Compositional Gradients in *Gramineae* Genes

Gane Ka-Shu Wong,[1,2,3,5] Jun Wang,[1,2,4] Lin Tao,[1,2] Jun Tan,[1,2] JianGuo Zhang,[2] Douglas A. Passey,[3] and Jun Yu[1–3]

[1]*Hangzhou Genomics Institute, Institute of Bioinformatics of Zhejiang University, Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China;* [2]*Beijing Genomics Institute, Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China;* [3]*University of Washington Genome Center, Department of Medicine, Seattle, Washington 98195, USA* [4]*College of Life Sciences, Peking University, Beijing 100871, China.*

In this study, we describe a property of *Gramineae* genes, and perhaps all monocot genes, that is not observed in eudicot genes. Along the direction of transcription, beginning at the junction of the 5′-UTR and the coding region, there are gradients in GC content, codon usage, and amino-acid usage. The magnitudes of these gradients are large enough to hinder the annotation of the rice genome and to confound the detection of protein homologies across the monocot–eudicot divide.

Extreme heterogeneity in local GC content is one of the most instantly recognizable characteristics of the human genome (Bernardi 2000; Eyre-Walker and Hurst 2001). Typically, the GC content varies from 26%–65%. This characteristic of the sequence is even more striking in light of the homogeneity in local AG content, which fluctuates by a only few percent about a mean of 50% (Forsdyke and Mortimer 2000). Most discussions of compositional properties have been centered on a description of the human genome as a mosaic of GC- and AT-rich isochores, which are found in warm-blooded vertebrates but not cold-blooded vertebrates. More recently, it was reported that there is an elevated GC content in *Gramineae* (grass) genomes, extending perhaps to monocot genomes but not to eudicot genomes (Carels and Bernardi 2000). It is not known how, or even if, this new phenomenon is related to isochores.

In this study, we focus on the monocot–eudicot phenomenon, using an analysis of the publicly accessible plant cDNA sequences (i.e., experimentally derived transcripts) supplemented by the recently completed genome sequences of rice (Yu et al. 2002) and of *arabidopsis* (The Arabidopsis Genome Initiative 2000). Although studies of plant genomic sequences have refuted Bernardi's interpretations (Dubcovsky et al. 2001; Meyers et al. 2001), the fact remains that *Gramineae* genes are more GC-rich than eudicot genes. The problem is the presumption that variations in GC content occur between genes, and that there are two classes of genes in *Gramineae*. We argue to the contrary that most of the variations in GC content occur within genes in the form of a negative gradient along the direction of transcription.

RESULTS

Our analysis was limited to flowering plants, routinely divided into monocots and eudicots. *Gramineae* are a subfamily of monocots, but as it is not possible to get large amounts of cDNA data for a monocot that is not also a *Gramineae*, it is not known if the gradients exist in all monocots. This is, however, the obvious extrapolation, based on evolutionary clads and the absence of gradients in eudicots. We studied four *Gramineae* (rice, maize, barley, and wheat) and six eudicots (*arabidopsis*, tobacco, tomato, soybean, pea, potato). For brevity, we depict only the two largest *Gramineae* data sets, *Oryza sativa* (rice) and *Zea mays* (maize), and the two largest eudicot data sets, *Arabidopsis thaliana* (*arabidopsis*) and *Nicotiana tabacum* (tobacco). The other data sets, however, did behave as we had expected, in line with their evolutionary clads.

GC Content

Initially, we wrote a program to display the GC content for individual genes as a function of position along the direction of transcription. After observing many hundreds of examples, it was clear that there is a gradient in the GC content of *Gramineae* genes, but not eudicot genes. The magnitude of the gradient varied from gene to gene, and even though zero gradients were sometimes observed, positive gradients were rarely observed. Typically, the 5′-ends of a *Gramineae* gene were up to 25% more rich in GC content than their 3′-ends. In Figure 1, we depict the best available homologs (possible orthologs) for four pairs of rice–*arabidopsis* genes. The gradients exhibited a finite range, extending ~1 kb from the 5′-end, before petering out. To quantify the effect for individual genes, we computed the slope of the trend in GC content across the first kilobase of the coding region. Figure 2 shows that the *Gramineae* distributions are skewed toward negative GC content gradients much more so than the eudicots.

