9 Harushima, Y. *et al.* (1998) A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics* 148, 479–494

10 Kurata, N. *et al.* (1997) Physical mapping of the rice genome with YAC clones. *Plant Mol. Biol.* 35, 101–113

11 Yamamoto, K. and Sasaki, T. (1997) Large-scale EST sequencing in rice. *Plant Mol. Biol.* 35, 135–144

12 Sasaki, T. and Burr, B. (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* 3, 138–141

13 Barry, G.F. (2001) The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* 125, 1164–1165

14 Goff, S.A. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296, 92–100

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

16 Leach, J. *et al.* (2002) Why finishing the rice genome matters. *Science* 296, 45

Universal inheritable barcodes for identifying organisms

Jonathan Gressel and Gal Ehrlich

The needs for recognition of novel conventional or transgenic organisms include protection of patented or Identity Preserved lines, detecting transgenics and tracing dispersal. We propose simple 'Biobarcodes™' using universal PCR primers to recognize the universal 'nonsense' recognition site of all biobarcodes, followed by a variable nonsense sequence. The proposed sequences are long enough to allow recognition in spite of mutations, have stop codons to prevent coding, and will not self anneal. Sequences of PCR-amplified biobarcodes can be compared to a universal database.

Published online: 31 October 2002

Jonathan Gressel
Plant Sciences, Weizmann Institute of Science, Rehovot 76100, Israel.
e-mail: jonathan.gressel@weizmann.ac.il

Gal Ehrlich
Ehrlich & Partners, 28 Bezalel Street, Ramat-Gan 52521, Israel.

These are a variety of needs for devising simpler recognition methods for organisms marketed in commerce or released in the environment; whether they are conventionally selected, mutant or transgenic bacteria, fungi, plants or animals. The needs include: (1) the need for protection for patented or other IP lines, where IP takes on the designation of either 'Intellectual Property' or 'Identity Preserved'. It is often hard to prove that a line has been 'miss-appropriated' by a competitor or illegally grown. (2) Labeling regulatory authorities and various consumer groups are demanding labeling of transgenic commodities. This is accomplished by segregating plants or their products and externally following them throughout their commercial life. Internal markers are an adjunct to such external markings. (3) The need to trace organisms in the environment. The use of biocontrol agents to control weed, bacterial, fungal or insect pests and the use of other live organisms as inoculants is increasing. Knowledge about their dispersal in plants and other organisms as well as in the environment is necessary, irrespective of whether the organisms are indigenous or transgenic. Many of the agents are closely related to known pathogens or pests and there are claims that an organism can change its host range and attack valuable species (with consequences of liability). There are also fears that organisms will introgress into other organisms, and there are needs to ascertain whether the new organism changed host range or whether an epidemic was the result of wild strains [1]. These issues will become more acute with transgenically enhanced biocontrol agents [2]. This can still be a problem even if failsafe mechanisms are transformed into biocontrol agents along with hypervirulence genes [3]. Labeling biocontrol agents with selectable markers (such as *gus*, *gfp* or antibiotic resistance) has been used for following the movement of biocontrol agents in the environment [4,5], but will not differentiate between them if the same few markers are continually used. RFLP has been used to follow and differentiate between one organism and closely related strains of the same pathogen [6].

There has been a considerable expenditure on identifying valuable organisms – transgenic and non-transgenic. Seed companies use AFLP and other molecular techniques to ascertain whether competitors have incorporated their genetic material, and the seed industry uses it to ascertain whether farmers have been storing and reusing transgenic seed off license. Consumer advocacy groups and regulators have been using similar techniques to ascertain whether transgenic products have been mixed with non-transgenic organisms. PCR amplification of the DNA in question, using a series of primers for typically used promoters, terminators, marker genes and the genes of importance, is often used to find trace amounts of transgenes in crops and commodities [7–11]. Some processed products might need tens of PCR reactions because of the mixing of crops and the numbers of possible transgenes that they might contain. This problem will be exacerbated as more transgenic crops and organisms reach the market. The multiple sampling and PCR reactions are time consuming and fraught with problems. For example, the commonly used 35S promoter can be found in almost any non-transgenic commodity containing DNA from a cole (*Brassica*) crop because a proportion of such crops is invariably infected by the ubiquitous cauliflower mosaic virus, the source of that promoter. The PCR results must be further verified by using either restriction endonucleases, Southern blotting or direct sequencing, or using nested PCR or quantitative competitive PCR reactions. Highly processed foods have their DNA fragmented to <400 bp. Detection can be hampered by a lack of availability of DNA reference material and of sequence information [7–11].

