Genetics and Molecular Biology, 24 (1-4), 85-92 (2001)

Molecular chaperone genes in the sugarcane expressed sequence database (SUCEST)

*Júlio C. Borges[1,2], Maria C. Peroto[1,2] and Carlos H.I. Ramos[1,2]**

Abstract

Some newly synthesized proteins require the assistance of molecular chaperones for their correct folding. Chaperones are also involved in the dissolution of protein aggregates making their study significant for both biotechnology and medicine and the identification of chaperones and stress-related protein sequences in different organisms is an important task. We used bioinformatic tools to investigate the information generated by the Sugarcane Expressed Sequence Tag (SUCEST) genome project in order to identify and annotate molecular chaperones. We considered that the SUCEST sequences belonged to this category of proteins when their E-values were lower than 1.0e-05. Our annotation shows that 4,164 of the 5' expressed sequence tag (EST) sequences were homologous to molecular chaperones, nearly 1.8% of all the 5' ESTs sequenced during the SUCEST project. About 43% of the chaperones which we found were Hsp70 chaperones and its co-chaperones, 10% were Hsp90 chaperones and 13% were peptidyl-prolyl *cis, trans* isomerase. Based on the annotation results we predicted 156 different chaperone gene subclasses in the sugarcane genome. Taken together, our results indicate that genes which encode chaperones were diverse and abundantly expressed in sugarcane cells, which emphasizes their biological importance.

INTRODUCTION

Large-scale cDNA sequencing has become a fast and satisfactory alternative to the complete sequencing of the entire genome of an organism. By generating and sequencing cDNAs researchers can obtain the coded information of the genes responsible for different mRNAs, and these partial cDNA sequences (known as expressed sequence tags or ESTs) can quickly provide information on the genetic profile of an organism by using bioinformatics to mine sequences in public databases and annotation the original ESTs by homology. This approach was used by us to identify chaperones in the Brazilian sugarcane expressed sequence tag (SUCEST) database (http://sucest.lad.ic.unicamp.br/en/).

Chaperones are proteins capable of assisting the folding of other proteins *in vivo* (Gething and Sambrook, 1992; Ellis and Hartl, 1996). Conditions of stress, such as temperature variation, can induce protein aggregation inside cells and, consequently, many chaperones were first identified as heat shock proteins (Hsps). However, in addition to induced expression under situations of stress, they are also constitutively expressed. Chaperones are able to bind to exposed hydrophobic residues in unfolded proteins (usually buried in the native state) and, according to Flynn *et al.* (1989), such a mechanism prevents the incorrect folding of the protein. Chaperones are also involved in a broad variety of functions other than protein folding (Ellis and Hartl, 1996; Lindquist and Craig, 1988; Fink, 1999). More than twenty chaperone families have been described according to their molecular weight and biochemical function (Cowan and Lewis, 1999), the most important families of chaperones and stress-related proteins being summarized in the next few paragraphs.

The chaperonin-like family, which assists in late-stage protein folding (Ellis and Hartl, 1996) is the most studied of the chaperone families, and is comprised of the Hsp60 (GroEL-like) family and its co-chaperone Hsp10 (GroES-like). The cytoplasmic form of the GroEL-like family is known as TriC (TCP-1 ring complex) which is not induced by heat-shock (Hendrick and Hartl, 1993; Kim *et al.*, 1994). The Hsp70 (DnaK-like) family and its co-chaperones GrpE and Hsp40 (DnaJ-like) form a system that binds to nascent peptide chains helping their folding (Fink, 1999). The Hsp90 family is highly expressed in response to conditions of stress and participates in the stabilization of several receptors, *e.g.* the steroid receptor (Buchner, 1999). The Hsp100 family is involved in protein disaggregation, acting together with the Hsp70 system (Schirmer *et al.*, 1996). Little is known about the function of the Small Hsp (sHsp) set of proteins, a family which is not highly conserved and whose cytoplasmic forms represent 1% of the total proteins present in a thermally stressed cell (Jakob and Buchner, 1994; Boston *et al.*, 1996). Peptidyl-prolyl *cis, trans* isomerase (PPIase) catalyses the *cis, trans* isomerization of Proline peptide bonds in proteins. It is divided into two classes, cyclophilins and FK506-binding proteins or immunophilins (Fischer *et al.*, 1984; Schreiber, 1991). Proline isomerization in the peptide is slow and it limits the

[1]*Centro de Biologia Molecular Estrutural, Laboratório Nacional de Luz Síncrotron. Caixa Postal 6192, 13084-971 Campinas, SP, Brazil.*
[2]*Dept. de Bioquímica, Instituto de Biologia, UNICAMP. Campinas, SP, Brazil.*
**Send correspondence to Carlos H.I. Ramos. E-mail: cramos@lnls.br.*

protein folding kinetic rate. Peptidyl-disulfide isomerase (PDIase, involved in disulfide bond rearrangement catalysis), the calnexin/calreticulin family, heat shock transcription factors (HSTF, involved in the regulation of chaperones expression), Bip, auxilin, Sec*B* and the trigger factor also belong to the chaperone protein category (Fink, 1999). PPIase, PDIase, Bip, and calnexin/calreticulin are involved in protein folding inside the endoplasmic reticulum.

