Genetics and Molecular Biology, 24 (1-4), 291-296 (2001)

Sugarcane phytocystatins: Identification, classification and expression pattern analysis

*Emerson Moreira Reis[1] and Rogério Margis[1,2]**

Abstract

The cystatins are tightly binding, but reversible, inhibitors of cysteine proteinases, which constitute a superfamily of evolutionary related proteins. They have been subdivided into three families: the cystatin family which contain two disulfide bonds, the stefin family which lack disulfide bonds, and the kininogen family composed of large glycoproteins containing three repeats similar to those found in the cystatin family. Members of the cystatin superfamily occurring in plants are currently known as phytocystatins, defined as proteins lacking disulfide bonds but possessing a conserved N-terminal amino acid sequence (L-A-R-[FY]-A-[VI]-X(3)-N). We have used the protein sequence deduced from seven phytocystatins (from the *Arabidopsis thaliana* genome project) and from the SUgarCane EST project (SUCEST) database to identify 25 possible sugarcane phytocystatins. Phylogenetic analysis has allowed us to cluster these phytocystatins into four distinct groups: (i) those with a characteristic N-terminal consensus, (ii) those with the same consensus plus a long C-terminal extension; (iii) those that lack the consensus but contain the highly conserved QxVxG motif found in all members of the superfamily and (iv) those that lack both the consensus and the QxVxG motif.

INTRODUCTION

The cystatin superfamily is composed of tightly-binding, reversibly competitive cysteine proteinases inhibitors (Nicklin and Barret, 1984) called cystatins, which exhibit similarities in their amino acid sequences and functions (Barret *et al*., 1986). The cystatins have been classified into three distinct families: the cystatin family consisting of groups of small proteins with two disulfide bonds, the stefin family of small proteins (~12 kDa) lacking disulfide bonds, and the kininogen family of large glycoproteins (60-120 kDa) containing three repeats similar to those found the cystatin family (Barret *et al*., 1986). The activity of cystatins has been explained by the presence of three contact points with the target proteinase. Studies with human cystatin C indicate that the first point of interaction involves the N-terminal glycyl-containing segment, which is substrate-like and accommodates the S1 and S2 enzyme sub-sites (Abrahamson *et al*., 1987). The other points of interaction occur on hairpin loops formed by antiparallel sheets, the first loop containing the QxVxV segment (a highly conserved sequence present in the cystatin superfamily) which stabilizes the complex by providing an area of extended contact with the proteinase binding site (Bode *et al*., 1988). The second loop corresponds to a tryptophan-containing segment that also interacts with the proteinase binding site (Nycander and Bjork, 1990).

Plant cystatins present structural peculiarities, genomic arrangements and intrinsic diversity which, when compared to animal cystatins, justify the creation of a new family, the phytocystatins (Kondo *et al*., 1991; Margis *et al*., 1998), within the cystatin superfamily. This is supported by the presence of an N-terminal helix consensus sequence ([LVI]-[AGT]-[RKE]-[FY]-[AS]-[VI]-X-[EDQV]-[HYFQ]-N) only found in phytocystatins, as well as by phylogenetic tree analysis which clusters them in a single branch statistically distinct from the other families and also their diversity in regard to the position and organization of their introns (Margis *et al*., 1998).

It is widely assumed that phytocystatins perform a defensive role in plants because of their effects on exogenous proteinases such as those produced by insects and nematodes (Liang *et al*., 1991; Zhao *et al*., 1996). The involvement of phytocystatins in plant defense is support by the observation that transgenic plants expressing them show enhanced resistance to phytophagous organisms (Leple *et al*., 1995; Urwin *et al*., 1995), as well as the fact that they can be induced in tomato and soybean by wounding or methyl-jasmonate (Bolter, 1993; Botella *et al*, 1996). In seeds, phytocystatin mRNAs show an expression pattern similar to that of major seed storage proteins (Abe *et al*., 1987, Abe *et al*., 1992), and this, together with the their ability to inhibit endogenous cysteine proteinases, has led to the proposal that they are involved in the regulation of

