J. Plant Biol. 41 (2) : 116~124 (1998)

Analysis of 176 Expressed Sequence Tags Generated from cDNA Clones of Hot Pepper by Single-pass Sequencing

Hong, Sung-Tae, Jae-Eun Chung[1], Gynheung An[1] and Seong-Ryong Kim*

Department of Life Science, Sogang University, Seoul 121-742, Korea

[1]Department of Life Science, Pohang University of Science and Technology, Pohang 790-784, Korea

As a part of the project to identify novel genes from the hot pepper (*Capsicum annuum* L. cv. Happy Dry), we have constructed several cDNA libraries and 176 randomly selected cDNA clones were partially sequenced. This expressed sequence tag (EST) analysis identified 95 clones (54.0%) that had a significant homology to a known protein sequence in the NCBI database. Of these clones, eighteen of them are related to genes not from the plant kingdom, indicating that 10.2% of the ESTs were newly identified in plants. Functional categorization of these clones revealed that the genes involved in metabolic pathways such as glycolysis and photosynthesis are most abundant, and genes in translational apparatus ranked next in abundance. Expression patterns of four ESTs were examined by RNA blot analysis. The CAN14 clone, which has a 58% identity to the potato patatin protein over a 117 amino acid overlap, was highly expressed in anther tissue but not in fruit tissues. The CFR2 clone, which is 88% identical to apospory-associated protein, showed relatively higher expression levels in seedlings and roots compared to other tissues. The transcript of the CFR11 clone, which shows a homology to γ-thionin, was abundant in every organ that was examined except roots. The CFR 29 clone which is 92% identical to the putative osmoprotectant from tomato root showed a root-preferential expression pattern.

Keywords: hot pepper, cDNA, expressed sequence tags, patatin, apomixis, thionin, osmoprotectant

Expressed Sequence Tag (EST) analysis has been known as an efficient way for gathering the information concerning the genome of an organism. ESTs are short sequences of a few hundred base pairs in length, which are derived by partial, single pass sequencing of randomly selected cDNA clones. As a part of the human genome project, Adams *et al.* (1991) advocated that sequencing of random cDNA clones is an efficient method in terms of both speed and cost. The database of EST (dbEST) has been a rapidly growing division of GenBank. However, random sequencing of cDNA clones has revealed that in many cases redundancy of clones could not be avoided. To reduce this problem, subtractive libraries or cDNA chips have been replaced instead of simple sequencing.

Currently more than 32,000 ESTs from *Arabidopsis thaliana* and 18,000 ESTs from rice have been deposited in databases. About 40% of the *Arabidopsis* ESTs have homology to known genes in the database. If there exist 20,000 genes in *Arabidopsis*, the number indicates that more than 60% of the genes has already been identified (Somerville and Somerville, 1997). The ESTs can be further used for comprehensive integration of expressed genes and for physical mapping to the genome (Rounsley *et al.*, 1996). Also, it has been shown that EST analysis may be a useful tool for analyzing the splicing patterns of previously characterized cDNAs, as a significant number of intron sequences in dbEST was found (Wolfsberg and Landsman, 1997).

Hot peppers are used as fresh vegetables and processed foods such as pastes and hot sauce in Korea. Hot peppers have a great amount of carotenoids, vitamins, and amino acids, some of which are quantitatively characterized by several groups (Kim *et al.*, 1997; Collins *et al.*, 1995). The plants are considered as one of the most important vegetable plants in Korea, where the annual market value of hot pepper-related industry is over 2 billion US dollar. Pepper seeds have become important exporting

*Corresponding author: Fax +82-2-704-3601

materials of Korean seed companies after the open-pollinated varieties have been replaced by F1 hybrid seeds (Park, 1992). The genome of the red pepper was investigated by analysis of reassociation kinetics. It was shown that the red pepper genome has a 1C DNA content of 1.25×10^9 bp (An *et al.*, 1996). Currently, a molecular linkage map of the hot pepper is being constructed by several groups in Korea (Kim *et al.*, 1997). Several groups of scientists are currently involved in study of pepper plants.

In this study, we have constructed cDNA libraries from four different organs of the hot pepper and partially sequenced 176 cDNA clones, among which 54% showed homology with GenBank database sequences. Based on the identification of the putative functions of the cDNA clones, analysis of gene expression patterns for some clones was performed.

MATERIALS AND METHODS

Plant Samples and Bacterial Strains

Field-grown hot pepper plants (*Capsicum annuum* L. cv. Happy Dry) at the flowering stage were harvested during the 1995 and 1996 seasons. Total floral buds and anthers at the final bud stage preceding anthesis were collected and stored in liquid nitrogen until used. The young fruits, length shorter than 2 cm, were used as fruit samples. For RNA blot analysis, hot pepper plants were grown to the flowering stage under greenhouse conditions.