The goal of this work was to identify molecular chaperone genes in sugarcane using homology comparison between sequences in the SUCEST database and the sequences available at other public databases. Nearly 1.8% of all the 5' ESTs sequenced during the SUCEST project are related to molecular chaperones. From this total, about 43% are related to the Hsp70 chaperone and its co-chaperones, 10% to the Hsp90 chaperone and 13% to peptidyl-prolyl *cis, trans* isomerase. If we consider that the sum of the 5' ESTs sequenced during the SUCEST project is in good accordance with the mRNA expression profile of sugarcane cells, it seems that the genes encoding chaperone proteins have significant expression in this organism. We have predicted 156 molecular chaperone gene subclasses in the sugarcane genome from our annotation, indicating the overall diversity of these proteins. We have also found that the genes related to cytoplasmic chaperone proteins have both higher expression and diversity than genes related to chaperones occurring in other cellular compartments.

MATERIAL AND METHODS

Data on how the EST libraries were prepared, sequencing, clustering and other important data are available at the SUCEST home page (http://sucest.lad.ic.unicamp.br/en/). The mining strategy used for molecular chaperone annotation is summarized in Figure 1. First, we chose translated amino acid sequences of specific mRNAs coding for chaperones and stress-related proteins deposited at public databases. The choice of a particular prototype protein sequence followed a hierarchical order of preference, *i.e.* the most preferred sequences were of plant origin (*e.g. Arabidopsis thaliana*, *Zea mays,* etc.), followed by eukaryotic origin and, finally, prokaryotic origin. After choosing the prototype protein sequences, we compared them with the first level cluster consensus generated from the SUCEST database using the fragment assembly program (Phrap) and the basic local alignment search tool tBLASTn program (Altschul *et al.*, 1997). This strategy compared amino acid sequences from public databases with nucleotide sequences in the SUCEST database, and the EST cluster which had a statistically significant chaperone match (*i.e.* an E-value lower than 1.0e-05) was considered to be a chaperone. To increase the accuracy of the prediction, the matches found in the first round of mining went through a second round of mining (or sometimes even more rounds) which allowed us to find better matches and/or new clusters (Figure 1). The

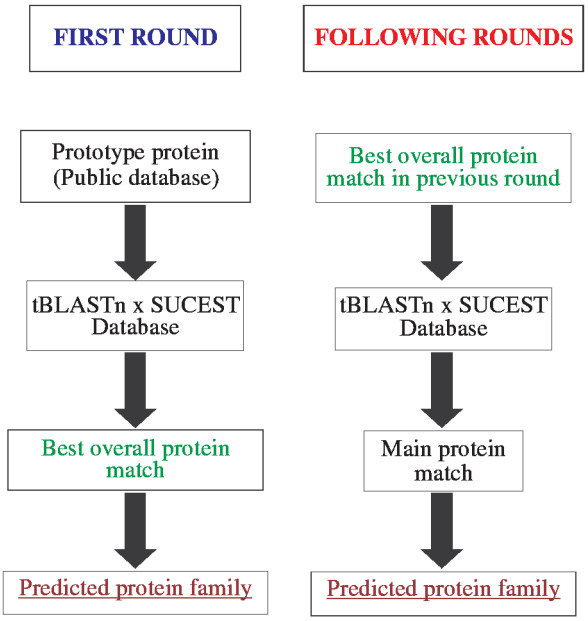

Figure 1 - General scheme for sequence mining and annotation strategy. Translated amino acid sequences of specifics mRNAs were chosen from chaperones and stress-related proteins deposited at public databases. The prototype sequences were compared with the first level cluster consensus generated from the sugarcane expressed sequence tag (SUCEST) database using the basic local alignment search tool tBLASTn program (Altschul *et al.*, 1997). The tBLASTn program was used to compare amino acid sequences from public databases with nucleotide sequences generated from the SUCEST database and the SUCEST cluster sequence that had a significant molecular chaperone match was considered to be a chaperone. To increase the accuracy of the prediction, the matches found in the first round of mining went through a second round (or more) of mining, allowing the location of better matches and/or new clusters.

homology between protein sequences was analyzed by LALIGN alignment (http://www.ch.embnet.org/software/LALIGN_form.html). In order to know the proportion of sequences associated with chaperone genes, the number of those sequences was calculated and expressed as a percentage of all SUCEST sequenced 5' ESTs (Table I) and all SUCEST sequences annotated as chaperone genes (Tables I and II).

