The mutual information theory for the certification of rice coding sequences

Nicolas Carels[a,*], Ramon Vidal[a], Ricardo Mansilla[b], Diego Frías[a]

[a] *Laboratório de Bioinformática, Universidade Estadual de Santa Cruz, Rodovia Ilhéus/Itabuna km. 16, Ilhéus, Bahia, Brazil*
[b] *Centro de Investigacíones Interdiciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México, Ciudad Universitaria, Del. Coyoacan, México, DF, Mexico*

Received 7 April 2004; accepted 13 May 2004

Available online 25 May 2004

Edited by Robert B. Russell

Abstract **We report here the use of the mutual information theory for the certification of annotated rice coding sequences of both GenBank and TIGR databases. Considering coding sequences larger than 600 bp, we successfully screened out genes with aberrant compositional features. We found that they represent about 10% of both datasets after cleaning for gene redundancy. Most of the rejected accessions showed a different trend in GC3% vs GC2% plot compared to the set of accessions that have been published in international journals. This suggests the existence of a bias in the pattern recognition algorithms used by gene prediction programs.**

Keywords: Rice; AMI; GenBank; TIGR; Bimodality

1. Introduction

The conclusions drawn by bioinformatics rely on data quality. However, the joint development of sequencing techniques and automated annotation is responsible for the vast majority of gene annotations by in silico techniques. A consequence of this is the apparition of a substantial number of annotations corresponding to uncharacterized proteins in the genomic databases. Such proteins are considered as ''hypothetical'', ''similar'' or ''putative'' when they are homologous to accessions with unknown function. When there is no homolog, the term used for classification is ''unique''. Protein families with such *fuzzy* annotations are natural targets for functional genomics. However, *fuzzy* annotations may represent false gene predictions [7]. Moreover, the actual practice of protein annotation relies mainly on previous annotations. If certification steps are missed, it may lead to a ''snow ball'' effect.

Most gene prediction methods use hidden Markov models (HMMs) [1,2,9,13,14] or neural networks [13,15,17,18]. In optimal conditions, the precision of those algorithms was claimed to be in the range of 92–95%. However, the efficiency of these algorithms mainly depends on the quality of the ''training dataset'' used to recognize the patterns of interest. The training set can be (i) contaminated with erroneous sequences, (ii) redundant with the validation set, and (iii) not representative of the true population. In addition, the error of any pattern recognition method increases rapidly as the sequence length decreases because of the increasing lack of statistical significance. As a consequence, the risk to draw conclusions from false positives or true negatives is real (see [11], for a discussion of these methods). This risk particularly applies to large scale analyses where whole genomes are taken into consideration. Systematic bias in gene determination may introduce 'original' behaviors and lead to publication of wrong conclusions (see [4]).

Recently, a new kind of method for coding sequence (CDS) identification has been proposed [6,10,16]. Those methods detect short range spatial correlations between nucleotides specific to CDS due to their codon structure. Such methods do not rely on training or learning steps. The most outstanding methods of this class are the Sequence Fourier Spectrum [16] and the Average Mutual Information (AMI) [6].

Based on a statistically extremely robust gene sample, we found that the distribution of AMI-certified CDS confirms that the relationship GC3% vs GC2% must be considered monotonous [4].

Rice gene number has been claimed to be in the range of ~50 000, about 50% of which do not have homology with those of *Arabidopsis* [19]. The careful examination of the distribution of those predicted genes for GC3% vs GC2% by Cruveiller et al. [4] showed that it is bi-univocal, since it has two linear trends with different slopes intercepting at low GC3 and GC2 levels. This distribution differs essentially from those observed for other known eukaryote species. It would have promoted massive changes in transcription, transcriptional regulation, or translation since the dicot–monocot divergence, however, experimental evidence of such phenomenon was not found. In this article, using AMI, we showed by reference to a statistically representative set of experimental genes that one of the two trends of the plot of rice gene for GC3% vs GC2% is effectively formed by false positives. It is the first time that an automatic procedure is proposed to eliminate the major part of these false positives. In addition, it also allows to infer rice gene number with higher confidence and to solve an ongoing polemic regarding weather or not the distribution of *Gramineae* genes is bimodal [20].

2. Materials and methods

We retrieved complete nuclear CDSs of *Oryza sativa* from GeneBank and TIGR. The GenBank (release 137–15 August 2003) CDS

* Corresponding author. Fax: +55-73-680-5226.
E-mail address: carels@uesc.br (N. Carels).