To capture the general trend as a function of position along the transcript, we performed ensemble averages across the set of cDNAs, with the origin at the junction of the 5′-UTR and coding region. For better signal-to-noise ratios, we used a 51-bp (or 17-codon) sliding window. At each position, we examined the sequence contents of cDNAs that extended out to that position and computed their mean values. In principle, this could have produced an artifactual gradient (e.g., suppose there is no gradient but small genes are GC-rich and large genes are AT-rich). In practice, we already know there is a gradient, however, and given that different genes exhibit different gradients, the reality is that the ensemble averages hide the true magnitudes of the gradients. For the *Gramineae* genes shown in Figure 3, the gradient in ensemble averaged GC content was only $-0.1\ \mathrm{kb}^{-1}$, or half the magnitude of the $-0.2\ \mathrm{kb}^{-1}$ per gene gradients shown in Figure 2.

Figure 4 depicts how much of the effect is attributable to

[5]Corresponding author.
E-MAIL gksw@u.washington.edu; FAX (206) 685-7344.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.189102.

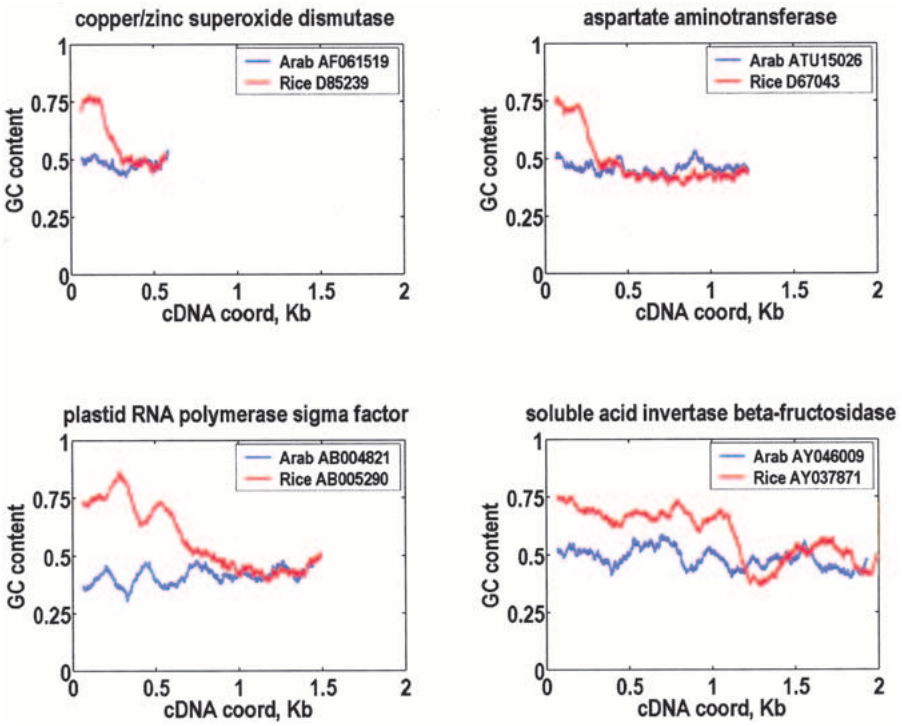

Figure 1 GC content as function of position from start of coding region for four pairs of best available *O. sativa* and *A. thaliana* homologs (possible orthologs). A 129-bp sliding window, equal to the median size of a rice exon, was used to filter out the fluctuations in the sequence.

each of the three codon positions: GC1, GC2, and GC3. Not surprisingly, most of it is attributable to GC3, as changes in GC3 are least likely to have phenotypic consequences. More precisely, at each of the three codon positions, there are $64 \times 3 = 192$ possible base substitutions. The numbers of substitutions that will not change the amino acid are 8, 2, and 128, respectively. Notice that this argument cannot explain why, for the *Gramineae* genes, the relative magnitudes were GC3>GC1>GC2, whereas for the eudicot genes, they were GC1>GC2≈GC3. We also note that this variation in GC content between the three codon positions was why we had to use a 51-bp sliding window in addition to the ensemble average. Had we not done so, the results would have appeared to be artifactually noisy, although a closer examination would have revealed that this noise was threefold periodic.