RESULTS

The list of the molecular chaperone families studied in this work, their cellular location and the mining results are shown in Table I. Results in Table I are expressed as the number of First Level Clusters and Annotated Sequences for each chaperone class. Table II shows the number of potential annotated chaperone genes, or subclasses, for each class of this protein category. The results in Table II are expressed as the number of annotated subclasses and the average identity between the classes.

The SUCEST database contains around 230,000 sequenced 5 ESTs (http://sucest.lad.ic.unicamp.br/en/; Telles *et al.*, 2001) and we annotated 4,164 as belonging to the molecular chaperone category (Table I). This value

Table I - The main chaperone families and classes, their cellular location and the mining results.

Molecular chaperones			Mining results		
Family	Class	Intracellular location	First level clusters	Annotated 5'EST sequences	Annotated 5'ESTs sequences (%) [Family only]
Clp	Hsp100/ClpB	Cytoplasm, mitochondria	14	33	3.3
	ClpA/C	Chloroplast	15	66	
	ClpX	-	11	37	
Hsp90	Hsp82	Cytoplasm	46	291	9.7
	Grp94	ER	18	77	
	cp-Hsp82	Chloroplast	7	17	
	Hsp90-like	-	7	20	
Hsp70	Hsp/Hsc70	Cytoplasm and nucleus	51	556	20.7
	Mt-Hsp70	Mitochondria	9	193	
	cp-Hsp70	chloroplast	10	24	
	Bip/Grp78	ER	13	56	
	Hsp110 homologue	-	11	34	
Hsp70 co-chaperones	Hsp40/DnaJ	Cytoplasm, mitochondria, ER	174	886	21.9
	Auxilin	-	6	16	
	GrpE	Mitochondria	5	10	
Hsp60/ Chaperonin-like	Hsp60	Mitochondria	5	31	3.1
	Cpn60	Chloroplast	16	98	
Chaperonin-like co-chaperone	Hsp10	Mitochondrion	5	38	1.6
	Cpn21	Chloroplast	6	29	
TriC	TCP-1	Cytoplasm	62	310	7.5
Small Hsp	SmHsp class I	Cytoplasm	30	115	4.7
	SmHsp class II	Cytoplasm	9	40	
	SmHsp class III	Mitochondria	8	28	
	SmHsp class IV	Chloroplast	6	12	
Calnexin/ calreticulin	-	ER	44	211	5.1
PPIase	Cyclophilin	Cytoplasm, ER, mitochondria, chloroplast	65	452	13.3
	FK506-BP	Cytoplasm, ER, mitochondria	30	100	
PDIase	-	ER	46	225	5.4
HSTF	-	-	29	98	2.4
Others	-	-	20	55	1.3
Total	**777**	**4,164 (1.8%)**	**777**	**4.164 (18%)**	**100 (4.164)**

Chaperones represented 1.8% of all 5' expressed sequence tags (ESTs) in the sugarcane expressed sequence tag (SUCEST) database. The table shows the main chaperone families and classes, their cellular location and the mining results. The strategy used for the annotation of molecular chaperones is summarized in Figure 1. The results are expressed as the number of first level clusters and annotated sequences for each class of chaperones. We found that the number of SUCEST sequences associated with molecular chaperones (4,164) represented 1.8% of all SUCEST sequenced 5' ESTs. The proportion of annotated sequences in each family is shown as a percentage of the sequences associated with molecular chaperones (4,164). ER = endoplasmic reticulum.

represents almost 1.8% of all the 5' ESTs sequenced during the SUCEST project. The proportion of annotated sequences in each family was also calculated and expressed as a percentage of the sequences associated with chaperones (4,164). The relative magnitudes being as follows: >20%: the Hsp70 family (21%) and its co-chaperones (22%); 10-20%: the PPIase (13.5%) and Hsp90 (10%) families; 5-10%: the TriC (7.5%), calnexin/calreticuim (5%) and PDIase (5%); <5%: the remaining families representing about 16% (Table I).

Table II shows that the predicted number of chaperone gene subclasses in the sugarcane genome is 156.

Table II - The number of potential annotated chaperone genes.

