Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 317 (2003) 137–140
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

The correlation between GC_3 and hydropathy in human genes

Kamel Jabbari[a], Stéphane Cruveiller[b], Oliver Clay[b], Giorgio Bernardi[a,b,*]

[a] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy*

Received 24 July 2002; received in revised form 1 October 2002; accepted 28 April 2003

Abstract

A positive correlation holds between the GC level of third codon positions of human genes (GC_3) and hydropathy of the encoded proteins. This correlation may appear counterintuitive, since it links a physical property of proteins to the base composition of 'synonymous' sites. We here establish the nontriviality of the correlation, which has recently been contested. In particular, the correlation cannot simply be a consequence of an analogous correlation for first and second codon positions, since no such correlation exists. More generally, for any explanation via two chained correlations, the intermediate property would need to be strongly correlated with hydrophobicity and/or GC_3.
© 2003 Elsevier B.V. All rights reserved.

Keywords: Hydrophobicity; Encoded proteins; Codon positions

Correlations between GC_1, GC_2 and GC_3 were reported several years ago, both inter- and intragenomically (Wada and Suyama, 1985; Bernardi and Bernardi, 1986; Sueoka, 1988; Aïssani et al., 1991; D'Onofrio and Bernardi, 1992; D'Onofrio et al., 1991, 1999). Correlations were also found to hold intergenomically between GC_3 and hydropathy; the nontriviality of this latter relation is important, because it links a compositional property of genes' 'silent' sites to a functionally relevant property of proteins (D'Onofrio et al., 1999). One would typically expect such properties to be reflected in the second rather than in the third codon position (see Chiusano et al., 1999 and references therein).

Correlations indicate clear relationships between variables that can often be characterized by straight regression lines. It is therefore easy to imagine that correlations should typically obey the principle of transitivity. Thus, from a line $y = ax + b$ and a second line $z = cy + d$ one can calculate the line relating z and x by a simple algebraic manipulation. This kind of transitivity does not follow for correlations; however. Significant correlations may exist between variables z and y, and between variables y and x, yet not between z and x.

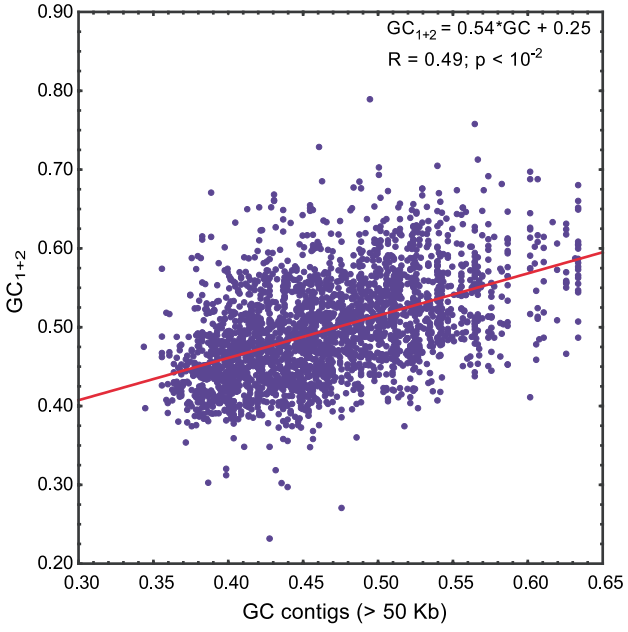

Fig. 1. Correlation between GC level of large (>50 kb) DNA sequences from human and GC_{1+2} of the embedded coding sequences.

Abbreviations: GC, molar ratio of guanine+cytosine; GC_1, GC_2, GC_3, guanine+cytosine at first, second and third codon positions.

* Corresponding author. Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy. Tel.: +39-81-583-3215; fax: +39-81-245-5807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

0378-1119/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00663-2

In a recent review article, Eyre-Walker and Hurst (2001) stated that "genes in the (G+C)-rich isochores yield proteins with different amino acid compositions (D'Onofrio et al., 1991) and hydropathies (D'Onofrio et al., 1999) to those in the (G+C)-poor isochores; both features seem to be a consequence of the correlation between isochore G+C content and GC_{12}". We would like to offer here three comments on this statement.

The first comment is that the authors' use of "isochore GC" invokes GC_3, the GC level of third codon positions, which is strongly correlated with isochore GC (Bernardi et al., 1985; Bernardi and Bernardi, 1986), R being 0.82 for human genes (Clay et al., 1996). Since what is under consideration here are amino acid composition and hydropathy, the problem must concern primarily the encoded proteins, and therefore the correlation between GC_3 and GC_{1+2} (the average GC levels of first+second codon positions). Incidentally, the correlation between isochore GC and GC_{1+2} also holds (Fig. 1).

