Pure & Appl. Chem., Vol. 65, No. 9, pp. 1947–1950, 1993.
Printed in Great Britain.
© 1993 IUPAC

Correlation between genome size, observed codon preference, and Gibbs energy of codon–anticodon interaction

H.H. Klump, Dept. Biochemistry, University of Cape Town, Rondebosch, 7700, South Africa

Abstract

The universal correspondence between the primary structure of a gene, characterized by a unique sequence of only four different nucleotide residues, and the primary structure of a protein, characterized by a given sequence of twenty different amino acid residues, is known as the "genetic code". This code is degenerated in the sense that a group of six, four or just two synonymous codons can code for a given amino acid. In this paper I will discuss only the special features of duet codons and in particular the effect of the purine/pyrimidine base proper in the third position, the observed substitution frequency, and the major determinant of the codon usage bias in prokaryots and in eukaryots.

INTRODUCTION

Large scale sequencing of DNA from a variety of sources has led to a new insight into the bias affecting the observed frequency of any of the 64 possible codons. We can still hold that the classic genetic code can be viewed as the universal language of all living beings. (1) It is obvious, however, from our knowledge as it stands today that many species speak their own dialect. Distinct codon strategies were found for different genomes. Yet a surprising consistency of choices was found among codons in genes of the same or of related genomes. (2) Within a given genome codon usage is conserved in most genes. This finding has been confirmed for numerous organisms. *E. coli* genes, coding for proteins with a wide variety of functions have now been sequenced, enclosing abundant proteins for ribosomes etc., as well as for proteins that are present in only a few copies per cell, such as lac repressor. There are two important points to make: First, there are similarities in codon choices among all different *E. coli* genes (3); Second, although all *E. coli* genes show this tendency to confirm to a specific codon usage pattern, there is a clear difference in the preferred codon set between highly and poorly expressed genes. (4)

It was found for many *E. coli* genes that the degree of bias in codon choice is directly related to the expression level of a particular gene. Using the codon usage data of these genes allows to deduce a set of favoured codons. It is interesting to note that the mean GC content (the percentage of G/C bases in either the first, second or third position of a set of codons for the 20 amino acids) of this biased set matches perfectly with the mean G/C content of the whole genome. (50 % /50% for *E. coli*). (5) It can be shown that this general rule holds for other prokaryots such as *B. subtilis* or *T. ferrooxidans* i.a. as well. Based on this rule one can propose a preferred for any prokaryot when the mean value of the G/C content of the genome is known. Synonymous codon choice patterns in other enterobacteriacaea such as *S. typhimurium, K. pneumoniae*, or *Erwinia amylovora* were found to resemble that of *E. coli*, confirming an earlier finding of Grantham who stated for prokaryots that taxonomically related organisms exhibit a related codon preference.

It was shown in the mean time that for eukaryots with a small genome size such as *S. cerevisiae* a similar relation between the G/C content (6) of the favoured codon set and the G/C content of the total genome applies. Again choices among synonymous codons are strongly biased, and a clear relationship between the degree of bias and the level of expression of the gene proper has been observed. The extend of this bias exceeds that found in *E. coli*. A comparison between the preferred codons for yeast with those in *E. coli* shows that these two organisms use completely different codons for five amino acids (glu, lys, pro, leu, and arg). (2)

The study of codon usage in higher eukaryots is complicated by the cellular differentiation. Examining the genes which code for the immunoglobulins, globulins, or peptide hormones it was found that mammalian genes use a codon strategy that must be very different from both bacteria and yeast genes. In general, A ending codons are avoided, C- and G- ending codons are predominant for twofold degenerated codons e.g. Various laboratories have examined codon usage in vertebrate muscle genes (avian alpha actin, myosine light/heavy chain i.a. and liver genes (rat albumin etc.) and have found a considerable degree of similarity of synonymous codon usage between the two groups of genes. Viewing the data of a large variety of genes from higher eukaryots one can conclude that for vertebrates a consensus codon usage pattern exists which is independent of taxonomic class (7) and tissues of expression. It should be noted, however, that minor but nevertheless important differences in codon usage can exist between highly and poorly expressed genes. It should also be added here that as in the case of *E. coli* and its phages the codon strategies of mammalian viruses (SV40, adenovirus, hepatitis v,) differ from those of their hosts. It is noteworthy that in all higher vertebrates NCG codons (N=A, T,C and G) in the serine,

proline, threonine and alanine codon quartets are almost completely absent. Obviously these CG combinations serve as important signals for other gene functions.