Molecular chaperones		Potential annotated genes		
Family	Class	Number of subclasses	Average identity (%)	Contribution (%) [Family only]
Clp	Hsp100/ClpB	4	60	5.8
	ClpA/C	4	64	
	ClpX	1	-	
Hsp90	Hsp82	2	87	3.8
	Grp94	1	-	
	CpHsp82	1	-	
	Hsp90-like	2	45	
Hsp70	Hsp/Hsc70	3	88	7.7
	mt-Hsp70	2	77	
	Bip/Grp78	2	41	
	cp-Hsp70	1	-	
	Hsp110 homologue	4	35	
Hsp70 co-chaperones	Hsp40/DnaJ	39	-	28.2
	Auxilin	2	88	
	GrpE	3	32	
Hsp60/Chaperonin-like	Hsp60	1	-	1.9
	Cpn60	2	46	
Chaperonin-like co-chaperone	Hsp10	2	23	2.6
	Cpn21	2	73	
TriC	TCP-1	9	30	5.8
Small Hsp	SmHsp class I	5	58	7.0
	SmHsp class II	3	54	
	SmHsp class III	2	75	
	SmHsp class IV	1	-	
Calnexin/ calreticulin	-	5	-	3.2
PPIase	Cyclophilin	15	-	17.3
	FK506-BP	12	-	
PDIase	-	7	18	4.5
HSTF	-	14	-	9.0
Others	-	5	-	3.2
Total	**-**	**156**	**-**	**100 (156)**

Predicted molecular chaperone genes are diverse in the sugarcane genome. The table shows the amount of potential annotated chaperone genes, or subclasses, for each class of this protein category. The strategy used for the annotation of molecular chaperones is summarized in Figure 1. The results are expressed as the number of subclasses, and the average identity between them. The average identity was calculated using the LALIGN software program. From the annotation results we predicted a total of 156 different chaperone gene subclasses in the sugarcane genome. The proportion of potential chaperone genes in each family is also shown as a percentage of all the predicted chaperone subclasses (156).

The proportion of potential chaperone genes in each family was also calculated and expressed as a percentage of all the predicted chaperone subclasses (*i.e.* 156). The relative magnitudes being as follows: >20%: the Hsp70 co-chaperones (28%); 10-20%: PPIase (17.5%); 5-10%: HSTF, Hsp70, small Hsp, Clp and TriC families (although <10% individually, in total these families made up about 35.5% of the total); <5%: the remaining families representing about 19% (Table II).

DISCUSSION

Genes encoding molecular chaperones are diverse and abundantly expressed.

The information available at SUCEST was searched for the annotation of molecular chaperone genes by homology comparison with public databases using bioinformatic tools. The results were used to predict gene expression

profiles based on the gross number of annotated 5' EST sequences. The degree of homology between public and SUCEST database sequences were used to predict gene function. We are aware that functional analysis based only on homology does not always generate clear-cut conclusions, hence great care was taken to annotate each sequence and further studies are underway to characterize the function of some of the proteins described here.

We found that a significant number of the sequenced 5' ESTs in the SUCEST database belonged to the chaperone gene category. Knowing that the number of ESTs in a reliable library can be related to the number of mRNAs produced by the cell and therefore to the gene expression profile of a cell, we concluded that chaperone genes have a remarkably high level of expression in sugarcane cells. The SUCEST database contains ESTs from cells of a variety of tissues which have been exposed to various environmental conditions, so we are confident that the data generated provides a reliable picture of the expression level and diversity of sugarcane genes. All these considerations have been discussed in detail by Bonaldo *et al*. (1996).

A high expression level of chaperone genes is in accordance with our current knowledge of this category of protein. The proportion of proteins in a cell that needs the help of chaperones to fold is unknown, but several estimates (Ellis and Hartl, 1996; Lorimer, 1996) state that this value is around 10%, perhaps rising to 30% under conditions of stress. A considerable number of chaperones have to be present in a cell to help proteins to fold. Our results suggest that about 1.8% of expressed sugarcane genes belong to the chaperone category, and our identification of a high level of chaperone expression confirms the important role played by chaperones in the cell.

The diversity of the genes that encode chaperones is shown by our prediction of 156 potential chaperone subclasses present in the sugarcane genome. The number of genes which we annotated as chaperones could not easily be related to the number of first level clusters. Most of these clusters were not full-length and/or had just minor variations. Further annotation of the final SUCEST clustering and analysis of gene expression by tissues (in progress) will be necessary for investigation into the connection between the number of clusters and the number of potential genes. Clusters that present small variation are more interesting since they can give information about allele variation in the sugarcane genome.

