Opinion TRENDS in Plant Science Vol.9 No.6 June 2004

The new genes of rice: a closer look

Kamel Jabbari[1,4], Stéphane Cruveiller[2,3], Oliver Clay[2], Jérôme Le Saux[3] and Giorgio Bernardi[1,2]

[1]Laboratoire de Génétique Moléculaire, Institut Jacques Monod, Tour 43, 4 place Jussieu, 75005 Paris, France
[2]Laboratoire di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italy
[3]Atelier de Génomique Comparative, Genoscope, Centre National de Séquençage, 2 rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France
[4]Present address: Laboratoire des Organismes Photosynthétiques et Environnement, Département de Biologie – ENS/CNRS FRE 2433, Ecole Normale Supérieure, 46 Rue D'Ulm, 75230 Paris Cedex 05, France

Reports accompanying draft or finished sequences of rice chromosomes and full-length cDNA libraries indicate that between a third and half of the (largely predicted) protein-coding genes of rice might have no identifiable homologs in *Arabidopsis* and/or other species. The set of apparent 'no-homolog' sequences are predicted to exhibit striking compositional and structural deviations from experimentally verified protein-coding sequences. Here we discuss evolutionary and other implications of the proposed gene set and argue that a more likely answer to the riddle is that a large proportion of the anomalous rice sequences are never translated into functional proteins *in vivo*, that is, they represent incorrect gene predictions.

On 5 April 2002, the first draft sequences of entire rice genomes were published, together with their initial annotations [1,2]. Perhaps the most striking discovery, reported by one of the two groups [1], was the unusual gene set they proposed, which was largely based on predictions. Roughly half of the protein-coding genes in rice had, apparently, no significant homologs in the previously sequenced genome of *Arabidopsis*, or, in most cases, in the sequenced DNA of any other taxa. Since then, sequences of individual rice chromosomes have been finished [3,4] and a database of full-length cDNA sequences has been constructed [5]. These sequences have been consistently accompanied by estimates of the proportion of protein-coding genes that have no detectable homologs in other flowering plants that range between a third and half of the sequences.

Two possible interpretations come to mind for such findings. The first interpretation is that most of the 'unmatched' gene predictions are incorrect and will never be confirmed, that is, if one eliminates all false predictions one will find that most of the remaining, true genes do have orthologs or close paralogs in other species. An alternative interpretation is that the gene predictions are largely correct, that is, between a third and half of all protein-coding genes have no homologs in any entirely or partially sequenced species. This second interpretation would suggest that ~15 000–20 000 genes were formed (or altered beyond recognition) during a relatively brief evolutionary time span of some 130 million–200 million years [6,7]. In other words, gene duplication followed by minor mutations or modifications was not the route to these extant genes, but instead *ex novo* genesis, or such dramatic reshuffling of DNA that the original fragments can no longer be reconstructed or recognized in dicots via any known method. Given the relatively short time span, presumably *ex novo* genesis of protein-coding genes would have typically occurred via promoters that fortuitously sequestered adjacent, randomly occurring stop codon-free DNA. Soon after such an event, the protein product would have needed to acquire functional status to prevent accidental erosion of the coding DNA: otherwise its alleles, unchecked by selection, would presumably soon become riddled with stops except in a few rare genes. This route to functionality becomes difficult to envisage when one considers the many differences between the 'matched' and 'unmatched' genes of rice.

Putative rice genes with no homologs in *Arabidopsis* ('NH genes') exhibit striking compositional and structural anomalies

