ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Cocoa EST database: Comprehensive database of Cocoa Expressed Sequence Tags (ESTs)

Naganeeswaran S.[1], Elain Apshara S. [2], Manimekalai R. [3,*], Amal Vasu[4] and Malhotra, S.K[5].

[1]Research Scholar, Crop Improvement Division, Central Plantation Crops Research Institute, Kasaragod, Kerala, India
[2]Principal Scientist, Crop Improvement Division, Central Plantation Crops Research Institute (RS), Vittal, Karnataka, India
[3]Senior Scientist, Crop Improvement Division, Sugarcane Breeding Institute, Coimbatore, Tamilnadu, India
[4]Project Trainee, Crop Improvement Division, Central Plantation Crops Research Institute, Kasaragod, Kerala, India
[5]Horticultural Commissioner, Department of Agriculture & Coop. Krishi Bhavan, New Delhi, India

* Corresponding Author

ABSTRACT: Cocoa EST database (CocoaESTdb) is a secondary EST database of cocoa. For database development, cocoa EST sequences were retrieved from public domains and classified into twenty major groups based on tissue source, developmental stage and influence of biotic / abiotic factors. CocoaESTdb presently contains 158739 good quality ESTs and corresponding annotations like blast similarity, gene ontology information, metabolic pathways and microsatellite based molecular data. CocoaESTdb is available at: http://cocoaestdb.cpcribioinformatics.in/.

KEYWORDS: Cocoa; database; EST; tool; pathway

I. INTRODUCTION

Chocolate tree Theobroma cacao L., (cocoa) commonly called as "food of the gods", is a small, evergreen, tropical tree, belonging to the family Malvaceae and originated from the tropical rain forest of South America [1]. Cocoa is the main raw material for chocolate, cocoa powder and cocoa liquor. Cocoa is a one of the most important cash crops in many tropical countries and many people depend on cocoa plantations for their income. In India, cocoa is mainly cultivated as intercrop in arecanut, coconut and oil palm gardens. At present, the demand of cocoa is much more than its total production. The sale of cocoa products in chocolate industry has increased because of the nutritional and therapeutic qualities of cocoa [2, 3]. Cocoa quality improvement and disease resistance are two important challenges for the cocoa growers [4].

Expressed Sequence Tags (ESTs) generation from cDNA is fast and cost-effective method for gene discovery and gene expression analysis. At present, more than 1,60,000 cocoa-EST sequences, isolated from various tissues and various condition are available in public EST database [5, 6, 7, 8]. These ESTs were annotated as a whole library. There is no exclusive tissue specific or biotic/ abiotic factor wise annotation. With the aim to identify the genes expressed in different tissues and under different environmental conditions, we analyzed the ESTs category wise and developed the user-friendly database "CocoaESTdb" (available at:http://cocoaestdb.cpcribioinformatics.in/). The database serves as source for cocoa ESTs annotation results, gene ontology, metabolic pathways and microsatellite data. This database aims accelerate the research in cocoa for identification of genes for quality improvement and stress response.

II. DATABASE DESIGN AND IMPLEMENTATION

Data source:

Totally 160058 cocoa EST sequences were retrieved from public databases (dbEST and ESTtik) [9, 5] and classified into 20 major groups based on the type of source tissue and the influence of biotic and abiotic factors (Table 1). Fifty six libraries of cocoa EST sequences, which include 149649 EST sequences were retrieved from ESTtik database and 10409 EST sequences were retrieved from NCBI-dbEST database. These ESTs are classified based on tissue type,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

developmental stage and influence of biotic / abiotic factors. TheUniProtKB protein information is retrieved and stored in MySQL database and used for gene ontology prediction.

EST pre processing and Assembly:

Poor quality and contaminated regions (vectors, organelles, and low-complexity regions) were trimmed/removed using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) (Parameter for SeqClean: –c 2, -n 2000, -l 100, -x 96 and –y 11) program embedded with NCBI's UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Each twenty EST datasets were processed individually using SeqClean tool and totally 158739 valid sequences were obtained (Table 1). Sequence redundancy was removed using CAP3 [10] de novo assembly program with default parameter.

Table 1: Cocoa EST information: source of cocoa EST sequences derived, total number of EST sequences and number of valid EST sequences after processing.