The number of chaperone subclasses and the degree of identity among them in each family are discussed below. Our results did not seem to follow a general trend, probably because they were dependent on protein function. A general observation, which seems to be independent of the chaperone family concerned, is that the gene classes of the chaperones found in the cytoplasm seem to have had both higher expression and higher diversity than the classes xpressed in other cellular compartments.

Chaperone related genes encoding the Hsp70 family and its co-chaperones are the most expressed and diverse

The importance of the Hsp70 family and its co-chaperones is confirmed by our annotation results which showed a high number of 5' EST sequences related to these proteins. We estimate that their genes alone are responsible for almost 1% of the entire gene expression in sugarcane cells.

The Hsp70 protein has ATPase activity and is involved in helping the folding of nascent proteins by recognizing and binding to unfolded peptides (Hartl, 1996). Since most of the nascent proteins that need chaperones to fold are targets for Hsp70 there are probably a high number of Hsp70 molecules in cells. It is known that other plants have multiple genes coding for the cytoplasmic form of Hsp70, such genes sharing about 90% identity with each other (Boston *et al*., 1996). This supports our observation that sugarcane cytoplasmic Hsp70 had the highest expression in the family and displays 3 related potential gene subclasses. The high expression and diversity of the cytoplasm form of Hsp70 is probably due to the fact that most of the nascent proteins are produced (and therefore have to fold) in this cellular compartment. Other important functions of Hsp70 also help to explain its high gene expression and diversity, for example the cooperation of Hsp70 with most of the other chaperone families. In this type of cooperation the proteins folded by Hsp70 are transferred to the Hsp60/Hsp10 system (or TriC, when in the cytoplasm), Hsp70 also cooperates with Hsp90 for receptor stabilization and Hsp70 can recognize aggregated proteins delivered by the Hsp100 family (Martin and Hartl, 1997; Fink, 1999). Hsp70 is also important for protein translocation, its isoform present at the lumen (named as either Bip or Kar2) is responsible for protein folding in this compartment (Hammond and Helenius, 1994).

The co-chaperone Hsp40 presents the nascent proteins to Hsp70 and can possibly works as a chaperone by itself. The expression gene profile which we found for this chaperone is a confirmation of its importance to the cell. These results are also interesting because they confirm previous results showing that DnaJ-like proteins from other organisms are highly diverse and are abundantly expressed in the cell (Cheethan and Caplan, 1998). This family is characterized by the presence of a highly conserved helical N-terminal domain, called the J-domain, which binds to Hsp70 stimulating its ATPase activity (Cheethan and Caplan, 1998; Fink, 1999). It is currently unknown if there is a specific interaction between Hsp70 and Hsp40 and whether or not the J-domain is involved (Cheetham and Caplan, 1998). We annotated a large number of 5' EST sequences as homologous to the J-domain, indicating that this domain probably has a widespread distribution. Hsp40 is composed of at least three more domains, not always conserved (Cheethan and Caplan, 1998; Fink, 1999), that are probably

responsible for the high number of related genes and the large amount of diversity found in this family. Of special interest are our results which show that 3 out of the 39 Hsp40 related genes are responsible for half of the 5' EST sequences annotated for this chaperone. These very abundant chaperone genes also present high identity and, due to their abundance, it is possible that the proteins coded for these genes are responsible for the majority of the Hsp40 functions in sugarcane.

GrpE had an unexpectedly lower expression level in comparison to Hsp70 and Hsp40. This co-chaperone stimulates Hsp70 ATPase activity and is thought to interact with Hsp70 in a molecular ratio of 3:1 (Schönfeld and Behlke, 1998). We can only conjecture the reasons for such different expression, one possibility being restricted compartimentalization, GrpE having been described in prokaryotes and in mitochondria whereas Hsp70 and Hsp40 are present in all cellular compartments (Netzer and Hartl, 1998). Another possibility is restricted functions, one reported function of GrpE is to bind to Hsp70 and stimulate some of its functions. Yet another possibility is transient function, in which GrpE participates in the folding reaction only during the release of ADP, whereas Hsp70 and Hsp40 remain bound to the folding protein during the whole process (Hartl, 1996).