First, as Jun Yu *et al.* [1] have noted, the putative genes with no detectable homologs in *Arabidopsis*, or 'NH' genes, have coding sequences that are, on average, only about half as long as those from genes that do have homologs. Second, they have twice as many introns as normal in the 0.2–2.0 kb range. Thus, associated splicing constraints would have to be met, in addition to the already heavy requirements for an *ex novo* gene to evolve to functionality (or even translatability) in 130 million–200 million years. Third, the distribution of GC levels from the putative gene set contains a hump of unusually GC-rich coding sequences [1]. Further inspection (compare Ref. [8] with Figure 1a) reveals that this hump is mainly because of 'unmatched' genes. Many of their sequences fail to show differences in GC among the three codon positions and have GC2 (second position levels) soaring to heights that are virtually unknown either in prokaryotic or eukaryotic coding regions [8]. In particular, such unmatched sequences deviate strongly from a linear relationship between GC2 and GC3 that is largely conserved (given enough intragenomic variability) from human to *E. coli* [8]. These observations make it unlikely that accidental sequestering of such extreme noncoding DNA by

Corresponding author: Giorgio Bernardi (bernardi@szn.it).

 1360-1385/$ - see front matter doi:10.1016/j.tplants.2004.04.006

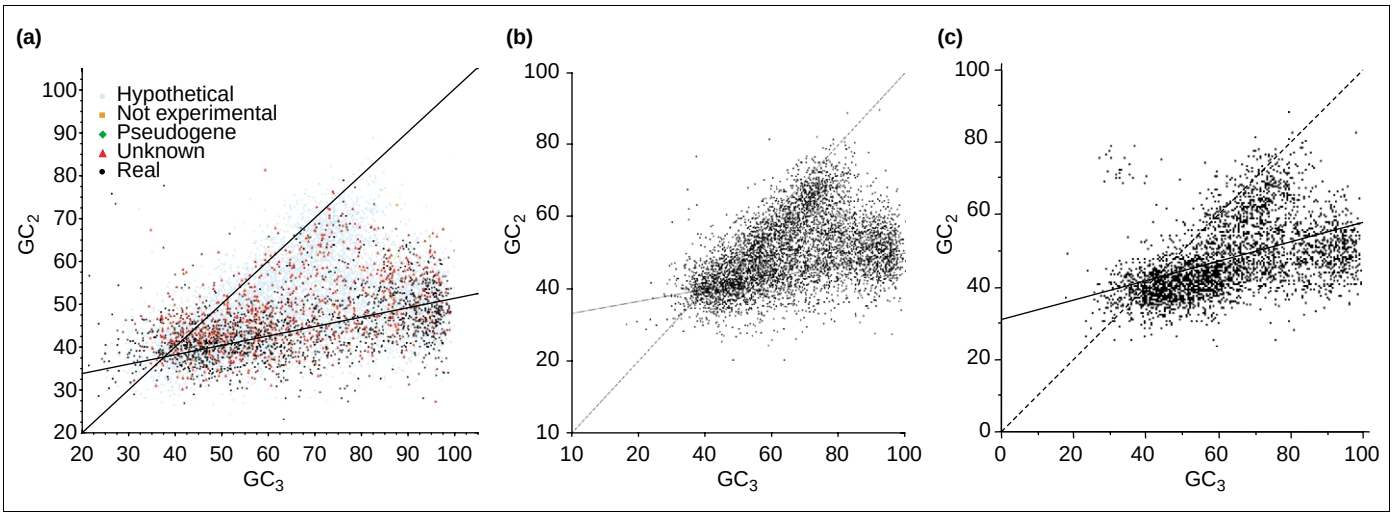

Figure 1. Scatterplots of GC2 versus GC3 for predicted sets of protein-coding sequences of rice. These scatterplots allow many compositionally anomalous, predicted genes to be quickly recognized and targeted for further study or verification. Known protein-coding sequences, in rice and other species (including human and *E. coli*), cluster along an evolutionarily conserved line (bottom arm of scatterplot). This line has a low slope (~1/6) and passes through the main diagonal at ~40%. In other words, if a real gene has a high GC level (% guanine and cytosine) in its third codon positions (GC3), it will almost certainly have a low GC in its second codon positions (GC2). This is not the case for incorrectly predicted 'genes' that are noncoding DNA: because they do not have real codons, there are no systematic GC differences among codon positions. The incorrect 'genes' can therefore extend upwards along the steeper diagonal of slope 1 (GC2 = GC3), forming an upper arm of the scatterplot. Chromosomes and sources: **(a)** Experimentally verified and predicted coding sequences present in GenBank when the first rice drafts were published (release 129, retrieved 31 May 2002; $N = 10\ 087$), after eliminating redundancy and classifying the sequences according to their annotations (real genes, not experimental, unknown, pseudogenes and hypothetical; modified from Ref. [31], with permission). **(b)** Chromosome 1 [3] (modified from Ref. [8], with permission). **(c)** Chromosome 10 [4].

