THE IDENTIFICATION OF SUGARCANE GENES BY RANDOM SEQUENCING OF cDNA CLONES

DL CARSON[1], BI HUCKETT[1] AND FC BOTHA[2]

[1]*SASA Experiment Station, Private Bag X02, Mount Edgecombe, 4300*
[2]*Institute for Plant Biotechnology, University of Stellenbosch, Private Bag X01, Matieland, 7602*

Abstract

Very little is known about sugarcane gene sequences. In this study, the aim is to identify sugarcane genes expressed in the leaf roll and the mature stem. Random DNA sequencing of cDNA clones allows the identification of genes and the development of Expressed Sequence Tags (ESTs). A total of 250 clones randomly selected from a leaf roll cDNA library and 112 from a mature stem library have been sequenced. Homology searches with sequences located in international databases have indicated a broad diversity of genes encoding proteins such as metabolic enzymes, structural proteins and regulatory factors. Preliminary analyses have indicated that although genes common to both the leaf roll and mature stem have been detected, several stress-related genes have been identified in the mature stem. Further studies will determine whether ESTs may be used to identify differentially expressed genes in sugarcane tissues.

Introduction

Sugarcane is not a simple plant on a genetic level, being a complex polyploid with chromosome numbers ranging between 90 and 120. Genetic research into sugarcane has discovered numerous agronomically important phenotypic traits; however, prior to the initiation of this project, only ten sugarcane gene sequences were published on the international databases. The isolation and functional identification of specific sugarcane genes could have important consequences for genetic engineering and sugarcane breeding.

The last decade has seen a proliferation of amino acid sequence information deduced from cloned genes of known function. It is now frequently possible to infer the probable function of a newly isolated gene solely on the basis of nucleotide or deduced amino acid sequence homology to genes of known function (Pearson, 1991). This has led to large-scale sequencing projects of anonymous cDNA clones for a variety of both plant and animal species. These projects have indicated that partial cDNA sequences, or Expressed Sequence Tags (ESTs), can be used successfully to identify putative clones for a wide range of gene products. In January 1998 the National Centre for Biotechnology Information (NCBI) database of Expressed Sequence Tags (dbEST) listed over 1 400 000 entries from 101 different organisms.

This paper reports on the initiation of an EST database for sugarcane, and presents the analysis of 250 anonymous cDNA clones from a cDNA library prepared from the leaf roll (meristematic region) of the commercial sugarcane variety NCo376.

Materials and methods

mRNA was isolated from total RNA extracted from the leaf roll of field-grown plants of the commercial variety NCo376. First-strand cDNA synthesis was performed using 1 µg of mRNA and the reaction catalysed by the RNase H⁻ M-Mulv (Moloney-Murine Leukemia Virus) reverse transcriptase enzyme (Stratagene). Second strand synthesis was performed directly following first-strand synthesis using standard protocols. Double-stranded cDNA was blunt-end ligated to an annealed amplification adaptor set (Jepson *et al.*, 1991). After ligation, cDNA was size fractionated through a Quick-Spin, Linkers 6 column (Boehringer Mannheim). Ligated, size fractionated cDNA was PCR amplified according to the procedure described in Jepson *et al.*, (1991). Amplified cDNA was restriction digested with EcoRI, cloned into the EcoRI site of the λZAP II cloning vector, and packaged according to the manufacturer's instructions (Stratagene).

Individual recombinant phages were randomly selected from plated out aliquots of the leaf roll cDNA library and stored in SM buffer (100 mM NaCl, 8 mM $MgSO_4.7H_2O$, 20 mM Tris-HCl pH 7,5, 0,01% gelatin) at 4°C. Template cDNA for sequencing was prepared in one of two ways. Recombinant phagemids were excised from individual phages using the ExAssist helper phage system according to the manufacturer's instructions. Phagemid DNA was isolated using a Rapid Plasmid Isolation Protocol (Holmes and Quigly, 1981). Alternatively, templates for DNA sequencing were prepared by specific PCR amplification of cDNA inserts directly from individual phage suspensions in SM buffer, using the M13 Reverse and T7 primers. Amplified inserts were purified using QIAquick spin columns (Qiagen) prior to sequencing.

Both phagemids and amplified cDNA inserts were sequenced by dye terminator cycle sequencing using the AmpliTaq DNA polymerase, FS ready reaction kit (Perkin Elmer). The M13 reverse primer was used to generate partial sequences for all cDNAs. Sequence analysis was performed on an ABI Prism 310 Genetic Analyzer.

EST sequences were compared to the nonredundant protein databases by using the BLASTX (Altschul *et al.*, 1990) e-mail server provided by the National Centre for Biotechnology Information, NCBI, (blast@ncbi.nlm.nih.gov). Sequences showing a PAM120 similarity score of over 80

were considered homologous proteins for the clones. The EST was identified as the protein showing the highest score among the candidate proteins.

Results and Discussion

A total of 250 clones have been randomly selected from the leaf roll cDNA library and sequenced. Table 1 indicates selected examples of sequenced cDNA clones identified by database homology searches. All identified clones have been deposited in the NCBI database of ESTs, dbEST, and complete results may be found under the Accession numbers AA080579-AA080668; AA269151-AA269177; AA269289-AA269300; AA525639-AA525698; AA577629-AA577669.