Chaperonin-like related genes

Our results indicate that 5' EST sequences corresponding to Hsp60 are present at twice the frequency of Hsp10 sequences. This data is consistent with the molecular ratio of 14:7 found when these chaperones interact to produce a barrel-like shape structure where protein folding takes place. Lorimer (1996) states that, in *Escherichia coli*, the average percentage protein mass of GroEL and GroES is 1.4 and 0.2%, respectively, values which differ somewhat from what we observed in sugarcane. Our data indicated that sugarcane GroEL-like and GroES-like related genes are responsible for 0.06 and 0.03% (respectively) of total gene expression in sugarcane. Great care has to be taken in the analysis of such data because it is hard to compare gene and protein expression due to the different lifetimes of mRNA and protein. Divergence in data can also occur because of the complexity of trying to compare eukaryotes with prokaryotes. Lorimer (1996) calculated that these two chaperones (GroEL and GroES) are able to help the folding of only 2-5% of all proteins in *E. coli*. Even though nearly 90% of the proteins in the cell do not need chaperones to fold, there is no doubt that these proteins are indispensable to cells because blocking or inhibiting them produces a lethal phenotype in plants (Boston *et al*., 1996).

We found that the genes belonging to the TCP-1 subfamily were highly expressed in certain compartments of sugarcane cells. The TCP-1 protein was found in the cytoplasm and we showed that its expression was almost twice as much in the cytoplasm as that of its mitochondrial (Hsp60) and chloroplast (cpn60) counterparts. Although the other members of this family have been widely studied, little is known about the function of TCP-1 and its protein targets (Kim *et al*., 1994). TCP-1 may be involved in cytoskeletal protein folding and it forms a high weight molecular hetero-oligomeric complex with several subunits (Fink, 1999). Eight different forms of this protein (sharing about 30% identity) have been described by Kubota *et al*. (1995), which is in good agreement with our results for sugarcane where we assigned gene expression and predicted potential genes for all of the known subunits. The high level of TCP-1 expression which we observed in sugarcane is an indication that this protein plays an important role in the cell, probably performing chaperonin-like functions in the cytoplasm.

Hsp90 related genes are abundantly expressed

In our study we found that the sugarcane genes encoding Hsp90 are highly expressed, as would be expected from its known abundance in eukaryotes, this protein being one of the most abundant cytoplasmic proteins, representing up to 2% of the total cellular soluble protein even in the absence of stress (Boston *et al*., 1996; Buchner, 1999). Caplan (1999) has suggested that the high expression of Hsp90 is probably due to its involvement in a broad variety of functions. Our results showing that Hsp90 related genes were highly expressed in sugarcane tissues but do not have high diversity are of great importance. We were able to show that only two potential genes (belonging to the cytoplasmic Hsp82 class and extremely similar) were responsible for most of the gene expression of this family.

The function of Hsp90 is not well established but it can modulate the functions of other proteins, participating in the transcription process by binding to transcription factors, binding to glucocorticoid receptors and interacting with several other chaperone families such as Hsp70, SmHsp and PPIases (Caplan, 1999). Our data indicating a high gene expression level in this family is a positive indication that Hsp90 chaperones are involved in several important cellular functions.

Genes of chaperones belonging to the Hsp100 family

Two highly conserved ATP binding site domains in each terminus (Schirmer *et al*., 1996) characterize the Hsp100, ClpA, B, C and D proteins. The Hsp100 family was only modestly expressed in sugarcane as verified by the small number of their 5' EST sequences which we annotated. Our results showed the presence of eight potential genes with high identity for ClpA, B and C, but none for ClpD. The high identity found was probably due to the conserved ATP binding site domains present in this family. According to Schirmer *et al.* (1996), the other subdivision

(containing the ClpM, N, X and Y proteins) of the Hsp100 family is smaller, having only one nucleotide binding site. We were able to find only one gene related to the ClpX protein and no related sequences were found belonging to the remaining three chaperones. The failure to identify the remaining Hsp100 genes is probably related to the low number of Hsp100 family chaperones described in eukaryotes. The data available about this family refer mainly to Hsp104 in *Saccharomyces cerevisiae,* cytoplasmic Hsp100 in *Arabidopsis thaliana,* ClpX in *Homo sapiens* and SKD3 in mouse and rat (Parsell *et al*., 1994; Perier *et al*., 1995; Boston *et al*., 1996). Besides their function in protein folding, the ClpA, C and X proteins participate in proteolytic processes, which probably explains why they have the highest levels of expression in the family.

SmHsp genes

The gene expression profile which we found suggests that the SmHsp chaperone family is not abundant in sugarcane. However, the number of predicted subclasses which we annotated for this family indicates its high diversity. These proteins are considered to be diverse and abundant in plants where they are highly induced by heat-shock (Waters *et al*., 1996) and are divided into five classes. We found classes I and II in the cytoplasm of sugarcane cells and our results show that these two classes are responsible for almost 80% of the SmHsp gene expression. Classes III and IV (located in mitochondria and chloroplasts) respond for the remaining 20%, since class V (located in the endoplasmic reticulum) was not identified in our annotation effort. Although sequence homology between proteins of different classes is low, our results showed high identity between SmHsp genes inside each class. For instance, we annotated five potential class I proteins which shared almost 60% identity. There is little information about this family and the sequence homology between its components. However, we were able to show that conservation between the four classes is low but that the proteins belonging to a specific class are conserved.