promoters would have resulted in a functional gene that is now translated *in vivo*. (The copy-number and homology searches conducted by Yu *et al.* [1] essentially rule out transposon, repetitive DNA or lateral transfer explanations for the predicted set of rice genes.)

The NH genes remain an enigma

After the initial reports that accompanied the draft genome sequences, we would have expected that the astonishing finding of ~20 000 unmatched genes in rice would have attracted the attention of at least two large groups of research workers. One group would have been those involved in exploiting the new information on genes and chromosomal sequences for practical purposes (e.g. plant breeding) and in setting up tools (e.g. microarrays or mutant libraries) to facilitate functional genomics studies on the rice genes. The other expected group would have included those fascinated by the apparently unprecedented evolutionary phenomenon of a recent boom of genes within a single taxon of monocot or cereal plants. Surprisingly, the literature published since April 2002 reveals no attention of either kind being given to the report of Yu *et al.*, even though by now draft and/or finished, annotated sequences of rice chromosomes have been published (including different subspecies, *japonica* which is more commonly grown in Japan and Korea, and *indica*, the most widely consumed rice in China and India), as well as a large (>28 000) collection of experimentally retrieved, sequenced full-length cDNAs for the *japonica* subspecies [5]. In the sequence of chromosome 1, reported by Takuji Sasaki *et al.* in 2002 [3], the authors confirmed that a large proportion of the putative genes in that chromosome remained unmatched. In the smaller chromosome 10, finished and described in June 2003 [4], roughly a third of the genes still have no homolog in *Arabidopsis*. In the cDNA collection presented in July 2003, the authors found that 36% of the putative coding regions had no homologs in *Arabidopsis*. Annotations in *Arabidopsis* have been revised or improved since the first draft of that genome and full-length cDNA databases are also available for this species [9,10], but current NH estimates would probably remain close to their original values using the same homology criteria. Moreover, marked compositional deviations (many genes with abnormally high GC2 levels, extending along the upper main diagonal of the GC3 versus GC2 scatterplots) are found in all the NH gene sets (Figure 1).

A continued, widespread acceptance of a protein-coding status for the NH genes could, if inappropriate, have practical consequences for post-genomics projects. Current projects include a planned resource center that will eventually hold mutants for 90% of the rice genes [11], or a recent augmentation of the original rice chip, representing 21 000 genes that were selected for reliability [12,13], to a full-genome chip representing more than twice as many genes [14]. The unquestioned protein-coding status of NH genes could affect gene predictions in other plants and thus delay the understanding of their genomes.

As Donald Kennedy stated when the initial draft sequences of rice were published [15], the full rice sequence now 'affords entry into the similar but larger genomes of the other cereal grains on which the world depends'; other statements pointing in the same direction are found in annual reports or web pages of companies focusing on cereal post-genomics. Such observations apply, in particular, to the genes in the cereal genomes. Genes that might play important roles in disease resistance, stress tolerance, plant development and yield enhancement are prime economic targets. A snowball effect could be triggered if unconfirmed gene predictions in one plant are routinely used to reinforce or support gene predictions, via sequence similarity, in another plant. If such an effect

escalates, it could jeopardize or delay the reliable annotation and mapping of protein-coding genes in other cereals. For example, if 30% of the putative genes for rice do not code for proteins *in vivo* but are generally believed to do so, there could be a snowball effect in which putative gene predictions in several species are recursively or autocatalytically built on strings of previous predictions [8,16]. Indeed, given the huge numbers of predicted genes, we have apparently no way of quickly purging annotations of fake genes by applying bulk experimental checks at the protein level.