[1]*Laboratório de Genética Molecular Vegetal, Departamento de Genética, Instituto de Biologia, CCS, Universidade Federal do Rio de Janeiro, Ilha do Fundão, 21944-970 Rio de Janeiro, RJ, Brazil.*
[2]*Departamento de Bioquímica, Instituto de Química, CCMN, Universidade Federal do Rio de Janeiro, Ilha do Fundão, 21944-970 Rio de Janeiro, RJ, Brazil.*
Send correspondence to Rogério Margis. E-mail: margisr@ufrj.br.

protein turnover during seed development. More recently, the involvement of phytocystatins as modulators in programmed cell death has been reported (Solomon *et al*., 1999).

In the study presented in this paper, as part of the SUgarCane EST (SUCEST) program to characterize sugarcane expressed sequence tags (ESTs), we have identified sugarcane phytocystatins and described their structural features, localization and putative functions.

MATERIAL AND METHODS

Sequence data, alignment and phylogenetic analysis

A T-Blast-n search (Altschul *et al*., 1997) was performed using *Arabidopsis thaliana* phytocystatins as bait sequences against the full SUCEST cDNA data bank. The acronyms and accession numbers of these sequences being Ath1 (GB: T00752), Ath2 (GB: H71431), Ath3 (GB: BAB09081), Ath4 (GB: CAA03929), Ath5 (GB: AAD15406), Ath6 (GB: AAG51028) and Ath7 (GB: BAB11533) from *A. thaliana,* and OS 02 (PIR: S13027) from *Oryza sativa* (rice).

The Multiple Alignment Program (MAP) was used to compute a multiple global alignment of sequences using the pairwise method, its alignment algorithm computing the best overlapping alignment between two sequences without penalizing terminal gaps. This program, designed in a space-efficient manner, allows long sequences to be aligned (Huang, 1994) and was the method used to align the proteins deduced from the SUCEST clusters and the *A. thaliana* phytocystatins.

Phylogenetic analysis was performed using the Molecular Evolutionary Genetics Analysis (MEGA) software, version 2.0 (Kumar *et al*., 2000), the pair-wise deletion option being used for the amino acid gaps in the sugarcane cysteine protease multiple alignment, and the phylogenetic tree obtained from neighbor-joining analysis using the p-distance method. For tree construction the confidence levels assigned at various nodes were determined after 5000 replications using the Interior Branch test (Sitnikova *et al*., 1995). The expression pattern tree was constructed using the unweighted pair group method with averages (UPGMA) number of site differences and pairwise deletion also present in the MEGA software.

Description of SUCEST cDNA libraries

All sugarcane sequences used in this work corresponding to sequenced reads and cluster consensus (Phrap option in data base) were obtained from the SUCEST project (http://sucest.lad.dcc.unicamp.br/en/) and derived from cDNA libraries specific for different sugarcane tissues, organs or growth conditions as follows: apical meristem from mature (AM1) and (AM2) immature plants; 1 cm (FL1) and 5 cm (FL3) flower base; 50 cm (FL4), 20 cm (FL5) and10 cm (FL8) flower stem; lateral buds (LB1 and LB2); large (LR1) and small (LR2) leaf-root insert libraries; etiolated leaves (LV1); the grouped data of two non-redundant libraries (NRn); the grouped data of three root libraries (RTn); the grouped data of three leaf-root transition zone libraries (RZn); stem bark (SB1); grouped data of two seed libraries of different insert sizes (SDn); grouped data of two stem libraries from the first and fourth internodes (STn); libraries derived from *calli* submitted to a 4-37 °C temperature change and three (CL3), four (CL4), six (CL6) and eight (CL7) hours of a light/dark cycle; plants infected with the bacteria *Gluconacetobacter diazotroficans* (AD1) and *Herbaspirillum rubrisubalbicans* (HR1).

