A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes

ABSTRACT

Our model suggests that GC content drives codon usage (rather than the converse). It unifies a large body of empirical evidence concerning relationships between GC content and amino-acid or codon usage in disparate systems. The relationship between GC content and codon and amino-acid usage is ahistorical; it is replicated independently in the three domains of living organisms, reinforcing the idea that genes and genomes at mutation/selection equilibrium reproduce a unique relationship between nucleic acid and protein composition. Thus, the model may be useful in predicting amino-acid or nucleotide sequences in poorly characterized taxa.

INTRODUCTION

Background Different organisms have idiosyncratic, and sometimes extremely biased, preferences for one synonymous codon over another. Although differences in codon usage among genes and species have been widely studied, general principles have been difficult to find. Although it has been known for some time that the frequencies of some codons and amino acids correlate with genome GC content, the causality has remained unclear: correlations could exist because selection for a particular codon or amino-acid usage produces a particular genome GC content, or because mutation towards a particular GC content determines codon and amino-acid usage according to combinatorial principles. Here we show that codon and amino-acid usage is consistent with forces acting on nucleotides, rather than on codons or amino acids, although both mutation and selection play important roles. Codon usage can be surprisingly biased in different species. For example, the amino acid lysine has two codons, AAA and AAG. Although some organisms, such as Lactobacillus acidophilus, use the two codons equally, others show extreme preferences: Streptomyces venezuelae uses AAA only 2.2% of the time, whereas Buchneria aphidicola uses it for 91% of lysine residues. Amino-acid usage also differs greatly among species: for instance, the amount of arginine varies almost ten-fold, from less than 1.5% of all amino-acid residues in species of Borrelia to 12.7% in Mycobacterium tuberculosis (data from). Because of these extreme biases, knowing an organism's preferred codon usage is of direct practical relevance in minimizing degeneracy of PCR primers and in maximizing the effectiveness of in vivo genetic manipulation. Trends in codon usage across species could also influence molecular phylogenetic reconstruction, and clarify the relative roles of neutral evolution and natural selection in determining nucleotide sequences. The evolutionary theory of synonymous codon usage began with two separate lines of research, both of which suggested that most substitutions were selectively neutral, but which explained different phenomena. The first line sought to explain interspecific variation in overall sequence composition, and noted correlations between GC content and amino acid content across different species. This suggested that genomes were at equilibrium with respect to mutation, and explained how directional mutation could affect the composition of coding sequences, although it does not explain why species with similar genome compositions have recognizably distinct sequences for individual genes. The second line sought to explain the origin and maintenance of sequence variation within populations, and the fixation of particular alleles between species. This relied on the concept of silent mutations and the relative power of selection

and drift in small populations. Different usage patterns of synonymous codons are invisible at the protein level: how can selection operate when the amino-acid meaning remains unchanged? However, without directional mutation pressure, the fixation of silent mutants would not lead to the extreme biases in synonymous codon usage actually observed. Subsequently, codon usage in a few species has been extensively characterized, and linked causally to a wide variety of both adaptive and nonadaptive factors including tRNA abundance, gene expression level, local compositional biases, rates and patterns of mutation, protein composition, protein structure, translation optimization (but see), gene length, and mRNA secondary structure. In contrast, trends across species have received far less attention. The genome GC content has been shown to correlate with cross-species differences in frequencies of codons and amino acids. Genome composition may even influence the structure and chemistry of proteins. Comparing different microbial genomes, codon usage in individual genes also correlates with estimated expression level and tRNA copy number. One important point is that these regressions are ahistorical: by predicting a relationship between gene and protein composition, these studies imply that the history of a gene or species is unimportant compared to its current state. This has important implications for species or genes that have uncertain phylogenetic relationships or differ greatly in composition from their close relatives. Although closely related organisms tend to have similar genome compositions, there are considerable exceptions (such as Mycoplasma pneumoniae versus M. genitalium). If distantly related species with similar GC contents have the same amino-acid or codon usages, we can conclude that phylogenetic constraints are relatively unimportant, and perhaps that genomes are at or near equilibrium with respect to mutation and selection (otherwise, different unrelated species would not attain the same amino-acid composition predicted from the nucleotide composition). Such ahistorical relationships are particularly useful in cases where the goal is prediction of the current state of a sequence (for example, for making PCR primers), rather than reconstruction of its history. Although regression lines have been fitted to relationships between GC content and codon and amino-acid usage empirically, permitting qualitative inferences, quantitative theoretical predictions relating these responses to each other have thus far had limited success. This can be remedied by taking into account the differential effect of selection on the different positions within codons. Here we present a simple model, based solely on purifying selection and mutation at the nucleotide level, that quantitatively predicts both codon and amino-acid usage trends across archaea, bacteria and eukaryotes on the basis of the genome GC content. The model also provides insights into the causality between genome composition and protein composition. Every nucleic acid sequence necessarily has an associated GC content, but there need be no similarities in codon usage between different species with the same GC content (for instance, any specified GC content could be obtained by mixing AAA and GGG codons in different ratios). If GC content were an artifact of selection for a particular codon or amino-acid usage, there would be many different ways of arranging the codon frequencies to get the same GC content. If, on the other hand, the codon and amino-acid usage is an artifact of mutation (or selection) towards a particular GC content, the responses of the three codon positions to directional nucleotide substitution predict a single codon or amino-acid usage for each GC content. Thus, if distantly related species fit the response curves predicted by the model, we can conclude either that forces at the nucleotide level drive codon and amino-acid usage, and there is nothing special about certain codons or amino acids, or that there is a unique spectrum of preferred codon and amino-acid usages that applies to all species, extends over a huge range

of compositions, and happens to match the predictions of the model by chance.