Various explanations have been offered in the literature for the observed codon usage bias. The many models proposed can be grouped according to their most important feature into four different categories. (8-10)

1.) Biased codon usage can only be effective when it is accompanied by a correspondingly biased tRNA profile or to phrase this in a different way: codon usage and tRNA availability are adapted to each other. (Translational efficiency)

2.) Pronounced mRNA secondary structure formation and stability will modify transcriptional efficiency.

3.) The strength of the codon/anticodon binding varies depending on the composition of the codon and the corresponding anticodon. The preference of certain codon/anticodon combinations of intermediate strength can limit the codon choice. (Transcriptional efficiency) It can be seen that intermediately stable codon/anticodon combinations are highly favoured in highly expressed genes from bacteria or from yeast.

4.) For higher vertebrates conclusions about coadaptation of codon usage and tRNA profile should be drawn with great care. Ribosomes, as can be shown, generally perform at their maximal speed. Different from yeast and bacterial genes the G/C content of the preferred codon set is (63-64%) almost 50% higher than the G/C content of the total genome (42%). High G/C content is equivalent to an exceedingly high interaction energy between a given codon and its anticodon. It also reflects a quest for high precision in transcription.

RESULTS

To understands the pronounced bias in codon preference of prokaryots, viruses, yeast, and higher eukaryots we will demonstrate for a subset of all codons, namely for the duet codons, how the observed codon usage frequency is correlated with the codon/anticodon interaction energy for each of the nine twofold degenerated codons. (5)

The Gibbs energy values for the codon/anticodon interaction were obtained experimentally as described previously with the help of adiabatic differential scanning calorimetry (DSC) applied to the helix coil transition of a set of carefully selected DNA sequences. (11) Figs 2 to 7 will exemplify for a variety of species, mammals, plants, chloroplasts, bacteria, and viruses, how in the case of the twofold degenerated codons the observed usage frequency and the Gibbs interaction energy are correlated and can serve to explain different strategies.

The twofold degenerated codons are chosen to simplify the picture. Thus we have only to take into account the two purines (G/A) and the two pyrimidines (C/T) respectively in the third position of each of the codons. A transition in that position will preserve the amino acid encoded, a transversion will only shift the coding to a substitution by an amino acid of similar character, i.e. size and/or charge.

Fig. 1

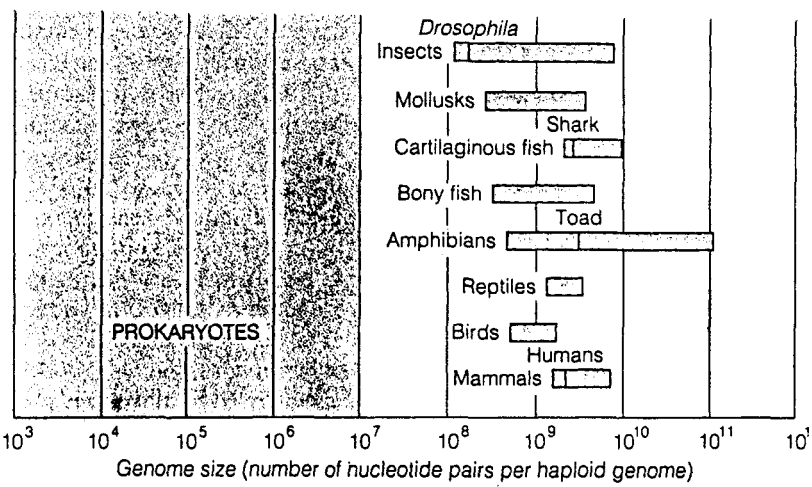

Fig. 1 shows an overview of a variety of genome sizes (one to ten million base pairs) for prokaryots in general and for a selected group of eukaryots (one to ten billion base pairs) to emphasize the large difference in size to understand the two conflicting goals eukaryots and prokaryots have to aim at to do best in their reproduction strategies.