RESULTS AND DISCUSSION

Identification of SUCEST phytocystatins

The sequences of seven *Arabidopsis thaliana* phytocystatins and one typical phytocystatin-II from *Oryza sativa* were used to find homologous SUCEST clusters with a real potential for presenting cysteine proteinase activity. The overall analysis, using the T-Blast-n algorithm with an e^{-10} cut-off value, allowed the identification of 25 different consensus clusters that separated into four distinct groups (Table I, Figures 1 and 2).

Table I - Frequency of sugarcane phytocystatin related reads and clusters in sugarcane cDNA libraries.

Acronym	Library	AD1	HR1	AM1	AM2	CLn	FL1	FL3	FL4	FL5	FL8	LB2	LRn	LV1	RTn	RZn	SB1	SDn	ST1	ST3	Total
	Cluster																				
Group I																					
Sc01	SCEQRT1025E03.g	1	1	-	-	-	-	-	-	-	-	2	2	1	1	1	-	12	-	-	21
Sc02	SCMCRT2103C04.g	1	1	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	3
Sc03	SCJFRT2053E01.g	-	-	1	-	-	-	-	-	-	-	-	-	-	6	-	-	-	-	-	7
Sc04	SCCCRZ2001G09.g	-	1	-	-	1	-	-	-	-	-	-	1	1	-	1	-	-	1	-	6
Sc05	SCEQAM1041A08.g	-	-	2	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	3
Sc06	SCRLAD1044F09.g	1	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	Total	3	3	3	-	1	2	-	-	-	-	2	3	2	8	2	-	12	1	-	42

Table I (cont).

Acronym / Cluster	Library	AD1	HR1	AM1	AM2	CLn	FL1	FL3	FL4	FL5	FL8	LB2	LRn	LV1	RTn	RZn	SB1	SDn	STn1	STn3	Total
Group II																					
Sc07	SCMCRZ3067B01.g	-	-	-	-	-	-	1	-	-	-	-	-	-	-	2	1	1	-	-	5
Sc08	SCBGAM1090E08.g	-	-	2	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	3
Sc09	SCEPLR1051C09.g	-	-	-	-	-	1	2	-	-	-	-	2	-	-	2	-	-	-	1	8
Sc10	SCJLRT2050B04.g	-	-	-	-	-	2	1	2	-	-	-	-	-	1	-	-	2	-	-	8
Sc11	SCMCRT2086B07.g	-	1	2	-	-	1	-	2	-	-	1	-	-	1	1	-	-	-	1	10
Sc12	SCCCST1003G05.g	-	1	-	-	-	-	1	-	-	-	1	-	-	-	-	-	-	1	-	4
Sc13	SCMCST1054D03.g	-	1	-	-	-	2	-	-	-	1	-	-	-	-	1	1	-	1	2	9
	Total	-	3	4	-	-	6	5	4	1	1	2	2	-	2	6	2	3	2	4	47
Group III																					
Sc14	SCVPAM1059C07.g	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
Sc15	SCAGCL6015G10.g	1	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
Sc16	SCCCRZ3002G10.g	-	-	-	-	2	2	-	-	-	-	-	-	-	-	1	-	-	-	-	5
Sc17	SCEZRT2018F03.g	-	-	-	-	-	-	-	-	-	-	1	-	-	3	-	-	-	-	-	4
Sc18	SCMCRT2087B02.g	-	-	-	-	2	-	-	-	-	-	-	-	-	1	-	-	1	-	-	4
Sc19	SCCCCL4014F01.g	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
	Total	1	1	1	-	7	2	-	-	-	-	1	-	-	4	1	-	1	-	-	19
Group IV																					
Sc20	SCEZRT2019D04.g	1	-	-	1	-	-	-	-	-	-	-	-	-	1	3	-	4	-	3	13
Sc21	SCCCLR1076E04.g	-	-	1	-	-	-	-	-	-	1	-	1	-	2	1	1	1	2	-	10
Sc22	SCSFST3081F04.g	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1
Sc23	SCEPRZ3047A01.g	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	1
Sc24	SCPRLB2030C11.g	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	2
Sc25	SCSGST3121C03.g	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	2
	Total	2	-	1	1	-	-	-	-	-	1	2	1	-	3	5	1	5	2	5	29

