Cocoa and Oil palm putative gene database

R. Manimekalai, V.H. Rakesh, S. Anoop and George V. Thomas

Central Plantation Crops Research Institute,
Kasaragod - 671 124, Kerala

(Manuscript Received: 04-01-10, Revised: 11-11-10, Accepted: 19-50-11)

Abstract

Oil palm and cocoa are important plantation crops and require bioinformatics tools to hasten the research output and aid in crop improvement programmes. The current work was undertaken to assign putative function to available Expressed Sequence Tags (EST's) of oil palm and cocoa. Annotated EST's of cocoa and oil palm were developed into searchable database. EST's of oil palm and cocoa were first retrieved from dbEST. FASTA formatted EST sequences were converted into contigs by running in CAP3. The contigs sequences were run in BLAST tool and their putative functions were predicted based on homology. A database of annotated ESTs was developed using MySQL and PHP programs. In this database, EST's of cocoa and oil palm, BLAST results and gene information were stored as different tables. The database homepage contains six menus namely 'Home', 'About database', 'Tool', 'Useful links', 'Site map' and 'Contact us'. The same page contains annotated gene information for cocoa and oil palm separately. For browsing the annotated ESTs of cocoa and oilpalm, separate text boxes are provided such as 'ESTs', 'blast results' and 'gene information'. The text box 'EST's' of oil palm has links to six different tables which stores information about six different tissues and cocoa contains seven different tables, which stores information about seven different tissues. The 'gene information' contains the 'contig number', 'similarities found in each organism', 'accession number', 'structure accession number' and 'gene function'. The cocoa and oil palm putative gene database - COPGENE is hosted at CPCRI bioinformatics website (www.bioinfcpcri.org).

Keywords: Cocoa, ESTs, database, oil palm

Introduction

Genome annotation is the process of marking genes and other biological features in a DNA sequence. Different software systems are used to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

Cocoa (*Theobroma cacao*) is a small (4–8 m tall) evergreen tree in the family Sterculiaceae (alternatively Malvaceae). Its beans are used in chocolate industry. Oil palm is a monocotyledonous plant of the palm family (Arecaceae). Oil palm is used in commercial agriculture in the production of palm oil. Enormous Expressed sequence Tags (ESTs) are available for both cocoa and oil palm.

A systematic study of genes expressed by means of EST analysis in oil palm has been undertaken (Jouannic *et al.*, 2005). Argout *et al.* (2008) generated 149 650 ESTs

of cocoa from various tissues under different conditions. Transcriptomic changes in oil palm upon the inoculation of *Bacillus sphaericus* showed up regulation of cyclophilin, ribosomal proteins and genes related to protein synthesis and processing and membrane transport (Lim *et al.*, 2010).

Annotated databases are the result of the applications of tools and software's for analysis of genomes and large amount of available scattered data. Annotated databases are designed to provide public access to ongoing research focused on the biotechnological aspects of both plants and animals.

The genome database GrainGenes (Gerard *et al.*, 1998) is a useful tool for collecting and displaying research information for the cereals community. The sequences were compared against NCBI genome data banks using BLAST search programs.

POGs/PlantRBP (http://plantrbp.uoregon.edu/) is a relational database (Walker *et al.*, 2006) that integrates data from rice, Arabidopsis and maize by placing the complete Arabidopsis and rice proteomes and available

maize sequences into 'putative orthologous groups' (POGs). Transport DB (http://www.membranetransport.org/) is a comprehensive database on cytoplasmic membrane transporters and outer membrane channels in organisms whose complete genome sequences are available (Ren *et al*., 2004). Repbase is a reference database (Kohany *et al*., 2006) of eukaryotic repetitive DNA, which includes prototypic sequences of repeats and basic information described in annotations. The MGOS (*Magnaporthe grisea Oryza sativa*) web-based database (Greer and Soderlund, 2009) contains data from *Oryza sativa* and the fungal pathogen *Magnaporthe grisea* interaction. Chloroplast Genome database (Cui *et al*., 2005) contains annotated chloroplast/plastid genomes from the NCBI Organelle Genomes section at NCBI. The cocoaGen DB at Tropgene database has genetic, genomic and phenotypic data of cocoa (Ruiz *et al*., 2004).