0014-5793/$22.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2004.05.026

(36 104) were retrieved, using the Infobiogen server (see http://www.infobiogen.fr) and the ACNUC/QUERY retrieval system [5] with the options: t = cds et no o = plastid et no o = mitochondrion et no k = partial et no k = est.

We used the bibliographical references reported under the field MEDLINE in the features to build a dataset of experimentally proven genes as follows: (i) the MEDLINE identification numbers were used to retrieve the abstracts of the corresponding genes from the NCBI server (PubMed), through a CGI interface (PERL); (ii) those abstracts were then screened to eliminate mitochondrial and chloroplast genes as well as (retro)transposons and references based on any kind of automatic in silico process. The remaining sample was 405 CDS.

To eliminate the redundancy from CDS samples, we looked for homology between sequences using BLASTN with the "-e" option equal to 0.0001. A cleaning procedure was then applied to the BLASTN file in order to eliminate the sequences implied in a homologous pair with the highest hit when it was above a given identity level. The identity level above which two sequences were considered redundant was set at 90% over 90% of the homologous regions with the shorter sequence of the pair.

The coding sequences of the protein genes were used to calculate their GC level in all three codon positions using the software CODONW [12].

In the case of TIGR, we downloaded the file "all.cds" from: ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_1.0/all_chrs. It contains 56 056 putative CDS identified among 358 Mb of non-overlapping sequences for all 12 chromosomes. We then eliminated the partial sequences and the sequences including the words: "tRNA", "transposon", "element", "plast" and "mitoch". The remaining file (52 309) CDS was further cleaned for redundancy as described above.

The histograms of gene distribution and the contour plots were obtained with Scilab-2.7 (http://scilabsoft.inria.fr/) with a class interval of 2% GC (except for the sample of published genes where class interval was set to 5% GC to have an image with better rendering). The orthogonal regression lines were calculated as described by Jolicoeur [8].

The AMI analyses were performed as described in [6] on sequences higher than 600 bp, because with such a criterion, the discrimination efficiency between coding and non-coding sequences is >98%.

We also established a statistic of the annotations. To do this, we considered basically two classes: (i) one that we called *fuzzy* with annotations such as: "hypothetical", "similar", "putative", "unknown" and "unnamed" CDS; (ii) the other that we called *consistent* because the annotations to which it refers was affirmative. We did this classification in order to find the proportion of these annotations between both CDS samples either certified or not by the AMI. To do this, we counted the lines of CDS definition containing these words using the function "egrep" in Linux in the CDS files either certified or not by the AMI.

3. Results

The non-redundant dataset of nuclear rice CDS (25 649) was found to be 71.04% of the complete set available from GenBank (36 104). The plot of gene with respect to GC2% and GC3% was found to be "bi-univocal" as described by Cruveiller et al. [4]. This was true for the total GenBank sample as well as for the sample cleaned for redundancy (Fig. 1A). In contrast, the experimentally proven published genes larger than 600 bp (405) display a univocal (monotonous) distribution of GC3% vs GC2% (Fig. 1B).

In the conditions of the gene sample of Fig. 1B we found, using the AMI, a prediction efficiency of 98% for true positives. With this error rate in mind, we analyzed the non-redundant dataset of nuclear rice CDS of GenBank and TIGR.

In GenBank, 17 069 (66.55%) non-redundant genes were larger than 600 bp. Among them, the putative "false positives" (2424) eliminated by the AMI are plotted in Fig. 1C and the putative "true positives" (14 645 = 86%) in Fig. 1D. We found