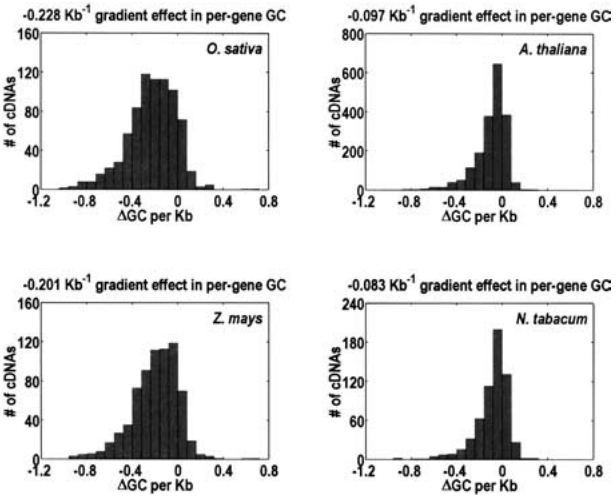

Figure 2 Distributions for "per-gene" GC content gradients in *O. sativa*, *Z. mays*, *A. thaliana*, and *N. tabacum*. The gradient is the slope of the trend in GC content versus position, defined only for the first kilobase of the coding region, to respect the finite extent of the gradient effect.

Codon Usage

The best ab initio gene-prediction programs for eukaryotic genomes (Rogic et al. 2001) employ codon usage statistics for the initial phase of exon detection. Some popular examples include `FGeneSH` (Salamov and Solovyev 2000) and `GenScan` (Burge and Karlin 1997). Codon usage has been studied for many decades (Powell and Moriyama 1997; Karlin et al. 1998; Duret and Mouchiroud 1999), but there is still no explanation that is consistent with all observations in all organisms. As a practical matter, it is not necessary to have a mechanistic understanding of codon usage to write a gene-prediction program, but it is important to have an accurate description of the phenomenon. If, as a result of the gradients in GC content, there is a gradient in codon usage, then that could be a problem. Programs like `GenScan` use different statistics for different genes, based on a regional GC content, but in their current incarnation, they do not allow for different statistics along the direction of transcription for a single gene.

It is reasonable to expect gradients in codon usage for *Gramineae* genes because, in at least one reported model, GC content determines codon usage (Knight et al. 2001). Codon usage can be quantified in different ways. A commonly used statistic is *Ne*, or the effective-number-of-codons (Wright 1990; Comeron and Aguade 1998). *Ne* is a number from 20 to 61, with a simple intuitive interpretation. It is 20 when one codon is used for each amino acid, meaning codon usage is maximally biased. It is 61 when every possible codon is equally likely to be used, meaning there is no bias. As originally defined, the *Ne* formula was applied to all of the codons in a single gene, but here, we apply the formula to all of the codons observed at a given position along the coding region, averaged across the ensemble of cDNAs. Figure 5 depicts this result. As expected, *Gramineae* genes exhibited a gradient in their effective-number-of-codons. We also examined codon usage patterns for specific amino acids. Some exhibited stronger gradients than others, but if there was any overall trend, it was that gradients were strongest for amino acids with the largest number of synonymous codons (data not shown).

The implications of these codon usage gradients were discussed in greater detail in our analyses of the rice genome (Yu et al. 2002). Most of the extant gene-prediction programs

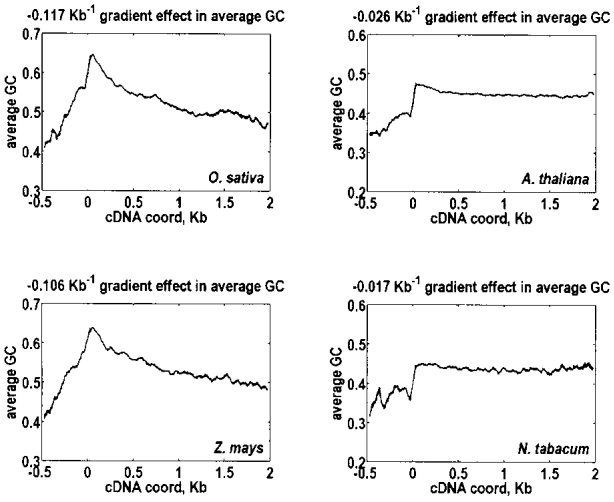

Figure 3 Overall GC content as a function of cDNA position, relative to the start of the coding region, and averaged over all cDNAs with a 51-bp sliding window. Shown here are *O. sativa*, *Z. mays*, *A. thaliana*, and *N. tabacum*. Negative coordinates are 5′-UTR. Positive coordinates are coding.