A comprehensive database is not available for oil palm and cocoa regarding the functional annotation of ESTs. The main objective of this work was genome analysis and annotation of oil palm and cocoa EST's using functional annotation methods and development of user friendly comprehensive database on annotated ESTs of cocoa and oil palms. The COPGENE annotated database provides information on contigs, EST's and predicted functions of genes for cocoa and oilpalm. Functional annotation will help in development of functional markers as well as prediction of important genes for biotic or abiotic resistance in oil palm and cocoa.

Materials and Methods

Sequence retrieval and categorization

Oil palm and cocoa EST's were retrieved from dbEST (Division of GenBank). It contains thousands of EST's of many organisms. These retrieved sequences were saved in text format. The tissues from which oil palm EST's developed are categorized as abnormal apex (998 ESTs), normal apex (313 ESTs), male inflorescence (625 ESTs), female inflorescence (349 ESTs) and immature zygotic embryos (126 ESTs). Cocoa ESTs are categorized into seven types; bean and leaves (4443 ESTs), defense related from cocoa leaves (2114 ESTs), differential display (6 ESTs), immature zygotic embryo (1 EST), mature zygotic embryo (4 ESTs), young red leaves (9 ESTs) and somatic embryo (4 ESTs).

Preparation of Contigs

EST sequences were converted to FASTA (Fast-Alignment Tool) format. FASTA formatted sequences were clustered and the conserved sequences were made into contigs by running in CAP3 (Contig Assembly Program 3). The CAP3 results were saved in text format.

Prediction of function

BLASTx was used to find similarity of unknown EST's against nr (non redundant) database of plants. The contigs in each tissue were run in BLASTx and BLASTn. In BLASTn the search was done against three organisms individually viz. *Arabidopsis thaliana, Oryza sativa* and *Zea mays.* In BLASTx, the search was done against Viridiplantae database. From the BLASTx and BLASTn results, the hits with high score were chosen.

From the above results organisms having high score were selected by comparing with BLASTx and BLASTn. Maximum identity and E value were also taken into account while comparing the score. The selected hit (organism) was saved in HTML format. Organisms with best score were identified for predicting their gene function based on homology. Comparative analyses were performed to predict gene functions and also the function of EST's.

Development of annotated database

MySQL and PHP programs were used for development of a database containing information on EST's, BLAST results and gene information of cocoa and oil palm. This information was stored in six different tables (three for cocoa and three for oil palm).

Results and Discussion

Expressed sequence Tags or EST's provide researchers with a quick and inexpensive route for discovering new genes, for obtaining data on gene expression and regulation and for constructing genome maps. The primary step of genome sequencing project is to retrieve the EST sequences of respective organisms. Identification of functions of genomic sequences is a challenging task. It varies among organisms and depends upon genome size as well as the presence or absence of introns.

Massive sequence databases provide protein and nucleic acid sequences used for genome annotation studies. Various useful annotated databases were created for retrieval and identification of functional genes in different organisms. The methodology behind all these databases were one and the same i.e. genome analysis (Gerard *et al*., 1998).

The present work focuses on genome analysis and annotation of oil palm and cocoa ESTs and to develop an annotated database that provides information on contigs, EST's, gene information and sequence similarity. The ESTs were prepared into contigs (Table 1). The contigs were subjected to BLAST analysis and the results were developed into searchable database.