Large-scale proteomics methods might eventually allow such routine checks, perhaps via chip-based antigen–antibody recognition or future methods extending mass spectrometry. At present, bulk experimental checks at the protein level still appear to be rare and cover limited subsets of the proteome. Recent studies have not yet confirmed the proportions of NH proteins among those detected *in vivo*. An experimental study of the rice proteome [17] directly identified 2528 unique proteins in the leaf, root and seed, but only 360 of them were reported to be unmatched proteins. This proportion (14.2%) remains well below 33–50%. A proteome-wide discrepancy of this kind would suggest that a large proportion of the NH proteins are non-functional and/or that NH proteins tend to be much less widely or frequently expressed than other rice proteins.

In view of the difficulties in experimentally checking the predicted protein set of rice, re-scrutiny of existing prediction algorithms, of the training and reference sequences used, and of possible artefacts and biases, becomes an essential task. The potential waste could be considerable, if funds and resources continue to be invested in expensive, experimental postgenomic analyses of 10 000–20 000 putative genes that might, later, turn out never to be expressed as proteins.

Are the NH genes really NH?

In view of its potential importance, it is surprising that during the past 18 months there has been almost no further reflection on the 'unmatched' genes. Few comments or follow-up studies have appeared, although one recent report [18] has addressed the question of whether some of the putative 'NH' rice genes possess homologs (orthologs or paralogs) in *Arabidopsis* or other eukaryotes. Some homologs could have escaped conventional homology searches but be revealed by more-sensitive methods. The finding of additional homologies would reduce the credibility burden of massive, recent *ex novo* gene formation in monocots and the need to explain such an unusual event. Lucjan Wyrwicz *et al.* [18] propose the use of structure-guided protein comparisons to uncover previously hidden homologies. They report that approaches based on recognizing local similarities in protein structure [19,20] do reveal homologies in other species for putative rice proteins with 'NH' domains. According to their report, at least a third of the 100 rice sequences (translated from transcripts of Ref. [5] and containing InterPro domains that apparently had no homologs in *Arabidopsis*) that they analyzed had sequence and structural similarity to *Arabidopsis* transcripts. This would suggest that more-sensitive measures might eventually detect homologs for other 'NH' genes. It will be interesting to watch further developments of such structure-guided sequence comparisons and to see how they could be extended to allow automated screening of entire predicted protein sets for rice. In principle, structure predictions could be used to guide sequence comparisons and error or significance tracking could be woven into automated, bulk searches for structural homologs. However, at present, it seems safest to consider the possibility that a large proportion of the 'NH' rice sequences might not have homologs, at the protein level, in *Arabidopsis*.

Are the NH genes a consequence of gene loss in *Arabidopsis*?

At least some genes were presumably lost during the evolution of the compact *Arabidopsis* genome. Because *Arabidopsis* is the only entirely sequenced plant with which rice can be exhaustively compared, one could ask if its particularly compact genome might explain much of the NH enigma. More precisely, could most of the NH genes have been lost in the lineage leading to *Arabidopsis*, but be present – although not yet sequenced – in other plants? This possibility cannot be completely ruled out, but it would require a remarkable bias in the genes that have been chosen so far for sequencing (in addition to other unlikely events discussed here).

It would be interesting to focus on *bona fide* proteins that exist in cereals but not in dicots, or that were lost in some dicots such as *Arabidopsis* but not in others. Only a small number (several hundred) of experimentally studied protein-coding genes, encoded in the nuclear genome, had been strategically sequenced in monocots other than rice, so it is unlikely that such databases would enable evolutionary tracking of rice-, cereal- or monocot-specific genes. Until more angiosperm genomes are sequenced, appropriate zooblot experiments, systematic sequencing of orthologs and/or microarray-based comparisons might help. For now, simple evolutionary considerations suggest that some of the NH genes without matches in *Arabidopsis* were present in the common ancestor of monocots and dicots, that some are common to cereals and other monocots but not to dicots, and that some are common to rice and other cereals such as wheat and/or barley, but not to other monocots. One would expect only a small remainder to be truly rice specific, even if the high initial estimates of the NH gene proportion in rice (a third to half) are correct.