The second comment is that the authors' corrected statement concerns two existing correlations, which are shown in Fig. 2, (1) the correlation between GC_3 and GC_{1+2} and (2) the correlation between GC_3 and hydropathy, which is also valid for intergenomic comparisons (D'Onofrio et al., 1999), as well as a tacit hypothesis, namely that the correlation between GC_3 and hydropathy holds for GC_{1+2}. The authors' argument indicates an implicit assumption, that the correlations under consideration are transitive (see Fig. 3).

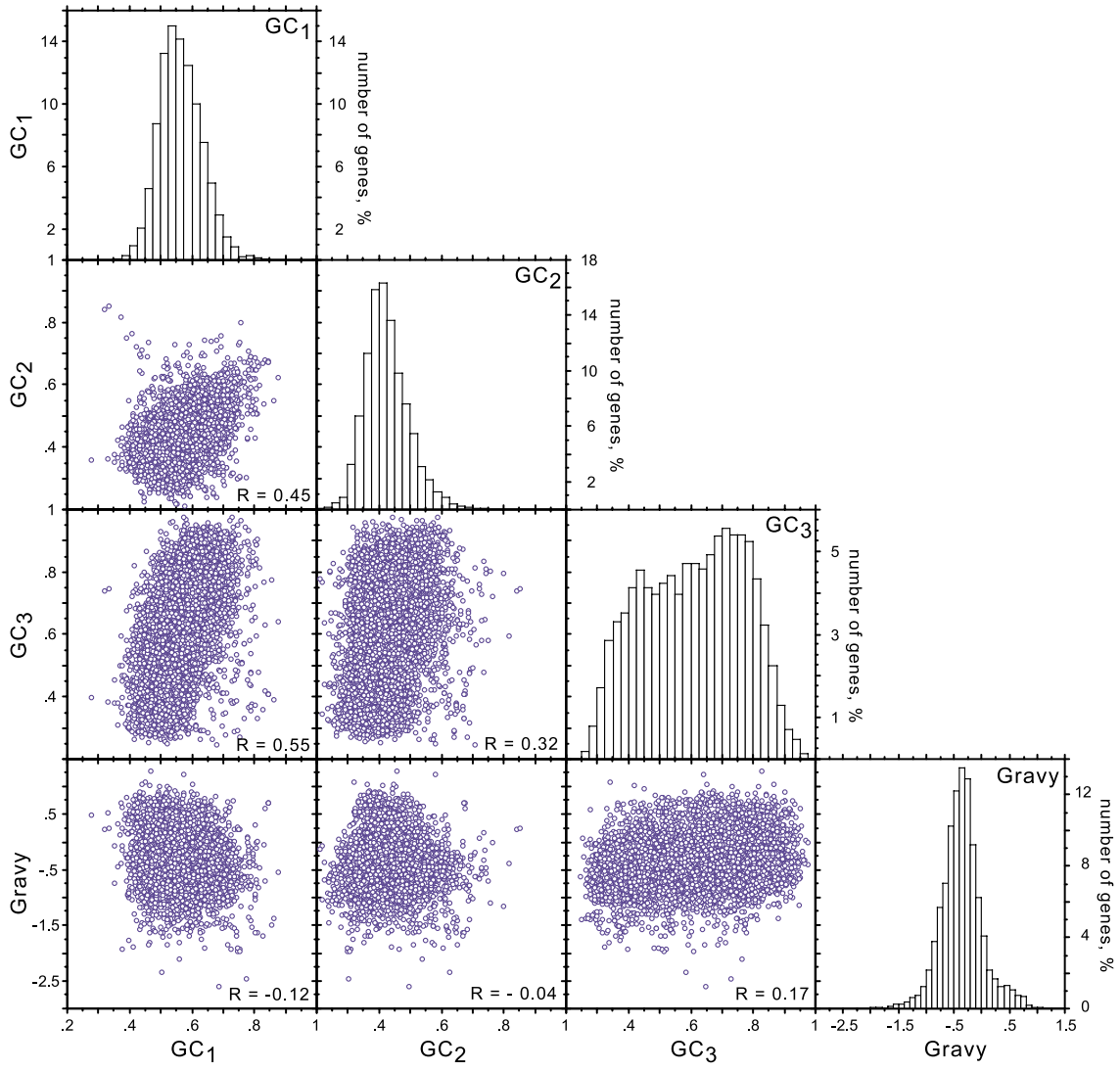

Fig. 2. Matrix plot representing all possible correlations involving GC_1, GC_2, GC_3 of human genes and hydrophobicity. Correlation coefficients (R) are indicated. Sample size is $N=20{,}148$. Well-annotated human genes were extracted from GenBank (Release 127, 15 December 2001) using the ACNUC retrieval system (Gouy et al. 1985). Hydropathy values of each encoded protein were calculated (Lobry and Gautier 1994) as $\Sigma\ (n_i/N)H_i$, where n_i is the number of occurrences of a given amino acid, N is the total number of amino acids in the protein and H_i is the assigned value of hydrophobicity according to the scale of Kyte and Doolittle (1982).

From a purely statistical point of view, transitivity of correlations is valid only if correlation coefficients are high enough. For example, if A and B are correlated with a positive correlation coefficient R, and if B and C are correlated with the same positive correlation coefficient R, then A is guaranteed to correlate positively with C only if $2R^2>1$ (Eq. (1)), i.e., only if $R>\sqrt{2}/2=0.707$. As has been emphasized by Kendall and Stuart (1976, pp. 424–428), such a relation "is by no means a trivial result". If R_{AB}, R_{BC} and R_{AC} describe the three correlations concerned, then the general relation that can be derived is $1+2R_{AB}R_{BC}R_{AC}-R_{AB}^2-R_{BC}^2-R_{AC}^2\geq 0$ (since the correlation matrix is non-negative definite, it has a non-negative determinant). In the case of two equal, positive correlations $R_{AB}=R_{BC}$, the