performed poorly on rice. For reasons that remained unclear, the best of the lot was FGeneSH. Of the 53,398 predicted rice genes with initial and terminal exons, half of them did not have a detectable homolog in *A. thaliana* (nor even in *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*). These no homolog genes were unusual in other ways too. They were on average half the size of those genes that did have a homolog. It was argued that FGeneSH failed to detect many exons for these genes because their GC content gradients were exceptionally pronounced. Nevertheless, a substantial fraction of these genes were confirmed by expressed sequence tags (ESTs). Therefore, without major revisions to the gene-prediction programs, half of the rice gene set will remain in limbo.

Amino-Acid Usage

Given the existence of gradients in codon usage, the obvious next question is, are there gradients in amino-acid usage? That the answer would be yes is surprising, as so much of the gradient was in GC3, which generally does not affect amino-acid sequence. Closer examination of the data, however, reveals that the overall GC1 and GC2 levels for *Gramineae* genes are almost 10% higher than for eudicot genes. We show in Figure 6 the usage patterns for the four most common amino acids: alanine, leucine, glycine, and serine. The most notable differences are the serine and alanine enrichments at the 5′-ends of the eudicot and *Gramineae* genes, respectively. These differences can be explained by a GC content change in the first codon position, from UCN in eudicots, to GCN in *Gramineae*. A closer examination of the best available rice–*arabidopsis* homologs (possible orthologs) reveals a frequent conversion of serine to alanine, as if the most important constraints on amino-acid sequence were side-chain volumes (Zamyatnin 1972) and accessible surface areas (Chothia 1976).

The gradient in amino acid usage was actually much stronger than is implied by Figure 6, because of the previously noted tendency of our ensemble averages to obscure the true magnitude of the gradients. For a more powerful demonstration of this effect, we computed the probability of a homologous match across the monocot–eudicot divide, as a function of position from the start of the coding region. Starting with the most recent set of cDNA sequences for either rice or *arabidopsis*, we searched for homologous matches in the complete genome sequences of *arabidopsis* or rice, respectively. We compared the protein sequences to all six reading frames of the target genome, using TblastN (Altschul and Gish 1996), so that when the homology search failed, it would not be because of a gene being missing from the annotation of the target genome. We demonstrate in Figure 7 how the probability of finding a TblastN hit dropped precipitously at the 5′-end of the genes. Far from the end, it was ~90%, but within the first few hundred bases of the end, it fell well below 50%. To show that gene size was not a factor, we divided the cDNAs into two classes, based on size of coding region. No differences were observed between small and large genes. The direction of analysis (i.e., whether we started with rice or *arabidopsis* cDNAs) did not make a difference either.

It is certainly possible that some of this effect is attributable to signal peptides, which tend to be poorly conserved, but given that the average eukaryotic signal peptide is 22.6