Among the 250 clones analysed, 36% showed a similarity score above 80 against the NCBI protein databases, indicating significant sequence homologies between the leaf roll cDNA clones and the database entries. Of the remaining 64% of the ESTs, 36% appear to be new genes not only in sugarcane but in all organisms, while the other 28% did not show significant homology to known genes in the database (i.e. similarity scores below 80). Of the 250 clones analysed, 26% showed significant amino acid sequence homology to previously identified plant genes such as enolase and sucrose synthase (Table 1). Of the 65 identified plant clones, 37% showed homology to monocotyledonous plant species. As expected, these proteins gave a high similarity score. High similarity scores were also observed for some dicotyledenous species but this was usually for fundamental proteins such as ribosomal proteins, where large conserved regions of the genome between species would be expected. One hundred and twenty-four ESTs (50%) showed homology to previously identified genes from species other than higher plants, for example, pyruvate dehydrogenase from *Synechocystis* spp. (Table 1). The targeted species were widely distributed from bacteria to humans.

Table 1. Selected examples of sugarcane ESTs from a leaf roll cDNA library.

EST#	Putative identification	ID	Similar	Overlap	Score	Organism
AA080655	Vacuolar H^+- ATPase	55	56	60	280	*Gossypium hirsutum*
AA080634	Sucrose synthase	55	61	68	301	*Triticum aestivum*
AA080585	Translation elongation factor eEF-1	28	33	36	151	*Arabidopsis thaliana*
AA080586	Enolase	97	101	105	511	*Zea mays*
AA080646	Glutathione S-transferase	31	43	82	138	*Arabidopsis thaliana*
AA269154	Pectin methylesterase	49	63	90	261	*Melandrium album*
AA269174	Pyruvate dehydrogenase E1 component	61	70	93	313	*Synechocystis* spp.
AA269290	Disease resistance protein RPM1	18	26	49	86	*Arabidopsis thaliana*
AA577653	Triose-phosphate isomerase	60	60	63	302	*Zea mays*
AA577666	Nucleolar histone deacetylase HD2	22	23	24	111	*Zea mays*

The EST# is the accession number assigned by dbEST. The numbers in the columns designated ID, Similar and Overlap refer to the number of identical (ID) or Similar (Similar) amino acids in a region of a particular length (Overlap). The column designated Organism refers to the source of the protein that exhibits homology to the sugarcane EST.

All identified leaf roll EST clones were categorised according to biochemical and metabolic functions (Figure 1).

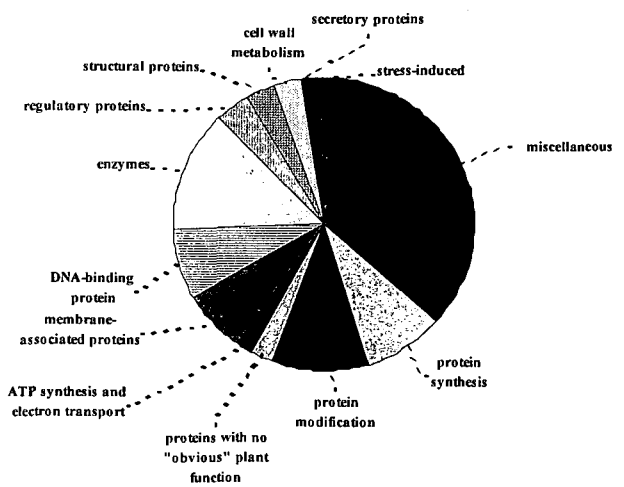

Figure 1. Classification of identified leaf roll cDNA clones according to biochemical and metabolic function.

The leaf roll cDNA clones exhibited homology to a broad diversity of genes, encoding metabolic enzymes and proteins associated with a wide range of cellular processes. The largest number of clones (37%) was found to encode many proteins as yet uncharacterised. Of the remaining genes that were identified, about 13,3% were metabolic enzymes, 10,4% involved in protein modification and 8% each involved in protein synthesis and DNA-binding. A strong representation of genes coding for proteins associated with protein modification, protein synthesis and DNA-binding is indicative of the vigorous growth state of the leaf roll tissue used for cDNA library construction. Fewer genes were involved in regulation (4%), structural proteins (2,9%), cell wall metabolism (2,9%), secretory proteins (2,9%) and ATP synthesis and electron transport (1,7%). One clone was identified as being stress-induced. A small percentage of clones (2,3%) were identified as being homologous to animal proteins with no obvious functions in plants.

It is anticipated that clones identified from the leaf roll cDNA library reflect the regulation of gene expression related to

differentiation, growth condition and environmental stress. Work is in progress to sequence 250 clones from a mature stem cDNA library. To date, 112 clones have been sequenced and preliminary analyses have indicated that although genes common to both the leaf roll and mature stem have been detected, several stress-related genes have been identified in the mature stem (data not shown). Comparison and analysis of the clones obtained from these two cDNA libraries might provide insights into the genes or proteins that play important roles in gene regulation during maturation of the sugarcane plant.

Acknowledgements

The authors thank Avril Harvey for performing the sequencing reactions.

REFERENCES

Altschul, SF, Gish, W, Miller, W, Myers, EW and Lipman, D (1990). Basic local alignment search tool. *J Mol Biol* 215: 403-410.

Holmes, DS and Quigley, M (1981). A rapid boiling method for the preparation of bacterial plasmids. *Anal Biochem* 114: 193-197.

Jepson, I, Bray, J, Jenkins, G, Schuch, W and Edwards, K (1991). A rapid procedure for the construction of PCR cDNA libraries from small amounts of plant tissue. *Plant Molecular Biology Reporter* 9(2): 131-138.

Pearson, WR (1991). Identifying distantly related protein sequences. *Curr Opinion Struct Biol* 1: 321-326.