Codon usage

1 Introduction
2 Chromosomes Topology & Counts
3 Genome size
4 Replichores and gene orientation
5 Chirochores
6 G+C content
7 Codon usage

Degeneracy of the genetic code

Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr UAA: Stop UAG: Stop	UGU, UGC: Cys UGA: Stop UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

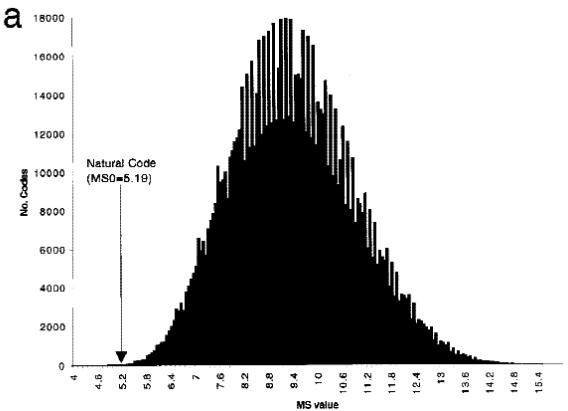
The genetic code is one in a million

The Genetic Code Is One in a Million

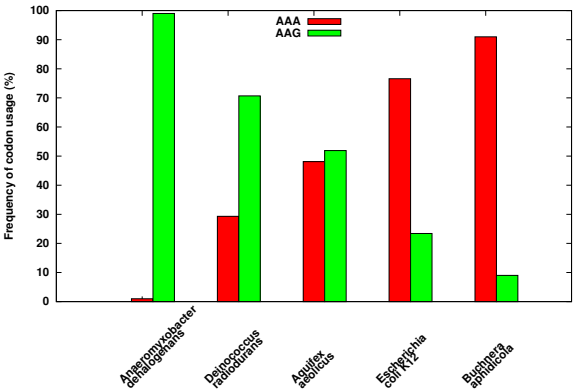
Stephen J. Freeland,[1] Laurence D. Hurst[2]

[1] Department of Genetics, Downing Street, Cambridge CB2 3EH, UK
[2] Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA27AY, UK

The genetic code is one in a million



Codon usage in Bacteria



Possible causes of codon usage in prokaryotes

- Random
- DNA composition biases : GC content, chirochores
- Optimization of translation

Random codon usage?

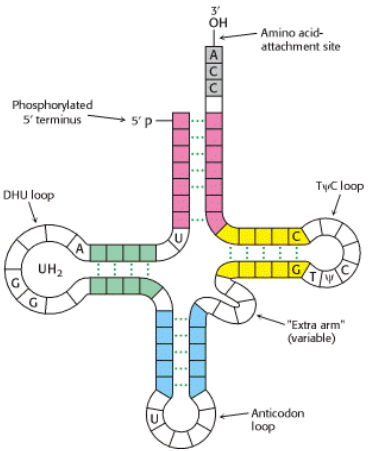
- If codon usage was evolving according to neutral evolution theory, each gene (and more generally each sequence fragment) should have a random usage
- This could not lead to a global, per organism, biased codon usage
- The organism level biased codon usage was called the *genome hypothesis*

Codon bias as a consequence of composition biases

- The global or local GC bias (and more generally the DNA composition) can affect codon usage, as 2-fold and 4-fold degenerated synonymous codons end either in A/C/T/G, in C/T or in A/G. The same could be thought off for 6-fold degenerated codons, but would not be so easily measurable. Third codon position composition biases are the mostly studied measure for this bias.
- The strand on which the gene is coded can be important, due to asymmetric DNA composition between strand (GC skew). This is well-seen in *B. burgdorferi*, whose genes have two distincts codon usages depending on their strand (leading or lagging).