http://plants.trends.com 1360-1385/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S1360-1385(02)02364-6

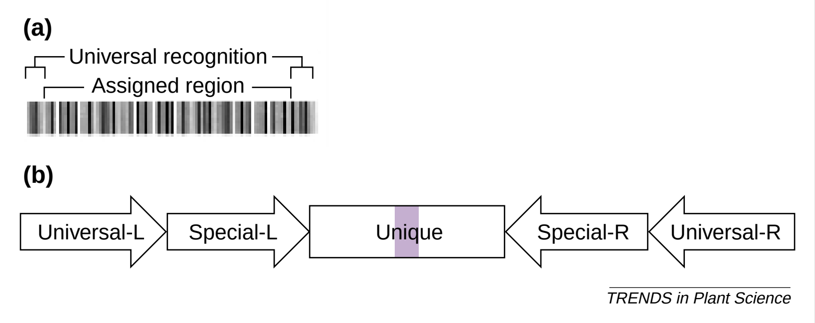

Fig. 1. (a) A standard barcode with universal recognition sites at either end defining orientation, and a variable sequence of bars of various widths and spacing that can be accessed from a database. (b) A schematic representation of a Biobarcode™ with fixed universal nonsense DNA sequences defining universal recognition sites at either end defining orientation, and a unique defined variable nonsense DNA sequence that can be accessed from a database containing a stop codon (purple) to prevent frameshifts from defining peptide coding. A special assigned sequence can be added for those groups that are interested in having further definition of typed products. Abbreviations: L, left; R, right.

Immunodiagnostic methods have been developed for many of the gene products. Again, each sample would have to be reacted with a series of antibodies when there is a possibility of many different gene products. Even one product can be hard to detect when a single mutation is introduced into a commonly occurring plant gene, such as those encoding many herbicide resistances. It was easier to develop an immunoassay to detect bacterial 5-enolpyruvylshikimate-phosphate synthase encoding resistance to the herbicide glyphosate that could distinguish it from the susceptible plant enzyme [12]. Even then, a triple sandwich technique was required.

Even when transgenics are discovered by such procedures or 'kits', there is no information as to source. Thus, regulatory authorities might wish to consider simple, common recognition sequences for detecting transgenic or other organisms in the market place. It is depressing to contemplate how much is being invested in detecting transgenic DNA in commodities (and no one has died from eating it) against how little is being expended towards detecting commonly occurring 'natural' bacterial contamination and mycotoxins in foodstuffs (which have negatively affected many lives). We propose a new recognition system, which, if imposed, would require far less expenditure in detecting DNA, allowing resources to be used for detecting truly dangerous contaminants in foodstuffs.

Barcodes and biobarcodes

The barcode system was developed to identify individuals rapidly, whether articles in a store, automobiles or test tubes. A simple barcode reader first identifies that there is a barcode to be read, and in what orientation, by seeing a pattern of bars that is otherwise too rare to be expected randomly. It does so by recognizing a universal sequence of spaced bars in an orientation-dependent manner that appear as universal recognition sites on either end of the barcode (Fig. 1a). In between the universal recognition sites is a variable series of bars of differing width and spacing. This assigned variable region is read by the reader into a computer and the database then identifies the object as being someone's car, candy bar or blood sample.