How could many NH genes have been incorrectly predicted?

Some algorithmic or training artefacts or sensitivities that could have led to large numbers of unmatched, compositionally anomalous rice gene predictions have been discussed by Stéphane Cruveiller *et al.* [8]. First, the training or reference sets of experimentally validated, functional nuclear cereal genes available for the initial predictions in rice must have been small because only a few hundred coding sequences were known. Preparing a representative reference sample of *bona fide* non-coding rice DNA would not have been an easy task. By contrast,

the known human genes that were used for predictions in the human genome contained up to 9000 or more coding sequences. Several long sequences had been studied in detail, with the help of experiments, to trap coding DNA and identify the non-coding DNA [21].

Second, in some cases, higher scores were given to rice gene candidates if the same prediction had already been made via a different method or by a different group. Such strategies could initiate a snowball effect in which false genes trigger the prediction of more false genes, either in the same species or in different species [8,16].

Third, the NH rice genes, as described by Yu *et al.*, exhibited a marked GC gradient, in which the GC level decreased from the 5′ to the 3′ end of the gene [1]. A tendency for GC- and/or CpG-richer 5′ ends of genes can be observed in many real genes in mammals and other taxa, for example where the promoter region is covered by a CpG island, but also where it is not [22,23]. It is not unreasonable to suspect that some criteria and/or hidden weighting schemes, derived via machine learning, could have effectively increased the scores of rice gene candidates if they exhibited GC gradients. In other words, in the genomes of mammals (and/or of other species that were initially used to develop or train gene prediction programs), GC gradients and/or other sequence features might be preferentially present in coding DNA, but in rice they might also be frequent in non-coding DNA. Some feature(s) of this type could, in principle, trick gene-recognition programs into mistakenly signaling the presence of protein-coding genes in non-coding rice DNA.

Are the NH genes transcribed *in vivo*?

Could many of the GC-rich NH open reading frames (ORFs) be transcribed and play a role(s) at the RNA level, even though they might never be exported and translated? In other words, even though the programs that detected the NH genes were presumably targeting pol II-transcribed, protein-coding genes, could those programs have accidentally picked up many GC- or 'GC2'-rich, *bona fide* transcripts that do not code for proteins? High GC levels, similar or higher to those of coding regions, might be frequent among some RNAs. For example, in many species, the genes for rRNA (rDNA) are GC-richer than genes coding for proteins. Furthermore, intact ORFs might stabilize transcripts even if they are not destined for translation because RNAs that are riddled with stop codons appear to be more readily targeted for degradation [24].

Recently, >28 000 full-length rice cDNA sequences have been compiled, together with the predicted or verified coding regions inside them and the conceptually translated protein sequences [5]. The authors reported that more than a third (36%) of the putative coding sequences embedded in the experimentally detected cDNAs had no homologs in *Arabidopsis*. We also found a substantial level of compositional anomalies in this dataset (before and after removing short sequences). This would suggest that many anomalous or NH sequences are transcribed, although the hybrid nature of the cDNA dataset and possible uncertainties in locating the correct ORF within some of the cDNAs should be kept in mind. Indeed, the total KOME (*Knowledge-based Oryza Molecular Biological Encyclopedia*, http://cdna01.dna.affrc.go.jp/cDNA) dataset (described in ref. [5] and its online supplement) is a compilation of cDNA sequences from two sources, FAIS (Foundation for Advancement of International Science) and RIKEN (Institute of Physical and Chemical Research), in almost equal proportions: the two institutes used different protocols to obtain the libraries and sequences. We found that the composition and length histograms of the predicted coding sequences differed between the two cDNA collections and that the FAIS set is responsible for most of the compositional anomalies we observed in the pooled cDNAs [i.e. the FAIS data are more compositionally anomalous when compared with experimentally verified genes, with genes that have homologs, or with the RIKEN data (12.3% of the sequences had a GC2 of >60% in the FAIS data compared with only 6.1% in the RIKEN data)]. The predicted ORFs in the FAIS cDNAs are also shorter (13% with <100 codons compared with 6.1% of RIKEN cDNAs and only 21.6% with >400 codons compared with 42% of RIKEN cDNAs). Careful examination of the two experimental protocols and of the strategies for finding the proposed coding regions in the cDNAs therefore might be needed before we can accurately assess which of the two cDNA sets (or which weighted combination of the two sets) is most representative of the transcriptome. In some sequences, the gene-finding strategy was apparently a simple search for the longest ORF in the cDNA, which also leads to some short and/or ambiguous gene predictions. At present, the ORFs sequenced by FAIS suggest that many of the anomalous, predicted protein-coding sequences are transcribed, whereas to date the ORFs sequenced by RIKEN do not strongly support such a conclusion.