Proline and disulfide isomerase genes

The enzymes implicated in proline and disulfide isomerization were responsible for almost one-fifth of the sequenced sugarcane 5' EST which we analyzed. These results show that these isomerases are very important to sugarcane cells and agree with the fact that these enzymes are present in several different organisms and cellular compartments. Proline and disulfide isomerases are involved in the kinetics of the protein folding process. During folding it is important that the hydrophobic regions of protein remain exposed for only a short time, preventing protein aggregation within cells. This is possibly the reason why a large number of isomerase molecules are probably needed in the cell. It is also probable that the widespread cellular distribution of these enzymes is the reason for the high genetic diversity of these enzymes which we found in sugarcane. In our study we found 27 related genes for PPIase, almost equally distributed between cyclophilin and the FK506-binding protein. We found that the average identity of the annotated cyclophilins was surprisingly high, with 15 related genes sharing nearly 50% identity.

FINAL CONSIDERATIONS

The study of chaperones and stress-related proteins is important for our understanding of the protein folding process, and to achieve advances in biotechnology and medicine. In general, plants are submitted to daily (and in some cases extreme) temperature variation, making the identification and characterization of their chaperones and stress-related proteins of special interest. The use of plants as hosts for heterologous protein expression for the production of large amounts of proteins and other metabolites has recently become important, and the study of plant chaperones can assist the enhancement of heterologous protein expression in many organisms, including sugarcane. The abundance and diversity of chaperones and stress-related proteins in sugarcane which we have revealed in this paper confirms the biological importance of this category of proteins and suggests that chaperones are probably involved in several different cell functions.

The data presented in this paper shows that chaperone and stress-related protein genes are abundantly expressed and have ample diversity in sugarcane, and this statement can almost certainly be applied to other, if not all, organisms. The protein classes present in the cytoplasm have both higher expression and diversity than the classes present in other cellular compartments. The most abundant and diverse molecular chaperone was Hsp70 and its co-chaperones, followed by PPIase and Hsp90. Small Hsp chaperones were present in sugarcane and the cytoplasmic homologues showed the highest identity. Our results also indicate that other chaperone families were present in sugarcane *e.g.* the Hsp100 family, heat shock transcription factors, calnexin/calreticulin, trigger-factor and the stress-induced protein sti1-p, although no Sec*B* homologous were found.

RESUMO

Algumas proteínas ao serem sintetizadas necessitam do auxílio de chaperones moleculares para seu correto enovelamento. Chaperones também estão envolvidas na dissolução de agregados protéicos, fazendo com que seu estudo seja de relevância biotecnológica e médica. Portanto, a identificação de seqüências de chaperones moleculares é uma tarefa importante. Nós usamos ferramentas de bioinformática para procurar informações geradas pelo sugarcane EST Genome Project (SUCEST) a fim de identificar e anotar chaperones e proteínas relacionas ao

estresse. As seqüências do SUCEST eram anotadas como pertencentes a uma categoria de proteínas se o E-value encontrado fosse menor que 1,0e-05. Nossas anotações mostram que 4.164 seqüências 5' EST são homólogas a chaperones moleculares, aproximadamente 1,8% de todos os 5' EST que foram seqüenciados pelo SUCEST. Deste total, cerca de 44% pertence à família Hsp70 e suas co-chaperones, 10% à família Hsp90 e 13% à peptidil *cis, trans* isomerase. Do resultado da anotação, nós predizemos a existência de 156 diferentes subclasses de genes relacionados a chaperones no genoma da cana-de-açúcar. Juntos, nossos resultados apontam que os genes que codificam para estas proteínas são diversos e abundantemente expressos, o que enfatiza sua importância biológica.

ACKNOWLEDGMENTS

We thank FAPESP and MCT/CNPq for financial support. We are in great debt to all the SUCEST participants for generating the database. We thank T. M. Santos and A. N. Capella for helpful discussions and G. Furtado for comments about the text.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, 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.

Bonaldo, M.F., Lennon, G. and **Soares, M.B.** (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res. 6*: 791-806.

Boston R.S., Viitanen P.V. and **Vierling E.** (1996). Molecular chaperones and protein folding in plants. *Plant. Mol. Biol. 32*: 191-222.

Buchner, **J.** (1999). Hsp90 & Co. - a holding for folding. *Trends Biochem. Sci. 24*: 136-141.