Table 1. The number of contigs obtained for oil palm

		Tissue Category	Contigs
1	Oil palm	Abnormal apex	86
2		Normal apex	13
3		Female inflorescence	16
4		Male inflorescence	44
5		Immature zygotic embryo	9
1	Cocoa	Bean and leaves	424
2		Defense related from cocoa leaves	359
3		Differential display	0
4		Immature zygotic embryo	0
5		Mature zygotic Embryo	0
6		Young red leaves	1
7		Somatic embryo	0

Home Page of the database

The home page contains the information about cocoa and oil palm ESTs. It has six menus for 'Home', 'About database', 'Tools used' in the database, 'Useful links', 'Site map' and 'Contact us'. The same page has annotated gene information for cocoa and oil palm separately. BLAST results of contigs and predicted gene information are provided using separate text boxes. For browsing the annotated ESTs in oil palms, three text boxes were provided such as 'ESTs', 'BLAST results' and 'gene information'. Fig 1 shows the view of home page of the COPGENE database.

Fig. 1. Database home page

Oil palm-EST results

EST's of oil palm contains five different tables, which stores information about five different tissues. EST's from each table contains information about EST's in FASTA format and number of contigs present in each tissue (abnormal apex, male inflorescence, female inflorescence, normal apex and immature zygotic embryo).

Cocoa-EST results

EST's of cocoa contains seven different tables for storing the information of seven different tissues (bean and leaves, defense related, differential display, mature zygotic embryo, immature zygotic embryo and somatic embryo).

Database structure of EST's

For browsing the oil palm /cocoa ESTs, a text box with label 'ESTs' is provided. With this text box a link is provided for different tables, which stores information on ESTs from different tissues. Fig. 2 shows the popup menu of the EST text box revealing the ESTs from abnormal apex, male inflorescence, female inflorescences, normal apex and immature zygotic embryos.
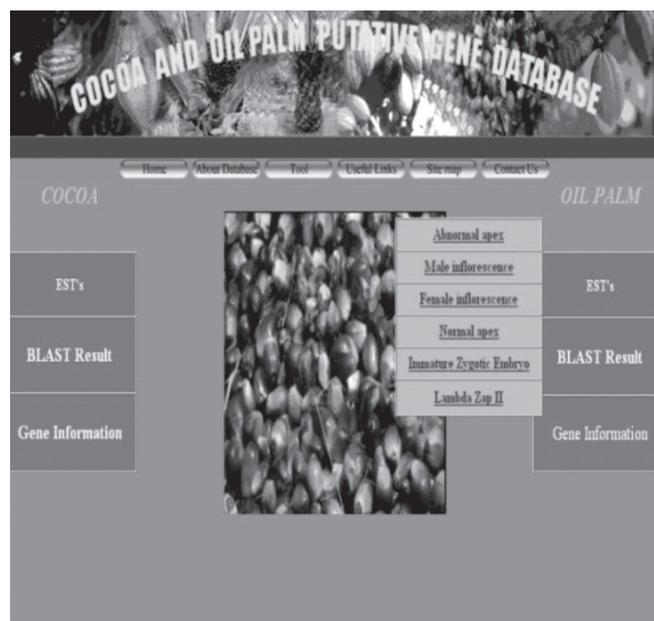

Fig. 2. Popup menu of oil palm EST's page

Architecture of contigs information – CAP 3 results

The results of CAP 3 (done for individual tissue's ESTs) are provided as number of contigs. The link for the number of contigs is given under the text box with tissue type labels. By pressing the tissue names, it will show the contigs sequences and fasta formatted EST sequences in text file. Each tissue contains information about number of contigs present and EST's (FASTA format).

For eg., if EST's button in oil palm is selected, it displays five tissue names. Selecting the option 'abnormal apex' tissue in turn shows number of contigs and EST's

in FASTA format (Fig. 3). If EST's button in cocoa is selected, it would show seven tissue names. Then by selecting the option 'bean and leaves' tissue in turn shows number of contigs and EST's in FASTA format. The same can be performed for all tissues giving the corresponding results (Fig. 4).

Fig. 3. Oilpalm abnormal apex EST page

Fig. 4. Cocoa bean and leaves EST page

Oil palm- BLAST result

BLAST result of oil palm contains five different tables, which stores information about EST's from five different tissues. Each tissue contains information about contig number and BLAST result.