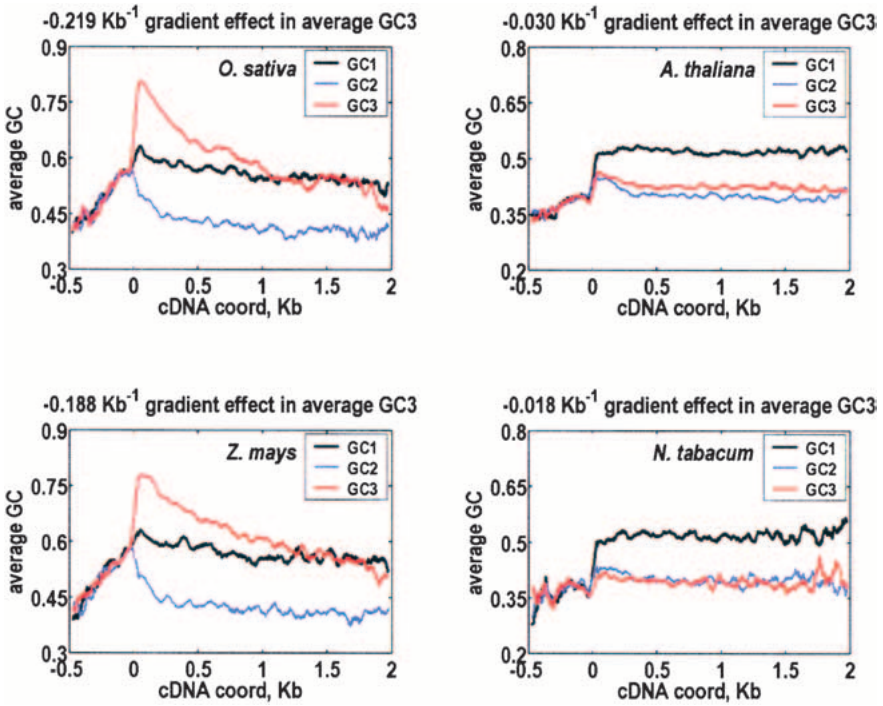

Figure 4 GC1, GC2, and GC3 content as a function of cDNA position, relative to the start of the coding region, and averaged over all cDNAs with a 51-bp sliding window. Shown here are *O. sativa*, *Z. mays*, *A. thaliana*, and *N. tabacum*. Phase information is extended into the 5′-UTR.

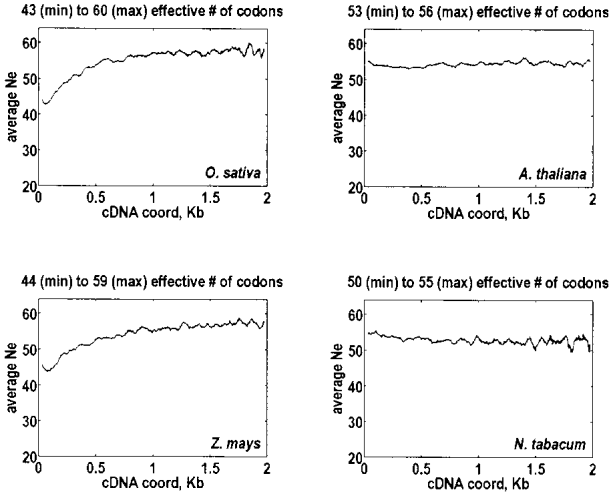

Figure 5 Effective-number-of-codons as a function of cDNA position, relative to the start of the coding region, and averaged over all cDNAs with a 51-bp sliding window. Shown here are *O. sativa*, *Z. mays*, *A. thaliana*, and *N. tabacum*. *Ne* is a statistical measure of codon bias. It is 20 if only one codon is likely to be used for each amino acid, implying maximal bias. It is 61 if every codon is equally likely to be used, implying minimal bias.

amino acids (http://www.cbs.dtu.dk/services/SignalP/sp_lengths. html) or 67.8 bp, differences in signal peptide sequences cannot explain the lack of homology. To confirm that this lack of homology is a monocot–eudicot problem, we performed a `TblastN` analysis for maize-to-rice, and no comparable effect was observed (data not shown). Parenthetically, there is also a 3′-end effect, much shorter in range than the 5′-end effect, but it is why in Figure 7 we truncated the analysis at the midpoint of each gene.

Intron Gradients

Another issue is, what happens to the introns? Is there an intron effect? It is not unreasonable to expect one if transcription is involved because the introns are transcribed along with the exons. Plant introns are more AT-rich than plant exons, by at least 10% on average. In computing ensemble averages, it is therefore important that we analyze exon and intron sequences separately, to avoid artifacts caused by the differences in the baselines for exons and introns. The origin is still at the junction of the 5′-UTR and coding region, but we must use genomic instead of cDNA coordinates. There is a subtle problem with this definition because there are no intron data at the origin. The start codons, however, are situated randomly with respect to the splice sites, and so, averaged across the ensemble of cDNAs, there is only a small window near the origin with no intron data. As shown in Figure 8, there is a GC content gradient in rice introns, but not in *arabidopsis* introns.

DISCUSSION

We have presented a description of the GC content, codon usage, and amino acid usage gradients in *Gramineae* genes. How it arose was of secondary concern because of the important practical implications for the annotation of the rice genome sequence, both for gene-prediction programs and for detection of protein homology across the monocot–eudicot divide. Considering the still unresolved issue of how GC content heterogeneities arise, it is outside the scope of this manuscript to review all extant models, especially as, to our knowledge, no model predicts a gradient. Nevertheless, it is worth pointing out the salient features that any successful model must have.

