Genetics and Molecular Biology, 24 (1-4), 131-140 (2001)

DNA repair-related genes in sugarcane expressed sequence tags (ESTs)

R.M.A. Costa, W.C. Lima, C.I.G. Vogel, C.M. Berra, D.D. Luche, R. Medina-Silva, R.S. Galhardo, C.F.M. Menck and V.R. Oliveira*

Abstract

There is much interest in the identification and characterization of genes involved in DNA repair because of their importance in the maintenance of the genome integrity. The high level of conservation of DNA repair genes means that these genetic elements may be used in phylogenetic studies as a source of information on the genetic origin and evolution of species. The mechanisms by which damaged DNA is repaired are well understood in bacteria, yeast and mammals, but much remains to be learned as regards plants. We identified genes involved in DNA repair mechanisms in sugarcane using a similarity search of the Brazilian Sugarcane Expressed Sequence Tag (SUCEST) database against known sequences deposited in other public databases (National Center of Biotechnology Information (NCBI) database and the Munich Information Center for Protein Sequences (MIPS) *Arabidopsis thaliana* database). This search revealed that most of the various proteins involved in DNA repair in sugarcane are similar to those found in other eukaryotes. However, we also identified certain intriguing features found only in plants, probably due to the independent evolution of this kingdom. The DNA repair mechanisms investigated include photoreactivation, base excision repair, nucleotide excision repair, mismatch repair, non-homologous end joining, homologous recombination repair and DNA lesion tolerance. We report the main differences found in the DNA repair machinery in plant cells as compared to other organisms. These differences point to potentially different strategies plants employ to deal with DNA damage, that deserve further investigation.

INTRODUCTION

Sugarcane, cultivated in tropical and subtropical areas, is one of the most important crops and there has been great interest in developing more resistant and efficient sugarcane cultivars by classical genetic studies. However, the application of conventional genetics and breeding techniques to this crop is difficult due to sugarcane's complex polyploid pattern of inheritance and almost exclusively vegetative reproduction. In order to obtain a considerable amount of genetic information, a Brazilian group has sequenced a large number of sugarcane cDNA expressed sequence tags (ESTs) from several plant tissues growing under different conditions. This project, the Sugarcane Expressed Sequence Tag (SUCEST) genome project, provides an extensive database of different EST libraries covering various tissues and conditions of growth.

The interest in the identification of genes involved in DNA repair is due to their importance in the maintenance of genomic integrity. Agents that may cause physical damage or modification to the genetic material continually threaten DNA. Such alterations include DNA strand breaks and base loss (resulting in abasic sites) and chemical modification of bases to form a miscoding or non-coding lesion. These events can disrupt essential cellular processes such as DNA transcription and replication and cause mutagenesis, aging and even cell death. During evolution efficient mechanisms that recognize and repair DNA lesions have been selected for. Several DNA repair mechanisms exist to deal with different kinds of lesions, *e.g.* double strand breaks are repaired by the homologous recombination pathway (Kanaar *et al.,* 1998) and small base modifications are removed by base excision repair (Memisoglu and Samson, 2000). It should not, however, be forgotten that the relationship between DNA damage and some DNA 'tolerance' mechanisms is also important in the creation of genetic diversity.

The studies of DNA repair mechanisms have focused largely on bacteria, yeast and human beings, and have shown a high level of conservation between these organisms. Although plants are continuously exposed to environmental damaging agents, due to their sessile life style, relatively little is known about DNA repair systems in these organisms. In *Arabidopsis* and tobacco plants, Ries *et al.* (2000) have reported that the exposure of plants to elevated levels of ultraviolet (UV) solar radiation increases the frequency of somatic homologous DNA rearrangements, which affect genome integrity. The accumulation of DNA lesions in seeds seems to be involved in the seed aging process and lower germination rates (Vonarx *et al.*, 1998; Costa *et al.*, 2001). Therefore, the identification of mechanisms that avoid the accumulation of DNA damage in plants could lead to agriculturally useful improvements in various crops, including sugarcane.

In this paper, we report the identification of genes involved in DNA repair mechanisms in sugarcane, using a

Depto. de Microbiologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP, Brazil.
Send correspondence to C.F.M. Menck. E-mail: cfmmenck@usp.br.

protein sequence similarity search of the SUCEST database. This search revealed that the majority of proteins involved in DNA repair in sugarcane are similar to those found in other eukaryotes. The protein sequences discovered by us represented mechanisms such as photoreactivation, base excision repair, nucleotide excision repair, mismatch repair, non-homologous end joining and DNA lesion tolerance. When we compared the expressed gene sequences of monocotyledonous sugarcane with the identified DNA repair genes of the dicotyledonous *Arabidopsis thaliana* we found high conservation of the DNA repair mechanisms in these two phylogenetically distant species. We also observed intriguing characteristics so far found only in plants: the absence of important known components of some of the DNA repair mechanisms. These features seem to be specific to plants and suggest a distinct evolutionary history for some DNA repair components in the plant kingdom. In agreement with the findings for *A. thaliana*, the majority of sugarcane repair genes displayed higher sequence similarity to mammalian than to yeast genes, based on BLAST analysis.

METHODOLOGY

Sugarcane EST clusters obtained by CAP3 clustering (Pimentel and da Silva, 2001) were compared to known genes related to DNA repair in the National Center of Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) and the Munich Information Center for Protein Sequences (MIPS, http://mips.gsf.de) *Arabidopsis thaliana* databases. The basic local alignment search tool (BLAST) program (Altschul *et al*., 1997) was used for similarity searches against both nucleotide and protein databases. For these searches an e-value lower than 1×10^{-5} was used as a threshold of similarity. The gene names in the tables correspond to homolog proteins from human or yeast.