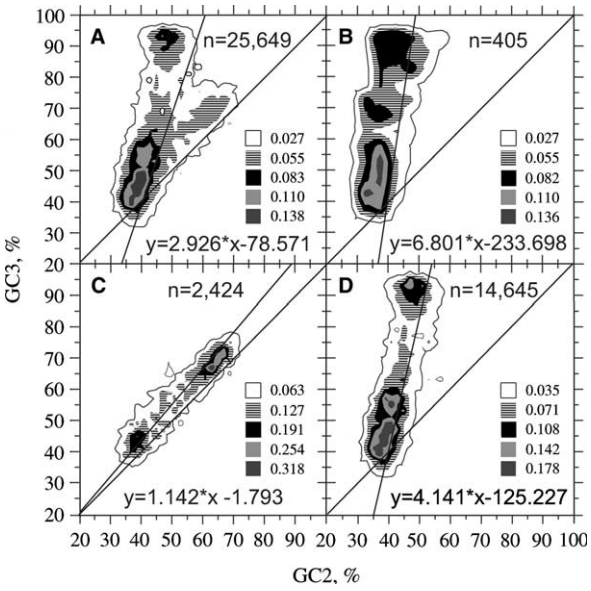

Fig. 1. Relationship between GC2 and GC3 in rice genes from: (A) the non-redundant dataset (25 649) of CDS >600 bp from GenBank, $r = 0.44$; (B) the non-redundant dataset (405) available from publications, $r = 0.35$; (C) the non-redundant dataset of B rejected by the AMI algorithm (2424), $r = 0.83$; and (D) the non-redundant dataset of B certified by the AMI algorithm (14 645), $r = 0.54$. The values for the contour lines indicate the relative density (%) of genes per unit area (%) of the plot. All correlation coefficients "r" were statistically significant at $P < 0.0001$. The different colored areas represent the different gene densities.

that Fig. 1D displays the same relationship as shown in Fig. 1B, with slight differences of the slopes of the regression lines probably due to the large sample size differences. In contrast, the slope of the regression line of Fig. 1C is close to the diagonal with the consequence that the GC2% of those "genes" should be approximately equal to GC3%. These sequences are classified as non-coding by the AMI and this is true with a rate of confidence of 98%.

In the GenBank dataset, we found 8353 (58%) and 1299 (54%) *fuzzy* annotations among the AMI-certified and AMI-rejected CDS, respectively.

In TIGR, the redundancy that we found using the criteria outlined above was about half (~14%) that of GenBank. We eliminated 7592 from a total of 52 309 CDS leaving 44 717 non-redundant genes. The redundancy in the TIGR dataset mainly affects the GC-poor genes (data not shown).

Among the 44 717 non-redundant genes, 27 979 (62.57%) were at least 600 bp; among which the AMI accepted 24 914 (89% = 24 914 of 27 979) and rejected 3065 (~11%) CDS. The contour plot of the 24 914 (Fig. 2B) and the CDS distribution for GC3% (Fig. 3A) of the TIGR dataset are similar to those of the 14 645 AMI-certified CDS from GenBank (Figs. 1D and 3B, respectively), leading to the confirmation that in GenBank as well as in TIGR the CDS distribution for GC3% is bimodal [3]. This bimodality of the CDS distribution tends to be hidden by the false positive genes as it was observed in GenBank and TIGR (Fig. 3).

Compared to the GenBank dataset (Fig. 1C), the profile of the rice CDS from TIGR rejected by the AMI (Fig. 2C) presents an interesting feature. The error of CDS prediction remains similar to that observed in GenBank (~10%). However,

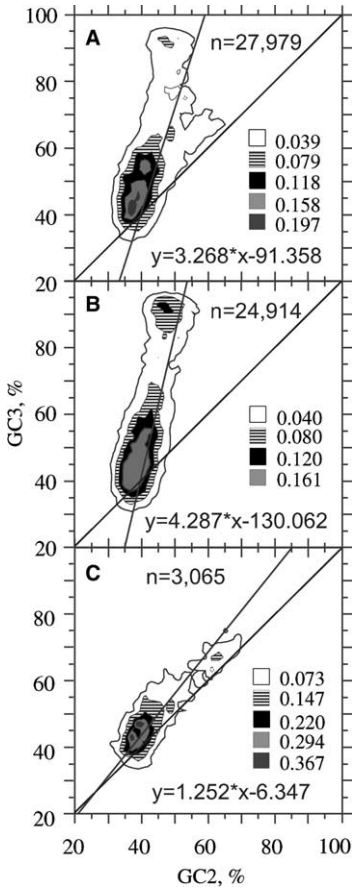

Fig. 2. Relationship between GC2 and GC3 in rice genes from: (A) the non-redundant dataset (27 979) of CDS >600 bp from TIGR, $r = 0.50$; (B) the non-redundant dataset of A certified by the AMI algorithm (24 914), $r = 0.58$; and (C) the non-redundant dataset of B rejected by the AMI algorithm (3065), $r = 0.78$. The values for the contour lines indicate the relative density (%) of genes per unit area (%) of the plot. All correlation coefficients "r" were statistically significant at $P < 0.0001$.

