Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 333 (2004) 187–188
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Incorrectly predicted genes in rice?☆

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

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

Received 12 January 2004; accepted 12 February 2004

Available online 23 April 2004

Abstract

Between one third and one half of the proposed rice genes appear to have no homologs in other species, including *Arabidopsis*. Compositional considerations, and a comparison of curated rice sequences with ex novo predictions, suggest that many or most of the putative genes without homologs may be false positive predictions, i.e., sequences that are never translated into functional proteins in vivo.

Keywords: Gene; Codon; Amino acid

In genes, second positions of codons are largely constrained by the amino acids they encode, whereas third positions reflect constraints in base composition. The scatterplot of the frequencies of GC base pairs in the second (GC_2) and third (GC_3) positions of genes from a given genome defines a correlation that is well conserved from prokaryotes to eukaryotes (D'Onofrio and Bernardi, 1992; D'Onofrio et al., 1999). In all species analyzed to date, the axis is far from the diagonal, and in a wide range of species (including human and *Escherichia coli*) it is close to the line $GC_3=6GC_2-2$ (D'Onofrio et al., 1999). We were therefore surprised to find that this conservation was apparently violated in the recently sequenced and annotated rice genome, which showed many genes aligning along the expected axis, but also many extending along the diagonal. Such behavior would usually indicate contamination of the data set by intergenic or other noncoding DNA. Furthermore, 50.6% of genes reported for rice had no orthologs in *Arabidopsis thaliana* (Yu et al., 2002). We labeled the rice genes in GenBank, which correspond essentially to the "cDNA" sequences analyzed by Yu et al. according to their annotations (see Fig. 1). Almost all the genes clustering along the diagonal were in fact annotated as predicted or putative, whereas the large majority of the experimentally determined genes line up along the axis that is expected for coding sequences. We conclude that many, if not most, of the points appearing along the main diagonal in Fig. 1 are likely to represent rice sequences that are not translated into proteins. This may have led to a considerable overestimate of the proportion of coding sequences that lack orthologs in *Arabidopsis*. Simple GC_2 versus GC_3 scatterplots can, therefore, serve as a quick check to identify computationally predicted or expressed sequence tag-based genes that are unlikely to code for proteins.

☆ This brief note was submitted to Science, in June 2002, in response to a paper by Yu et al. of April 2002 describing a draft sequence of the rice genome. In December 2002 we were informed of the journal's intent to publish our note without further changes; the proofs (edited copy) were soon sent to us and returned. After waiting for 1 year, we were notified on January 9, 2004 that Science was not willing to publish our comment in its present form. In the same week another note appeared (L.S. Wyrwicz et al., Science 303, 168) on the problem of rice genes without homologs, also suggesting that there will not be large numbers of 'no-homolog' genes in rice, although for a different reason: where these genes do encode proteins, appropriate structure-guided alignments often reveal homologs in other species. We here reproduce our note verbatim (with an abstract), since it remains pertinent to the more recently finished chromosome sequences of rice, and since it has already been cited elsewhere, following standard procedure, as 'in press'.

* Corresponding author. Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa comunale, 80121 Naples, Italy. Tel.: +39-81-5833300; fax: +39-81-245-5807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.02.039

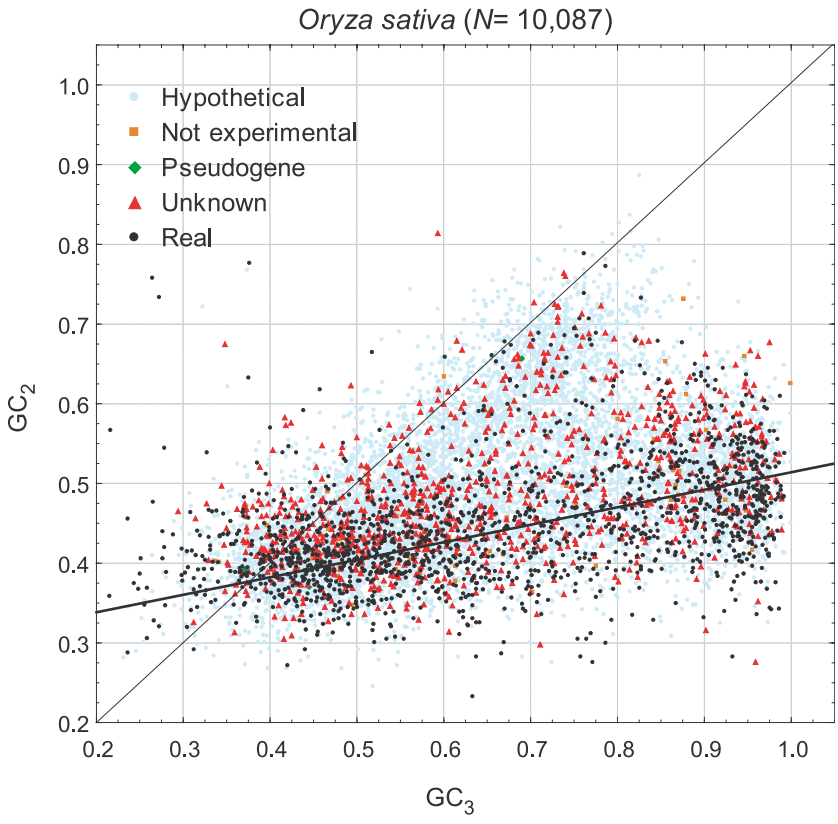

Fig. 1. Scatterplot of GC_2 vs. GC_3 levels in predicted and experimentally identified rice genes. The diagonal (GC_2=GC_3) is indicated. Complete coding sequences from *Oryza sativa* were extracted from GenBank (release 129; retrieved 31 May 2002) using ACNUC software (Gouy et al., 1985). Redundancies were removed on the basis of protein alignments using as a cutoff 90% identity for an overlap of 90%. The resulting gene set (*N*=10,087) was partitioned into five classes according to the annotations (real genes, not experimental, unknown, pseudogenes and hypothetical) in the informative fields product, gene name, evidence, and note, using a script written in Perl.

References

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

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.

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.

Yu, J., Hu, S., Wang, J., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296, 79–92.