A basal level of functional RNA transcripts, not destined for translation but fortuitously containing ORFs, would also seem compatible with the earlier results of Yu *et al.* [1] who found that 15.4% of their predicted NH genes matched ESTs from UniGene clusters.

Under what conditions could the NH genes be expressed as functional proteins?

To help to assess whether the compositionally anomalous NH sequences are likely to be translated into functional proteins, one can follow at least two paths. One path could be to experimentally characterize the differential gene expression of the NH mRNAs. For example, it will be interesting to study the first microarray publications using the new whole-genome chip: how often will NH genes be among confirmed up- or down-regulated genes in rice? Comparison with microarray results for *Arabidopsis* [10] could also be useful in this context. Even a few examples of clear up- or down-regulation of anomalous NH genes would provide a helpful, first glimpse into their possible expression contexts and thus into their possible roles. Another path is to characterize the structural and functional properties of the putative proteins they encode.

We conducted a pilot study using structure programs and domain searches to see how the proteins encoded by compositionally deviant predicted rice genes (GC2 approximately equal to GC3 when GC3 is high) would differ from those that exhibit the usual intra-codon

contrast (GC2 ≪ GC3 when GC3 is high). When plotted in the GC2 versus GC3 plane, the normal genes are scattered closely around a well-documented, evolutionarily conserved crest (approximately GC3 = 6 GC2 − 200%) that departs strongly from the main diagonal (Figures 1a–c); only rarely do they have high GC2 levels. 'Matched' rice genes generally have far fewer compositionally deviant genes than 'unmatched' genes do. To give an example, we found that 27.5% of the NH genes published by Yu *et al.* [1], but only 6.5% of the matched genes (with-homolog or WH), had GC2 levels above 55% and that 7.9% of the NH genes, but only 0.3% of the WH genes, had GC2 levels above 65% [8].

We compared a set of obvious deviants (~1000 genes with GC2 levels of >65%, a level that is rare among normal genes) and a similarly sized set of obvious conformers (GC2 levels of <37%) at the protein level. We used the SOPMA and PHD structure prediction programs [25,26] available on the PBIL web server [27] (SOPMA, but not PHD, distinguishes coils from turns). Protein domains were determined locally using the InterProScan software [28,29]. In the GC2-rich deviants, almost all of which are of the 'unmatched' type, coil structures were essentially twice as frequent, and helices half as frequent, compared to the compositionally normal group; only the turn and sheet proportions were similar in both protein sets. This is in agreement with the idea that certain residues that are over-represented in the GC2-rich proteins (in particular, proline, glycine and tyrosine) can interrupt or 'break' a helix more easily. As one would expect from a set of incorrectly predicted genes, protein domains from the Pfam database [30] were rarely found in the conceptual translations of anomalous genes (i.e. they were found in 30.3% of the anomalous proteins compared with 86.5% of the normal proteins). Those that were present in the anomalous set were half proline- or glycine-rich domains and half alanine-rich domains.

