KMITL Sci. Tech. J. Vol. 7 No. 1 Jan. - Jun. 2007

STUDY ON BIO-CONTROL RELATED GENES OF *Trichoderma* sp. AND *Chaetomium* spp.

Yang Qian*, Cong Hua, Zhang Haiyan, Yao Lin, Liu Pigang and Jin Hongxing

Department of Life Science and Engineering, Harbin Institute of Technology,
Harbin, 150001 P R China

ABSTRACT

In order to identify the bio-control related genes from *Trichoderma harzianum, Chaetomium cuperium* and *C. globosum*, the expressed sequence tags (ESTs) of these plant disease bio-control fungi were determined. The cDNA libraries of those fungi were constructed from their mycelia and induced cultures. ESTs of the fungi were sequenced. Analysis of those sequences showed that more than 100 genes from the fungi are plant disease bio-control related genes although more possible bio-control related genes are still to be investigated. The biocontrol related genes are very important for new bio-fungicide development in the future.

KEYWORDS: Bio-control related genes, expressed sequence tags, *Trichoderma harzianum*, *Chaetomium cupreum*, *Chaetomium globosum*

1. INTRODUCTION

Trichoderma harzianum, *Chaetomium cupreum* and *C. globosum.* are fungi with a world wide distribution. Their potential in the biological control of plant disease are well known. Their bio-control mechanisms include producing antibiotics and ergosterols compounds that can suppress different plant pathogens, especially those soil bone plant pathogenic fungi, stimulate growth of plants and induce resistance to the diseases [1-3]. They can be used to control many soil borne diseases of vegetable and fruits. However, although they are very effective plant disease bio-control microorganisms, since they are very sensitive to environmental condition, so that they are hardly used in the field where the conditions are not suitable for their survival.

In order to develop new technique of applying these fungi in plant disease bio-control, studying their genetic background especially their genome are essential. Therefore the EST of the fungi were sequenced and analyzed in this study. The information achieved in the study will be very useful for developing new techniques of using the bio-control related gene products to produce a new generation of bio-fungicides. In this way the bio-control program of plant diseases with these fungi could be widely applied to ensure organic food production and the sustainable development of agriculture all over the world.

*Corresponding author.: Tel: +86-451-86414342 Fax: +86-451-86412952
E-mail : yangq@hit.edu.cn

2. MATERIALS AND METHODS

2.1 Materials

T. harzianum and *Chaetomium* spp. (*C. cupreum* and *C. globosum*) were collected from Heilongjiang Province, P R China and supplied by Dr. Kasem Soytong, KMITL Thailand respectively. They were stored at 4℃ in the laboratory of Life Science and Engineering Department, Harbin Institute of Technology, P R China.

2.2 RNA isolation and cDNA library construction of *T. harzianum, C. cupreum* and *C. globosum.*

T. harzianum, C. cupreum and *C. globosum* were cultured in Potato Dextrose (PD) medium for 36h, 60h and 3 days respectively, then the mycelia were collected and ground under liquid nitrogen using a mortar and pestle. Total RNA was extracted by using guanidium-isothiocyanate method, polyATtract® mRNA isolation System (Promega). First-trand cDNA was synthesized using an oligo-dT linker-primer with a *Xho*I cloning site. After the other end of each cDNA was ligated to an adaptor with an *EcoR*I-compatible overhang, cDNA was ligated directionally into the *EcoR*I and *Xho*I sites of the pBluescript II sk(+) vector (Stratagene).