E. coli strains MC1000 [*araD139*, (*araABC-leu*) *7679*], *galU*, *galK*, (*lac*)*X74*, *thi-*, *rpsL*(Str^r)] and XL-1 Blue MRF [(F'::Tn*10 proAB*, *lacI^{q}ZΔM15*) *Δ*(*mcrA*) *183*, *Δ*(*mcrCB-hsdSMR-mrr*) *recA1*, *endA1*, *gyrA96* (Nal^r), *thi-1*, *hsdR17* ($r_K^-m_K^+$), *supE44*, *relA1*, *lac*] were used as hosts for molecular cloning. The f1 helper phage, R408, was used for *in vivo* excision of the pBluescript plasmid vector from the λ ZapII phage (Stratagene).

Construction of cDNA Libraries and Analysis of ESTs

Total RNA isolation and cDNA library construction was performed as described (Kim *et al.*, 1996). For EST analysis, about 40 to 50 cDNA clones from each library for leaf, floral bud, anther and fruit were randomly selected and their 5' ends were sequenced. Template DNAs were prepared by the alkaline lysis method using the Wizard SV column (Promega). Dideoxynucleotide chain termination sequencing (Sanger *et al.*, 1977) was conducted using Sequenase Version 2.0 or The thermosequenase cycle DNA sequencing kit (Amersham). Computer softwares, DNAsis and PROsis (Hitachi), were used for sequence analysis. GenBank, EMBL and SwissProt databases were searched for sequence homology using a BLAST algorithm program (Altschul *et al.*, 1990).

RNA Analysis and Preparation of Labeled Probes

Ten μg of total RNA were resolved on a 1.2% formaldehyde agarose gel, blotted onto a nylon membrane, and hybridized with a radioactively labeled probe (Sambrook *et al.*, 1989). RNA hybridization was performed for 16 to 24 hr at 42°C. The membrane was washed in 2×SSC, 0.1% SDS at room temperature. If necessary, the membrane was further washed in 0.1×SSC, 0.1% SDS, at 65°C and exposed to a Kodak XAR-5 film or phosphoimage plate (Fuji BAS 1500, Japan). DNA fragments for hybridization were purified by electroelution and radioactively labeled using [α-^{32}P] dCTP (3000 ci/mmole) (Dupont) by the random priming method (Feinverg and Vogelstein, 1983). Unincorporated nucleotides were removed by G-50 Sephadex column chromatography.

RESULTS AND DISCUSSION

The Hot Pepper cDNA Library Construction and Generation of ESTs

The strategy for EST analysis was the same as that used for human brain ESTs (Adams *et al.*, 1991). This involves partial sequencing of cDNA clones by only single-pass sequencing and homology-searching by either in nucleic acid databases or in protein databases after translation of EST into peptide sequences.

cDNA libraries from leaves (CLF), floral buds (CFB), anthers (CAN), and fruit tissues (CFR) were generated. Oligo (dT) was used as a primer for the synthesis of the first-strand cDNA, and then double-stranded cDNA was ligated into a λ ZAPII or UNI-ZAP XR vector (Stratagene). As shown in Table 1, the initial plaque forming unit (pfu) of each library was between 4.2×10^5 (fruit) to 3.5×10^6 (leaf). Insertion efficiency analyzed either by white/blue selection or by restriction analysis was more than 80%.

Table 1. Hot pepper cDNA libraries and EST characterization

cDNA library	Leaf	Floral bud	Anther	Fruit
Original plaque forming unit	3.5×10^6	6.5×10^5	6.9×10^5	4.2×10^5
Average insert size	0.6 kb	1.0 kb	1.0 kb	1.2 kb
Insertional efficiency	96%	95%	80%	80%
Number of ESTs	42	57	38	39
Database match (%)	24 (57.1)	34 (59.6)	19 (50.0)	18 (46.2)
Plant	23	30	11	14
Solanaceae	11	9	6	10
Brassicaceae	1	13	4	4
Other kingdom	1	4	8	4

Each library was converted *en masse* to pBluescript plasmids and the clones containing cDNA inserts longer than 0.4 kb were selected after agarose gel electrophoresis. The average insert sizes for each of the four different libraries are shown in Table 1.

Characterization of the Pepper ESTs by Database Search

We have generated 176 ESTs from the hot pepper libraries. The ESTs are composed of 42 ESTs from the leaf library, 57 from the floral bud library, 38 from the anther library, and 39 from the fruit library (Table 1). Most of the clones from cDNA libraries contained cDNA inserts longer than 0.4 kb, and these inserts were subjected to the single pass sequencing. The size distribution of the inserts was 0.3 kb to 2.2 kb, with a mean of 0.9 kb and with 46 clones being larger than 1.0 kb. Eighteen ESTs (CAN1 to CAN18) of the anther cDNA library were sequenced from both ends using Sequenase Ver. 2.0 (Amersham), and the remaining 158 ESTs were sequenced by a cycle sequencing method using Thermosequenase (Amersham). The average sequenced length of the ESTs was 310 nucleotides for the conventional sequencing and 328 nucleotides for the cycle sequencing.