S.No	Source of Cocoa EST Library	Number of Sequence	Valid Sequence
1	Bark tissue	3478	3461
2	Callus tissue	3434	3428
4	Chemical treatment	2121	1854
5	Cortex tissue	8913	8901
6	Development stage	16913	16478
7	Drought stress	5451	5427
8	Germination seed	16515	16460
9	Moniliophthora perniciosa affected	12285	12182
10	Moniliophthora roreri affected	6233	6191
11	Phytophthora megakarya infected	6373	6349
12	Phytophthora palmivora infected	14831	14741
13	Pollinations	25905	25817
14	Root	3567	3564
15	Sahlbergella singularis	4724	4711
16	Seed	14059	13988
17	Stem	4938	4923
18	Testa	4005	3993
19	Tricoderma	32	32
20	Young cushions	2849	2824

Similarity search:

UniProtKB/Swiss-Prot protein sequences were retrieved from UniProt database [11] and stand alone UniprotKB blast database was created using formatdb program. BlastX [12] search was carried out for each dataset of assembled ESTs against (e-value e^{-10}) uniprot protein sequence database and the results were saved in tab separated table format. Likewise each EST datasets were Blast searched against with available cocoa genome information.

Gene ontology analysis:

Gene ontology was predicted using locally developed stand alone Gene Ontology Analysis Tool (GOAT) using (e-value e^{-20}) standalone UniProtKB protein information database. Graphical representations of results (pie chart and bar diagram) were generated using Python-Matplot module (Fig. 1c, d, e).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

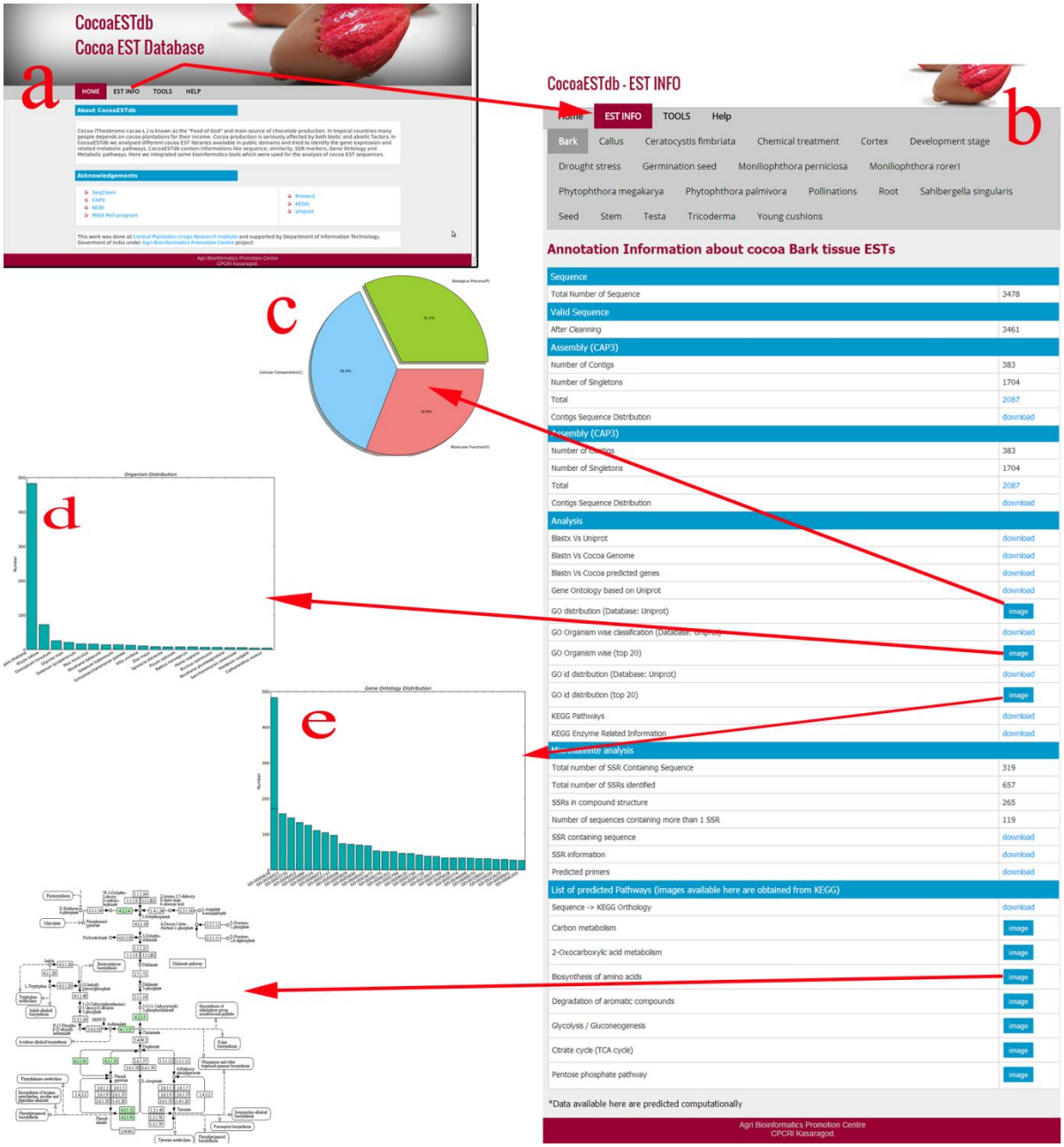

Fig. 1: Overview structure of CocoaESTdb. a) cocoaESTdb front page; b) Annotation information of bark tissue ESTs; c, d, e) Grapical representation of gene ontology result; f) KEGG based pathway map

Metabolic pathway analysis:

Each dataset of EST was searched against KEGG metabolic pathway database [13, 14] and the resulted pathway gene information were stored in MySQL database management system . User can view metabolic pathways (Fig. 1f) and download the KEGG orthology for the individual EST dataset.