RESULTS AND DISCUSSION

Reversal of DNA damage

In principle, reversal of a DNA lesion is the simplest mechanism by which damaged DNA can be repaired. The biochemical mechanism is based on a one-step reaction, where one specific enzyme recognizes and reverts the lesion to the original molecular configuration in an error-free manner. Two of the most studied reversal mechanisms, photoreactivation and alkylation repair, are discussed below.

Photoreactivation

Photoreactivation is a light-dependent pathway that acts upon lesions induced by UV irradiation, which produces either cyclobutane pyrimidine dimers (CPD) or 6-4 photoproducts (Friedberg *et al.*, 1995). Photoreactivation is accomplished by a large number of enzymes, collectively called photolyases, whose homologues are found in many prokaryotes and eukaryotes. Two different types of photolyases were discovered, one for each type of DNA damage (*i.e.* CPD-photolyase or (6-4) photolyase). Despite the different substrates involved, the repair mechanisms both involve a single enzymatic step that has a similar mechanism (Sancar, 2000; Vonarx, *et al.*, 1998). The proteins of the photolyase family display a considerable structural similarity, and phylogenetic analysis suggests that an ancestral gene for CPD photolyase was duplicated before the divergence of prokaryotes and eukaryotes and that each copy has evolved independently (Todo, 1999), one to become Class I CPD photolyase (PhrI) in prokaryotes and the other Class II CPD photolyase (PhrII) in eukaryotes. It appears that the PhrI gene has been transmitted to eukaryotes and became functionally divergent, producing one gene whose product is the (6-4) photolyase and another gene that codes for a structurally related protein family (with a different function), the cryptochromes. These proteins, first identified in *Arabidopsis thaliana* (cryptocrome 1 and cryptocrome 2 apoproteins), are found in animals and plants where they are photoreceptors that mediate light dpendent responses and do not act in DNA repair (Eisen and Hanawalt, 1999).

Our analysis of sugarcane ESTs revealed the presence of homologues for Class I and Class II photolyases, as well as for the blue light photoreceptors Cry proteins (Table I). These proteins showed high sequence similarity with the corresponding *A. thaliana* proteins. This was to be expected because, by necessity, plants are often exposed to high light. The several genes from the photolyase family that were found are probably important for the maintenance of integrity of the plant genome.

Table I - Direct repair related proteins and their homologues in different organisms.

Enzyme (prototype)	Yeast	Human	*A. thaliana*	Sugarcane cluster (e-value)[a]
Photoreactivation				
6-4 photolyase	-	-	+	SCACCL6007E01.g (3e-50)
Phr1	+	+	+	SCCCST2002A06.g (e-115)
Phr2	-	-	+	SCJFRT2053C12.g (3e-24)
Photoreceptors				
Cry1	-	+	+	SCAGST3138B05.g (e-115)
Cry2	-	+	+	SCRFST1042F05.g (3e-81)
Alkylation reversal				
Ogt (Mgmt)	+	+	-	-
Ada	-	-	-	-

[a]tBLASTn e-values for the best hit obtained probing with *A. thaliana* proteins.

Alkylation repair

Another type of lesion reversal occurs with respect to the damage caused by alkyl groups covalently linked to DNA bases. Alkyltransferase proteins reverse alkylation by transferring the alkyl group from the DNA to itself, in a suicide process. One alkyltransferase, Ogt (O6-methyl guanine transferase), is found in many (but not all) organisms. Homologues of the Ogt protein are found in at least some species from each of the major groups, and it has been proposed that this protein is ancient and was present in the last common ancestor of all organisms (Eisen and Hanawalt, 1999). However, our search for *OGT* sugarcane homologues was unsuccessful (Table I). Comparison of our data with the *A. thaliana* genome revealed that both plants have no homologues of these genes. Since these plants are not very closely related, and no other *OGT* sequence from plants was identified in the databases, it is reasonable to propose that the common plant ancestor lost the alkyltransferase genes. This implies that, in plants, this type of base damage may be removed from DNA by different analogues or, most probably, by a different DNA repair pathway.

Excision repair

Although efficient, the direct reversal of lesions is very limited due to its high substrate/enzyme specificity, so organisms have evolved a more general DNA repair mode in which damaged bases are excised from the genome and replaced by a normal nucleotide sequence. This type of cellular response to DNA damage is called excision repair and its different modalities are described below.

Base excision repair

Base excision repair (BER) is responsible for removing a wide range of DNA lesions, including imidazole open rings and deaminated, alkylated, oxidized and absent bases (Memisoglu and Samson, 2000). Glycosylases are able to recognize and excise the damaged bases from the sugar phosphate backbone, the initial step of BER. This excision results in an abasic site (or AP, apurinic/apyrimidinic), which is recognized by another group of enzymes, the AP-endonucleases, that make an incision at the 5' or 3' phosphodiester of the AP site, generate a nucleotide gap, and then filled by polymerization and ligation of a new nucleotide to the DNA sequence (Cadet *et al.*, 2000).

As expected, we found many base excision repair enzymes in the SUCEST database (Table II), the AP endonucleases of the *E. coli* Exonuclease III family being widespread among eukaryotes, and sugarcane is no exception. In humans and *Arabidopsis thaliana*, more than one homologue has been found (Hadi and Wilson III, 2000), and that may also be the case for sugarcane. Another class of base excision enzymes, the AP endonucleases of the *E. coli* Endonuclease IV family, is more restricted among eukaryotes, although the presence of an *A. thaliana* protein (of unknown function) in the database with a very low similarity value has been reported (Hadi and Wilson III, 2000). We did not find any EST homologue to the Endonuclease IV family protein in sugarcane, which may have been due to low expression of the mRNA for this gene. However, given its absence in most eukaryotes, we speculate that it is also absent from sugarcane.

We found many DNA glycosylases such as *E. coli* Udg and Nth homologues in sugarcane, which were also present in *A. thaliana* and other eukaryotes. A noteworthy absence was the enzyme MutY, which protects DNA

Table II - Base excision repair related proteins and their homologues in different organisms.