*The libraries were are as follows: apical meristem from mature (AM1) and (AM2) immature plants; 1cm (FL1) and 5 cm (FL3) flower base; 50 cm (FL4), 20 cm (FL5) and10 cm (FL8) flower stem; lateral buds (LB1 and LB2); large and small leaf-root insert (LRn); etiolated leaves (LV1); grouped data of three root libraries (RTn); grouped data of three leaf-root transition zone libraries (RZn); stem bark (SB1); grouped data of two seed libraries of different insert sizes (SDn); grouped data of two stem libraries from the first and fourth internodes (STn); grouped data of libraries derived from *calli* submitted to a 4-37 °C temperature change and three to eight hours of a light/dark cycle (CLn); plants infected with the bacteria *Gluconsugarcane diazotroficans* (AD1) and *Herbaspirillum rubrisubalbicans* (HR1).

Relationship of sugarcane phytocystatins

Phylogenetic tree of sugarcane phytocystatin clusters shows that they can be classified into four groups: I) those with the characteristic L-A-R-[FY]-A-[VI]-X(3)-N consensus forming part of the N-terminal helix, II) those with the same consensus plus a long C-terminal extension, III) those that lack the consensus but contain the highly conserved QxVxG motif found in all members of the superfamily and IV) those without the conserved QxVxG motif (Figures 1 and 2).

Group I, which includes the rice phytocystatin-II and Ath1 and Ath4, has the typical phytocystatin structural organization of the helix and at turns I and II (Margis *et al*., 1998). The absence of the L-A-R-[FY]-A-[VI]-X(3)-N consensus normally present in the helix has important constraints for the maintenance of the three-dimensional organization of these inhibitors because it may introduce conformational changes in members of this group, including possible change of their target proteinases. The C-terminal extension (of as yet unknown function) present in the members of this group has also been found in other phytocystatins (Margis *et al*., 1998). Members of group II present the same features but with an extended C-terminal end.

Group III is composed of Ath2, Ath3, Ath5 and sugarcane clusters SC14 to SC19, the principal characteristic of this group being the presence of an extended sequence just after the region corresponding to the N-terminal helix. The genomic analysis of several phytocystatins has demonstrated that this region corresponds to a splicing junction