the shape of the contour plot of Fig. 2C deserves several comments. The proportion of miss-identified genes favors GC-poor genes. Moreover, the trend of the relationship between GC2% and GC3% appears similar to that of "true positives" (Fig. 2B) until 55% GC2–GC3. Then, the trend changes and becomes parallel to the diagonal. Compared to GenBank it is likely that the TIGR gene predictor implements a discrimination function sensible to the ratio GC3% vs GC2%. However, as far as we know for genomes such as *Arabidopsis*, where the GC3% range is about the same as that of GC2% (4% difference on the average), the prediction error (~10%, unpublished data) carried out by the gene predictor cannot be eliminated by methods other than those based on mutual information and Fourier transform.

In the TIGR dataset, we found 20 167 (81%) and 2754 (90%) *fuzzy* annotations among the AMI-certified and AMI-rejected CDS, respectively.

4. Discussion

The higher redundancy that we found in GenBank (30%) compared to TIGR (14%) is probably due to the fact that

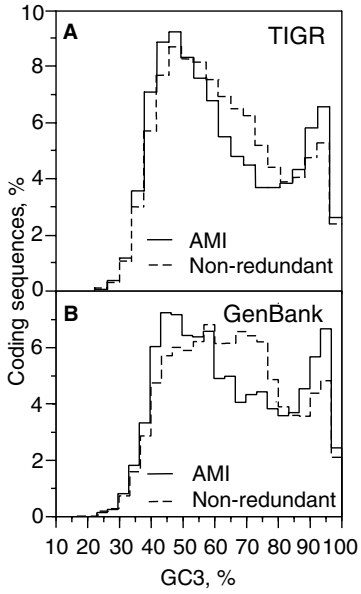

Fig. 3. Gene distribution with concern to GC3. The dotted line represents the non-redundant dataset and the solid line represents the AMI-certified CDS. (A) TIGR; (B) GenBank.

GenBank is a public repository. Therefore, numbers of genes can be entered independently as different alleles by different authors. Such a redundancy does not have biological meaning. In contrast, since only one group manages the TIGR database, it is expected that the 14% redundancy has a biological meaning and reflects paralogous genes in this genome. Those paralogous genes are probably more numerous, but the counting is relative and depends on the criteria that we applied to BLASTN for detection. Here, we were only interested in normalizing both datasets before any statistical treatment.

Based on coding sequences (CDS) larger than 600 bp, we found that the rate of wrongly annotated genes of rice in GenBank and TIGR was about ~10% of the non-redundant datasets. Therefore, the use of AMI appears to be a complementary method to HMM and neural networks, since it succeeded to eliminate genes that are obvious false positives by comparison to coding sequences that were decrypted by laboratory techniques. Those false positives display a random distribution along the line of identity between GC2% and GC3% (diagonal of Figs. 1C and 2C) and form a suspicious trend in the plot of rice genes for GC3% vs GC2% that is not present in the plot of the experimental gene set used as reference.

We found that in TIGR *fuzzy* annotations (~80%) are almost double that of GenBank (~50%). TIGR is a member of the IRGSP consortium (International Rice Genome Sequencing Project), which justifies that TIGR annotations are more complete than those of GenBank. It is likely that because of individual contribution much of the coding sequences to GenBank are simply not annotated. The proportion of *fuzzy* to *consistent* annotations in AMI-certified and AMI-rejected CDS appears to be roughly the same in GenBank and TIGR. It is interesting to conclude here that the process of annotation itself does not seem to be reliable, since a similar proportion of *fuzzy* and *consistent* annotations is found in AMI-certified as well as in AMI-rejected CDS (and this independently of the

database). This suggests that, in addition to the *fuzzy* annotation, the *consistent* genes identified as false positives by the AMI are, indeed, miss-annotated. Our results suggest that AMI should be introduced in the IRGSP annotation protocol (see: http://rgp.dna.affrc.go.jp/genomicdata/AnnSystem.html) to take this problem into account.

The claim, on the basis of predicted genes, that *Gramineae* could have up to ~50% novel genes compared to *Arabidopsis* [19] remains to be confirmed. Considering only the 62.5% of non-redundant CDS of TIGR larger than 600 bp, we found that rice contains only 13.5% more CDS than *Arabidopsis*. Therefore, the ~26.5% genes that make up the difference between rice and *Arabidopsis*, under the Yu et al. [19] statement, most probably have CDS <600 bp. Since the error rate of gene prediction increases below this threshold, the consistency of the above claim remains to be confirmed.