SSR analysis and Primer designing:

All the twenty EST datasets are analysed for SSRs. EST - SSRs were identified using MISA (MIcro Satellite Analysis tool) (http://pgrc.ipk-gatersleben.de/misa/) programme. Monomers with at least 10 repetitions, dimers with

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

minimum six repetitions and trimers, tetramers, pentamers and hexamers with minimum five repetitions were considered as valid SSRs. SSR specific primers were designed using Primer3 software [15]. User can download the result of the SSR containing sequences. SSR information and SSR primer in particular EST dataset.

Database development:

The front end of the database was created using HTML, CSS and Javascript (Fig. 1a). The back end programs were developed using PHP, Perl and Python. MySQL database management system was used as back end database. This database is updated twice in a year or updated based on the availability of cocoa EST sequences in public domain.

III. CONCLUSION

CocoaESTdb provides cocoa EST specific annotation information like similarity, gene-ontology, metabolic pathway and microsatellite based molecular data. CocoaESTdb stands as a comprehensive source of cocoa specific ESTs and corresponding annotation information for to accelerate the plant breeding programs for cocoa crop improvement. CocoaESTdb is available for user access at: http://cocoaestdb.cpcribioinformatics.in/.

REFERENCES

1. Cheesman, E. E., "Notes on the nomenclature, classification possible and relationships of cocoa populations", Trop Agric, Vol.21, pp.144-159, 1944.
2. Rusconi, M., and Conti, A., "*Theobroma cacao* L., the food of the gods: A scientific approach beyond myths and claims", Pharmacol Res, Vol.61, pp.5-13, 2010.
3. Gu, L., House, S. E., Wu, X., Ou, B., and Prior, R. L., "Procyanidin and catechin contents and antioxidant capacity of cocoa and chocolate products", J Agric Food Chem, Vol.54, pp.4057- 4061, 2006.
4. Bowers, J. H., Bailey, B. A., Hebbar, P. K., Sanogo, S., and Lumsden, R. D., "The impact of plant diseases on world chocolate production", Plant Health Progress, 2001.
5. Argout, X., Fouet, O., Wincker, P., Gramacho, K., Legavre, T., Sabau, X., et al,. "Towards the understanding of the cocoa transcriptome: production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* generated from various tissues and under various conditions", BMC Genomics, Vol.3, pp.512, 2008.
6. Arnold, A. E., Mejía, L. C., Kyllo, D., Rojas, E. I., Maynard, Z., Robbins, N., et al., "Fungal endophytes limit pathogen damage in a tropical tree", PNAS, Vol.100, pp.15649–15654, 2003.
7. Hanada, R. E., Pomella, A. W. V., Costa, H. S., Bezerra, J. L., Loguercio, L. L., and Pereira, J. O., "Endophytic fungal diversity in *Theobroma cacao* (cacao) and T. grandiflorum (cupuaçu) trees and their potential for growth promotion and biocontrol of black-pod disease", Fungal Biol, Vol.114, pp.901–910, 2010.
8. Gesteira, A.S., Fabienne, M., Carela, N., Da-Silva, A. C., Gramacho, K. P., Schuster, I., and Macedo, A. N., "Comparative Analysis of Expressed Genes from Cacao Meristems Infected by *Moniliophthora perniciosa",* Ann Bot, Vol.100, pp.129–140, 2007.
9. Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M., "dbEST--database for "expressed sequence tags"", Nat Genet, Vol.4, pp.332-333, 1993.
10. Huang, X., and Madan, A., "CAP3: A DNA Sequence Assembly Program", Genome Res, Vol.9, pp.868-877, 1999.
11. Magrane, M., and The UniProt consortium., "UniProt Knowledgebase: a hub of integrated protein data", Database, 2011.
12. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res, Vol.25, pp.3389-3402, 1997.
13. Kanehisa, M., and Goto, S., "KEGG: Kyoto Encyclopedia of Genes and Genomes", Nucleic Acids Res, Vol.28, pp.27-30, 2000.
14. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A., and Kanehisa, M., "KAAS: an automatic genome annotation and pathway reconstruction server", Nucleic Acids Res, Vol.35, pp.W182-W185, 2007.
15. Rozen, S., and Skaletsky, H., "Primer3 on the WWW for general users and for biologist programmers", Methods Mol Biol, Vol.132, pp.365-386, 2000.