Enzyme (prototype)	Yeast	Human	*A. thaliana*	Sugarcane cluster (e-value[a])
AP endonuclease (*E. coli* Exonuclease III)	Apn2 (Eth1)	Ape1 (RefI, HapI), Ape2	Arp, two other homologues	SCEZAM2032F04.g (e-96)
AP endonuclease (*E. coli* Endonuclease IV)	Apn1	-	-	-
Uracil DNA glycosylase (*E. coli* Ung)	Ung1	Udg1	Ung1	SCEQFL5048B07.g (e-42)
3-Methyladenine glycosylases (*E. coli* TagI and AlkA, human Aag)	Mag (AlkA family)	Aag	Aag, Mag, several TagI homologues	SCJLRT2049G09.g (Mag family, e-14), SCMCRT2104F02 (TagI family, e-37[b])
8-oxoguanine/formamidopyrimidine glycosylases (*E. coli* MutM, yeast Ogg1)	Ogg1	Ogg1	Ogg1, MutM1, MutM2	SCVPRZ2036B05.g (Ogg1family, e-66) SCCCLR2C01B12 (MutM family, e-111)
8-oxoguanine : adenine mispair glycosylase (*E. coli* MutY)	-	Myh	MutY	-
Thymine glycol glycosylases (*E. coli* Endonuclease III)	Ntg1, Ntg2	Ntg1	Nth1, other homologue	SCACSB1036F12 (e-85)
G:T mismatch glycosylase (human Tdg)	-	Tdg	-	-

[a]tBLASTn e-values for the best hit obtained probing with *A. thaliana* proteins.
[b]There are several Tag homologues in *A. thaliana*, as well as several clusters in sugarcane. Therefore, the e-value shown was obtained probing with *E. coli* TagI.

against the highly mutagenic effects of the oxidized base 8-oxoguanine by excising adenines mispaired with this lesion from DNA. It is interesting to note that this enzyme is absent from the complete genome sequence of *Saccharomyces cerevisiae*, and it will be of interest to determine whether its absence in the sugarcane database was due to very low representation of the EST for this mRNA, or it is in fact absent from the sugarcane genome as is the case in *Saccharomyces*.

Regarding 8-oxoguanine glycosylases, both *E. coli* MutM (Fpg) and yeast Ogg1 are normally present in plants, with two MutM homologues (MutM1 and MutM2) being present in *A. thaliana*. We found several sugarcane clusters that were very similar to MutM, some being clearly closer to MutM1, whereas some others could not be identified as MutM1 or MutM2.

As far as 3-Methyladenine glycosylases are concerned, it is known that there are two of these enzymes in *E. coli* encoded by the genes *alkA* and *tagI*. The *alkA* gene is specific for 3-methyladenine while the *tagI* gene product has a broader substrate specificity, excising other alkylated bases from DNA. In human cells, a third type of 3-methyladenine glycosylase, that bears no significant sequence similarity to the *E. coli* enzymes, has been identified, and the encoding gene was called *AAG*, while in yeast one *alkA* homologue, *MAG*, is found (Table II). In contrast, in *A. thaliana* all three classes of homologues (*alkA*, *tagI, AAG*) are represented, and in sugarcane both *alkA* (MAG) and *tagI* homologues can be found. By searching the *A. thaliana* database we identified at least 6 genes encoding *tagI* homologues dispersed in chromosomes 1, 3 and 5. In sugarcane, we identified several different EST clusters, also indicating the presence of multiple copies of this gene in sugarcane. This suggests that plants are in contact with a large amount of alkylating agents, either from environmental sources or as a consequence of their metabolism. Not surprisingly, Tdg mismatch glycosylase is found only in humans, since this enzyme removes thymine residues mispaired with guanine that arise in DNA due to deamination of 5-methylcytosine, which is abundant only in mammalian DNA.

Nucleotide excision repair

Nucleotide excision repair (NER) is one of the most flexible and general DNA repair pathways, removing a large spectrum of structurally unrelated DNA lesions that generate a considerable helical distortion in DNA. The basic repair mechanism is conserved from bacteria to humans, and consists of removing a single-stranded segment containing the lesion by dual incision of the damaged strand and subsequent gap-filling (de Laat *et al.*, 1999), but it appears that prokaryotic and eukaryotic nucleotide excision repair pathways have diverged significantly, since the proteins involved display no sequence homology. In eukaryotes, nucleotide excision repair proteins sharing high sequence similarity have been identified in yeast and animal cells, and the publication of the complete *A. thaliana* genome sequence (The *Arabidopsis* Genome Initiative, 2000) has revealed that most of the genes for these proteins are also found in this plant.

Our analysis of the sugarcane cDNAs confirmed that most of the nucleotide excision repair components are present in this plant too and that, as in *Arabidopsis*, they are very similar to the animal genes (Table III). However, more detailed comparison of the *Arabidopsis* and sugarcane sequences revealed features of the nucleotide excision repair pathway that is exclusive to plants. Some genes, which are essential for the nucleotide excision repair machinery in yeast and animal cells, seem to be absent in these plants because no corresponding genomic or cDNA sequences were identified either in *Arabidopsis* or sugarcane. This was the case for the Xpa protein, which plays a crucial role at an early stage of nucleotide excision repair as shown by the fact that mutations in the homologues of the Xpa gene found in human and yeast cells make the cells extremely sensitive to UV irradiation. Curiously, we found that both *Arabidopsis* and sugarcane contain genes similar to those that code for proteins (Xab1 and Xab2) that bind to Xpa. We also found that, in plants, the Rpa protein differs from its yeast and animal cell counterparts (which also act in recombination and replication) due to the absence of a third small 14 kD subunit (Rpa3).