Of the 176 ESTs generated, 95 (53.9%) sequence tags carried cDNA with significant amino acid sequence similarities to previously identified genes registered in protein databases. The rather high percentage of database matches may be due to the less stringent cut off score (greater than BLASTX score of 80 or 40% identity) and the increasing information of known genes in the database. Also, it has been known that DNA sequencing from the 5' end of cDNA is significantly more informative (Shen *et al.*, 1994). We used a less stringent cut-off score to get more clones that may have a conserved domain, motif, or a common protein structure. It was previously reported that the percentage of significant matches to known genes was 32% for *Arabidopsis* (Höfte *et al.*, 1993) and 48% for *Brassica* (Lim *et al.*, 1996). When the criteria of the BLASTX score greater than 80 was applied, the number of the high similarity EST clones was reduced to 64 (36.4%). All the sequences were automatically translated in the six open reading frames and were compared with the protein sequence database in Genbank using the subroutine BLASTX of Gapped Basic Local Alignment Search Tool (Gapped BLAST). If no significant homology was found, the sequences were compared at the nucleotide level using BLASTN. In this way, we found that two EST clones (CLF10 and CFR17) had similarity with ribosomal RNA genes. BLASTN analysis was found to be useful in detecting additional high similarities overlooked by the BLASTX, as BLASTN subroutine searches the database, dbESTs, and random sequence data of cDNA entries at the nucleotide level. Gapped BLAST is the 2.0 version of BLAST 1.4 and includes significant performance enhancements: the addition of 'gapping' routines, position-specific-iterated BLAST as well as extensive changes to the text report, and the format of the databases. The 'gapping' routine allows the introduction of deletions and insertions into alignments. With the gapped alignment tool, homologues do not have to be broken into several segments. Also, the scoring of gapped results tends to be more biologically meaningful (Altschul *et al.*, 1997).

Listed in Table 2 are the ESTs of *C. annuum* that show significant similarity to the sequence in the databases. Among these, 18 ESTs (10.2%) showed sequence homology with non-plant genes. Considering the possibility of finding new genes in plants by EST analysis, these 18 non-plant matched clones may be valuable for further examination. We observed that 77 ESTs encoded proteins previously identified in other plant species, and only 3 ESTs matched registered genes from the *Capsicum* species. Among the other plant gene-homologues, 34 ESTs (44.1%) shared sequence homology with genes from family *Solanaceae*, 22 ESTs (28.6%) with *Brassicaceae*, and 4 ESTs (5.2%) with *Poaceae* (Table 1). Of the 176 ESTs, 79 sequence tags (44.9%) did not show homology with any sequences in the databases and thus may represent previously uni-