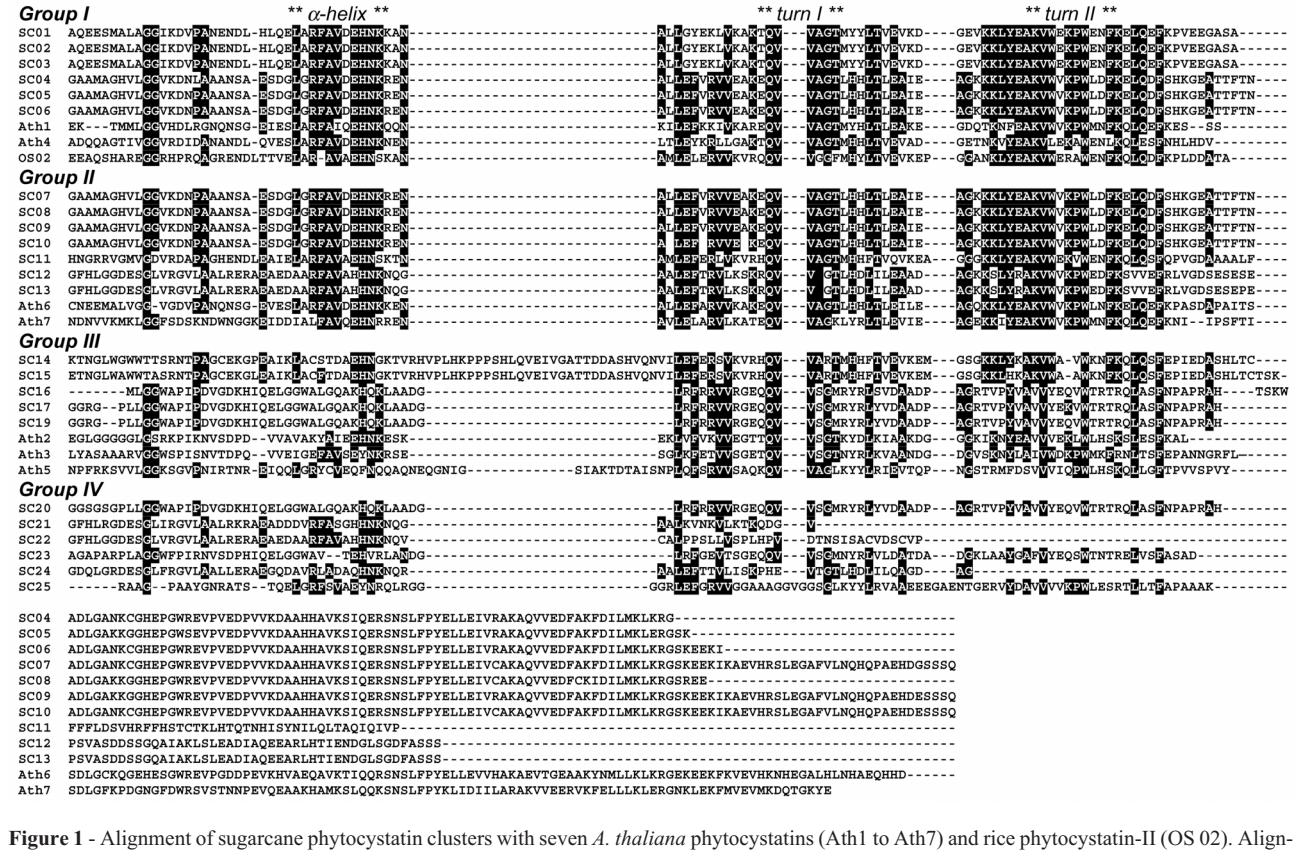

Figure 1 - Alignment of sugarcane phytocystatin clusters with seven *A. thaliana* phytocystatins (Ath1 to Ath7) and rice phytocystatin-II (OS 02). Alignment was performed using the MAP algorithm. Phytocystatins were divided into four major groups according to different sequence characteristics. Regions corresponding to putative phytocystatin N-terminal alpha-helix and to beta-turns I and II are emphasized.

