Evolution of Relative Reading Frame Bias in Unidirectional Prokaryotic Gene Overlaps

Peter J. A. Cock*1,2 and David E. Whitworth3

1 MOAC Doctoral Training Centre, University of Warwick, Coventry, United Kingdom
2 Plant Pathology, Scottish Crop Research Institute, Invergowrie, Dundee, United Kingdom
3 Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Ceredigion, United Kingdom

*Corresponding author: E-mail: peter.cock@scri.ac.uk.
Associate editor: James McInerney

Abstract

Pairs of unidirectional (same strand) genes can overlap in one of two phases (relative reading frames). There is a striking bias in the relative abundance of prokaryotic gene overlaps in the two possible phases. A simple model is presented based on unidirectional gene overlaps evolving from nonoverlapping gene pairs, through the adoption of alternative start codons by the downstream genes. Potential alternative start codons within upstream gene sequences were found to occur at greater frequencies in one phase, corresponding to the most prevalent phase of gene overlaps. We therefore suggest that the phase bias of overlapping genes is primarily a consequence of the N-terminal extension of downstream genes through adoption of new start codons.

Key words: annotation, overlaps, start codon, phase, reading frame.

Overlapping genes are ubiquitous in viruses, and common in prokaryotes, but are also found in eukaryotes (Normark et al. 1983; Rogozin et al. 2002; Makalowska et al. 2005). In overlapping coding sequences, the DNA simultaneously encodes portions of two separate polypeptides. Consequently, mutation of the overlapping region has implications for both proteins, and it is therefore nontrivial to characterize the selective pressures at work on such sequences. Neighboring genes on opposite strands of the DNA can be convergent (tail-to-tail, →←) or divergent (head-to-head, ←→). Same strand genes (head-to-tail, →→) are unidirectional, with an upstream and a downstream gene. In general, phases +0, +1, and +2 can be defined by the gene separation modulo 3 (Kingsford et al. 2007) or equivalently by overlaps of 3i, 3i − 1, and 3i − 2 (for an integer i). For unidirectional genes, this phase is also the relative reading frame (Cock and Whitworth 2007).

Unidirectional (same strand) overlaps are the most common overlap orientation in prokaryotes (Fukuda et al. 2003). Only overlaps in phases +1 and +2 are considered because in-phase overlaps can be regarded as a single gene with alternative initiation sites. Most unidirectional gene overlaps are of 1 or 4 bp (phase +2), but for longer overlaps, phase +1 is much more common than phase +2 (Eyre-Walker 1996; Borodovsky et al. 1999; Johnson and Chisholm 2004; Cock and Whitworth 2007; Lillo and Krakauer 2007).

We previously highlighted this interesting phase bias in prokaryotic unidirectional overlaps (Cock and Whitworth 2007) and attempted to explain it with a mutual constraint argument (akin to that of Rogozin et al. 2002 for convergent overlaps). Although this mechanism could explain the phase bias, it did not make testable predictions.

Kingsford et al. (2007) provided a simple model to explain the phase bias in convergent overlaps based on the observed frequencies at which alternative stop codons are found in the reverse complement sequence of nonoverlapping convergent genes. Herein, we extend this idea to unidirectional overlaps evolving from nonoverlapping unidirectional genes. First, C-terminal extension of the upstream gene is considered by looking at the frequencies of alternative stop codons within the downstream gene. The mechanism in mind is that the existing stop codon of a nonoverlapping upstream gene is lost through a point mutation or indel, and the gene therefore is extended to the next stop codon, which may cause an overlap with the downstream gene. Second, N-terminal extension of the downstream gene by the adoption of a new start codon is considered by looking at the frequencies of alternative start codons within the upstream gene. Here, the creation of overlaps from unidirectional neighboring genes is by the adoption of a new start codon for a downstream gene (e.g., due to an indel or an accumulation of point mutations, perhaps in association with the loss of the original start codon).

We demonstrate that C-terminal extension of an upstream gene does not give the observed phase bias, but that N-terminal extension of a downstream gene does. Coupled with an exponential fitness cost to the overlap length, this model reproduces the general character of the observed distribution of overlaps.