2.3 EST sequencing

The plasmid library of the cDNA from *T. harzianum, C. cupreum* and *C. globosum* were used to transform *Escherichia coli*. The bacteria were grown for 60 min at 37^{o}C and then plated out onto solid Yeast Tryptone (YT) medium and single plaque randomly picked and stored in YT buffer at 37^{o}C. The insert sizes of individual recombinant plaques were examined by specific PCR amplification by means of the T_3 reverse and T_7 primers followed by 1.5% (w/v) agarose gel electrophoresis. Templates for the ESTs from the mycelium library were removed and inoculated into a 10 ml overnight culture of Luria broth containing 50 mg/ml ampicillin. Plasmid DNA was isolated from a 5 ml aliquot of the overnight culture using rapid plasmid isolation protocol (Qiagen) and purified through QIAquick spin columns (Qiagen). Templates for DNA sequencing were prepared by specific PCR amplification of cDNA inserts directly by means of the T_3 primers. Amplified inserts were purified with QIAquick spin columns (Qiagen) prior to sequencing. Amplified insert cDNA were sequenced by MegaBASE1000 (Amersham Pharmacia Biotech) according to the manufacturer's instructions.

2.4 EST sequence analysis

The sequences of the 3 fungi were first screened by *Phred* in combination with a manual inspection to ensure quality. The sequences were then processed by *Crossmatch* to filter repetitive and vector sequences. Sequences representing rRNA, mitochondrial DNA and shorter than 100 bp were discarded. Through comparison of the deduced amino acid sequences with public protein databases such as non-redundant peptide and Swiss-Prot, a gene classification was assigned by using a criterion of homology of >30% identity in a sequence of 100 amino acids as well as a cut-off of E-value lower than 10 in a Blast search [4]. This step was followed by a nucleotide homology search in the EST database to give the expression profile of each gene. ESTs with known gene matches were categorized into different functional groups according to categories [5]. Relative levels of gene expression were computed by summing the number of ESTs matching that particular gene and dividing the sum by the total number of ESTs that match known genes. The ESTs were clustered on the basis of sequence similarity by using *Phrap*. Parameters were set so that ESTs were connected only with a minimum of 95% nucleotide identity in an overlap region of 40 nucleotides. The EST sequence of *T. harzianum, C. cupreum* and *C. globosum* were deposited in the GenBank and DDBJ databases respectively.

3. RESULTS

3.1 RNA isolation and characteristics of the constructed mycelium cDNA library

Mycelia of the 3 fungi were used for mRNA isolation before they were squeezed to complete dryness, since the high density of the mycelia made it difficult to remove the polysaccharide from the cell walls, which could affect purification of the mRNA. The titers of the constructed mycelium cDNA library of *T. harzianum, C. cupreum* and *C. globosum* were 1.2×10^6 pfu/mL (unamplified), 0.93×10^6 pfu/mL and 2.25×10^4 pfu/mL (unamplified) respectively. These titers were considered to be reasonable and sufficiently representative for analysis of the expressed genes presenting in the mycelia. Blue/white plaque selection following incubation of an aliquot of the library revealed 93-96.4% recombinant plaques. The quality of the library was assessed by examining the insert sizes of 32 randomly selected recombinant plaques from *T. harzianum*, of 32 randomly selected recombinant plaques from *C. cupreum* and of the 384 selected plaques from *C. globosum* by specific PCR amplification with T_7 and T_3 reverse primers.

Of the 32 selected plaques from *T. harzianum*, only one without inserts from *T. harzianum* was found. Insert sizes ranged between 800 and 4000 bp with an average insert size of 1200 bp. All the plaques from *C. cupreum* had inserts ranging in size from 500 bp to 3000 bp, most were above 1.2kb. Of the 384 selected plaques from *C. globosum*, insert sizes ranged between 400 and 3000 bp with an average insert size of 1000 bp. The results provided that the cDNA clones accurately represent the mycelia mRNA population, the complexity of the library should reflect that of the mycelium transcriptome at this stage and should be sufficient for identifying all expressed genes of the 2 fungi in mycelium stage.