An analogous DNA-encoded system with universal recognition sites binding a variable region is proposed akin to the barcode system. This biobarcode™ system would enable material to be assayed in one PCR-sequencing run using universal primers that identify all 'biobarcoded' biological materials. In this system, open codes are to be designed and assigned by a single repository. They begin and end with the common universal 'barcode analogous' orientation-dependent nucleotide sequences that are recognized by the pair of universal recognition PCR primers, which, under PCR conditions, duplicate the whole intervening sequence (Fig. 1b). The assigned code is transformed into the target organism, as part of a transgenic construct, or it can be transformed directly into an otherwise non-transgenic organism.

The universal 'bar code recognition' nonsense (non-coding) nucleotide sequence is designed to be long enough that a few mutational changes will still allow it to be recognized by a PCR primer set. The initial universal recognition code is followed by a designed nonsense (non-coding) nucleotide sequence that is long enough to allow tens of millions of different such sequences to be generated, and again allow for some mutational changes. Neither the initial common recognition sequence, nor the particular individual strain sequence, should even vaguely resemble nonsense sequences reported in any sequence database. Frame-shift mutations should not render any part of the barcode sequence as an open reading frame long enough to allow significant polypeptides to be made. Stop codons are thus inserted into the assigned sequences in all reading frames to prevent a frame-shift mutation from becoming a genetic coding sequence.

Some people might not care whether there is any biobarcode in an organism, just whether their propriety DNA is found in it. Such people could have a special sequence pair assigned to them that would follow the universal recognition sequence in a biobarcode. Thus, they could use a primer that recognizes this sequence, and start the PCR reaction from there. Others wishing to know whose and what DNA is in the same organism or sample can probe with the universal primer pair.

A considerable amount of computational power with appropriate algorithms is needed to generate the common universal code sequence and the following variable individual strain sequence. The algorithms used to generate the sequences are being designed to exclude sequences that could self anneal, preventing the taq polymerase from amplifying the biobarcoded DNA. The biobarcode can be inserted in tandem with the genes of choice for transformation, or it can be co-transformed with the gene of choice in cases where the organism is transformed with a 'sense' transformation. In other cases, an excisable selectable marker will be needed.

If a product contains more than one biobarcode (i.e. a foodstuff concocted from different transgenic crops bearing different transgenes), there will be more than one band on the PCR gel to be sequenced. The sequenced bands are then compared to the public database. This is a positive method, with results from all biobarcode-labeled species, versus the guessing at what transgenes might be present, as occurs using the present technologies.

Regulation

It is envisaged that a single body would assign the biobarcodes and maintain a public database listing the biobarcode sequences of organisms that have been released. It is clearly advantageous to industry, regulators and taxpayers that such a system be instituted because of the amount of resources saved and the protection provided. Because of the savings, it is in the public interest that such a system be universally instituted.

In Canada, for example, the insertion of a biobarcode would probably not come under regulatory scrutiny if biobarcodes are introduced into plants because they are not considered to be 'plants with novel traits' if they are non-protein producing. In Europe, any introduced sequence (except antisense) seems to come under regulatory scrutiny, but after due risk assessment it is hoped that a blanket approval could be obtained for biobarcodes that meet the specific criteria listed above (and possibly others).

Acknowledgements
Useful discussions with Eithan Rubin of the Weizmann Bioinformatics Group are gratefully acknowledged. This research is supported in part by the Levin Fund. J.G. is the Gilbert de Botton Professor of Plant Sciences.

References

1 Gressel, J. (2002) *Molecular Biology of Weed Control*, Taylor & Francis

2 Vurro, M. *et al.*, eds (2001) *Enhancing Biocontrol Agents and Handling Risks*, IOS Press, Amsterdam

3 Gressel, J. (2001) Potential failsafe mechanisms against the spread and introgression of transgenic hypervirulent biocontrol fungi. *Trends Biotechnol.* 19, 149–154

4 Eparvier, A. and Alabouvette, C. (1994) Use of ELISA and GUS transformed strains to study competition between pathogenic and nonpathogenic *Fusarium oxysporum* for root colonization. *Biocontrol Sci. Technol.* 4, 35–47