In total, 3, 153, 393 gene pairs were considered: 460, 065 divergent, 460, 751 convergent, and 2, 232, 577 unidirectional; a split of 14.6%, 14.6%, and 70.8%, respectively (3 significant figures [sf]). As noted by Lillo and Krakauer (2007), the number of divergent and convergent pairs are expected to be almost equal. Within each orientation, the proportion of overlapping gene pairs varies considerably. Only 3.1% of divergent genes are annotated as overlapping compared with 13% of convergent pairs and 21% (2 sf) of
FIG. 1. Observed unidirectional overlaps, showing the clear +1 phase bias for overlaps of 7 bp or more. This figure includes 28 overlaps of length 5, which all use the alternative start codon ATT. These ratios are in good agreement with published results (Fukuda et al. 2003; Cock and Whitworth 2007; Kingsford et al. 2007). For the rest of this paper, we focus on unidirectional overlaps. Figure 1 shows the observed distribution, with the clear phase +1 bias in overlaps of 7 bp or more. This bias persists even for genomes grouped by GC% (data not shown).

Unidirectional pairs. These ratios are in good agreement with published results (Fukuda et al. 2003; Cock and Whitworth 2007; Kingsford et al. 2007). For the rest of this paper, we focus on unidirectional overlaps. Figure 1 shows the observed distribution, with the clear phase +1 bias in overlaps of 7 bp or more. This bias persists even for genomes grouped by GC% (data not shown).

Figure 2 shows overlaps generated from the last out-of-frame stop codon within each nonoverlapped downstream gene, that is, consideration of C-terminal extension of an upstream gene by adoption of a new stop codon. This shows only a small phase bias, but it is opposite to that observed in the annotated genomes.

Figure 3 shows overlaps generated from the first out-of-frame start codons within each nonoverlapped upstream gene, that is, consideration of N-terminal extension of a downstream gene by adoption of a new start codon. This shows the same phase bias observed in the annotated genomes (fig. 1), although the distribution of short overlaps is very different. In particular, a large number of potential overlaps of length 5 bp have been identified, all of which use the alternative start codon ATT. However, only a handful of such overlaps were found in the original survey, and this analysis is overly simplistic as the different start codons are not necessarily equally likely. Indeed, some of the start codons defined in the relevant NCBI translation tables (11 and 4) are only used in a small subset of the prokaryotes.

It would be possible to extend the above analysis with an acceptance weighting based on observed start codon frequencies. Instead, figure 4 shows the more straightforward approach of only considering the three most common start codons (ATG, GTG, or TTG). Again there is a strong phase +1 bias in the longer overlaps, and now the ratios of overlaps 1 and 4 are much closer to those observed (fig. 1).

+1 bias in the longer overlaps, and now the ratios of overlaps 1 and 4 are much closer to those observed (fig. 1).

One marked difference between the observed distribution (fig. 1) and figures 3 and 4 is the very different decay rates. Kingsford et al. (2007) observed a similar phenomenon in their analysis of convergent overlaps and
resolved this with a two-stage model, whereby there is a phase bias at overlap creation due to codon frequencies, but with overlap length subject to an exponential fitness, which can be determined empirically to match the observed data. Figure 5 show the data normalized with an exponential decay of 0.0931 (least squares difference fitting for overlaps up to 200 bp). There is good agreement with figure 1, but the phase bias is slightly less pronounced.

Two mechanisms for neighboring unidirectional genes to become overlapped were considered. C-terminal extension of the upstream gene (fig. 2) does not explain the observed pattern in unidirectional gene overlaps (fig. 1). However, considering N-terminal extension of the downstream gene by looking for alternative start codons does predict the observed phase bias (figs. 3 and 4), largely explained as due to the relative frequencies of alternative start codons in the two reading frames. Together with an exponential fitness criteria on the overlap length, this predicts a distribution close to that observed (fig. 5). This proposed exponential fitness cost could be due to the metabolic burden of making a longer protein, the increased likelihood of problems with protein misfolding/aggregation, or a combination of these or other effects. This model is compatible with mutual sequence constraint arguments as in Cock and Whitworth (2007) but provides a much clearer explanation.