3.2 Generation and annotation of expressed sequence tags

For generation of the ESTs, only clones with an insert sequence larger than 500 bp of *T. harzianum* were selected for sequencing from 5′ends of the cDNA. To date, 4128 clones have been subjected to single-run partial sequencing by specific PCR amplification of insert DNA using T_3 primers and 3298 high quality sequences with a minimum of 100 bp of continuous sequence were achieved for further analysis after removing low-quality sequences and sequences representing ribosomal, mitochondrial and repetitive sequences. The 3298 high-quality sequences were integrated into 1740 clusters, of which 964 clusters (2174 sequences) represented known genes by comparison to the GenBank non-redundant protein database, 225 clusters (451 sequences) matched dbEST entries and 551 clusters (673 sequences) showed no homology with nucleotide sequences deposited in the public databases and thus were considered as novel ESTs (Table 1). Of all known genes, 362 clusters (410 sequences) and 278 clusters (319 sequences) were homologous to *N. crassa* and *A. nidulans* respectively. Among all the clusters, 1277 were singlets while the other 463 were contigs composed of multiple ESTs ranging from 2 (210 clusters) to 85 (one cluster). The redundancy rate for protein encoding ESTs was 36.3%. Of these contigs, about 69% (320 contigs) consisted of only 2 or 3 EST clones and only 9 contained more than 20 EST clones. The high percentage of low redundancy sequences indicated the complexity of the cDNA library, suggesting a relatively good representation of the library. Since the library was neither subtracted nor normalized in any way, the number of clones derived from an individual gene could approximate the expression level of that gene in the mycelium stage. High redundancy of a specific cDNA sequence among ESTs is likely to be correlated with a higher expression level of the corresponding gene. Therefore, the contigs containing the highest number of ESTs were listed in Table 2 as highly expressed genes. The genes most frequently explored among the ESTs were the cell wall protein QID3 precursor, the Woronin body major protein, DNA damage-responsive protein and oxidoreductase ywfD as described in Table 2. Elongation factor 1-alpha and eukaryotic translation initiation factor 5 were also highly expressed, which may represent a potential for these cells to enter the cell cycle, leading to differentiation in response to a physiological requirement. However, the gene number represented by the 3298 ESTs was less than

the number of individual cluster groups. This is because ESTs from different regions of the same gene did not fall into the same cluster, such as contig426 and contig451 shown in Table 2. In addition, alternatively spliced forms of a single gene could also be grouped into different clusters.

Table 1 *Trichoderma harzianum* EST sequence analysis

Items of the analysis	Number of the analyzed ESTs
ESTs submitted to GenBank	3298
Total reading length	1748 kb
Average length per EST	446 bp
G +C content	53.16%
Homology to nr database	2174
Known ESTs	451
Unknown ESTs	673
Unigene number	1740
Contig number	463
Singlet number	1277
ESTs/contig	4.37
Known unigenes	964

Table 2 Known high redundancy genes

Contig no.	Redundancy	Putative function
Contig447	85	Cell wall protein QID3 precursor
Contig452	75	Hypothetical oxidoreductase
Contig451a	58	Woronin body major protein
Contig448	58	DNA damage-responsive protein
Contig450	47	Norsolorinic acid reductase
Contig445	25	Elongation factor 1-alpha
Contig444	23	Retinol dehydrogenase 11
Contig442	20	ADP, ATP carrier protein
Contig439	17	Acyl-CoA desaturase 1
Contig462	17	Mucin 2 precursor
Contig437	16	Thiazole biosynthetic enzyme
Contig433	13	Superoxide dismutase
Contig431	12	Eukaryotic translation initiation
Contig429	11	Peptidyl-prolyl cis isomerase
Contig430	11	IPC-B hydroxylase
Contig423	10	3-Ketoacyl-acyl carrier protein reductase
Contig426b	10	Woronin body major protein
Contig428	10	Retinol dehydrogenase

a, b mean Contig451 and contig426 match to slightly difference regions of the Woronin body major protein, indicating these contigs may represent the same gene.