The data indicate that there is transcription-related mutation bias and translation-related selection. The transcription-related bias is manifested in the strict directionality of the gradient, with GC contents always decreasing from 5′- to 3′-ends. This would predict that there is a gradient effect in the introns because they are also transcribed, and as we were able to show, there most certainly are GC content gradients in rice but not in *arabidopsis* introns. We can only speculate on the molecular mechanisms, but our leading hypothesis is that it would include elements of transcription-coupled DNA repair (Thoma 1999; Svejstrup 2002), coupled to the process of transcription initiation, elongation, and termination (Kim et al. 2001). The overall bias toward higher GC content is attrib-

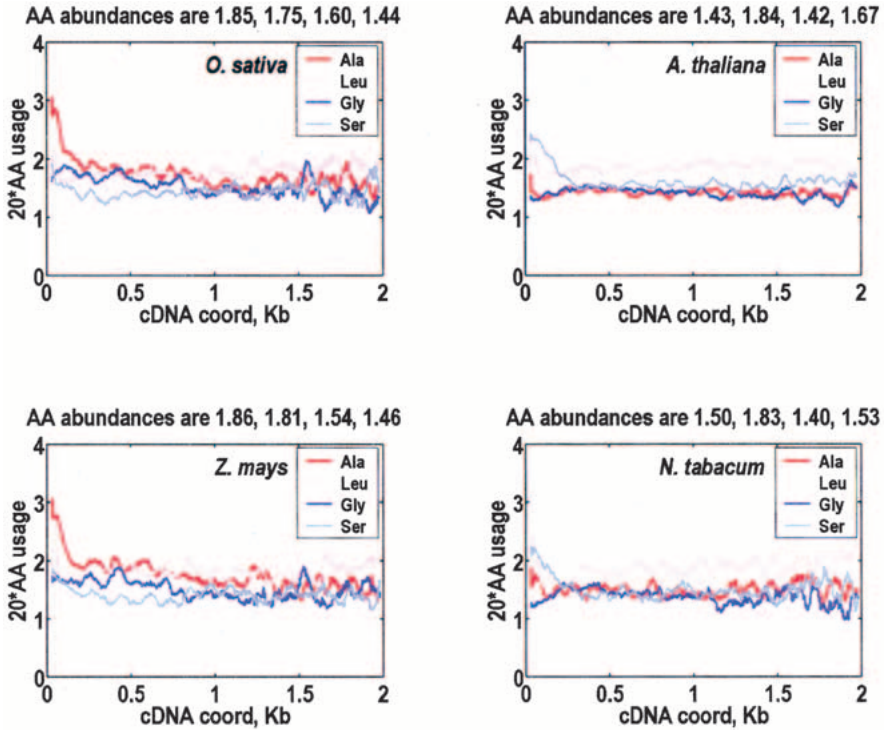

Figure 6 Frequency of occurrence for amino acids alanine, leucine, glycine, and serine, as function of cDNA position, relative to the start of the coding region, and averaged over all cDNAs with a 51-bp sliding window. Shown here are *O. sativa*, *Z. mays*, *A. thaliana*, and *N. tabacum*. When all 20 amino acids occur with equal probability, the normalized frequencies are 1.

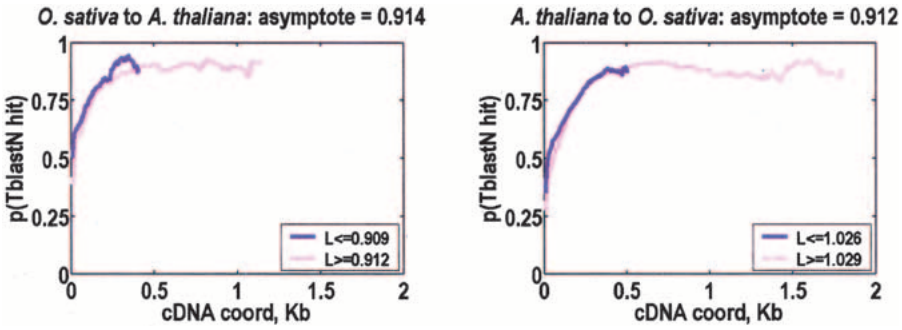

Figure 7 Probability of a homologous match across the monocot–eudicot divide, as a function of cDNA position, relative to the start of the coding region. `TblastN` searches were conducted in both directions, rice-to-*arabidopsis* and *arabidopsis*-to-rice. The cDNA data were divided into two equally populated groups, based on the size of the coding region, to emphasize that the reduced probability at the 5′-end is a position effect, not a gene size effect.

utable to the low-fidelity polymerases that facilitate replicative bypass (Cleaver et al. 2001). A gradient would arise when the repair process aborts or bypasses the lesions to be repaired more frequently than transcription itself.