Table 2. Hot pepper ESTs putatively identified by the database search

Clone	Putative identification	Organisms	LC[a]	% Id[b]	Acc. No.[c]	DB[d]
Metabolism						
CFB36	ADP-Ribosylation factor 1	*Arabidopsis thaliana*	66	79	P36397	SP
CLF19	ATP synthase gamma chain	*Ipomoea batatas*	86	92	P26360	SP
CFB55	Biotin carboxyl carrier protein	*Glycine max*	85	47	U40666	GB
CLF11	Biotin carboxylase precursor	*Glycine max*	110	91	AF007100	GB
CFR1	Carboxypeptidase I	*Hordeum vulgare*	83	73	1314177B	PRF
CLF5	Chl. a-b binding protein type I precursor	*Lycopersicon esculentum*	77	86	S06329	PIR
CLF2	Chl. a-b binding protein LHC I, type III precursor	*Lycopersicon esculentum*	47	93	S04125	PIR
CAN2	Chlorophyll magnesium chelatase	*Glycine max*	37	95	JC4312	PIR
CAN24	Cyt. C oxidase chain I	*Lycopersicon esculentum*	104	81	S65346	PIR
CFB14	Cyt. P450 hydroxylase	*Zea mays*	77	42	X81829	EMBL
CLF38	Glyceraldehyde-3-phosphate dehydrogenase	*Nicotiana tabacum*	105	72	Z72488	EMBL
CFB12	Decarboxylase homolog	*Arabidopsis thaliana*	46	80	Z97341	EMBL
CFR6	Glycosyl transferase	*Arabidopsis thaliana*	52	58	AF001308	GB
CFB40	Inorganic pyrophosphatase	*Nicotiana tabacum*	39	79	S54173	PIR
CFB34	Isopropylmalate dehydrogenase	*Brassica napus*	63	68	P29102	SP
CLF9	Methyltransferase	*Prunus armeniaca*	71	69	U82011	GB
CAN22	Phosphoglycerate dehydrogenase	*Arabidopsis thaliana*	81	73	AB003280	DDBJ
CLF1	Photosystem I subunit II precursor	*Cucumis sativus*	130	68	P32869	SP
CLF8	Photosystem I subunit	*Nicotiana sylvestris*	137	73	Q41228	SP
CLF30	Photosystem II 5kD protein	*Gossypium hirsutum*	105	51	P31336	SP
CFB2	Polyphenol oxidase B precousor	*Solanum tuberosum*	79	87	Q06355	SP
CAN29	Polyphenol oxidase B precursor	*Lycopersicon esculentum*	34	88	Q08304	SP
CFB3	Polyphonol oxidase precursor	*Lycopersicon esculentum*	64	75	S22970	PIR
CFB39	Protein phosphatase type I	*Nicotiana tabacum*	70	96	Z93770	EMBL
CLF36	RuBP carboxylase/oxygenase	*Lycopersicon esculentum*	109	81	X05983	EMBL
CLF20	Pyruvate dehydrogenase E1 subunit B	*Pisum sativum*	126	91	P52904	SP
CFB52	Ribonucleotide reductase 1	*Arabidopsis thaliana*	68	90	Y07746	EMBL
CFR35	Serine carboxylase	*Arabidopsis thaliana*	129	72	AC002332	GB
CAN9	Serine carboxypeptidase precursor	*Arabidopsis thaliana*	103	79	AC002332	GB
CFB29*	Glutathione S-transferase	*Homo sapiens*	42	50	U80819	GB
CFR39*	Membrane transporter	*Bacillus subtilis*	93	45	Z99107	EMBL
CAN25*	NADPH quinone oxidoreductase	*Homo sapiens*	101	56	AF010309	GB
CAN28*	Peroxisome biosynthesis protein PAS1	*Homo sapiens*	96	52	P46463	SP
CLF41*	Sorbitol dehydrogenase	*Homo sapiens*	113	50	Q00796	SP
CFR13*	Udp N-acetylglucosamine O-acyltransferase	*Escherichia coli*	129	34	P10440	SP
Stress / Resistance						
CFR4	L-ascorbate peroxidase	*Capsicum annuum*	109	94	X81376	EMBL
CFR32	Clp-like energy-dependent protease	*Solanum lycopesicum*	119	81	L38581	GB
CLF17	Endochitinase precursor	*Pisum sativum*	77	53	P36907	SP
CLF15	Endochitinase precursor	*Solanum tuberosum*	44	57	P52405	SP
CLF35	Endochitinase	*Castanea sativa*	60	63	U48687	GB
CFB18	Heatshock cognate protein	*Solanum commersonii*	94	94	AF002667	GB
CAN33	Heatshock cognate protein 80	*Lycopersicon esculentum*	107	96	P36181	SP
CFB7	Heatshock protein 83	*Arabidopsis thaliana*	63	87	P27323	SP
CFB33	Heatshock protein	*Arabidopsis thaliana*	64	86	1908431B	PRF
CLF7	Pectinesterase	*Lycopersicon esculentum*	63	78	Z94058	EMBL
CLF3	Prohibitin	*Nicotiana tabacum*	133	87	U69154	GB
CAN38	Proteinase inhibitor II	*Arabidopsis thaliana*	48	79	S30578	PIR
CLF16	Putative HR-like lesion inducing protein	*Nicotiana tabacum*	63	57	U66271	GB
CFR29	Putative osmoprotectant	*Lycopersicon esculentum*	80	92	Z46654	EMBL
CLF21	SA induced mRNA protein product	*Nicotiana tabacum*	90	41	M97194	GB
CFR19	Gamma-Thionin	*Nicotiana tabacum*	49	39	P32026	SP
CFR11	Gamma-Thionin	*Nicotiana tabacum*	84	37	P32026	SP
CFR3	Gamma-Thionin	*Nicotiana tabacum*	35	40	P32026	SP