With *C. cupreum*, the cDNA clones with insert size more than 700 bp were selected for sequencing using T3 primers. 3069 high quality sequences with a minimum of 100 bp were achieved from 3647 cDNA clones for further analysis after removing sequences representing ribosomal, mitochondrial and vector sequences. Minimum, average, maximum lengths of high

quality sequences were 102, 501, 773bp respectively. Average G+C content was 54.76% with the range from 9.83% to 75.44% which was similar to other filamentous fungi. Using Phrap and Phred program, 3069 ESTs were assembled into 1471 Unigenes with 392 contigs and 1079 singlets. BlastX analysis of the high quality sequences revealed that 874 (59.4%) ESTs could be assigned a putative identity based on strong sequence homology to protein in the GenBank. Since non-normalized primary cDNA libraries were used, the number of cDNA clones derived from mRNA could reflect the expression level of that gene in the mycelial growth stage. The contigs containing 10 or more ESTs were listed in Table 3.

To achieve high quality of the ESTs, only the clones with an insert longer than 400 bp of *C. globosum* were selected for sequencing. The BLAST result of each comparison was screened manually. Sequences shorter than 100 bp were excluded. 3141 EST fragments were selected after the screening. Putative identification of the EST fragments was attributed to each EST fragment. 3141 of them were then pieced together and 1381 sequences were created, in which 387 were contigs and 994 were singlets. 513 sequences were annotated, indicating that they were all known genes, after removing the repeated sequences 466 genes left; the 868 sequences un-annotated were new genes.

Table 3 Assembled contigs containing more than 10 ESTs from *C. cupreum*

Contig no.	***Redundancy***	***Putative Function***
Contig428	109	glyceraldehyde-3-phosphate dehydrogenase
Contig63	72	coproporphyrinogen oxidase
Contig30	56	predicted protein
Contig433	54	C-4 sterol methyl oxidase
Contig313	36	hypothetical protein
Contig2	27	pyruvate decarboxylase
Contig39	23	hypothetical protein
Contig390	21	xylulose-5-phosphate/fructose-6-phosphate phosphoketolase
Contig392	19	EF1-alpha translation elongation factor
Contig425	19	stress-inducible protein sti35（Thiazole biosynthetic enzyme）
Contig421	17	Actin
Contig420	17	glutamine synthase
Contig422	16	aspartic protease
Contig371	16	ammonium transporter
Contig419	16	ATP synthase protein 9
Contig6	15	ADH3_EMENI alcohol dehydrogenase
Contig417	15	hypothetical protein
Contig47	14	β-1,3-exoglucanase
Contig317	14	ATP citrate lyase
Contig412	13	histone H2B
Contig298	13	pyruvate kinase
Contig410	11	aspartate aminotransferase
Contig408	11	heat shock protein 30
Contig411	11	ADP-ATP translocase
Contig407	11	cyclophilin

All the ESTs of the fungi from this study have been submitted to the GenBank and DDBJ respectively.

3.3 Functional identification of the EST genes

Taking the clustering results as a guide, the ESTs of *T. harzianum* in the database were annotated and classified based on their known or inferred function. During this process, clusters representing the same gene were merged into a single group. All identified ESTs were categorized into three groups, the cell component, molecular function and biological process respectively. The mycelium cDNA clones exhibited homology to a broad diversity of genes, including enzymes and proteins. The largest numbers of clones (38.5%) were associated with biological processes including physiological, developmental and cellular processes. Of the remaining clones that were identified, 25.9% was found to encode various proteins such as cell component, which implies a great need for materials to produce hyphae for the mycelium stage. A further 35.6% of clones were involved in molecular function, which included 286 clones of catalytic activity, 148 binding proteins and a variety of activity proteins such as toxin activity, defense/immunity protein activity, cell adhesion molecule activity and antioxidant activity. Interestingly, some obsolete function genes were found among these categories, which indicated a potential function maintained by *T. harzianum,* and these data might be beneficial for exploring the evolutionary process.