Compositional bimodality of plant genes has been first reported by Carels and Bernardi [3]. In the case of *Gramineae* genes, the compositional heterogeneity is so large that it can be easily found just by looking at the profile of the gene distribution for GC3%. This evidence was denied by Meyers et al. [20] using maize ESTs. However, Meyers et al. [20] were wrong in their methodology. They did not find the bimodality, because they did not take EST redundancy and GC3% into consideration. With a coding sequence sample statistically extremely robust, Fig. 3 clearly shows that the rice coding sequence distribution for GC3% is, indeed, bimodal.

As a conclusion, the mis-understandings that occurred regarding: (i) the compositional bimodality of *Gramineae* genes, (ii) the bi-univocal gene distribution for GC3% vs GC2%, and (iii) the gene number in the rice genome are three different consequences of a same problem: mis-annotation. The certification of coding sequences of rice using AMI is a clear improvement to the present day knowledge about rice gene annotation. Preliminary results of coding sequences certification by AMI suggest that the method will also be very useful for the annotation of other genomes such as *Arabidopsis*, for instance.

Acknowledgements: We thank Carter Miller for the manuscript revision. This research was supported by the Brazilian CNPq agency providing researcher fellowship to N. Carels and student fellowship to R. Vidal.

References

[1] Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for predicition of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29, 2607–2618.

[2] Borodovsky, M. and McIninch, J.D. (1993) GeneMark: parallel gene recognition for both DNA strands. Comp. Chem. 17, 123–133.

[3] Carels, N. and Bernardi, G. (2000) Two classes of genes in plants. Genetics 154, 1819–1825.

[4] Cruveiller, S., Jabbari, K., Clay, O. and Bernardi, G. (2003) Compositional features of eukaryotic genomes for checking predicted genes. Brief Bioinform. 4, 43–52.

[5] Gouy, M., Gautier, C., Attimonelli, N., Lanave, C. and Di Paola, G. (1985) CABIOS 1, 167–172.

[6] Grosse, I., Herzel, H., Buldyrev, S.V. and Stanley, H.E. (2000) Species independence of mutual information in coding and noncoding DNA. Phys. Rev. E 61, 5624–5629.

[7] Iliopoulos, I., Tsoka, S., Andrade, M.A., Janssen, P., Audit, B., Tramontano, A., Valencia, A., Leroy, C., Sander, C. and Ouzounis, C.A. (2000) Genome sequences and great expectations. Genome Biol. 2, 1.1–1.3.

[8] Jolicoeur, P. (1963) The generalization of the allometry equation. Biometrics 19, 497–499.

[9] Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. Proc. Int. Conf. Intelligent. Syst. Mol. Biol. 4, 134–142.

[10] Li, W., Marr, T.G. and Kaneko, K. (1994) Understanding long-range correlation in DNA sequences. Physica D 75, 392–416.

[11] Mathé, C., Sagot, M.F., Schiex, Th. and Rouzé, P. (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30, 4103–4117.

[12] Peden, J. (1999). Analysis of codon usage, Ph.D. Thesis, University of Nottingham, UK. Available from <http://www.mol-biol.ox.ac.uk/cu/>.

[13] Reese, M.G., Kulp, D., Tammana, H. and Haussler, D. (2000) Genie-Gene finding in *Drosophila melanogaster*. Genome Res. 10, 529–538.

[14] Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26, 544–548.

[15] Sherriff, A. and Ott, J. (2001) Applications of neural networks for gene finding. Adv. Genet. 42, 287–297.

[16] Tiwary, S., Ramchandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. (1997) Prediction of probably genes by Fourier analysis of genomic sequences. CABIOS 13, 263–270.

[17] Uberbacher, E.C. and Mural, R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl. Acad. Sci. USA 88, 11261–11265.

[18] Xu, Y., Mural, R.J., Einstein, J.R., Shah, M.B. and Uberbacher, E.C. (1996) Grail: a multiagent neural network system for gene identification. Proc. IEEE 84, 1544–1551.

[19] Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y. and Zhang, X., et al. (2002) A draft sequence of the rice genome *Oryza sativa* L. ssp. *indica*). Science 296, 79–92.

[20] Meyers, B.C., Tingey, S.V. and Morgante, M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11, 1660–1676.