5 Cohen, B. *et al.* (2002) Transgenically enhanced expression of indole-3-acetic acid (IAA) confers hypervirulence to plant pathogens. *Phytopathology* 92, 590–596

6 Hintz, W.E. *et al.* (2001) Development of genetic markers for risk assessment of biological control agents. *Can. J. Plant Pathol.* 23, 13–18

7 Meyer, R. (1999) Development and application of DNA analytical methods for the detection of GMOs in food. *Food Control* 10, 391–399

8 Huebner, P. *et al.* (1999) Quantitative competitive PCR for the detection of genetically modified organisms in food. *Food Control* 10, 353–358

9 Wurz, A. *et al.* (1999) Quantitative analysis of genetically modified organisms (GMO) in processed food by PCR-based methods. *Food Control* 10, 385–389

10 Lipp, M. *et al.* (2001) Validation of a method based on polymerase chain reaction for the detection of genetically modified organisms in various processed foodstuffs. *Eur. Food Res. Technol.* 212, 497–504

11 Anklam, E. *et al.* (2002) Analytical methods for detection and determination of genetically modified organisms in agricultural crops and plant derived food products. *Eur. Food Res. Technol.* 214, 3–26

12 Rogan, G.J. *et al.* (1999) Immunodiagnostic methods for detection of 5-enolpyruvylshikimate-phosphate synthase in Roundup-Ready soybeans. *Food Control* 10, 407–414

Domains as functional building blocks of plant proteins

Bernard C-H. Lam and Eduardo Blumwald

Emerging evidence in eukaryotic systems suggests that many proteins of diverse cellular processes are made up of protein domains that are well defined in both sequence and structure. This article updates the identification of many 'classic' eukaryotic protein domains in various plant cellular processes, with particular emphasis on the non-catalytic categories. We discuss the importance of domains to plant-protein functions and cellular networking, and the emergence of plant-specific domains.

Bernard C-H. Lam
Dept of Molecular Biology, Massachusetts General Hospital, and Dept of Genetics, Harvard Medical School, Boston, MA 02114, USA.

Eduardo Blumwald
Dept of Pomology, University of California, One Shields Ave, Davis, CA 95616, USA.
e-mail: eblumwald@ucdavis.edu

The completion of the genome sequence of the model plant *Arabidopsis* [1] has allowed many crucial questions in plant science to be investigated much more quickly. The identification of genes responsible for key *Arabidopsis* mutants has been simplified and comparisons between *Arabidopsis* and fungal and animal genomes has allowed the identification of the genetic bases of the differences between plants and other eukaryotic organisms. At the biochemical level, the study of the changes of the complete expression patterns in response to developmental or environmental signals at the RNA and protein levels have been made possible by the development of DNA microarrays and mass-spectrometry techniques, respectively.

However, these advances facilitated by the *Arabidopsis* genome development are also generating many more questions to be studied. Notably, fewer than 10% of the predicted *Arabidopsis* genes have been studied experimentally in an extensive manner. Moreover, more than 25% of the genes cannot be classified according to function(s) by mere sequence comparison to proteins of known roles in other organisms [1]. Although functional-genomic approaches such as the large-scale generation of T-DNA insertional mutant lines have been initiated [2], complementary approaches will probably be needed until the function of every gene product of *Arabidopsis* and other model plants is known.

Interestingly, in addition to the holistic or 'top-down' approach through mutant identification and characterization, there is currently a renaissance of plant studies starting at the level of the gene product itself (a minimalist or 'bottom-up' approach). For example, many key proteins involved in processes such as membrane transport [3] and signal transduction [4] were first identified *in silico* by searching the databases for gene products with sequence similarity to their counterparts (which were well characterized in many cases) in other model organisms such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Caenorhabditis elegans* and various mammalian species. Function prediction by amino acid sequence analysis or 'homology search' has become a common starting point of experimental design of plant research in the post-genome era.

PII: S1360-1385(02)02337-3