Sabath et al. (2008) describe a related analysis of unidirectional overlaps, which also found that the phase bias in longer unidirectional overlaps could be explained in terms of the relative abundance of alternative start codons within an upstream gene and rejected the complementary explanation of the adoption of alternative stop codons within a downstream gene. Rather than observing alternative start and stop codons directly from real gene sequences, their frequencies were inferred from di-codon frequencies taken as the product of observed codon frequencies (assuming di-codon frequencies are independent). This cannot capture any differences in codon bias within genes, for example, between the 5′ and 3′ terminal regions of a gene. Also, a much smaller data set was used, drawing on annotated overlaps from only 167 genomes. Nevertheless, their work is supportive of the results given here.

Translational coupling provides a biological reason for short unidirectional gene overlaps and thus, the very large number of overlaps of 1 or 4 bp. This may also apply to 5 bp overlaps, which could be tested in vitro, suggesting that the high number of unidirectional overlaps generated using the "rare" alternative start codon ATT (fig. 3) may be of biological relevance, with the handful of cases annotated being just the tip of the iceberg. Eyre-Walker (1996) noted skewed ratios of alternative stop codons in short overlaps of 1 or 4 bp, so atypical start codon usage in this context is not unreasonable.

Although the genetic code itself appears to induce these phase biases in longer overlaps, without searching for Shine–Dalgarno translation initiation sites or direct experimental evidence, it is not clear how many of these annotated long unidirectional overlaps are biologically relevant. Although translational coupling provides a biological reason for short unidirectional gene overlaps, it may not apply to the longer overlaps reported. A recent analysis by Pallejà et al. (2008) concluded all unidirectional overlaps over 60 bp in their data set of 338 prokaryotic genomes were mis-annotations, but did identify some "real overlaps." However, the phase bias is still found when only genes with annotated
functionality are considered (Cock and Whitworth 2007; Lillo and Krakauer 2007; Sabath et al. 2008), and 98.1% (3 \text{ sf}) of annotated unidirectional overlaps are 60 bp or less.

The phase patterns in overlapping genes have no immediately apparent evolutionary function, but rather are inherently linked to the genetic code itself, which has evolved under various pressures. Itzkovitz and Alon (2007) explored a range of hypothetical genetic codes and concluded that those observed in nature are near optimal for encoding additional information within a protein sequence. This work did not specifically mention nucleotide sequences simultaneously encoding two proteins, but rather arbitrary (short) sequences representing possible DNA-binding regions or other motifs. Furthermore, the presence of alternative out-of-frame stop codons within a gene (hidden stop codons) has been looked at from the point of view of robustness to translational frameshift errors (Seligmam and Pollock 2004). The standard genetic code was found to terminate erroneous reads sooner than hypothetical genetic codes, which is advantageous as less resources are wasted constructing and degrading nonfunctional proteins. It seems reasonable that functions like double coding and hidden stop codons may have shaped the genetic code and thus indirectly contributed to the overlap phase patterns observed.

Methods

Separation/overlap frequencies were tabulated for divergent, convergent, and unidirectional adjacent annotated genes in 1,800 GenBank files for the 962 bacterial or archaeal species available from the NCBI as of 7 September 2009. For simplicity, any genes with nonexact locations, ambiguous sequences, internal stop codons, invalid start or stop codons (as verified using the declared genetic code), or special cases with noncontinuous coding sequences (e.g., from ribosomal slippages) were excluded, as were cases where one gene was entirely within another (fully overlapped). Separated gene pairs with any ambiguous sequence between them were also excluded.

For each nonoverlapping unidirectional gene pair, the downstream gene was searched for the first out-of-frame stop codon, whose location determined the length of a hypothetical unidirectional gene overlap. Additionally, the upstream gene was searched for the last out-of-frame start codon, the location of which also determined a hypothetical unidirectional gene overlap. If, due to the presence of a nearby in-frame stop codon, this hypothetical gene would be encoded within the upstream gene, it was rejected. Two variants of the start codon analysis were performed, first looking for any valid potential start codon in the genetic code declared for that organism and second looking only for the most typically used start codons (ATG, GTG, or TTG).

The analysis was written in Python using Biopython (Cock et al. 2009). Figures were drawn with R (R Development Core Team 2007).

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council via the MOAC Doctoral Training Centre (studentship to P.J.A.C.).

References