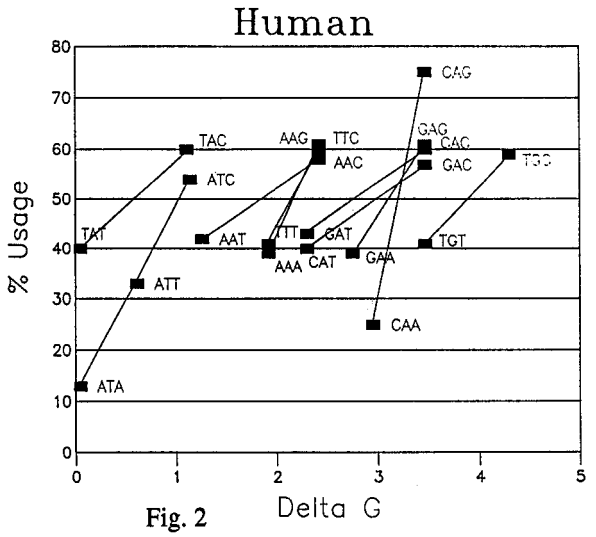

Fig. 2

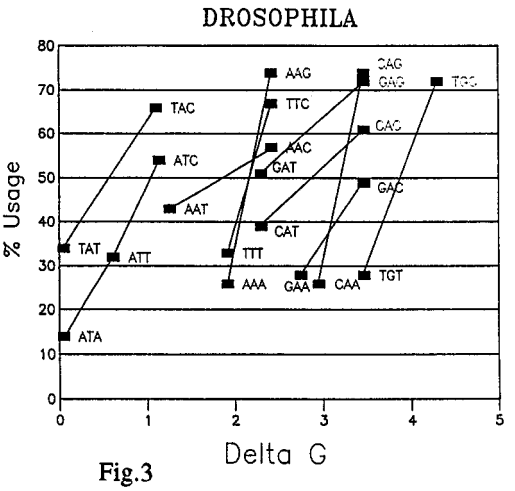

Fig.3

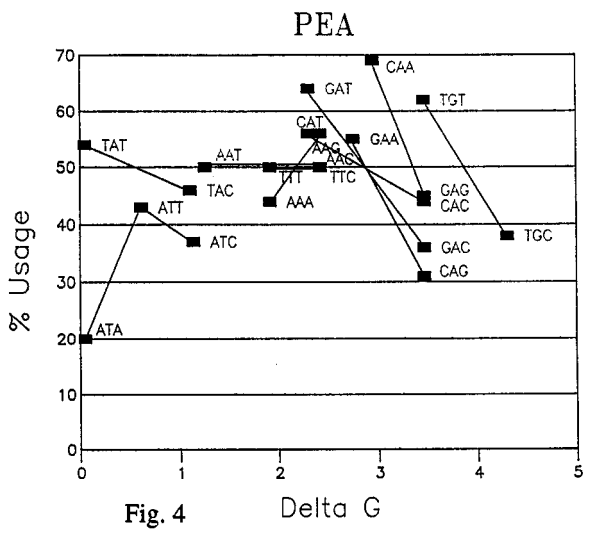

Fig. 4

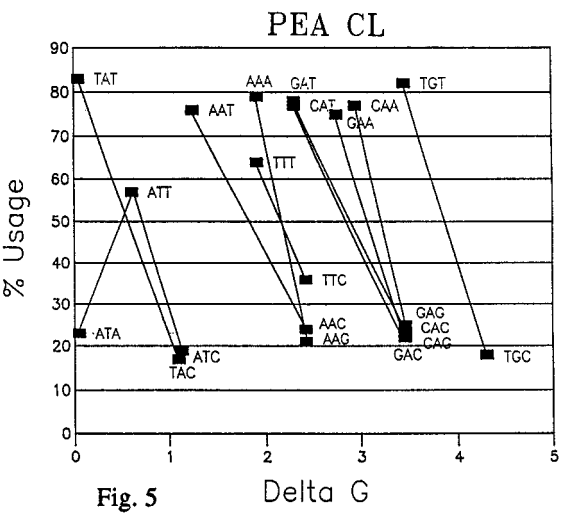

Fig. 5

Vertebrates have to go for precision more then for anything else to protect the integrity of their very large genome, (5) while prokaryots, which live always at the brink of starvation in their natural habitat, have to go for speed in the first place when they have managed to find an abundant food source, sacrificing the precision. In terms of thermodynamics, precision correlates with maximal Gibbs interaction energy whereas speed is reflected in an intermediate level of interaction energy to allow for fast and smooth reading by the DNA dependent RNA polymerase. So plotting the Gibbs energy for codon/anticodon interaction energy vs. the observed frequency (%) of a set of twofold degenerated codons for a given species will reveal its strategic aim. The units for the Gibbs energy are kcal per mole triplet codons. 1 kcal equals 4. 18 kJ. The codon usage is normalized to 100%. Figures 2-7 shall exemplify the correlation between the codon usage and the interaction free energy for men (Fig. 2), drosophila (Fig. 3), peas (Fig. 4), pea chloroplasts (Fig. 5), bacteria (Fig. 6) and viruses (Fig. 7).

Fig. 2 shows this plot for the human duet codons enclosing the threefold degenerated codons for ile. The corresponding two codons are connected by a straight line (ATC and ATT or AAA and AAG e.g.). The third position in each codon set determines the level of stability, adding an extra H-bond when C/G is present instead of T/A, which adds about 4 kJ to the molar Gibbs energy per codon.