general relation factorises to yield $R_{AC}\geq 2R_{AB}^2-1$, from which Eq. (1) follows. In the more general case of two arbitrary positive correlations R_{AB}, R_{BC}, transitivity can be invoked if $R_{AB}^2+R_{BC}^2>1$. However, the correlation coefficient of GC_3 vs. GC_{1+2} is 0.42 ($p<10^{-4}$; Clay et al., 1996), that of GC_3 vs. hydropathy is 0.17, and that of GC_{1+2} vs. hydropathy is -0.09 (Fig. 3). Thus transitivity cannot be assumed in the case under consideration.

As a third comment, Fig. 2 shows that while GC_3 is positively correlated with hydropathy, both GC_1 and GC_2 show slightly negative correlations with hydropathy, despite the fact that the correlations among the three codon positions are positive. Therefore, the statement that the positive correlation observed between GC_3 and hydropathy should be the result of the positive correlation between GC_3 and GC_{1+2} is incorrect.

As just shown, the correlation between GC_3 and hydrophobicity does not depend upon the correlation between GC_3 and GC_2 or GC_{1+2}. On the other hand, obviously the hydrophobicity of proteins depends upon the hydrophobicity of amino acids. The question then should be asked about the reason(s) for the correlation between GC_3 and hydrophobicity. The main reason is that quartets and duets are positively and negatively correlated with GC_3, respectively (D'Onofrio et al., 1999). Indeed, quartets comprise three abundant hydrophobic amino acids, duets six abundant hydrophilic amino acids. The resulting positive correlation between hydrophobicity and GC_3 is reduced in magnitude only by the opposing effect of the strongly hydrophobic triplet, isoleucine, and a strongly hydrophilic sextet, arginine (see Appendix A).

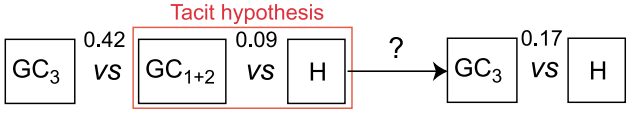

Fig. 3. Scheme showing the essence of the argument in Eyre-Walker and Hurst (2001): the transitivity of correlations (R values are indicated in boldface), and the assumption that a correlation exists between GC_{1+2} and hydropathy (outer box).

Acknowledgements

We thank an anonymous referee for helpful comments.