One interesting feature of nucleotide excision repair is the participation of the basal transcription factor, TFII-H, in opening the DNA around the lesion. TFII-H is a nine-subunit protein complex with multiple enzymatic activities, acting as an ATPase, helicase and kinase (de Laat *et al*, 1999). The kinase sub-complex is made up of three subunits (Cdk7, Cyclin H and MAT1), which have been described as being absent from the *A. thaliana* genome (The *Arabidopsis* Genome Initiative, 2000). However, our comparison of the sugarcane and *A. thaliana* sequences suggests that two of the kinase subunits, Cdk7 and Cyclin H, are present in both plants (Table III). We also found that the p52 subunit, which has been described (The *Arabidopsis* Genome Initiative, 2000) as being absent from the *A. thaliana* genome, was present in sugarcane, and, furthermore, our comparison of *Arabidopsis* and sugarcane revealed the presence of a small open reading frame (ORF) possibly encoding for this protein in *A. thaliana*. This disagreement with previously published work may have been the result of a program error during the annotation of the *Arabidopsis* genome. We found a similar annotation problem with *Csa* paralogues in the *Arabidopsis* genome.

We also found two bacterial nucleotide excision repair homologous genes (*MFD* and *UVRD*) in *Arabidopsis* and sugarcane databases. In *E. coli,* these genes participate in the nucleotide excision repair pathway carried out by UvrABC proteins, which are not found in plants. The function of these two genes in plants is difficult to predict, but

Table III - Nucleotide excision repair related proteins and their homologues in different organisms.

Eukaryotic nucleotide excision repair factor		Potential homologues			
Name	Function	Yeast	Mammals	*A. thaliana*	Sugarcane cluster (e-value[a])
Xpc/hHr23b	DNA damage sensor				
		Rad 4	Xpc	Xpc	Xpc: SCJFRT2055B09.g (e-53)[b]
		Rad 23	hHr23A	hHr23A	hHr23A: SCJLLR1033C10.g (e-40)[b]
			hHr23B	hHr23B	hHr23B: SCCCCL4006H12.g (5e-30)[b]
TFII-H	Opening of the double helix	Rad 25	Xpb	Xpb	Xpb: SCEQRT1025F02 (0.0)[b]
		Rad 3	Xpd	Xpd	Xpd: SCCCCL6003E12 (e-90)[b]
		Ssl1	p34	p34	p34: SCVPRZ3027B07 (6e-20)[a]
			p44	p44	p44: SCEQLR1093F02 (e-75)[b]
		Tfb4			
		Tfb1	p62	p62	p62: SCEZRT2023E03 (2e-72)[a]
		Tfb2	p52	p52*	p52: SCMCRT2086C05 (e-72)[b]
		Cak (Tfb3, Kin28, Ccl1)	Cak (Mat1, Cdk7, CyclinH)	Cak (Cdk7+*, CyclinH*)	Cak (Cdk7+, CyclinH): SCQGRT1042H04 (e-82)[b]
Xpa	Affinity to the damaged DNA	Rad14	Xpa	-	-
Rpa	Stabilizes the complex of opened DNA and positionates the nucleases	Rfa1	Rpa1	Rpa1	Rpa1: SCEQAM2037D06 (e-34)[b]
		Rfa2	Rpa2	Rpa2	Rpa2: SCCCLR1075B01 (e-21)[b]
		Rfa3	Rpa3	-	-
Xpg	Catalyzes the 3'incision	Rad2	Xpg	Xpg	SCEPRZ1008D03 (e-12)[b]**
Xpe	Repair protein	-	Xpe	Xpe	SCCCST1005A09 (e-179)[b]
Xpf	Catalyzes the 5'incision	Rad 10	Ercc1	Ercc1	Ercc1: SCCCLR1072A08 (e-67)[b]
		Rad 1	Xpf	Xpf	Xpf: SCCCLR1001D06 (e-177)[b]
Csa	Transcription coupled repair	Rad 28	Csa	Csa	SCRLFL1013H08.g (3e-44)[b]
Csb	Transcription coupled repair	Rad 26	Csb	Csb	SCCCLR1C02D02 (e-75)[b]**
Rad 7	Repair of inactive DNA	Rad 7	-	-	-
Rad 16	Repair of inactive DNA	Rad 16	-	Rad 16	SCEPRZ1011E02 (e-80)[c]

*It was published as absent (The *Arabidopsis* Genome Initiative, 2000), but our recent analysis proposes its presence (+). The high sequence homology with its human and yeast homologue does not cover the whole protein.
**These clusters are probably relatated to proteins in the same family.
[a]tBLASTn e-values for the best hit obtained probing with *A. thaliana* proteins.
[b]tBLASTn e-values for the best hit obtained probing with human proteins.

they probably are acting on the maintenance of organellar genome stability, since both contain a chloroplast localization signal.

Mismatch repair

DNA mismatch repair (MMR) corrects DNA mispairs and loops introduced by replication errors. It also plays important roles in transcription-coupled repair, meiosis and recombination (reviewed by Buermeyer *et al*., 1999). In humans, MMR deficiency is associated with hereditary cancer.

Escherichia coli has been used as a model of DNA mismatch repair, and in this organism four important Mut proteins participate in this pathway: MutS which can bind base mismatches and DNA loops; MutH, an endonuclease that discriminates a newly synthesized strand from the template by recognizing methylated sequences in the template; MutL, an ATPase which is important in various MMR steps (*e.g.* MutS scanning of DNA and stimulation of MutH endonuclease activity) and MutU (UvrD) which is a helicase that facilitates the removal of the DNA patch of a newly replicated strand containing a mismatch.

Although useful, the *E. coli* model has been shown to be insufficient to explain all the features of MMR. In eukaryotes and in many prokaryotes, there are no MutH homologues, raising the possibility that another mechanism exists for strand discrimination in the MMR pathway (Eisen and Hanawalt, 1999).