Table 2. Continued

Clone	Putative identification	Organisms	LC[a]	% Id[b]	Acc. No.[c]	DB[d]
CAN17*	Apoptosis inhibotor IAP	*Drosophila melanogaster*	53	47	Q24306	SP
Transcription / Translation						
CFB56	26S proteasome subunit	*Arabidopsis thaliana*	83	86	U54560	GB
CFB6	40S ribosomal protein S15	*Arabidopsis thaliana*	79	92	Q08112	SP
CFB10	50S ribosomal protein L11	*Spinacia oleracea*	68	81	P31164	SP
CLF42	50S ribosomal protein L40	*Spinacia oleracea*	105	40	P27684	SP
CFR9	60S ribosomal protein L27A	*Arabidopsis thaliana*	82	87	Z17767	EMBL
CFB23	60S ribosomal protein L39	*Zea mays*	51	94	P51425	SP
CFB15	60S ribosomal protein 37A	*Brassica rapa*	64	83	P43209	SP
CFB4	60S ribosomal protein L23	*Nicotiana tabacum*	69	75	Q07760	SP
CFR31	60S ribosomal protein L31	*Nicotiana glutinosa*	120	90	P46290	SP
CFB21	Chloroplast mRNA-binding protein CSP41	*Spinacia oleracea*	75	40	U49442	GB
CLF34	Plastid RNA polymerase σ subunit	*Arabidopsis thaliana*	101	66	AB004820	DDBJ
CFB31	Transcription factor	*Vicia faba*	54	69	X97907	EMBL
CFB24	Threonyl-tRNA synthase	*Arabidopsis thaliana*	87	72	AF007270	GB
CFB48*	40S ribosomal protein S7	*Fugu rubripes*	58	43	P50894	SP
CAN23*	Polypyrimidine tract binding protein	*Mus musculus*	48	40	P17225	SP
CFR37*	Ribosomal protein L7	*Saccharomyces cerevisiae*	84	48	P32495	SP
CFB43*	Transcription factor FKH-4	*Mus musculus*	56	54	Q64733	SP
CFR21*	Transcription factor BAF1	*Kluyveromyces marxianus*	97	44	P33293	SP
Signal transduction						
CFR30	Calmodulin	*Capsicum annuum*	95	97	U83402	GB
CAN21	GTP-binding protein	*Arabidopsis thaliana*	44	75	D89824	DDBJ
CFB19	Protein kinase	*Glycine max*	49	76	S29851	PIR
Cytoskeletal / Structural						
CFB5	Actin depolymerizing factor 1	*Arabidopsis thaliana*	59	61	U48938	GB
CFB16	Cell wall protein	*Lycopersicon esculentum*	61	54	X77373	EMBL
CLF4	Cell wall protein	*Lycopersicon esculentum*	83	56	X77373	EMBL
CFR38	Kinesin-like protein A	*Arabidopsis thaliana*	48	83	Q07970	SP
CFB13*	Su(var)3-9 protein	*Drosophila melanogaster*	42	45	P45975	SP
Others						
CLF10	18S rRNA gene	*Hydrangea macrophylla*	321 bp	94	U42781	GB
CFR17	25S rRNA gene	*Solanum tuberosum*	311 bp	97	X66471	EMBL
CFB53	AP2 domain containing protein	*Arabidopsis thaliana*	92	76	AF003098	GB
CFR2	Apospory-associated protein	*Pennisetum ciliare*	52	88	U13149	GB
CLF39	Basic 7S globulin	*Glycine max*	69	46	D16107	DDBJ
CFB32	Cell wall plasma membrane linker protein	*Brassica napus*	65	64	X94976	EMB
CFR14	GAST1 protein precursor	*Lycopersicon esculentum*	55	98	P27057	SP
CFB46	High mobility group-Y related protein A	*Glycine max*	71	38	Q00423	SP
CFR20	Histone H1	*Nicotiana tabacum*	79	57	S53502	PIR
CFB17	Histone H2B	*Capsicum annuum*	93	90	AF038386	GB
CAN19	Histone H4	*Lycopersicon esculentum*	103	98	P35057	SP
CFB47	Histone H4	*Lycopersicon esculentum*	75	97	P35057	SP
CLF6	Nodulin	*Glycine max*	73	63	Q02121	SP
CAN14	Patatin	*Solanum tuberosum*	117	58	Z27221	EMBL
CAN15	Pollen specific protein NTP 303 precursor	*Nicotiana tabacum*	151	66	P29162	SP
CAN13*	Putative RNA directed RNA polymerase	*Pepper mild motle virus*	215	94	M81413	GB
CAN18*	Putative RNA directed RNA polymerase	*Pepper mild motle virus*	107	88	M81413	GB
CAN5*	Putative RNA directed RNA polymerase	*Pepper mild motle virus*	175	83	M81413	GB
CAN7*	Putative RNA directed RNA polymerase	*Pepper mild motle virus*	212	91	M81413	GB
CFB30*	26S proteasome-associated pad1 homolog	*Mus musculus*	61	80	Y13071	EMBL

*Indicates non-plant matched EST clones. [a]'LC': Length Compared indicates the number of amino acid residues between a query sequence and its matched protein sequence. [b]'% ID: percentage identity at the peptide level. [c]Accession No: accession number of the matched sequences. [d]DB: Database. Database abbreviations: SP, SwissProt; PIR, Protein Identification Resource Data Bank; GB, GenBank.

dentified plant genes. Alternatively, sequencing of the 5' or 3' untranslated region (UTR) may not find homologous genes in database. To overcome this possibility, it may be necessary to run the sequencing gel long enough for targeting the inside of the open reading frame. An estimate of the number of full-length or putative start codon-containing clones can be made using the genes that has previously been known from other organisms, although we did not directly compare the lengths of individual inserts with those of the corresponding mRNA. It was expected that the following 23 ESTs have the full-length open reading frame or the expected first ATG codon based on the sequence alignment: CFB5, CFB15, CFB23, CFB33, CFB36, CFB48, CLF3, CLF4, CLF6, CLF8, CLF15, CLF16, CLF17, CLF 21, CLF30, CLF35, CLF42, CFR7, CFR9, CFR29, CFR31, CAN9, and CAN19.