The best evidence for a translation-related selection is the sharp transition in GC content at the start of the coding region, in both exons and introns. This makes sense if, in addition to the mutational biases, selection acts on the coding sequences, and the adjacent noncoding sequences, UTRs and introns, are inherited along with them. The influence of translation on codon usage can be explained by the relative abundance of different tRNAs (Percudani 2001). A closer examination shows that the maximal GC content and minimal effective-number-of-codons are located almost 50 bp after the start of the coding region. This is not an artifact of our 51-bp averaging windows, as we could reproduce the result without averaging windows. It is more likely to be a reflection of the physical dimensions of the underlying translation machinery.

Although GC content heterogeneities likely involve basic cellular processes, such as transcription and translation, the fact that these mechanisms are highly conserved does not imply that they are exactly the same in all organisms. Our data emphasize that there are enough differences, just within plants, to justify studying monocots and eudicots in isolation, much as complex diseases are often more readily studied by first segregating the phenotypes into subclasses.

METHODS

Figures 2–6 were based on cDNA sequences from GenBank Release 123, dated April 15th, 2001. Obvious redundancies were eliminated up front by removing any cDNA that was 90% contained in another. We required that the coding regions be annotated clearly, which eliminated many lower-quality cDNAs. The four largest *Gramineae* data sets were *O. sativa* (rice), *Z. mays* (maize), *Hordeum vulgare* (barley), and *Triticum aestivum* (wheat), with 776, 710, 345, and 339 cDNAs, respectively. The six largest eudicots data sets were *A. thaliana* (*arabidopsis*), *N. tabacum* (tobacco), *Lycopersicon esculentum* (tomato), *Glycine max* (soybean), *Pisum sativum* (pea), and *Solanum tuberosum* (potato), with 1904, 612, 493, 399, 382, and 276 cDNAs, respectively. Other figures, featuring only rice and *arabidopsis*, were created later, using GenBank 125, dated August 15th, 2001. This got us 813 and 4729 cDNAs for *O. sativa* (rice) and *A. thaliana* (*arabidopsis*), respectively. All the intron sequences came from an alignment of cDNA to genomic sequence. For rice, we used the Beijing data (Yu et al. 2002), and for *arabidopsis*, we used GenBank 125. The cDNA alignments favored quality over quantity. They required a 98% sequence identity, and consensus splice sites GT–AG were confirmed in almost all cases.

To compute the probability of a homologous match across the monocot–eudicot divide, we created two arrays, Ptop() and Pbot(), with one element for each base position relative to the start of the coding region. The number of cDNAs at each base position was recorded in Pbot(). We converted the cDNA sequence to protein sequence and compared it with all six reading frames of the target genome, using `TblastN` (Altschul and Gish 1996). Wherever a `TblastN` hit could be found, we incremented Ptop() by 1, but no more than 1, regardless of the number of hits. To avoid picking up any 3′-end effect, only the first half of each gene was used. Because positions at the end of the coding region could only get hits from one side, and positions farther from the end could get hits from both sides, there was a small "edge effect." It was corrected by performing an *arabidopsis*-to-*arabidopsis* comparison and using that to normalize the *arabidopsis*-to-rice comparison (and similarly for rice-to-*arabidopsis*). Finally, the probability of a homologous match was computed as the ratio of Ptop() divided by Pbot().