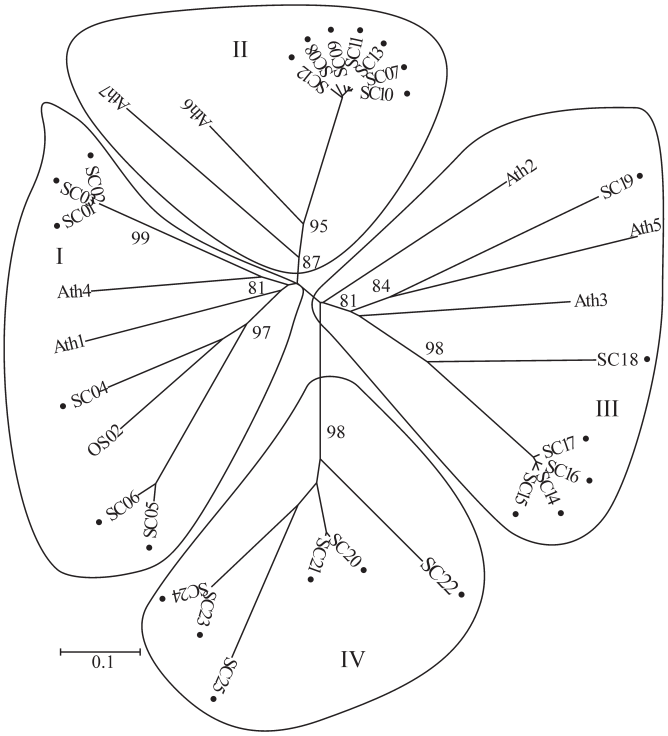

Figure 2 - Phytocystatin phylogenetic tree. The tree was constructed using MEGA software with the following parameters: p-distance, neighbor-joining, pair-wise deletion and internal branch test with 5000 replications. The analysis was performed based on the MAP alignment of 25 sugarcane clusters, seven *A. thaliana* phytocystatins (Ath1 to Ath7) and rice phytocystatin-II (OS 02). Four major groups with different phytocystatin forms were identified. The values obtained from the internal branch replication (80%) are indicated in the main branch points.

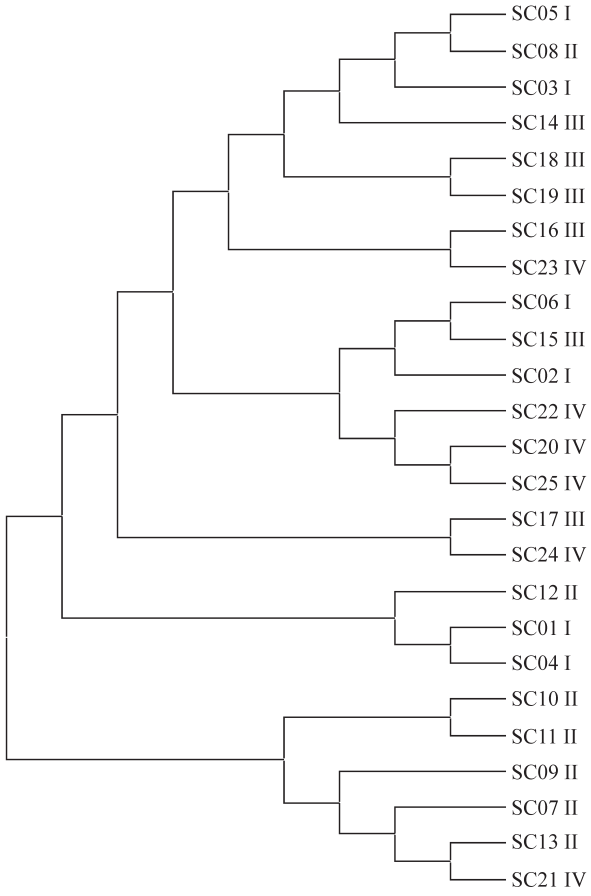

Figure 3 - Dendrogram of sugarcane phytocystatin expression pattern. The tree was constructed using the unweighted pair group method with averages (UPGMA) number of site differences and pairwise deletion. MEGA software was used for the analysis.

for phytocystins that present introns (Margis *et al*., 1998). The presence of this extra-sequence may be due to differential mRNA processing of SC14 and SC15. The real implications of this sequence on the inhibitory activity of phytocystatins is still unknown because no comparative biochemical studies have been carried out either with Aths1, Ath5 or any other phytocystatins presenting this structural peculiarity.

Group-IV, except cluster SC20, is composed of clusters not presenting the classical QxVxG signature/motif at turn-I. The detection of these clusters in several SUCEST libraries indicate that they are transcribed, but the conservation of the biological function of these clusters may be questioned since this motif is strictly related to the binding of these inhibitors to the active site of the cysteine proteinases and is indispensable for inhibition.

Analysis of sugarcane phytocystatin expression pattern

A preliminary analysis of sugarcane cysteine proteinases expression patterns was made by direct correlation of the reading frequency of each cluster in the different SUCEST cDNA libraries. The expression pattern analysis of the 25 phytocystatins (Table I and Figure 3) revealed that the best represented clusters were SC01, SC09, SC10, SC11, SC13, SC20 and SC21, with eight to 21 reads present in almost all libraries. Of the four sugarcane phytocystatin groups, group-III was the least represented, with only 19 reads distributed in nine different libraries. Curiously, group-III clusters have a high presence in *calli*, with all clusters presenting at least one read in this library.