Of the ESTs from *C. cupreum,* the most prevalent contig consist of 109 ESTs and had significant homology to glyceraldehyde-3-phosphate dehydrogenase. The second most abundant contig (63 ESTs) was homologous to coproporphyrinogen oxidase. However, the third most abundant contig showed no homology with any known genes in the database representing predicted protein. Contigs had significant homology to fungal genes in the GenBank and they encode proteins with interesting functions containing C-4 sterol methyl oxidase (Contig433, 54 copies), aspartic protease (Contig422, 16 copies), β-1,3-exoglucanase (Contig47, 14 copies), which could play significant role in antagonisms of plant pathogens. Glyceraldehyde-3-phosphate dehydrogenase, pyruvate kinase, pyruvate decarboxylase and EF1-alpha translation elongation factor were also high expressed which indicated that mycelia was in the process of metabolism fastigium. Of the unigenes, approximately 597 (40.6%) ESTs showed no or low similarity to protein sequences in the NCBI database according to BlastX algorithm. The isolation of expressed ESTs with no annotated function is potentially an important find since abundantly expressed genes are likely to have a key role in biocontrol function.

Results from the analysis of the 1334 sequences from *C. globosum* revealed that 27.2% of them were homologous to peptide sequences and nucleotide sequences present in the NCBI protein and nucleotide databases. Of the rest of 72.8%, 65.1% of the sequences were not similar to any sequence on the database according to the search criteria used in this study, and thus were interpreted as possibly representing new genes.

4. DISCUSSION

T. harzianum, C. cupreum and *C. globosum* mycelium cDNA librarys were constructed and the complexities of the librarys are over 1.2×10^6 PFU/mL, 0.93×10^6 pfu/mL and 2.25×10^4 pfu/mL (unamplified) respectively. 4128 clones from the cDNA library of *T. harzianum* were randomly sequenced and thereby identified 1740 putative genes of the fungus. It has been estimated that there are 5900 genes encoded by the *S. cerevisiae* genome, 8100 by the *A. nidulans* genome and 9200 in the *N. crassa* genome [6-7]. *T. harzianum*, belonging to the Hypocreales order of the Ascomyceta division, is most closely related to the Pyrenomycete *N. crassa,* and assuming similar genome size and conservation of gene number, the ESTs generated in this study represent approximately one quarter of the genes encoded by the *T. harzianum* genome. Although closely related gene families may not be distinguished by separate contigs, numbers of genes predicted by contig and singlet analysis are usually overestimated because non-overlapping sections of genes may be sequenced [8]. Genomics-based approaches such as the EST analysis reported here provide an efficient means of gene discovery, particularly for organisms such as *T. harzianum, C.*

cupreum and *C. globosum*, for which there have been very few molecular and genetic studies before this study. Consistent with the databases of other fungal EST studies [9-10], in the current study, nearly half of the translated cDNA sequences from the EST collection showed no significant similarity to known protein sequences. It is therefore likely that many new fungal genes may be discovered from some of these sequences.

The strong representation of ESTs with similarity to *N. crassa* (37.6% of known unigenes) undoubtedly reflects the large amount of sequence information available for *N. crassa* compared to that for filamentous fungi. Nearly one quarter of ESTs encoded homologues of proteins involved in the cellular component, which indicates active cell growth and division in the mycelium stage.

Plant disease bio-control related genes were achieved from the mycelia of *T. harzianum, C. cupreum* and *C. globosum*. In the mycelium collection of EST sequences of *C. globosum, T. harzianum* and *C. cupreum*, 8, 55, 142 bio-control related genes have been isolated respectively. The results indicated that the technique of EST analysis influence the functional gene, especially the bio-control related genes isolation a lot. Since none of the antibiotics production related genes has been carried out, there will be more bio-control related genes being found in the near future.