The putative 95 genes identified by homology search can be classified by their expected functions as shown in Fig. 1. Genes involved in metabolic pathways such as glycolysis and photosynthesis were most abundant, and genes in translational apparatus ranked next in abundance. Various ribosomal protein genes were especially abundant in floral buds and fruits, suggesting that cells in floral buds and young fruits are metabolically active. This observation is consistent with observations previously made in *Arabidopsis* (Höfte *et al.*, 1993), rice (Uchimiya *et al.*, 1992; Sasaki *et al.*, 1994), and *Brassica* (Lim *et al.*, 1996). Eleven resistance- or stress-related genes were identified. These sequences include proteases, endochitinases, heat-shock proteins, thionin, proteinase inhibitor, pectinesterase,

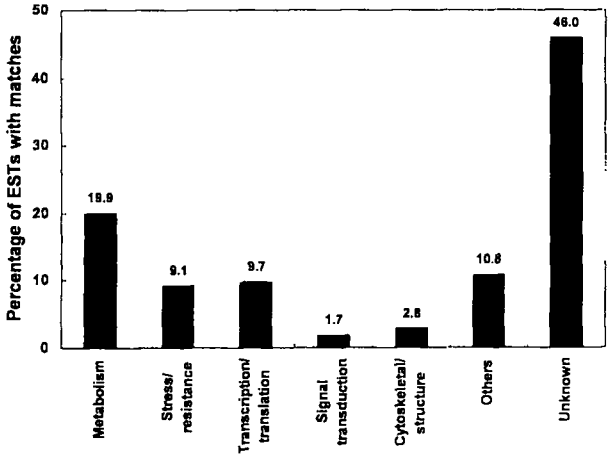

Fig. 1. Functional categorization of the putatively identified ESTs.

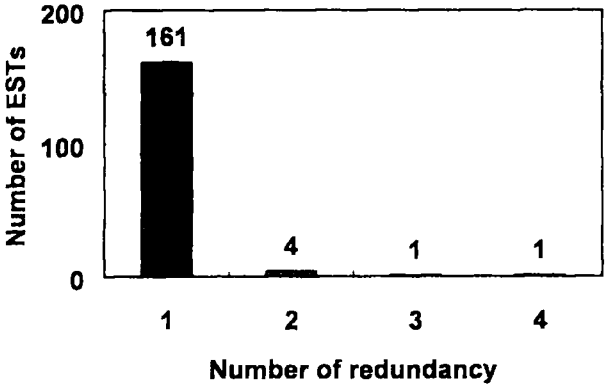

Fig. 2. Redundancy of the hot pepper EST sequences. The numbers on the bars indicate the number of the EST clones in each redundancy class. The EST clones with a nucleotide sequence identity of 90% or more on a 50 nucleotide stretch were considered as redundant clones.

prohibitin, and osmoprotectant. These results may indicate that the plant system expresses many resistance- or stress-related genes to cope with the intensive and various environmental stimuli. We found four redundant EST clones that encode putative RNA directed RNA polymerase of the pepper mild mottle virus (PMMV). It will be necessary to define the source of the mRNA to determine whether the anther tissue was contaminated by the virus. The CFR30 clone was very similar to the previously identified calmodulin cDNA, *CCM1* from *C. annuum* (Kim *et al.*, 1996). Comparison of CFR30 with the *CCM1* clone indicates that CFR30 is a member of calmodulin gene family in the hot pepper. The putative function of some ESTs identified by homology searches will be further evaluated by several methods, including full-sequencing of the clones, *in vitro* assay of protein product, and heterologous expression.

Expression Analysis of EST Clones

To further characterize the EST clones, the expression pattern of the four clones was examined by RNA blot analysis (Fig. 3). The CAN14 clone, which has 58% identity to the potato patatin protein, over a 117 amino acid overlap, was highly expressed in anther tissues but not in fruit tissues. The mRNA was also present in floral buds. These results agree with the previous result that a tobacco patatin gene was highly expressed in petals and anthers (Drews *et al.*, 1992). The CFR2 clone, which has 88% identity to the apospory-associated protein from buffelgrass (*Pennisetum ciliare*), shows relatively higher expression levels in seedlings and roots than

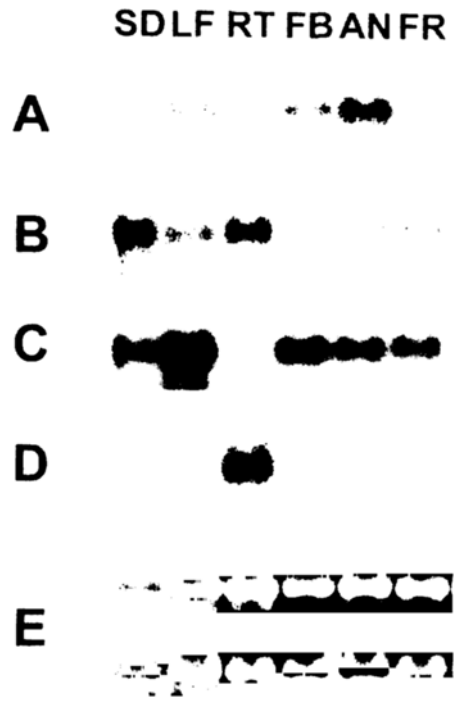

Fig. 3. Northern blot analysis of four ESTs. Ten μg of total RNAs extracted from seedlings (SD), leaves (LF), roots (RT), floral buds (FB), anthers (AN), and fruits (FR) were separated on an agarose-formamide gel. Radiolabeled probes of each clone were hybridized to the RNA blot and exposed to the phosphoimage plate. Exposure times varied for each gel blot. A, CAN14 (putative patatin); B, CFR2 (putative apospory-associated protein); C, CFR11 (putative γ-thionin); D, CFR29 (putative osmoprotectant). E, photograph of EtBr stained rRNA bands.