In case of the human codon set, all lines are sloped to the right indicating that the more stable codon is favoured over the lesser stable codon (60%/40%). This tendency is generally observed for all vertebrates. In case of the fly it is even more pronounced (70%/30%). This pattern changes completely when we depict this correlation for plants, plant organelles, bacteria or viruses. Depending on the C/G content of the genome the lines are sloping to the right (high G/C content) or to the left (low G/C content) indicating

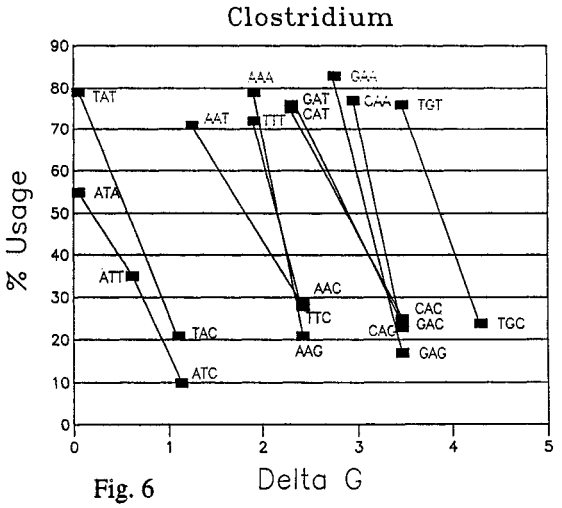

Fig. 6

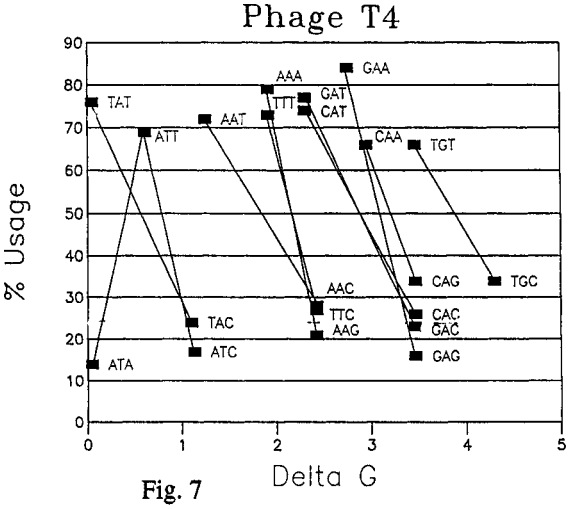

Fig. 7

that now the less stable codon set is favoured (80%/20%) as can be seen for chloroplasts, *phage T4*, or *B. subtilis*. It is difficult to explain some of the patterns like the pattern of the codon set for peas. We have not looked in sufficient depth into the plant case to come up with a rational explanation for the complicated pattern, but it may relate to the polyploidy of some of the plants.

CONCLUSIONS

Due to the limited space it is not possible to discuss any of the patterns in more detail, and to show the plots of the fourfold or the sixfold degenerated codons vs. the Gibbs energy. It is, however, obvious from this first brief overview that the Gibbs energy of codon/anticodon interaction correlates well with the observed codon bias and that these plots of Gibbs energy vs usage reveal the basic strategy of a species, the quest for precision or the urge for speed.

REFERENCES

1. J. Watson, in Molecular Biology of the Gene, Benjamin/Cummings, 4th ed., Menlo Park, USA (1987).
2. H. de Boer and R. Kastellin, in Maximizing Gene Expression (W. Reznokoff and L. Gold, eds.), pp. 225-282, Butterworth, London (and references therein) (1988).
3. H. Grosjean and W. Fries, Gene 18, 199-209 (1982).
4. M. Gouy and C. Gautier, Nucl. Acid Res. 10, 7055-7074 (1982).
5. H. Klump and D. Maeder, Pure & Appl. Chem 63, 1357-1366 (1991).
6. J. Bennetzen and B. Hall, J. Biol. Chem. 257, 3026-3031 (1982).
7. K. Hastings and C. Emerson, J. Biol. Evol. 19, 214-218 (1983).
8. H. Pfizinger, P. Guillemant, J. Weil, and D. Pillay, Nucl. Acid Res. 15, 1377-1386 (1987).
9. A. Wada and A. Suyama, FEBS Letters 188, 291-294 (1985).
10. R. Grantham, C. Grautier, M. Gouy, M. Jacobzone, and R. Mercier, Nucl. Acid Res. 9, r42-r74 (1981).
11. H. Klump, in Biochemical Thermodynamics 2nd ed. (M. Jones ed.), pp. 100-144, Elsevier, Amsterdam (1988).