Translation optimization and tRNAs

- tRNAs are small RNAs that link an amino-acid to the peptide sequence
- They have a palindromic structure
- They are amino acid specific AND codon "specific" (wooble)
- They differ greatly in number in the cell (from ∼100 to ∼5000)



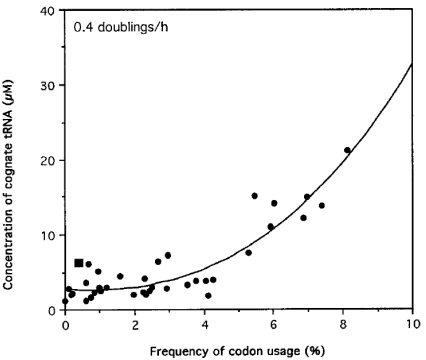
Wooble rules

Some tRNAs are able to recognize more than one codons. *E. coli* has 86 tRNAs but only recognize 39 anticodons out of 61. The matching rule is known as "wooble rule":

Anticodon 34	Complementary base
A	Unemployed
I	U,C,A
C	G
G	C,U
U	A,G
U_{modif}	G or A,G or A,G,U,C

Wooble rules

Relation between tRNA concentration and codon bias



Codon usage is related to tRNA content in cells. This could be due to:

- Energy of tRNA-codon link optimization – *dismissed because of diversity of codon usage*
- Ribosome waiting time optimization:
  - Translation accuracy
  - Translation speed

A simple model to explain this relation

Suppose that the time necessary to translate a given codon i of frequency f_i^a is inversely proportional to the concentration t_i^a of the tRNAs able to read it (which makes sense by diffusion laws). Then the time to decode a mRNA is of the form:

$$\sum_{j=1}^{61} f_i^a \frac{1}{t_j^a} \tag{1}$$

One can try to minimize this total time, by globally increasing the concentration of all tRNAs. At a fixed global concentration $T = \sum_i t_i^a$, one can derive the value of each t_i^a. This value is such as, for synonymous codons i and j:

$$\sqrt{\frac{f_i^a}{f_j^a}} = \frac{t_i^a}{t_j^a} \tag{2}$$

One model, many theories

This simple model hides a wealth of complex possibilities. A brief list is:

- The accuracy of translation of each codon could be considered as inversely porportional to the waiting time at the ribosome (the more you wait, the more probable you insert a wrong tRNA). Therefore, you could optimize *accuracy* or *speed* with the same model.
- Due to wooble rules, some tRNA decodes more than one codon, and *decoding is not as accurate or as fast in all cases* (factor 6 in speed).
- Translation speed can be limited by the waiting time of tRNAs at the ribosome (elongation models), *or* by the waiting time of ribosomes for each mRNA (loading). If loading is more limiting, only codon bias at the very beginning of the mRNA could be relevant.

Measures of codon bias

- Many measures of codon bias exist, and are mainly used to compare genes inside an organism
- Some measures need an *a priori* knowledge of the major codons, other the knowledge of a set of reference genes such as highly expressed genes
- Many measures are highly correlated, but only some of them are used commonly, for historical reasons

Effective number of codons N_c

This measure computes how many codons are effectively in use in a gene, from 20 (strictly one per amino acid) to 61 (random usage). The measure is, for the classical genetic code:

$$\hat{N_c} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}, \tag{3}$$

with

$$F_n = \frac{1}{Q_n} \sum_{a=1}^{Q_n} \frac{n_a \sum_{j=1}^{n} \left(f_j^a\right)^2 - 1}{n_a - 1} \tag{4}$$

Q_n being the number of amino acid degenerated n times, n_a the total number of codons for this aa, p_j the frequency of codon j relative to its synonymous.

Fraction of optimal codons

Given a predefined set of "preferred codons", or "major codons" – often the codons used preferentially by the ribosomal proteins – one can compute in each sequence:

$$F_{op} = \frac{N_{maj}}{N_{tot}} \tag{5}$$

where N_{maj} is the number of major codons in the gene, and N_{tot} the total number of codons.

Codon Adaptation Index

Given a predefined set of reference genes, one can compute the *adaptativity* on this set, as:

$$w_i = \frac{f_i^a}{\max_i f_i^a} \tag{6}$$

with f_i^a the frequency of codon i relative to all codons for aa a. Then, the CAI is:

$$CAI = \left(\prod_{k=1}^{N_{tot}} w_k \right)^{1/N_{tot}} \tag{7}$$

The CAI then measures a distance to the codon usage of the set of reference genes, usually the highly expressed genes.