Regarding the yeast Msh homologues of bacterial MutS, we found Msh2-Msh6 homologues in the human database but we were unable to find Msh4 homologues in the SUCEST or *Arabidopsis* databases. We also found the *Arabidopsis* Msh7 (also called Msh6-2) protein (Ade *et al*, 1999) in the SUCEST database, suggesting that this sequence may be present only in plants, although other plant genomes will have to be searched to confirm this. A guanine:thymine (G:T) mismatch binding protein (Gtbp) was present in the human and *A. thaliana* databases but absent from the sugarcane and yeast databases, which we searched. It is hard to know whether this absence is relevant

because we found other unclassified MutS homologues, which may substitute for the missing proteins, in the SUCEST and *Arabidopsis* databases. We also found a homologue of the bacterial MutS2 protein (which recognizes loops in DNA) in both these databases.

Regarding MutL homologues, there is some confusion concerning the names, since both Pms1 and Pms2 from plants and humans are homologues of yeast Pms1. In mammals, Pms1, Pms2 and Mlh1 form heterodimers that act in mismatch correction. From our search of the relevant databases, it seems that sugarcane and *Arabidopsis* are quite similar in such that they both lack the second yeast Pms1 homologue. No yeast Mlh2 homologue was found in the human or plant databases. As expected, we found no MutH homologues in sugarcane, although, interestingly, we did find MutU (UvrD)-like proteins in both the *Arabidopsis* and in sugarcane databases. However, other helicases may function in MMR as well, since humans and yeast lack the homologues to these enzymes. The participation of the MutU homologues in MMR remains to be experimentally determined.

DNA repair by recombination

Homologous genetic recombination is required for a variety of DNA repair and related activities. This process is involved in the successful segregation of chromosomes during cell division, contributes to the generation of organism diversity and is required for DNA repair by recombination. One kind of damage that can be repaired by homologous recombination is the double-strand break (DSB), which is the major threat to the genomic integrity of cells. Double-strand breaks are created by normal cellular processes such as meiotic recombination, restriction enzymes and V (D) J recombination (responsible for the generation of immunological diversity in mammals) as well as by endogenous and exogenous agents which damage DNA. Double-strand breaks can also result in chromosomal fragmentation, translocation and deletions. If the damage persists, or the repair is incorrect, the result can be genomic instability that, in some cells, can lead to carcinogenesis. Ries *et al.* (2000) have reported that elevated UV-B radiation reduces genome stability in plants, and that homologous repair pathways involving recombination might be implicated in eliminating UV-B-induced DNA lesions. There are two pathways involved in the repair of double-strand breaks, homologous recombination (HR) that ensures accurate repair and non-homologous end-joining (NHEJ), which is error-prone. In *E. coli*, yeast and mammals both pathways contribute to the repair of double-strand breaks, but their relative contribution can vary, *e.g.* in *E. coli* and yeast homologous recombination repairs most double-strand breaks, whereas in humans such breaks are mainly repaired by non-homologous end-joining (Kanaar *et al.*, 1998; Vonarx *et al.*, 1998; Pastink and Lohman, 1999).

As expected, we found many genes involved in repair by recombination in the SUCEST database (Table V). In eukaryotes, the strongly conserved *RAD52* epistasis group of genes (*RAD50-52, RAD54-55, RAD57, RAD59, MRE11, XRS2*) is important in mitotic and meiotic recombination as well as double-strand break repair. We found that the *RAD52*/*RAD59*, *RAD55*/*RAD57* and *XRS2* genes were absent from the *Arabidopsis* and SUCEST databases. In yeasts, the Xrs2 protein forms an exonucleolytic complex in conjunction with two other proteins (Mre11 and Rad50) which is involved in the initial step of double-strand break repair, irrespective of whether the repair is carried out by homologous recombination or non-homologous end-joining. We found no Xrs2 homologues in human or plant databases.

In humans, it is the Nbs1 protein which interacts with the Mre11 and Rad50 proteins to form a protein complex with nuclease activity which is involved in the initiation step of double-strand break repair. Nbs1 is the protein re-

Table IV - Mismatch repair proteins of *Escherichia coli* and their homologues in different organisms.

Enzyme (prototype)	Yeast	Human	*A. thaliana*	Sugarcane cluster (e-value[a])
E. coli MutS	Msh1-Msh6	Msh2-Msh6	Msh2- Msh3, Msh5, Msh7, MutS2, other	SCJFRZ1007F06 (Msh2, e-95) SCQSFL1127F02 (Msh3, e-66) SCAGFL1089G01 (Msh5, e-42[b]) SCACAD1038A12 (Msh6, e-77) SCUTFL1056E04 (Msh7, e-36) SCSBFL5017E02 (MutS2, e-28)
-	-	Gtbp	Gtbp	-
E. coli MutL	Mlh1-Mlh3, Pms1	Mlh1, Mlh3, Pms1, Pms2	Mlh1, Mlh3, Pms1	SCJLFL1047E04 (Mlh1, e-95) SCRLAM1009F02 (Mlh3, e-10[b]) SCSFAD1107G07 (Pms1, e-16[b])
E. coli MutH	-	-	-	-

[a]tBLASTn e-values for the best hit obtained probing with *A. thaliana* proteins. [b]Although the *A. thaliana* protein exists, the annotation is incomplete. Therefore, the e-value shown was obtained probing with human genes. - Absence of this protein.

sponsible for the Nijmegen Breakage Syndrome, a chromosomal instability disorder, characterized in part by cellular hypersensitivity to ionizing radiation. It is interesting that we found *NBS1* in both the *Arabidopsis* and SUCEST databases.

The *XRCC3* gene (a member of the *RAD51* gene family) has been related to chromosomal stability and double-strand break repair in mammalian cells, and although we found this gene in the *Arabidopsis* database we did not find it in the SUCEST database. It will be interesting to discover if its absence is due to very low representation of the EST in the SUCEST database or if it is indeed absent and sugarcane differs from *A. thaliana* in this respect.