In summary, the sets of protein-coding genes that are proposed for the rice genome are still enigmatic at the DNA, RNA and protein levels. During the next year or two, we should be able to determine if we are on the verge of discovering a new type(s) of anomalous but functional proteins, possibly specific to *Gramineae* or monocots, with highly unusual compositional and structural features, containing few of the domains that are known to date. Alternatively, if it turns out that a third or more of the published *ab initio* coding predictions for rice were incorrect, as we suspect, then this insight should point the way for improvements and future caution, for example, when one trains programs with small sequence sets. It would also validate a new criterion we propose to help in gene recognition. If anomalous behavior in the GC3 versus GC2 plots is also a reliable indicator of noncoding DNA in plants, as we claim, then this would add another parameter that can be included in prediction programs, or a simple plot that can be drawn in the ultimate validation step. Either way, the annotation and interpretation of the rice genome sequences remain, together, an unfinished and challenging task.

Acknowledgements
This paper is dedicated to the memory of Jeff Schell (1935–2003).

References

1 Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296, 79–92
2 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100
3 Sasaki, T. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature* 420, 312–316
4 The Rice Chromosome 10 Sequencing Consortium, (2003) In-depth view of structure, activity, and evolution of rice. *Science* 300, 1566–1569
5 Kikuchi, S. *et al.* (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* 301, 376–379
6 Stewart, W.N. and Rothwell, G.W. (1993) *Paleobotany and the Evolution of Plants*, Cambridge University Press
7 Whitelegge, J.P. (2002) Plant proteomics: BLASTing out of a MudPIT. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11564–11566
8 Cruveiller, S. *et al.* (2003) Compositional features of eukaryotic genomes for checking predicted genes. *Brief. Bioinform.* 4, 43–52
9 Seki, M. *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* 296, 141–145
10 Yamada, K. *et al.* (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302, 842–846
11 Cyranoski, D.A. (2003) Recipe for revolution? *Nature* 422, 796–798
12 Zhu, T. *et al.* (2003) Transcriptional control of nutrient partitioning during rice grain filling. *Plant Biotechnol. J.* 1, 59–70
13 Anderson, M. *et al.* (2003) Identification of nutrient partitioning genes participating in rice grain filling by singular value decomposition (SVD) of genome expression data. *BMC Genomics* 4, 26
14 Vogel, G. (2002) Retreat from Torrey Mesa: a chill wind in Ag research. *Science* 298, 2106
15 Kennedy, D. (2002) The importance of rice. *Science* 296, 13
16 Rinner, O. and Morgenstern, B. (2002) AGenDA: gene prediction by comparative sequence analysis. *In Silico Biol.* 2, 195–205
17 Koller, A. *et al.* (2002) Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11969–11974
18 Wyrwicz, L.S. *et al.* (2004) How unique is the rice transcriptome? *Science* 303, 168
19 Ginalski, K. *et al.* (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19, 1015–1018
20 Ginalski, K. and Rychlewski, L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res.* 31, 3291–3292
21 Chen, E.Y. *et al.* (1996) Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. *Hum. Mol. Genet.* 5, 659–668
22 Mizuno, M. and Kanehisa, M. (1994) Distribution profiles of GC content around the translation initiation site in different species. *FEBS Lett.* 352, 7–10
23 Shimizu, T.S. *et al.* (1997) CpG distribution patterns in methylated and non-methylated species. *Gene* 205, 103–107
24 Bühler, M. *et al.* (2002) Intranuclear degradation of nonsense codon-containing mRNA. *EMBO Rep.* 3, 646–651
25 Geourjon, C. and Déleage, G. (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.* 11, 681–684
26 Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232, 584–599
27 Perrière, G. *et al.* (2003) Integrated databanks access and sequence/structure analysis services at the PBIL. *Nucleic Acids Res.* 31, 3393–3399
28 Mulder, N.J. *et al.* (2003) The InterPro database brings increased coverage and new features. *Nucleic Acids Res.* 31, 315–318
29 Zdobnov, E.M. and Apweiler, R. (2001) InterProScan — an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848
30 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280
31 Cruveiller, S. *et al*. Incorrectly predicted genes in rice? *Gene* (in press)