Caplan, A.J. (1999). Hsp90s secrets unfold: new insights from structural and functional studies. *Trends Cell Biol. 9*: 262-268.

Cheetham, M.E. and **Caplan, A.J.** (1998). Structure, function and evolution of DnaJ: conservation and adaptation of chaperone function. *Cell Stress Chaperones 3*: 28-36.

Cowan, N.J. and **Lewis, S.A.** (1999). A chaperone with a hydrophilic surface. *Nat. Struct. Biol. 6*: 990-991.

Ellis, R.J. and **Hartl, F.U.** (1996). Protein folding in the cell: competing models of chaperonin function. *FASEB J. 10*: 20-26.

Fink, A.L. (1999). Chaperone-mediated protein folding. *Physiol. Rev. 79*: 425-449.

Fischer, G., Bang, H. and **Mech, C.** (1984). Determination of enzymatic catalysis for the cis-trans-isomerization of peptide binding in proline-containing peptides. *Biomed. Biochim. Acta 43*: 1101-1111.

Flynn, G.C., Chappell, T.G. and **Rothman, J.E.** (1989). Peptide binding and release by proteins implicated as catalysts of protein assembly. *Science 245*: 385-390.

Gething, M.J. and **Sambrook J.** (1992). Protein folding in the cell. *Nature 355*: 33-45.

Hammond, C. and **Helenius, A.** (1994). Folding of VSG G protein: sequential interaction with Bip and calnexin. *Science 266*: 456-458.

Hartl, F.U. (1996). Molecular chaperones in cellular protein folding. *Nature 381*: 571-580.

Hendrick, J.P. and **Hartl, F.U.** (1993). Molecular chaperone functions of heat-shock proteins. *Annu. Rev. Biochem*. *62*: 349-384.

Jakob, U. and **Buchner J.** (1994). Assisting spontaneity: The role of Hsp90 and small Hsps as molecular chaperones. *Trends Biochem. Sci. 19*: 205-211.

Kim, S., Willison, K.R. and **Horwich AL.** (1994). Cystosolic chaperonin subunits have a conserved ATPase domain but diverged polypeptide-binding domains. *Trends Biochem. Sci. 19*: 543-8.

Kubota, H., Hynes, G. and **Willison, K.** (1995).The chaperonin containing t-complex polypeptide 1 (TCP-1). Multisubunity machinery assisting in protein folding and assembly in the eukaryotic cytosol. *Eur. J. Biochem. 230*: 3-16.

Lindquist, S. and **Craig, E.A.** (1988). The heat-shock proteins. *Annu. Rev. Genet*. *22*: 631-677.

Lorimer, G.H. (1996). A quantitative assessment of the role of the chaperonin proteins in protein folding *in vivo*. *FASEB J.* 10: 5-9.

Martin, J. and **Hartl, F.U.** (1997). Chaperone-assisted protein folding. *Curr. Opin. Struct. Biol. 7*: 41-52.

Netzer, W.J. and **Hartl, F.U.** (1998). Protein folding in the cytosol: chaperonin-dependent and -independent mechanisms. *Trends Biochem. Sci. 23*: 68-73.

Parsell, D.A., Kowall, A.S. and **Lindquist, S.** (1994). *Saccharomyces cerevisiae* Hsp104 protein, purification and characterization of ATP-induced structural changes. *J. Biol. Chem. 269*: 4480-4487.

Perier, F., Radeke, C.M., Raab-Graham, K.F. and **Vanderberg, C.A.** (1995). Expression of a putative ATPase suppresses the growth defect of a yeast potassium transport mutant: identification of a mammalian member of the Clp/HSP104 family. *Gene 152*: 157-163.

Schirmer, E.C., Glover, J.R., Singer, M.A. and **Lindquist, S**. (1996). HSP100/Clp proteins: a common mechanism explains diverse functions. *Trends Biochem. Sci*. *21*: 289-296.

Schönfeld, H.J. and **Behlke, J.** (1998). Molecular chaperones and their interactions investigated by analytical ultracentrifugation and other methodologies. *Methods in Enzimol. 290*: 269-96.

Schreiber, S.L. (1991). Chemistry and biology of the immunophilins and their immunosuppressive ligands. *Science. 251*: 283-287.

Telles, G.P., Braga, M.D.V., Dias, Z., Lin, T.-L., Quitzau, J.A.A., da Silva, F.R. and **Meidanis, J.** (2001). Bioinformatics of the sugarcane EST project *Genetics and Molecular Biology 24* (1-4): 9-15.

Waters, E.R., Lee, G.J. and **Vierling, E.** (1996). Evolution, structure and function of the small heat shock proteins in plants. *J. Exp. Bot. 47*: 325-338.