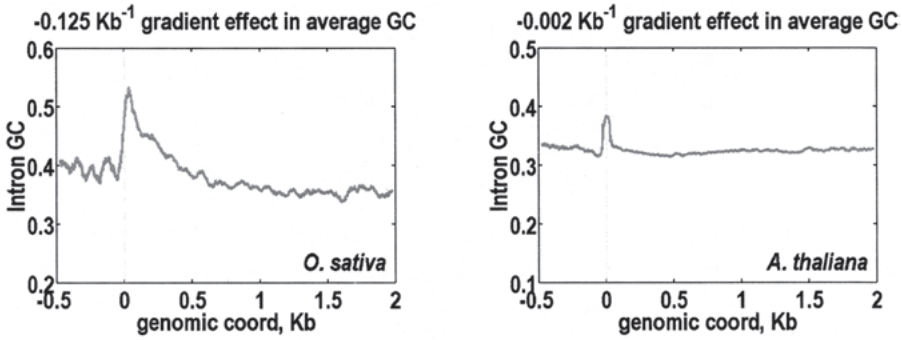

Figure 8 GC content of intron DNA, as a function of genomic position, relative to the start codon, and averaged over a 51-bp sliding window. Every intron is from a cDNA-to-genomic alignment. Only *O. sativa* and *A. thaliana* had enough genomic data to be so analyzed. Note that, unlike the previous figures, this abscissa is in genomic instead of cDNA coordinates.

ACKNOWLEDGMENTS

We thank Drs. Lars Bolund and Maynard Olson for comments and suggestions. We thank Amersham Pharmacia Biotech (China) Ltd., SUN Microsystems (China) Inc., Digital China Ltd., and Dawning Computer Corp. for their continuous support and excellent service. This work was jointly sponsored by the Chinese Academy of Sciences, Commission for Economy Planning, Ministry of Science and Technology, National Natural Science Foundation of China, Beijing Municipal Government, Zhejiang Provincial Government, and Hangzhou Municipal Government. This work was also supported by a grant

from National Institute of Environmental Health Sciences (1 RO1 ES09909).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F. and Gish, W. 1996. Local alignment statistics. *Methods Enzymol.* **266:** 460–480.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241:** 3–17.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Carels, N. and Bernardi, G. 2000. Two classes of genes in plants. *Genetics* **154:** 1819–1825.

Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105:** 1–12.

Cleaver, J.E., Karplus, K., Kashani-Sabet, M., and Limoli, C.L. 2001. Nucleotide excision repair "a legacy of creativity." *Mutat. Res.* **485:** 23–36.

Comeron, J.M., and Aguade, M. 1998. An evaluation of measures of synonmous codon usage bias. *J. Mol. Evol.* **47:** 268–274.

Dubcovsky, J., Ramakrishna, W., SanMiguel, P.J., Busso, C.S., Yan, L., Shiloff, B.A., and Bennetzen, J.L. 2001. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol.* **125:** 1342–1353.

Duret, L. and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci.* **96:** 4482–4487.

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

Forsdyke, D.R. and Mortimer, J.R. 2000. Chargaff's legacy. *Gene* **261:** 127–137.

Karlin, S., Campbell, A.M., and Mrazek, J. 1998. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32:** 185–225.

Kim, D.K., Yamaguchi, Y., Wada, T., and Handa, H. 2001. The regulation of elongation by eukaryotic RNA polymerase II: A recent view. *Mol. Cells* **11:** 267–274.

Knight, R.D., Freeland, S.J., and Landweber, L.F. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2:** research0010.1–0010.13.

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.

Percudani, R. 2001. Restricted wobble rules for eukaryotic genomes. *Trends Genet.* **17:** 133–135.

Powell, J.R. and Moriyama, E.N. 1997. Evolution of codon usage bias in Drosophila. *Proc. Natl. Acad. Sci.* **94:** 7784–7790.

Rogic, S., Mackworth, A.K., and Ouellette, F.B. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11:** 817–832.

Salamov, A.A. and Solovyev, V.V. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10:** 516–522.

Svejstrup, J.Q. 2002. Mechanisms of transcription-coupled DNA repair. *Nat. Rev. Mol. Cell Biol.* **3:** 21–29.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Thoma, F. 1999. Light and dark in chromatin repair: Repair of UV-induced DNA lesions by photolyase and nucleotide excision repair. *EMBO J.* **18:** 6585–6598.

Wright, F. 1990. The effective-number-of-codons used in a gene. *Gene* **87:** 23–29.

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

Zamyatnin, A.A. 1972. Protein volume in solution. *Prog. Biophys. Mol. Biol.* **24:** 107–123.

WEB SITE REFERENCES

http://www.cbs.dtu.dk/services/SignalP/sp_lengths.html.

Received February 14, 2002; accepted in revised form April 3, 2002.