Appendix A. Amino acid contributions to the correlation between hydrophobicity and GC_3

We begin with the definition of the correlation in question ($h\equiv$ hydrophobicity, $x\equiv GC_3$, $<\ldots>\equiv$ intergenomic or intergenic mean, $\sigma_{\ldots}\equiv$ standard deviation, $R_{\ldots}\equiv$ correlation coefficient):

$$R_{hx}=\frac{<(h-<h>)(x-<x>)>}{\sigma_h\sigma_x}. \quad (A.1)$$

The hydrophobicity can be expressed as the sum

$$h=h_{\mathrm{Ala}}f_{\mathrm{Ala}}+\ldots+h_{\mathrm{Val}}f_{\mathrm{Val}}=\sum_{i=1}^{20}h_if_i,$$

where f_i denotes the frequency of the i-th amino acid expressed as a fraction of 1, i.e., $\sum f_i=1$ (in principle, this normalization constrains the 20 frequencies f_i, although in practice, $f_i\ll 1$ and the constraint on each amino acid is weak).

We can exploit the fact that the hydrophobicity of any alanine amino acid (for example) in any genome is always the same ($+1.8$ in the Kyte–Doolittle scale), i.e., all the h_i are constants with zero variance. Therefore, $<h>=<\sum h_if_i>=\sum<h_if_i>=\sum h_i<f_i>$; it follows that $h-<h>=\sum h_i(f_i-<f_i>)$. Inserting into Eq. (A.1) above, we obtain

$$R_{hx}=\sum h_i\frac{\sigma_{f_i}}{\sigma_h}\frac{<(f_i-<f_i>)(x-<x>)>}{\sigma_{f_i}\sigma_x}.$$

The last factor is just the correlation R_{f_ix} between the frequency of the i-th amino acid and GC_3, so that

$$R_{hx}=\frac{1}{\sigma_h}\sum_{i=1}^{20}h_i\sigma_{f_i}R_{f_ix}. \quad (A.2)$$

Eq. (A.2) splits the correlation coefficient R_{hx} of hydrophobicity vs. GC_3 into its 20 constituent contributions, one from each amino acid. The intergenomic correlation R_{hx} for prokaryotes and eukaryotes is 0.44 and 0.27, respectively (D'Onofrio et al., 1999); the corresponding intergenic correlation in human is 0.17 (see Fig. 2). The relative contributions ($h_i\times\sigma_{fi}\times R_{f_ix}/\sigma_h$) to the total, positive correlation can be calculated from the data in each of these three cases: the only large negative contributions are consistently from Leu (a triplet) and Arg (a sextet). The major contributors are, in order of decreasing magnitude: Lys, Ile, Arg, Asn, Ala, Val for the intergenomic study of prokaryotes, Ile, Lys, Asn, Arg, Ala for the intergenomic

study of eukaryotes (D'Onofrio et al., 1999), and Lys, Ile, Arg, Leu, Asn, Ala, Glu, Asp for the intragenomic study of human genes (Fig. 2).

Eq. (A.2) tells us that an amino acid's contribution depends not only on its hydrophobicity and on the strength of its correlation with GC_3, but also on the variability of the amino acid's frequency among genomes (or among genes, in intragenomic studies). This rather technical dependence can be rephrased in simpler terms, using the slope s_i of the orthogonal regression line of amino acid frequency f_i vs. GC_3. As a rough but reasonable approximation, we obtain

$$R_{hx} \approx \frac{\sigma_x}{\sigma_h} \sum_{i=1}^{20} h_i s_i. \quad (A.3)$$

The approximation (A.3) again yields the major contributors listed above. It is also intuitively easier to grasp than Eq. (A.2): the relative contribution of each amino acid now depends (almost) only on its hydrophobicity and slope.

References

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gautier, C., Bernardi, G., 1991. The compositional properties of human genes. J. Mol. Evol. 32, 497–503.

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Chiusano, M., D'Onofrio, G., Alvarez-Valin, F., Jabbari, K., Colonna, G., Bernardi, G., 1999. Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure. Gene 238, 23–31.

Clay, O., Caccio, S., Zoubak, S., Mouchiroud, D., Bernardi, G., 1996. Human coding and noncoding DNA: compositional correlations. Mol. Phylogenet. Evol. 5, 2–12.

D'Onofrio, G., Bernardi, G., 1992. A universal compositional correlation among codon positions. Gene 110, 81–88.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, G., 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. 32, 504–510.

D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, G., 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene 238, 3–14.

Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev., Genet. 2, 549–555.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., di Paola, G., 1985. ACNUC—a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. 1, 167–172.

Kendall, M., Stuart, A., 1976. The Advanced Theory of Statistics, vol. 3, 3rd ed. Charles Griffin, London.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Lobry, J.R., Gautier, C., 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res. 22, 3174–3180.

Sueoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. U. S. A. 85, 2653–2657.

Wada, A., Suyama, A., 1985. Third letters in codons counterbalance the (G·C)-content of their first and second letters. FEBS Lett. 188, 291–294.