in other tissues. Whether the CFR2 protein is indeed involved in apomixis remains to be elucidated. A putative γ-thionin (CFR11) mRNA was abundantly present in all the tissues examined except in roots. Thionins have been identified from various tissues of different plant species (Florack and Stiekema, 1994). The *in vitro* toxicity of thionin to plant pathogen has been focused for improving the plant defense system. The CFR29 clone, which is 92% identical to Lemmi9, a putative osmoprotectant induced in tomato roots by nematode attack (Eyeken *et al.*, 1996). RNA blot analysis with this clone showed highly root-preferential expression pattern and barely detected in fruit tissues. This expression pattern of CFR29 is interesting since the clone was obtained from the fruit library.

Redundancy of the EST Clones and Conclusion

To examine the redundancy of the ESTs, the sequences of the 176 EST clones were compared to one another. EST clones with greater than 90% identity over a 50 nucleotide stretch were considered as redundant clones (Kwak *et al.*, 1996). It was shown that 158 ESTs (89.8%) were non-redundant. The relatively low percentage of redundancy in this study may be due to the coverage of several different tissues. If a specific tissue or organ was analyzed, the level of redundancy would have been higher.

The putative homologues of cell wall proteins from the tomato, histone H4, polyphenol oxidase, and endochitinase precursor were represented two times, and γ-thionin three times. It was interesting that although the putative thionin homologues were found to be redundant, two of the ESTs (CFR11 and CFR19) were failed to have significant homology to thionin. This indicates that either the database search algorithm has to be improved or the uniform application of the category may not be sufficient. The most redundant clones (CAN5, CAN7, CAN15, and CAN18) appeared 4 times and showed homology to the putative RNA directed RNA polymerase of PMMV. It should be noted that the clones were obtained exclusively in the anther library.

We could not find amino acid sequence similarities for 81 ESTs (46.1%) to previously registered proteins in database. Functional characterizations of the ESTs will require both biochemical and genetic studies, including full sequencing of the ESTs, analysis of expression pattern using the RNA blot analysis, and expression of the sense or antisense transcript.

The generation and sequence analysis of *C. annuum* ESTs have the goal of producing a resource with significantly enriched information concerning genes that are expressed in a localized fashion. Random nucleotide sequencing of cDNA libraries provided us with an opportunity to isolate various novel genes. EST sequences can be utilized as STSs (sequence-tagged sites), valuable resources for genetic mapping. STSs are genetic markers with nucleotide sequences of 200 to 500 bp which are unique in the genome of an organism (Olson *et al.*, 1989). ESTs can also be new resources for identifying candidate genes on RFLP maps for breeding pepper plants. Moreover, mapped ESTs offer an opportunity to reveal new aspects of regulatory mechanisms in plant gene expression (Park *et al.*, 1993, Umeda *et al.*, 1994). We are currently identifying the functions of some reproductive organ ESTs by *in vitro* and *in vivo* analysis. The results obtained by the analysis will contribute for deepening

the knowledge of the organ development in pepper plants.

The EST sequence data reported will appear in GenBank under the accession numbers of AA 840628-AA840811 and AA842818-AA842826. All the clones and libraries described here are available upon request to S.-R. Kim (E-mail: sungkim@ccs.sogang.ac.kr).

ACKNOWLEDGEMENT

We thank Charm An for critical reading of the manuscript. This research was supported by a grant (G7 08-04-19) from the Ministry of Science and Technology to G. An and from the Ministry of Education (Grant BSRI-96-4413) to S.-R. Kim.

LITERATURE CITED

Adams, M.D., J.M. Kelly, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombieand and J.C. Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**: 1651-1656.

Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelly, T.R. Utterback, J.W. Nagle and C. Fields. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355**: 632-634.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

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

An, C.S., S.C. Kim and S.L. Go. 1996. Analysis of red pepper (*Capsicum annuum*) genome. *J. Plant Biol.* **39**: 57-62.

Collins, M.D., L.M. Wasmund and P.W. Bosland. 1995. Improved method for quantifying capsaicinoids in Capsicum using HPLC. *HortSci.* **30**: 137-139.

Drews, G.N., T.P. Beals, A.Q. Bui and R.G. Goldberg. 1992. Regional and cell-specific gene expression patterns during petal development. *Plant Cell* **4**: 1383-1404.

Eycken, W.V. der, J. de A. Engler, D. Inze, M.V. Montagu and G. Gheysen. 1996. A molecular study of root-knot nematode-induced feeding sites. *Plant J.* **9**: 45-54.

Feinberg, A. and B. Vogelstein. 1983. A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132**: 6-13.

Florack, D.E.A. and W.J. Stiekema. 1994. Thionins: properties, possible biological roles and mechanisms of action. *Plant Mol. Biol.* **26**: 25-37.

