding to ~1 change per 10,000 sites. Approximately, 70% of the changes were non-synonymous and, as non-synonymous sites are three times more common than synonymous sites, the dN/dS ratio is therefore approaching parity (~0.8).

A surprisingly high dN/dS ratio between very similar sequences has been noted in other taxa. In their comparative genomics analysis of four bacterial species, King Jordan et al. noted that the average dN/dS was unusually high between two very closely related genomes of *Chlamydia pneumoniae* (Jordan et al., 2002). Read et al. (2003) compared the genome sequence of *B. anthracis* isolated from a victim of the 2001 anthrax attack in Florida, USA, with the sequence from a reference strain. These authors also noted that the few single nucleotide changes detected tended to be non-synonymous. The same effect can be observed using multi-locus sequence data