Phytocystatins were originally identified in rice and maize seeds (Kondo *et al*., 1991). Their direct implication with storage protein mobilization and oryzain regulation has been described. Here, we have found clusters of all four phytocystatin groups in the seed library, but cluster SC01 was definitely the most abundant and probably the sugarcane equivalent to rice cystatins I and II, which are involved in the control of protein mobilization during germination.

In order to identify any correlation between the expression pattern of the clusters in the libraries and the group to which they belong, a dendrogram was produced from the data presented in Table I. A matrix of libraries versus the presence/absence of clusters was generated and this matrix was used to produce the dendrogram of correlation of sugarcane expression pattern (Figure 3). A clear clustering of expression pattern was observed with group-II members (SC07, SC09, SC10, SC11 and SC13) and with three group-IV clusters (SC20, SC222 and SC25).

In the present work we have identified 25 different phytocystatin related clusters in the SUCEST data bank. The analysis of these SUCEST sequences and their expression pattern show that sugarcane possesses phytocystatins with new structural features and with a differentiated expression pattern not only related to seeds but also to other organs and tissues.

RESUMO

As cistatinas são inibidores competitivos e reversíveis de proteinases cisteínicas, compondo uma superfamília de proteínas evolutivamente relacionadas. Esta superfamília está dividida em três famílias: a primeira, das estefinas, é composta de proteínas destituídas de pontes de enxofre; a segunda, das cistatinas, agrupa proteínas que possuem pontes de sulfeto; a terceira família, dos kininogênios, se caracteriza por ser formada por glicoproteínas de alto peso molecular com três domínios repetidos similares aos da família das cistatinas. As fitocistatinas constituem uma família de cistatinas que possui representantes exclusivamente em plantas. Elas são proteínas que não possuem pontes de enxofre e contém uma seqüência conservada L-A-R-[FY]-A-[VI]-X3-N na região N-terminal. Neste trabalho nos usamos seqüências derivadas de sete fitocistatinas, identificadas no projeto genoma de *Arabidopsis*, para identificar possíveis membros da superfamília das cistatinas presentes no banco de dados de cana-de-açúcar do SUCEST. Foram identificadas vinte e cinco possíveis fitocistatinas de cana-de-açúcar. Uma análise filogenética

permitiu agrupar as fitocistatinas em quatro grupos: (i) contendo o consenso N-terminal ; (ii) contendo o consenso N-terminal e mais uma extensão C-terminal; (iii) destituído do consenso N-terminal e (iv) destituído do consenso N-terminal e do motivo conservado QxVxG encontrado em todos os membros da superfamília.

REFERENCES

Abe, K., Emori, Y., Kondo, H., Suzuki, K. and **Arai, S.** (1987). Molecular cloning of a cysteine proteinase inhibitor of rice. *J.Biol. Chem*. *262*: 16793-16797.

Abe, M., Abe, K., Kuroda, M. and **Arai, S.** (1992). Corn kernel cysteine proteinase inhibitor as a novel cystatin superfamily member of plant origin. *Eur. J. Biochem*. *209*: 933-937.

Abrahamson, M., Ritonja, A., Brown. M.A., Grubb. A., Machleidt, W. and **Barret, A.J.** (1987). Identification of the probable inhibitory reactive sites of the cysteine proteinase inhibitors human cystatin C and chicken cystatin. *J. Biol. Chem*. *262l*: 9688-9694.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and **Lipman, D.J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. *25*: 3389-3402.

Barret, A.J., Rawlings, N.D., Cavies, M.E. Machleidt, W., Salvesen, G. and **Turk, V.** (1986). Cysteine proteinase inhibitors of the cystatin super family. In: *Proteinase inhibitors* (Barrett, A.J. and Salvesen, G., eds.) pp. 515-569. Elsevier. Amsterdam.