CONCLUSION

Conclusions We have shown that the GC content of individual codons and amino acids is the primary determinant of their response to biases in sequence composition, both among and (to a lesser extent) within genomes. Although the literature contains many examples of correlations between GC content and the frequency of particular codons and amino acids, our model is able to recapture quantitatively the behavior of essentially all codons and amino acids by invoking forces that act only on the level of individual nucleotides. This is likely to be due to a combination of mutation and selection: mutation can act in parallel across an entire genome, changing many sites simultaneously; however, this process is limited by the consequences of error at each position. The simplest hypothesis, that codon usage depends solely on codon GC content, fits the data poorly (compare orange lines with red, green and blue lines in Figure 4). One can, however, explain most of the variance in the response of both codons and amino acids by taking into account the fact that the three codon positions change at different rates, and that the four nucleotides are not evenly distributed among the sites that are functionally constrained. Additionally, accounting for the fact that the four nucleotides change at different rates allows some further improvement, which ranges from minimal to drastic depending on the exact circumstances. This supports the basic principle of neutral evolution, the idea that most change in nucleotide and protein sequences is driven by mutation and limited by purifying selection that varies for different sites and molecules (reviewed in). Within this context, it supports the idea that most of this neutral change is driven by directional mutation, which thus explains differences in nucleotide composition among species. Although the conclusion that amino acids with GC-rich codon doublets are more frequent in GC-rich genomes, and that those with AU-rich codon doublets are more frequent in AT-rich genomes, is neither new nor surprising, our model accurately and quantitatively predicts these responses for essentially all codons and amino acids by invoking forces acting on individual nucleotides. The genetic code constrains which codons and which amino acids can respond to biases in nucleotide composition, in part because mixed codons necessarily respond more slowly to forces acting on particular types of bases than do homogeneous codons. Thus, although GC content only explains the variance in usage of some codons and some amino acids, we can accurately predict which codons and amino acids will show clear responses and, for those that do show clear responses, accurately predict their frequencies in particular genomes (for example, Figure 1 shows an example of a codon for which 85% of the variance in usage is explained by genome GC content, and an amino acid for which 79% of the variance is explained). Thus, especially for species with few close relatives, variable sites may even be more useful for predicting PCR primer sequences than conserved sites, although this will depend on the particular sequence and genome composition. We have focused on codon usage at the level of whole genomes (or samples of genes where whole genomes are not available), an area that has received relatively little attention. This large-scale view does not consider the selective factors influencing individual genes, and the fact that the model provides much better fit across genomes than within them may reflect local adaptation to factors such as expression level. What remains surprising is that our simple model can explain so much of the variance in codon and amino-acid response to GC content in these different systems. Identifying deviations from the

predictions based on nucleotide composition may identify genes that are under unusual selection pressures, whether for a particular amino-acid composition or for a specific pattern or degree of codon bias. The fact that both amino-acid and codon usage are so closely entwined with genome composition has important practical implications. For phylogenetic analysis, the fact that some amino acids (such as arginine) change rapidly and predictably with GC content slightly undermines the idea that amino-acid sequences are more stable than nucleotide sequences: pairs of species with convergent GC contents might also evolve convergent protein sequences, especially at functionally unconstrained positions. For example, the frequencies of both lysine and arginine are highly (but oppositely) correlated with GC content, and lysine and arginine can easily substitute for one another in proteins. Each of the three domains of life has explored a wide range of genome GC contents, and organisms at the extremes of the range but with different evolutionary histories may share more convergent amino-acid substitutions than currently recognized. For sequence analysis, the prospects are more promising: given very limited information about a species (the GC content), it may be possible to estimate the codon usage and therefore minimize the degeneracy of PCR primers, even if no closely related species have been characterized. Organisms with extreme genome compositions, or with genome compositions that differ markedly from their close relatives (such as Mycoplasma pneumoniae versus other mycoplasmas) should be particularly accessible. This should be especially useful in developmental genetics and in environmental applications where model systems are not available. The fact that the model holds independently for different lineages of organisms (for example, bacteria and eukaryotes), and, to a lesser extent, for individual genes within species, strongly suggests that the trends are ahistorical. Given rates of change for each nucleotide at each codon position, determined jointly by selection, mutation, and the genetic code structure, we can predict the codon and amino-acid composition of a particular sequence from its overall compositional properties, without reference to related sequences. Interestingly, the history of a sequence seems relatively important in determining its codon and amino-acid usage. This fact is likely to be particularly important in cases where a species diverges greatly in GC content from its closest relatives: knowing its GC content will allow much better prediction of specific gene sequences than simple comparison with conserved sites in related sequences (which may in some cases be similar because of shared genome composition rather than functional constraint). Finally, our model explains many of the details of individual codon and amino-acid responses over the wide range of genome compositions found in nature. Perhaps surprisingly, individual amino acids with specific structural or functional roles within proteins (such as and arginine) respond to GC content no differently than the rest, and their frequencies can be very sensitive to genome composition despite the effects this might have on the properties of the translated products. This ability of amino-acid frequencies to vary so widely implies that functional proteins may be less constrained by sequence (and therefore easier to evolve) than previously imagined.