Höfte, H., T. Desprez, J. Amselem, H. Chiapello, M. Caboche, A. Moisan, M.F. Jourjon, J.L. Charpenteau , P. Berthomjeu, D. Guerrier, J. Giraudat, F. Quigley, F. Thomas, D.Y. Yu, R. Mache, M. Raynal, R. Cooke, F. Grellet, M. Delsenu, Y. Parmentier, G.D. Matcjillac, C. Gigot, J. Fleck, G. Philipps, M. Axelos, C. Bardet, D. Tremousaygue and B. Lescure. 1993. An inventory of 1,152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. *Plant J.* **4**: 1051-1061.

Kim, B.D., B.C. Kang, S.H. Nam, B.S. Kim, N.S. Kim, M.H. Lee and K.S. Ha. 1997. Construction of a molecular linkage map and development of a molecular breeding technique. *J. Plant Biol.* **40**: 156-163.

Kim, S., Y.H. Kang, Z.-W. Lee, B.-D. Kim and K.S. Ha. 1997. Analysis of chemical constituents in fruits of red pepper (*Capsicum annuum* L. cv. Bugang). *J. Kor. Soc. Hort. Sci.* **38**: 384-390.

Kim, S.R., S.A. Kim and H.J. Kwak. 1996. Isolation and characterization of a hot pepper calmodulin cDNA clone. *Mol. Cells* **6**: 753-758.

Kwak, J.M., S.A. Kim, M.S. Soh, Y.S. Park, E.S. Shin, Y.J Kim, I.C. Kwun and H.G. Nam. 1996. Characterization of 457 expressed sequence tags generated from root cDNA clones of *Brassica napus* by single-pass sequencing. *Mol. Cells* **6**: 563-570.

Lim, C.O., H.Y. Kim, M.G. Kim, S.I. Lee, W.S. Chung, S.H. Park, I.H. Hwang and M.J. Cho. 1996. Expressed sequence tags of Chinese cabbage flower bud cDNA. *Plant Physiol.* **111**: 577-588.

Olson, M., L. Hoo, C. Cantor and D. Botstein. 1989. A common language for physical mapping of the human genome. *Science* **245**: 1434-1435.

Park, Y.S., J.M. Kwak, O.Y. Kwon, Y.S. Kim, D.S. Lee, M.J. Cho, H.H. Lee and H.G. Nam. 1993. Generation of expressed sequence tags of random root cDNA clones of *Brassica napus* by single-run partial sequencing. *Plant Physiol.* **103**: 359-370.

Park, H.G. 1992. Current status, problems, and prospects of hot pepper industry in Korea. *J. Kor. Capsicum Res. Coop.* **1**: 1-12.

Rounsley, S.D., A. Glodek, G. Sutton, M.D. Adams, C. Somerville and J.C. Venter and A.R. Kerlvage. 1996. The construction of *Arabidopsis* expressed sequence tags assemblies: a new resource to facilitate gene identification. *Plant Physiol.* **112**: 1177-1183.

Sambrook, J., E.F. Fritsch and T. Maniatis. 1989. Molecular cloning: A laboratory Manual. 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Sanger, F., S. Nicklen and A.R. Coulson. 1977. DNA sequence with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**: 5463-5467.

Sasaki, T., J. Song, Y. Koga-Ban, E. Matsui, F. Fang, H. Higo, H. Nagasaki, M. Hori, M. Miya, E. Murayama-Kayano, T. Takiguchi, A. Takasuga, T.

Niki, K. Ishimaru, H. Ikeda, Y. Yamamoto, Y. Mukai, I. Ohta, N. Miyadera. I. Havakkala and Y. Minobe. 1994. Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNAs from a callus cDNA library. *Plant J.* **6**: 615-624.

Shen, B., N. Carneiro, I.T. Jerez, B. Stevenson, T. McCreery, T. Hellentjaris, C. Baysdorfer, E. Almira, R.J. Ferl, J.E. Habben and B. Larkins. 1994. Partial sequencing and mapping of clones from two maize cDNA libraries. *Plant Mol. Biol.* **26**: 1084-1101.

Somerville, C. and S. Somerville. 1997. The *Arabidopsis* genome project. 5th ISPMB meeting, Singapore.

Uchimiya, H., S. Kidow, T. Shimazaki, S. Aotsuka, S. Takamatsu, R. Nishi, H. Hashimoto, Y. Matsubayashi, N. Kidou, M. Umeda and A. Kato. 1992. A random sequencing of cDNA libraries reveals a variety of expressed genes in cultures cells of rice (*Oriza sativa* L.). *Plant J.* **2**: 1005-1009.

Umeda, D., C. Hara, Y. Matsubayashi, H.H. Li, Q. Liu, F. Tadokora, S. Aotsuka and H. Uchimiya. 1994. Expressed sequence tags from cultured cells of rice (*Oriza sativa* L.) under stressed conditions: analysis of transcripts of gene engaged in ATP-generating pathways. *Plant Mol. Biol.* **25**: 469-478.

Wolfsberg, T.G. and D. Landsman. 1997. A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res.* **25**: 1626-1632.

Received March 16, 1998
Accepted April 7, 1998