In humans, the Xrcc4-7 proteins are specifically required for non-homologous end-joining, and we found Xrcc4, 5 and 6 in both the *A. thaliana* and sugarcane databases, although we found no *XRCC7* (DNA-PKcs) in plants. Since *XRCC4-7* homologues are not found in Archaea or bacteria, it appears that this pathway evolved only in eukaryotes, and it has been reported by Eisen and Hanawalt (1999) that sequence similarity between yeast and mammalian *XRCC4-7* proteins is very limited, indicating that there may be many functional differences between the *XRCC4-7* proteins from these two groups. We also found several different EST clusters in sugarcane which suggest the presence of multiple copies for the *recA* and *recQ* families, as has been reported for *A. thaliana* (The *Arabidopsis* Genome Initiative, 2000).

DNA replication and damage tolerance pathways

When removing lesions, the action of some DNA repair mechanisms results in gaps or breaks, which need to be filled. This DNA polymerization step uses enzymes involved in the normal replication of DNA, *e.g.* as DNA polymerases, DNA primase, PCNA (Proliferating cell nuclear antigen), RFC (Replication factor C) and DNA ligases (Budd and Campbell, 2000).

Our search for DNA polymerase homologues in the SUCEST database revealed that, as in *A. thaliana*, the majority of known DNA polymerases are present in sugarcane (Table VI). One sugarcane cluster displayed a high level of similarity to Pol β, a polymerase that acts as a dRPase (excises the deoxyribose-phosphate backbone from the DNA molecule) and synthesizes short patches across abasic sites in the base excision repair pathway. However, this cluster

Table V - Presence (+) or absence (-) of Recombinational Repair Proteins and their homologues in different organisms.

Protein	Yeast	Human	*A. thaliana*	Sugarcane name (e-value[a])
Homologous Recombination (HR)				
RecF	-	-	-	-
RecQ family	+	+	+	SCJLFL3018G05.g (e-73)
Dmc1	+	+	+	SCCCLR1069C03.g (e-163)
RecA	+	+	+	SCSBLB1034B03.g (e-32)
Rad51	+	+	+	SCCCLB1004H09.g (e-65)
Rad51B	-	+	+	SCMCST1050E11.g (e-51)
Rad51C	-	+	+	SCJFRZ2015C03 (e-52)
Rad52/Rad59	+	+	-	-
Rad54	+	+	+	SCACCL6011A06.g (e-114)
Rad55/57	+	+	-	-
Xrcc1	-	+	+	SCEZRZ1015B02.g (e-87)
Xrcc2	-	+	+	SCACSB1123A07.g (e-11)
Xrcc3	+	+	+	-
Rad50	+	+	+	SCRLAM1005E06.g (e-62)
Mre11/Rad32	+	+	+	SCQGSB1082E07.g (e-101)
Xrs2	+	-	-	-
Nbs1	-	+	+	SCCCLR1076C04.g (e-15)
HR and *HNEJ*[1]				
Ku70	+	+	+	SCCCRZ1C02A05.g (e-84)
Ku80/86	+	+	+	SCQGAM1045F09.g (e-163)
DNA-PKcs	-	+	-	-
Xrcc4	+	+	+	SCEZFL5087C02.g (e-56)
DNA ligase IV	+	+	+	SCQGLB1038C07.g (3e-11)[2]

[a]tBLASTn e-values for the best hit obtained probing with *A. thaliana* proteins. [1]NHEJ: Non-homologous end joining. [2]This cluster codes for a hypothetical protein that is correlated to another DNA ligase.

was closer to the human DNA Pol λ (Table VI), a polymerase linked to translesion synthesis (see below). It seems that in plants other proteins from the same family may perform the same functions as Pol β, although this will have to be confirmed experimentally.

Another polymerase, which we did not find in the plant databases, was the homologue to Pol γ, which is responsible for the replication of the genomes of organelles. However, we found, in SUCEST and *Arabidopsis* databases, homologues to bacterial polymerases, with possible transit peptide signals targeting to organelles. Therefore, we propose that plants may have evolved organellar polymerases different to those present in animals or yeast, a phenomenon that deserves further investigation.

During evolution, living organisms have developed mechanisms to minimize the deleterious effects of damage to DNA. These mechanisms include tolerance pathways in which the cell is able to deal with the lesions remaining in the genome during DNA replication (Bayton and Fuchs, 2000). When the normal replication machinery is arrested at a lesion on the template strand the replication fork is displaced and single-strand structures are formed. Specialized 'bypass polymerases' reconstruct the missing strand by to polymerizing across a lesion in a mechanism called translesion synthesis (TLS), although these polymerases generally exhibit low fidelity (*i.e.* are error-prone).

We found some translesion synthesis polymerases in the *Arabidopsis* and SUCEST databases, and comparing them to the yeast and human enzymes we were able to detect two important differences. One difference was the absence of the Rev7 accessory subunit of Pol ζ (Table VI), although the catalytic Rev3 subunit was clearly represented in plants. The Rev7 protein seems to have a regulatory role on the action of Pol ζ (Murakumo *et al*, 2001), and the putative absence of Rev7 in plants indicates that another protein may act in its place, suggesting a distinct strategy for plants in this type of DNA repair. Another interesting feature that we observed in the SUCEST and *Arabidopsis* databases was the presence of Pol λ, since homologues for this protein were not found in yeast cells but were present in humans. We also noticed that several of the new TLS DNA polymerases found in humans, described in recent years, were not found in the yeast or plant databases (Table VI), suggesting that these polymerases may be specific to animals. In this respect, the similarity between sugarcane and *Arabidopsis* suggests that the gene organization of the translesion synthesis pathway and DNA polymerases might be very conserved in plants, and probably endowed with different characteristics as compared to that which occurs in the genomes of yeasts and humans.

