Isochores, GC content, and Codon Bias

1. Isochores and GC Content

%GC= percent of G and C nucleotides in the genome (G=C, A=T, but GC may not = AT)

%GC varies among species (particularly bacteria, plants, invertebrates; little variation among vertebrates).

However, vertebrates show much more %GC heterogeneity within their genomes
The vertebrate genome can be divided into isochores = long stretches (100s of kb) of DNA with uniform %GC. With completion of the human genome, it was found that isochores could span >10 Mb

High %GC = heavy isochores (H)
Low %GC = light isochores (L)

Traditional isochore classification of human genome:
L1 (<37% GC), L2 (37-42%), H1 (42-47%), H2(47-52%), H3(>52%)
Often L1 and L2 are grouped together as a single L isochore

Heavy isochores are found only in warm-blooded vertebrates (mammals, birds), not in cold-blooded vertebrates (fish).

Why Isochores?

1) Selectionist hypothesis – GC-pairing is stronger than AT-pairing (3 vs. 2 hydrogen bonds) and may stabilize DNA at higher temperatures. This hypothesis is supported by the observation that heavy isochores are found in warm-blooded vertebrates. Furthermore, heavy isochores are gene-rich.

2) Mutationist hypothesis – The pool of available nucleotides changes over replication (which takes 8 or more hours for mammals). There is more GC available early in replication, so mutations will be biased towards G or C. Over time, regions of the genome that replicate early become GC-rich. In general, GC-rich regions have been observed to replicate early.

More recent analysis of complete human genome:
See: Cohen *et al.* (2005) Mol. Biol. Evol. 22: 1260-1272.

By doing a sliding window analysis of the human genome DNA sequence and defining isochores as segments >300 kb with distinct %GC and low heterogeneity, the authors found that isochores covered 41% of the human genome, and most had low %GC. They suggest a four-family model with mean GC contents of 35%, 38%, 41%, and 48% and conclude:

"These findings undermine the utility of the isochore theory and seem to indicate that the theory may have reached the limits of its usefulness as a description of genomic compositional structures"

On a much smaller scale, %GC varies among different regions of a gene. Typically,

Coding regions > introns > 5' flanking regions > 3' flanking regions

However, there is a strong correlation in %GC among all regions of a gene.

2. Codon Bias

An analysis of many protein-coding sequences from a species indicates that all of the synonymous codons for a particular amino acid are not used with equal frequency as would be expected at random. This phenomenon is known as "codon bias". Certain codons are "preferred" and are used much more frequently than "unpreferred" codons. For example, Leucine is an amino acid that shows very high codon bias. Leu can be encoded by 6 different codons, CTG, CTA, CTC, CTT, TTG, TTA. At random, we would expect each codon to be used about 1/6 (17%) of the time. However, in highly expressed *E. coli* genes, CTG is used ≈90% of the time. In yeast, TTG is used ≈90% of the time. The preferred codons correspond to the most abundant tRNA in each species, suggesting that selection favors the use of codons that increase the level of gene expression. Note that the preferred codons may differ from species to species.

The level of codon bias in a gene can be described by many measures. Two commonly used statistics are:

ENC = effective number of codons, the average number of codons that are used to encode the 20 amino acids. The minimum is 20 (one codon per a.a.) the maximum is 61 (all codons except the 3 stop codons). **Low ENC = high codon bias**.

Fop = frequency of optimal codons, the frequency with which the "optimal" codon is used for each amino acid. Optimal codons are defined as those used with the highest frequency in highly expressed genes. **High Fop = high codon bias**.

Fop is species-specific and requires that optimal codons are known. For this, one must have many gene sequences and expression information.

ENC can be applied to any species without prior knowledge of expression or codon usage.

Some observed patterns of codon bias:

a) higher in highly-expressed genes
b) higher in short genes than in long genes
c) higher in female-expressed than in male-expressed genes

Why is there codon bias?

Two explanations:

a) Selection – natural selection favors the use of optimal codons (those that correspond to the most abundant tRNA) to make translation faster and more accurate. Codon bias could be used as a way to regulate gene expression post-transcriptionally.

Evidence:
Highly expressed genes have higher codon bias

Conserved protein motifs, such as DNA binding domains, have higher bias than other protein regions. This suggests selection for accuracy of translation.

Experimental replacement of optimal codons with non-optimal codons reduces the level of protein. Example: Drosophila Alcohol dehydrogenase (ADH)

Replacement of optimal leucine codons with non-optimal codons leads to a lower level of ADH protein *in vivo* and reduces ethanol tolerance in adult flies

Wa-F = wild-type, optimal leucine codons
1 leu = 1 luecine codon changed from optimal to non-optimal
6 leu = 6 leucine codons changed from optimal to non-optimal
10 leu = 10 luecine codons changed from optimal to non-optimal

In a comparison of ADH protein concentration, it was found that:

Wa-F > 1 leu > 6 leu > 10 leu

The wild-type flies were also more tolerant to ethanol that the mutant flies. The LD50 (the ethanol concentration at which 50% of the flies were killed within 24 hours) was 9% for wild-type flies. It was only 7.5% for 10 leu flies.

Replacing sub-optimal leucine codons in the *Adh* gene with optimal codons increases ADH enzymatic activity in larvae, but decreases it in adults. This suggests that there may be a trade-off between optimal codon usage (or other factors) in different developmental stages. Codon bias is highest for larval-expressed genes.

b) Mutation (neutral) – there may be a bias in mutation. For example, if mutations from A or T to G or C are more frequent and there is no selection on synonymous sites, then these sites should become GC-rich. The reverse bias in mutation would lead to synonymous sites being AT-rich. In both cases, this would lead to non-random codon usage.

Most optimal codons end in G or C. Thus, the mutational hypothesis would require a mutational bias towards G or C. However, most observations indicate that the mutational bias is towards A or T, which runs counter to the prediction. However, there does appear to be biased mismatch repair in favor of G or C. This tends to be strongest in regions of high recombination (where most of the highly codon-biased genes are located). So this process could explain at least part of the observed codon bias. The GC content at third codon positions is correlated with local genomic GC content, suggesting an effect of mutation of codon usage.

The strength of selection acting on a particular codon is expected to be very weak (on the order of $Ns = 1$, where N is the population size and s is the selection coefficient). This means that it is often very difficult to distinguish between selective and neutral explanations for codon bias.

Typically, there is evidence for selection affecting codon usage in organisms with small genomes and large population sizes (bacteria, yeast, Drosophila). The evidence is much weaker in humans and other vertebrates, where it appears that mutational/repair biases can explain the observed codon usage.