Cocoa- BLAST results

BLAST result of cocoa contains three different tables, which stores information about three different tissues. Each tissue contains information about contig number and BLAST result.

Architecture of BLAST results

BLAST results are stored in different tables each for different tissues ESTs of oil palm (abnormal apex, male inflorescence, female inflorescences, normal apex and immature zygotic embryos) and cocoa (bean and leaves, defense related and young red leaves) and are provided as popup window. By selecting the button 'Blast result' will show the different tissues. Fig. 5 shows the popup menu of the 'Blast result' button. By clicking the individual tissue will show the information about contig number and BLASTx result. For eg. by selecting 'normal apex' tissue will show the page containing 13 contigs and link to the respective Blast result of the each contig. When the link to respective Blast is selected in turn will show the Blastx page. Fig. 6 shows the result page by selecting the tissue button 'normal apex'. For cocoa, selecting the 'bean and leaves' tissue will show the page containing 452 contigs and link to the respective blast result of each contig (Figs. 7 and 8).

Fig. 5. Popup menu of oil palm blast page

Oil palm - gene information result

'Gene information' of oil palm contains five different tables, which stores information about EST's from five different tissues. Each tissue contains information about contig number, similarities found in each organism, accession number, structure accession number and gene function of each organism.

314

Fig. 6. Oil palm normal apex blast result page

Fig. 7. Cocoa EST page (bean and leaves)

Cocoa - gene information result

'Gene information' of cocoa contains three different tables, which stores information about three different tissues (bean and leaves, defense related, young red leaves). Each tissue contains information about contig number, similarities found in each organism, accession number, structure accession number and gene function of each organism.

Architecture of gene information result

Gene information is stored in different tables each for different tissues ESTs (abnormal apex, male inflorescence, female inflorescences, normal apex and immature zygotic embryos) and provided as popup

Fig. 8. Cocoa blast result page

window. By selecting the button 'Gene information' will show the five different tissues. Fig. 9 shows the popup menu of the 'Gene information' button. By selecting the individual tissue will show the information about contigs, similarity detected with other organisms, accession number for the best-hit sequence in the respective organism in Blastn and structure accession number of the respective organism and gene function. For eg. by selecting immature zygotic embryo tissue will show the page containing gene information for 9 contigs. Link to the BlastN result page is provided under accession number. When the accession number is selected, it will show the Blastn page. Fig. 10 shows the result page by selecting the button 'immature zygotic embryo'. For cocoa, gene information is provided for three tissues (Fig. 11). By selecting 'young red leaves' tissue will show

Fig. 9. Oil palm Gene information page

315

Fig. 10. Oil palm gene information for immature zygotic embryos

the page containing gene information for 5 contigs (Fig. 12). Link to the Blastn result page is provided under accession number. When the accession number is selected, it will show the BlastN page.

The functional annotation revealed that some of the genes present in *Elaeis guineensis* show high similarity with functional genes of *Arabidopsis thaliana, Oryza sativa, Zea mays* and many other plants. The important contigs were identified using CAP3 tool in oil palm and cocoa. The BLAST result revealed that the contigs were highly similar with sequences of other plants like *Arabidopsis thaliana* and *Oryza sativa*. Using these methods some of the important genes like chlorophyll binding, glutathione peroxidase, glyceraldehyde-3-

Fig. 11. Cocoa Gene information page

Fig. 12. Cocoa EST page (bean and leaves)

phosphate dehydrogenase, UDP-glucuronate decarboxylase, amino acid transporter, drought-responsive and various others were identified for oil palm. More number of ESTs are found for heat shock protein in immature zygotic embryo tissue of oil palm. Heat shock protein is a molecular chaperone and essential for viability under all conditions in eukaryotes. Presence of heat shock protein is essential for the survival of the embryos. More numbers of ESTs are found for peroxidase in male inflorescence tissue of oil palm. Peroxidases has diverse functions in plant such as defense against pathogen, cross-linking of cell wall components, formation of lignin and suberin, auxin catabolism, and defense. Table 2 provides the annotated ESTs derived from abnormal apex tissue of oil palm. Other annotated ESTs can be browsed in the database.