Other genes related to genetic stability

In humans, various proteins involved in DNA repair are also related to hereditary diseases characterized by high chromosomal instability or a high incidence of cancer, or both (Machado and Menck, 1997). We looked for the presence of these proteins in the SUCEST database (Table VII) and found homologues for Ataxia Telangiectasia (Atm), Bloom (Blm) and Werner (Wrn) syndromes in agreement with previous findings in *A. thaliana* (The *Arabidopsis* Ge-

Table VI - Comparative analysis of polymerases in different organisms.

Protein	Yeast	Human	*A. thaliana*	Sugarcane cluster (e-value)
		Replicative polymerases		
Pol α	+	+	+	SCQGST1030C04.g (6e-36)[a]
Pol δ	+	+	+	SCCCST2003E02.g (1e-102)[a]
Pol ε	+	+	+	SCEZLB1012F06.g (2e-26)[a]
Pol β	+	+	~	-
Pol γ	+	+	~	~SCJCRZ1025F02.g (1e-163)[b]
		TLS polymerases		
Pol ζ (cat.sub.)	+	+	+	SCJLRT1021B10.g (8e-73)[a]
Rev7	+	+	-	-
Rev1	+	+	+	SCVPRZ3031E07.g (3e-32)[a]
Pol η	+	+	+	SCCCST1006H11.g (9e-35)[a]
Pol ι	-	+	-	-
Pol κ	-	+	-	-
Pol μ	-	+	-	-
Pol λ	-	+	+	SCRUFL1016F09.g (3e-90)[a]
Pol θ	-	+	-	-

[a]tBLASTn e-values for the best hit obtained probing with Human proteins. [b]tBLASTn e-values for the best hit obtained probing with *Arabidopsis* proteins. + present; - absent; ~ like-enzyme.

Table VII - Presence (+) or absence (-) of human hereditary disease proteins in different organisms.

Protein	Yeast	Human	*A. thaliana*	Sugarcane cluster (e-value)
Wrn	+	+	+	SCVPLB1020E11.g (4e-51)
Blm	+	+	+	SCVPLB1020E11.g (7e-72)
Atm	+	+	+	SCSBAM1088F01.g (3e-57)
Faa	-	+	-	-
Fac	-	+	-	-
Fad	-	+	-	-
Fae	-	+	-	-
Faf	-	+	-	-
Fag	-	+	-	-
P53	-	+	-	-
Brca-1	-	+	+	?
Brca-2	-	+	+	?

nome Initiative, 2000). We also found homologues to the Brca-1 and Brca-2 proteins (both related to hereditary breast cancer) in *A. thaliana*, although in sugarcane some clusters displayed high homology only with the N-terminal region of the Brca proteins. Since these are large proteins, it is difficult to conclude whether these sequences corresponded to homologous genes or simply displayed a certain domain similarity. In plants, we found no homologues to proteins related to Fanconi Anemia (the Faa, Fac, Fad, Fae, Faf and Fag proteins) or Li Fraumeni syndrome (the p53 protein), which indicates important differences in DNA metabolism between plants and humans.

CONCLUDING REMARKS

DNA repair mechanisms are well conserved among living organisms due to their importance in maintaining genome integrity. Cells have different DNA repair mechanisms which are either specific for some kinds of lesions (*e.g.* the photoreactivation mechanism) or have very high versatility, as is the case in nucleotide excision repair. Our search for components of these pathways in the SUCEST database revealed that most known DNA repair mechanisms are present in the sugarcane genome. Moreover, the majority of sugarcane sequences displayed a higher similarity to animal than to yeast proteins, as has also been found for *A. thaliana* repair genes (The *Arabidopsis* Genome Initiative, 2000). This is a surprising finding, since phylogenetic studies suggest that plants are more closely related to fungi than to animals.

Another interesting point is that certain DNA repair genes essential to other eukaryotes may be absent in sugarcane. It is important to determine whether the absence of these proteins is due to their ESTs having a very low frequency or if this absence is a specific characteristic of the sugarcane genome. Some genes did not present any homologues in the complete genome of *Arabidopsis*, suggesting that they have been lost during evolution. As *A. thaliana* and sugarcane are not closely related, being dicotyledonous and monocotyledonous plants respectively, we may speculate that the observed gene loss may have occurred early in the evolution of the plant kingdom. Among the genes absent from plants, some are very surprising and may indicate important differences in DNA repair mechanisms. The absence of a protein homologue to the alkyltransferases found in other eukaryotes indicates that other strategies are used in the plant cells to deal with the kind of lesions normally removed by these enzymes. Nucleotide excision repair is apparently complete, but the absence of a Xpa homologue is intriguing, and a substitute protein performing its function may yet be found. Another interesting feature observed in the SUCEST and *Arabidopsis* databases was the presence of gene products homologous to bacterial proteins (*e.g.* the nucleotide excision repair genes *mfd* and *UVRD*) and which were not found in other eukaryotes. Homologues for these proteins are also present in the *Arabidopsis* database (The *Arabidopsis* Genome Initiative, 2000) making it more probable that, among eukaryotes, these proteins are unique to plants. These plant homologues show high sequence similarity with corresponding cyanobacterial genes, and in *A. thaliana* chloroplast target peptide signals have been identified at the N-terminal region of the putative Mfd and UvrD proteins, suggesting that the genes coding for these proteins are remnants of the original genome from the cyanobacterial lineage that colonized plant cells and derived chloroplasts.

The genomic approach described in this work has revealed many interesting features of the processes plants employ to deal with DNA damage. However, this approach is limited and a more complete analysis, based on the complete sugarcane sequence and phylogeny, as well as biochemical and genetic studies may help us to understand the full functions of the SUCEST clusters reported in this paper.