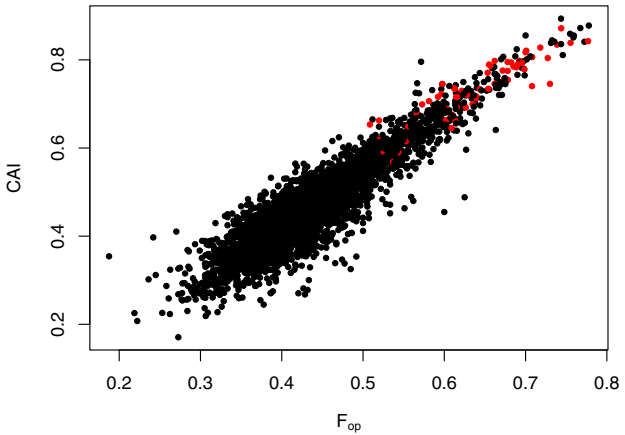
Why measuring the codon bias?

- In many cases, and in particular the CAI, codon bias is an indicator fo the expressivity of genes in bacteria, where translational control is more important than transcriptional control.
- Good measures of codon bias allow to predict the level of expression or the function of genes from the sequence, as genes sharing the same function and expression pattern tend to have the same codon usage.
- Codon usage can be an indicator of the origin of genes.
- Codon usage can indicate if a sequence is submitted to selection or not.

Do it yourself

- Get the file `http://pbil.univ-lyon1.fr/members/mbailly/AMIG/data/ecoli.gene.codons`
- Compute the F_{op} and CAI of each gene (using what you did for the GC can save a lot of time)
- Compare F_{op} and CAI

What you should get



What you should get (details)

```
> head(fop[1,])

     thrL      thrA      thrB      thrC      yaaX      yaaA
0.6190476 0.4390244 0.4258065 0.4556075 0.4489796 0.4031008

> head(cai)

     thrL      thrA      thrB      thrC      yaaX      yaaA
0.6986528 0.4633089 0.4628072 0.5042948 0.4596805 0.4736422

> cor.test(fop[1,],cai)

Pearson's product-moment correlation

data:  fop[1, ] and cai
t = 150.8992, df = 4297, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9123168 0.9218154
sample estimates:
      cor
0.9171964
```

A bit of help

- You have to find which genes code for ribosomal proteins. The names of these genes start by *rp*, and the third letter is either o,l,s or m[3]. The `grep` function and *regular expressions* can help you to find them.
- Compute first the F_{op}.
- To compute the CAI, you should not take into account codons with counts equal to 0, and first compute the logarithm of the CAI.

[3] Don't ask me why...

A solution (1)

A solution (1 bis)

A solution (1 ter)

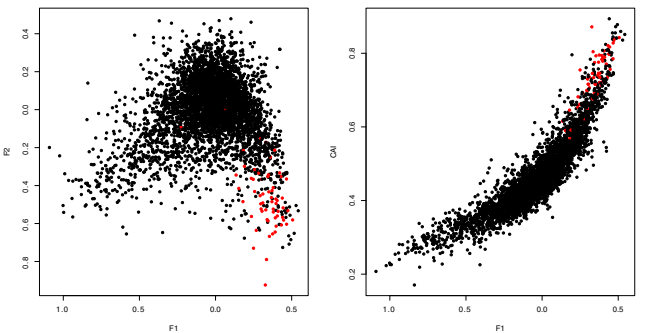
A solution (1 four)

A solution (1 five)

A solution (2)

Do you think you are done?

Now, on the same dataset, realize a Factorial Correspondance Analysis, and look at the correlation between the first axis of the FCA and *CAI*.



A solution

You are still there?

If you have time, you can look at another point of view about codon usage at
`http://pbil.univ-lyon1.fr/R/pdf/tdr623.pdf`