Table 2. Annotated EST's derived from abnormal apex of oil palm

Annotated Contig Number	Molecular function
11	Glutathione peroxidase activity
31	Elongation factor 1-alpha
40	Structural constituent of ribosome
42	Polyubiquitin
54	Adenosyl homocysteinase activity
57	Nucleic acid binding
58	Chalcone synthase activity
60	Ribosomal protein
64	Ribulose-bisphosphate carboxylase activity
65	Drought-responsive
70	Permease activity
71	Amino acid transporter activity
76	UDP-glucuronate decarboxylase activity
78	Glyceraldehyde-3-phosphate dehydrogenase
80	Glycine-rich RNA binding protein
81	Ribonuclease P activity
86	Expansin-related

The ESTs of cocoa were functionally annotated and revealed that EST's related to genes involved in DNA binding, chlorophyll binding, dehydrogenase and antioxidant activity. Many EST's from bean and leaves and defense related processes were assigned with chlorophyll binding function. This function is related to photosystem II proteins. Antioxidant activity was predicted in EST's from bean and leaves and defense related processes. Antioxidants are compounds that protect cells against the damaging effects of reactive oxygen species, such as singlet oxygen, superoxide, peroxyl radicals, hydroxyl radicals and peroxynitrite. Table 3 provides the annotated cocoa ESTs (defense related). Other annotated ESTs are provided in the COPGENE database.

Table 3. Annotated EST's of cocoa derived during defense related processes in leaves

Annotated Contig Number	Molecular function
5	Glucosidase activity
6	Pyruvate dehydrogenase activity
14	EAD-box helicase activity
15	Aminopeptidase activity
17	Manganese superoxide dismutase
19	RNA polymerase activity
24	Rac GTPase activator activity
29	Sorbitol transporter activity
36	Alpha-amylase activity
38	Shikimate kinase activity
40	DNA binding
43	40S ribosomal protein S4
52	Cinnamoyl-CoA reductase activity
53	Translation elongation factors
55	RuBisCO subunit binding-protein beta subunit
56	Glucosyltransferase activity
61	Ribosomal protein S7
69	Lipoxygenase activity
70	Homolog to co-chaperone p23
73	Chlorophyll binding
74	Chlorophyll binding
75	Nucleic acid binding
83	Chlorophyll binding
89	Glyceraldehyde-3-phosphate dehydrogenase
95	Ribosomal protein L7
99	Translation elongation factor activity
101	Glyceraldehyde-3-phosphate dehydrogenase activity
105	Dehydration responsive element binding protein
106	Lipid transfer protein
114	Photosystem I reaction
123	Antioxidant activity
126	UDP-glucuronate decarboxylase activity
132	Chitinase
133	Polyubiquitin (Protein binding)
136	Acyl-CoA oxidase activity
140	Photosystem I reaction
142	Chorismate synthase activity
147	Lycopene beta cyclase activity
149	Endopeptidase Clp activity
152	Subtilase activity
162	Monosaccharide transporter
166	Transferase activity
170	Glutathione reductase activity
172	Binding-protein
174	Rac GTPase activator activity
181	Amino oxidase
183	Ascorbate peroxidase activity
191	Chlorophyll binding
194	Fructose-bisphosphate aldolase activity
195	Ribosomal protein
196	Glyceraldehyde-3-phosphate dehydrogenase activity
199	Heat shock protein
200	Oxygen evolving
207	Chlorophyll binding
209	Arginase activity
213	3-dehydroquinate synthase activity
217	Glyceraldehyde-3-phosphate dehydrogenase activity
224	Acyl-ACP thioesterase activity
230	DNA binding
240	Fructose-bisphosphate aldolase activity
242	Chlorophyll binding protein
243	Histone deacetylase family
246	Binding-protein
248	Carbonic anhydrase
249	Nucleasome/chromatin assembly factor D protein
251	Histidine kinase 2
253	Chlorophyll binding
258	Translation elongation factor activity
266	Superoxide dismutase
267	Ribosomal protein
271	Serine carboxypeptidase activity
276	Alpha tubulin