RESUMO

A identificação e caracterização de genes envolvidos com reparo de DNA são de grande interesse, dada a sua

importância na manutenção da integridade genômica. Além disso, a alta conservação dos genes de reparo de DNA faz com que possam ser utilizados como fonte de informação no que diz respeito à origem e evolução das espécies. Os mecanismos relacionados à remoção de danos pelo reparo de DNA, bem como suas conseqüências biológicas, já são bem conhecidas em bactérias, leveduras e animais. Entretanto, no que diz respeito a organismos vegetais, ainda há muito a ser investigado. No presente trabalho, apresentamos a identificação dos genes envolvidos nas principais vias de reparo de DNA em cana-de-açúcar, através de uma análise de similaridade do banco de dados do projeto brasileiro Sugarcane Expressed Sequence Tag (SUCEST) com seqüências protéicas conhecidas disponíveis em outros bancos de dados públicos (National Center of Biotechnology Information (NCBI) e Munich Information Center for Protein Sequences (MIPS) *Arabidopsis thaliana*). Esta busca revelou que a gama de proteínas envolvidas no reparo de DNA em cana-de-açúcar é similar a de outros eucariotos. Mesmo assim, foi possível identificar algumas características interessantes encontradas apenas em vegetais, provavelmente em função do seu processo evolutivo independente. As vias de reparo de DNA aqui representadas incluem fotorreativação, reparo excisão de bases, reparo excisão de nucleotídeos, reparo mismatch, end-joinning não homólogo, reparo por recombinação homóloga e tolerância a lesões. Este trabalho descreve as principais diferenças encontradas na maquinaria de reparo de DNA de células vegetais em relação àquela de organismos nos quais encontra-se bem descrita. Tais diferenças chamam a atenção para um potencial de mecanismos distintos em vegetais, que merecem futuras investigações.

ACKNOWLEDGMENTS

This work was supported by the Brazilian agency FAPESP (São Paulo, SP, Brazil. proc. 99/02841-3, 98/11119-7).

REFERENCES

Ade, J., Belzile, F., Philippe, H. and **Doutriaux, M.P.** (1999). Four mismatch repair paralogues coexist in *Arabidopsis thaliana*: AtMSH2, AtMSH3, AtMSH6-1 and AtMSH6-2. *Mol. Gen. Genet*. *262* (2): 239-249.

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* (17): 3389-3402.

Bayton, K. and **Fuchs, R.P.** (2000). Lesions in DNA: hurdles for polymerases. *Trends Biochem. Sci. 25*: 74-79.

Budd, M.E. and **Campbell, J.L.** (2000). Interrelationships between DNA repair and DNA replication. *Review*. *Mut. Res. 451*: 241-255.

Buermeyer, A.B., Deschênes, S.M., Baker, S.M. and **Liskay, R.M.** (1999). Mammalian DNA mismatch repair. *Annu. Rev. Genet. 33*: 533-564.

Cadet, J., Bourdat, A.G., D'Ham, C., Duarte, V., Gasparutto, D., Romieu, A. and **Ravanat, J.L.** (2000). Oxidative base damage to DNA: specificity of base excision repair enzymes. *Mut. Res. 462* (2-3): 121-128.

Costa, R.M.A., Morgante, P.G., Berra, C.M., Nakabashi, M., Bruneau, D., Bouchez, K.S., Van Sluys, M. and **Menck, C.F.M.** (2001). The participation of *AtXPB1*, the *XPB/RAD25* homologue gene from *Arabidopsis thaliana*, in DNA repair and plant development. *Plant J. 28* (4): 385-395.

De Laat, W.L., Jaspers, N.G. and **Hoeijmakers, J.H.** (1999). Molecular mechanisms of nucleotide excision repair. *Genes Dev. 13* (7): 768-85.

Eisen, J.A. and **Hanawalt, P.C.** (1999). A phylogenomic study of DNA repair genes, proteins and processes. *Mutat. Res. 435*: 171-213.

Friedberg, E.C., Walker, G.C. and **Siede, W.** (1995). DNA Repair and Mutagenesis. ASM Press, Washington DC.

Hadi, M.Z. and **Wilson III, D.M.** (2000). Second human protein with homology to the *Escherichia coli* abasic site endonuclease Exonuclease III. *Environ. Mol. Mutagen. 36* (4): 312-324.

Kanaar, R., Hoeijmakers, J.H.J. and **van Gent, D.C.** (1998). Molecular mechanisms of DNA double-strand break repair. *Trends in Cell Biology 8*: 483-489.

Machado, C.R. and **Menck, C.F.M.** (1997). Human DNA repair diseases: from genome instability to cancer. *Brazilian Journal of Genetics 20* (4): 755-762.

Memisoglu, A. and **Samson, L**. (2000). Base excision repair in yeast and mammals. *Mutat. Res. 451*: 39-51.

Murakumo, Y., Ogura, Y., Ishii, H., Numata, S., Ichihara, M., Croce, C.M., Fishel, R. and **Takahashi, M.** (2001). Interactions in the error-prone postreplication repair proteins hREV1, hREV3, and hREV7. *J Biol Chem 276* (38): 35644-51.

Pastinik, A. and **Lohman, P.H**. (1999). Repair and consequences of double strand breaks in DNA. *Mutat. Res. 428* (1-2): 141-146.

Ries, G., Heller, W., Puchta, H., Sandermann, H., Seidlitz, H. and **Hohn, B.** (2000). Elevated UV-B radiation reduces genome stability in plants. *Nature 46*: 98-101.

Sancar, G.B. (2000). Enzymatic photoreactivation: 50 years and counting. *Mutat. Res*. *451*: 25-37.

Telles, G.P. and **da Silva, F.R.** (2001). Trimming and clustering sugarcanes ESTs. *Genetics and Molecular Biology 24* (1-4): 17-23.

The *Arabidopsis* Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature 408*: 796-815.

Todo, T. (1999). Functional diversity of the DNA photolyase/blue light receptor family. *Mutat. Res*. *434* (2): 89-97.

Vonarx, E.J., Mitchell, H.L., Karthikeyan, R., Chatterjee, I. and **Kunz, BA.** (1998). DNA repair in higher plants. *Mutat. Res. 400*: 187-200.