The use of an integrated biological control mechanism, in which several tactics are employed to combat the same pathogen, is a promising approach to improve the disease control and the consistency of the biological treatment. Myco-parasitism is a complex process including release of hydrolase, degradation of the cell wall and further penetration into the host mycelium. These enzymes can hydrolyze the substrate by diverse mechanisms. Chitinases are hydrolytic enzymes that are responsible for the degradation of the chitin of those plant pathogenic fungi to its monomer *N*-acetyl-D-glucosamine. In microorganisms, chitinases are known to be involved in autolysis, spore germination, branching and mycelia development, hyphal growth, cell separation, nutrition and parasitism [11]. Chitinases are used extensively in biological research for the generation of fungal protoplasts due to its ability to degrade the fungal cell wall. Class V chitinase and glycoprotein gp2, as glycosyl hydrolases, can effectively hydrolyze the 1,4-beta-linkages of *N*-acetyl-D-glucosamine polymers of chitin.

In order to further study all of these potential biocontrol genes and therefore utilize them in practice, full-length cDNA clones of some genes have been achieved by rapid amplification of cDNA ends. The results of this analysis provide a first look at global gene expression in *T. harzianum* mycelium. In an analysis of more than 3298 ESTs from the mycelium cDNA library, more than 55 biocontrol-associated sequences were definitively identified. The cDNA libraries generated from the mycelium stage, and corresponding EST collections, will therefore be important resources for further investigations aimed at gaining a better understanding of the molecular events that occur during mycelium development in *T. harzianum C. cupreum* and *C. globosum.*

This EST studies on these plant disease bio-control fungi supplied huge amount of gene resources for promoting the bio-control mechanisms study and new bio-fungicides development. In the near future, the study on verification of differential expression and functional characterization of selected plant disease bio-control related genes will be carried out, in order to have those fungi genetic resources applied in the practice of plant disease bio-control in the field instead of staying as laboratory curiosities. The work on the technique of formulation of those bio-control related gene products will also be carried out in future.

5. ACKNOWLEDGMENT

The gratitude is expressed to the High Tech Research and Development Program of China (863 Program) for it's financial support.

REFERENCES

[1] Soytong K, Soytong K. **1997** Chaetomium as a new broad spectrum mycofungicide. Proc. of First International Symposium on Biopesticide. 124–132.

[2] Yang Q, Song J Z, Liu L Q, *et al*. **2000** A study on biocontrol mechanism of *Chaetomium* spp. Advanced study on plant pest biological control. Heilongjiang science and technology press. 110–115.

[3] Dipietro A, Gut-rella M, Pachlatko J P, *et al*. **1992** Role of antibiotics produced by *Chaetomium globosum* in biocontrol of Pythiumultimum, a casual agent of damping-off. Phytopathology, 182(2): 131–135.

[4] Altschul S F, T L Madden, *et al*. **1997** Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res. 25: 3389–3402.

[5] Ashburner M, *et al*. **2000** Gene ontology: Tool for the unification of biology, the gene ontology consortium, Nat. Genet. 25: 25–29.

[6] Kelkar H S, J Grifith, M E Case. 2001. The Neurospora crassa genome: Cosmid libraries sorted by chromosome. Genetics. 157:979–990.

[7] Kupfer D M, C A Reece, S W Clifton, B A Roe, R A Prade. **1997** Multicellular ascomycetous fungal genomes contain more than 8000 genes, Fungal Genet. Biol. 21: 364–372.

[8] Fernandes J, V Brendel. **2002** Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and microarray hybridization. Plant Physiol. 128: 896–910.

[9] Keon J, A Bailey, J Hargreaves. **2000** A group of expressed cDNA sequences from the wheat fungal leaf blotch pathogen, *Mycosphaerella graminicola* (*Septoria trici*), Fung. Genet. Biol. 29: 118–133.

[10] LI Yan-song, Yang Qian. **2006** Establishment of bioinformation analysis database of Chaetomium globosum EST. China Journal of Bioinformatics. 22-25.

[11] Bidochka M J, S Burke. **1999** Extracellular hydrolytic enzymes in the fungal genus *Verticillium*: Adaptations for pathogenesis, Can. J. Microbiol. 45: 856–864.