The current work was undertaken to assign putative function of available EST's of oil palm and cocoa and to develop a database. Two hundred and five ESTs were annotated for cocoa using homology based gene function prediction method. One hundred sixty eight contigs were identified for oilpalm and the functions of 34 EST were predicted. Oil palm and cocoa are important plantation crops and since lot of researchers are working in this field, as the information of putative function of oil palm and cocoa EST's are scarce in public database, the present database would provide information on ESTs of cocoa and oil palm. The database can be updated with new ESTs from the data source.

References

Argout, X., Fouet, O., Wincker, P., Gramacho, K., Legavre, T., Sabau, X., Risterucci, A., Da Silva, C., Cascardo, J., Allegre, M., Kuhn, D., Verica, J., Courtois, B., Loor, G., Babin, R., Sounigo, O., Ducamp, M., Guiltinan, M.J., Ruiz, M., Alemanno, L., Machado, R., Phillips, W., Schnell, R., Gilmour, M., Rosenquist, E., Butler, D., Maximova, S., and Lanaud, C. 2008. Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under

various conditions. *BMC Genomics* **9**: 512 doi:10.1186/1471-2164-9-512.

Cui, L., Narayanan, V., Alexander, R., Kerr, W., Robert, K. J., Leebens-Mack, J., Izabela, M. and Claude, W. 2005. ChloroplastDB: the Chloroplast Genome Database. *Nuc. Acids Res.* **34**: D692-D696.

Gerard, R.L., Lance, A.L., Cheryl, C.H., Kent, F.M., Mark, E.S., David, E.M., Melinda, A., Nancy, A.F. and Olin, D.A. 1998. Assignation Putative Gene Functions to mapped probe loci in the Graingenes genome Database and Sequencing of Wheat Endosperm cDNAs. *Plant & Animal Genome VI Conference.* Sandiego CA. Jan 18-22, 1998. P5.

Greer, K. and Soderlund, C. 2009. *Magnaporthe grisea Oryza sativa* (MGOS) Interaction Database: Community Annotation. In Advances in Genetics, Genomics and Control of Rice Blast Disease. Part V, 405-415, DOI: 10.1007/978-1-4020-9500-9-38.

Jouannic, S, Argout X., Lechauve, F., Fizames. C., Borgel, A., Morcillo, F., Aberlenc-Bertossi, F., Duval, Y. and Treger, J.

2005. Analysis of expressed sequences tags from oil palm (*Elaeis guineensis*). *FEBS* **579**: 2709 - 2714.

Kohany, O., Andrew, J.G., Lukasz, H., and Jerzy, J. 2006. Annotation, submission and screening of repetitive elements in Repbase: Repbase Submitter and Censor. *BMC Bioinformatics* **7**: 474 - 484.

Lim, K., Shamsuddin Z. and Ho, C. 2010. Transcriptomic changes in the root of oil palm (*Elaeis guineensis* Jacq.) upon inoculation with *Bacillus sphaericus* UPMB10. *Tree Genet. Genomes* **6**:793-800.

Ren, Q., Katherine, H. K., and Ian, T. P. 2004. TransportDB: a relational database of cellular membrane transport systems. *Nuc. Acids Res.* **32**: D284-D288.

Ruiz, M., Rouard, M., Raboin, L.M., Lartaud, M., Lagoda, P., Courtois, B. 2004. TropGENE-DB, a multi-tropical crop information system. *Nuc. Acids Res.* **32**: D364-367.

Walker, N. S., Nicholas, S. and Alice, B. 2006. POGs / PlantRBP: a resource for comparative genomics in plants. *Nuc. Acids Res.* **35**: D852-D856.