Belenghi, B., Delledonne, M., Menachem, E. and **Levine, A.** (1999). Involvement of cysteine proteases and protease inhibitor genes in the regulation of programmed cell death in plants. *Plant Cell*. *11*: 431- 443.

Bode, W., Engh, R., Musil, D., Thiele, U., Huber, R., Karshikov, A. Brzin, J., Kos, J. and **Turk, V.** (1988). A X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. *EMBO J*. *7*: 2593-2599.

Bolter, C. (1993). Methyl jasmonate induces papain inhibitors in tomato leaves. *Plant Physiol*. *103*: 1347-1353.

Botella, M.A., Xu, Y., Prabha, T.N., Zhao, Y., Narasimhan, M.L., Wilson K.A., Nielsen, S.S., Bressan, R.A. and **Hasegawa, P.M.** (1996). Differential expression of soybean cysteine proteinase inhibitor genes during development and in response to wounding and methyl jasmonate. *Plant Physiol*. *112*: 1201-1210.

Huang, X. (1994). On Global Sequence Alignment. Computer Applications in the *Biosciences 10*: 227-235.

Kondo, H., Abe, K., Emori, Y. and **Arai, S.** (1991). Gene organization of oryzacystatin-II, a new cystatin super family member of plant origin, is closely related to that of oryzacystatin-I but different from those of the animal cystatins. *FEBS Lett*. *278*: 87-90.

Kumar, S., Tamura, K., Jacobsen, I. and **Nei, M.** (2000). MEGA2: Molecular Evolutionary Genetics Analysis, version 2.0. Pennsylvania and Arizona State Universities, University Park, Pennsylvania and Tempe, Arizona.

Leple, J.C., Bonade-Bottino, M., Augustin, S., Pilate, G., Le Tan, V.D., Delplanque, A., Cornu, D. and **Jonanin, L.** (1995). Toxicity of *Chrysomila tremulae* (Coleoptera: Chrysomelidae) of transgenie poplars expressing a cysteine proteinase inhibitor. *Mol. Breed 1*: 319-328.

Liang, C., Brookhart G., Feng, G.H., Reeck, G.R. and **Kramer, K.J.** (1991). Inhibition of digestive proteinases of stored grain coleoptera by oryzacystatin, a cysteine proteinase inhibitor from rice seed, *FEBS Lett*. *278*: 139-142.

Margis, R., Reis, E.M. and **Villeret, V.** (1998). Strutural and phylogenetic relationships among plant and animal cystatins. *Arch. Biochem. Biophys*. *359*: 24-30.

Nicklin, M.J.H. and **Barrett, A.J.** (1984). Inhibition of cysteine proteinases and dipeptidyl peptidase I by egg-white cystatin. *Biochem. J. 223*: 245-253.

Nycander, M. and **Bjork, I.** (1990). Evidence by chemical modification that tryptophan-104 of the cysteine-proteinase inhibitor chicken cystatin is located in or near the proteinase-binding site. *Biochem. J. 271*: 281-284.

Sitnikova, T., Rzhetsky, A. and **Nei, M.** (1995). Interior-branch and bootstrap tests of phylogenetic trees. *Mol. Biol. Evolution 12*: 319-333.

Solomon, M., Urwin, P., Atkinson, H.J., Waller, D.A. and **McPherson, M.J.** (1995). Engineered oryzacystatin-I expressed in transgenic hairy roots confers resistance to *Globodera pallida*. *Plant J. 8*: 121-131.

Zhao, Y., Botella, M.A., Subramanian, L., Niu, X., Nielsen, S.S., Bressan, R.A. and **Hasegawa, P.M.** (1996). Two wound-inducible soybean cysteine proteinase inhibitors have greater insect digestive proteinase inhibitory activities than a constitutive homologue. *Plant Physiol*. *111*: 1299-1306.