# ANALYSIS OF DROUGHT INDUCED EXPRESSED SEQUENCE TAGS (EST'S) LIBRARY AND IDENTIFICATION OF METABOLIC PATHWAYS IN COCOA

**S. Naganeeswaran[1]and S. Elain Apshara[2]**

[1]Bioinformatics Centre, Central Plantation Crops Research Institute, Kasaragod, Kerala

[2]Central Plantation Crops Research Institute, Regional Station, Vittal, Karnataka

## Introduction

The chocolate tree, cocoa is an important cash crop of tropical countries (Wood and Lass, 2001). Cocoa beans are the primary raw material for chocolate industry and the quality of beans is very important as they have antioxidant, antibacterial, antidiabetic, anticarcinogenic and cardio protective properties. In the cropping system, response of trees to biotic and abiotic stress is also a major concern in the newly introduced environment (Bowers *et al.,* 2001, Chanliau and Cross, 1996). Drought is an important abiotic factor affects the cocoa production in rainfed areas. Identifying and understanding the molecular mechanism of drought stress will help to overcome this problem. With the objective of identifying the functional genes expressed in stress condition, we have analyzed EST sequences retrieved from public domain.

## Materials and Methods

### Primary sequence source

Two drought induced cocoa cDNA libraries isolated from leaf and root tissue belong to the genotype of Scavina 6 were used in this study. Totally 5451 redundant EST sequences were downloaded from ESTtik database (Xavier Argout, 2008).

### EST Analysis

EST analysis procedure includes the following steps: 1) EST pre-processing, 2) EST assembly and 3) functional annotation. Figure 1 describes the currently implemented steps.

### EST pre-processing and assembly

EST processing like removal of vector contamination, trimming poly A/ T tail and low complexity region, removal of linker and adaptor sequence were performed using SeqClean (Gene Index project) tool. Vector contamination database UniVec was configured with local Blast (Altschul *et al.,* 1997) and used in SeqClean tool. Repeatmasker (http://www.repeatmasker.org) tool was used to remove the low complexity regions from the EST sequences. Clustering and assembling were done by adapting CAP3 (Huang and Madan, 1999) tool.

### Functional annotation

Non redundant EST sequences were subjected to blastx (Altschul *et al.,* 1997) similarity search and further those homologous sequences which are having E-value below e-10 were selected. Gene Ontology (Camon *et al.,* 2003) search, Enzyme search, Interproscan and KEGG mapping were done using Blast2go (Conesa *et al.,* 2005) tool.

### Database design

The information which was obtained from the processing and annotation of the EST sequences were

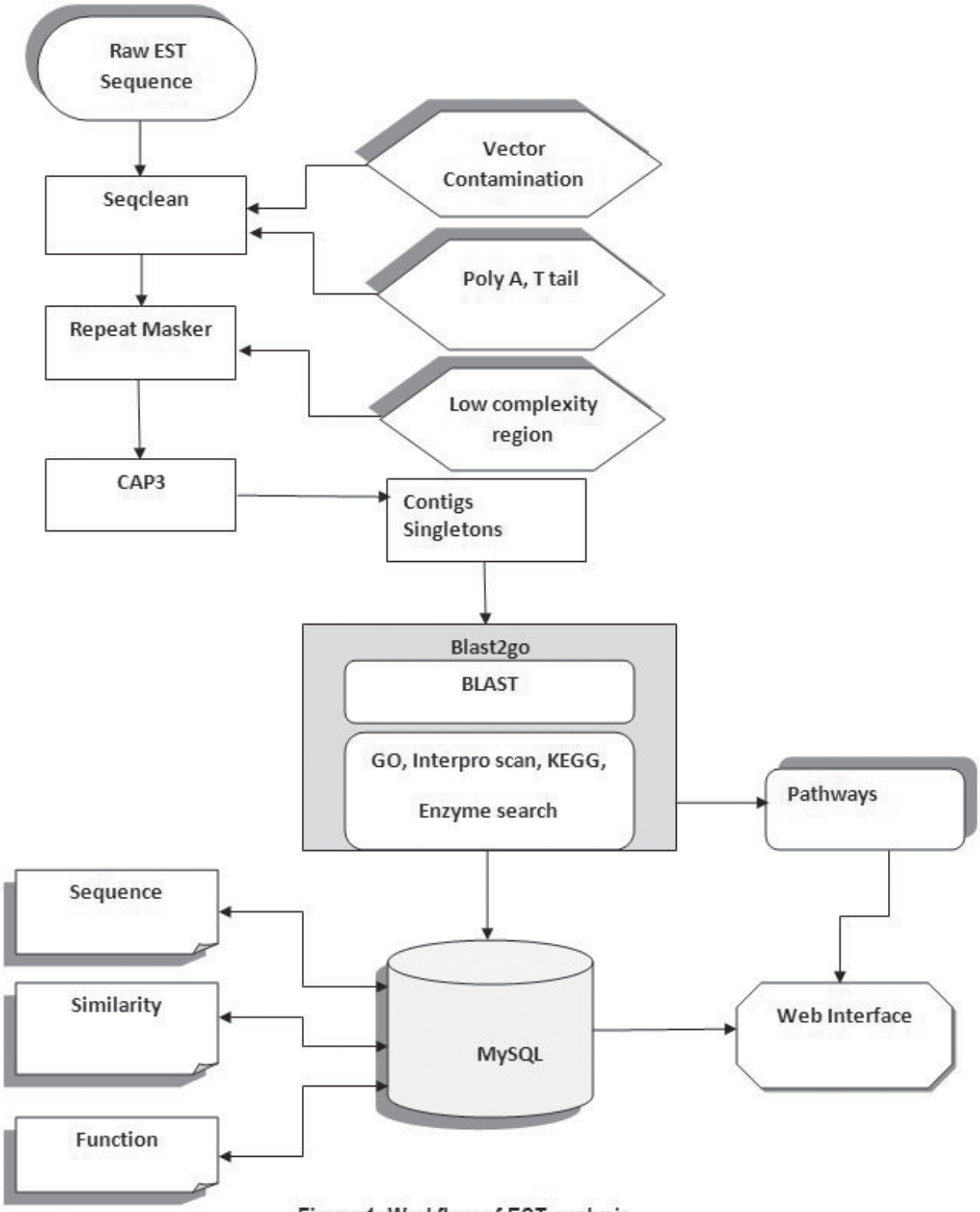


Figure 1: Workflow of EST analysis

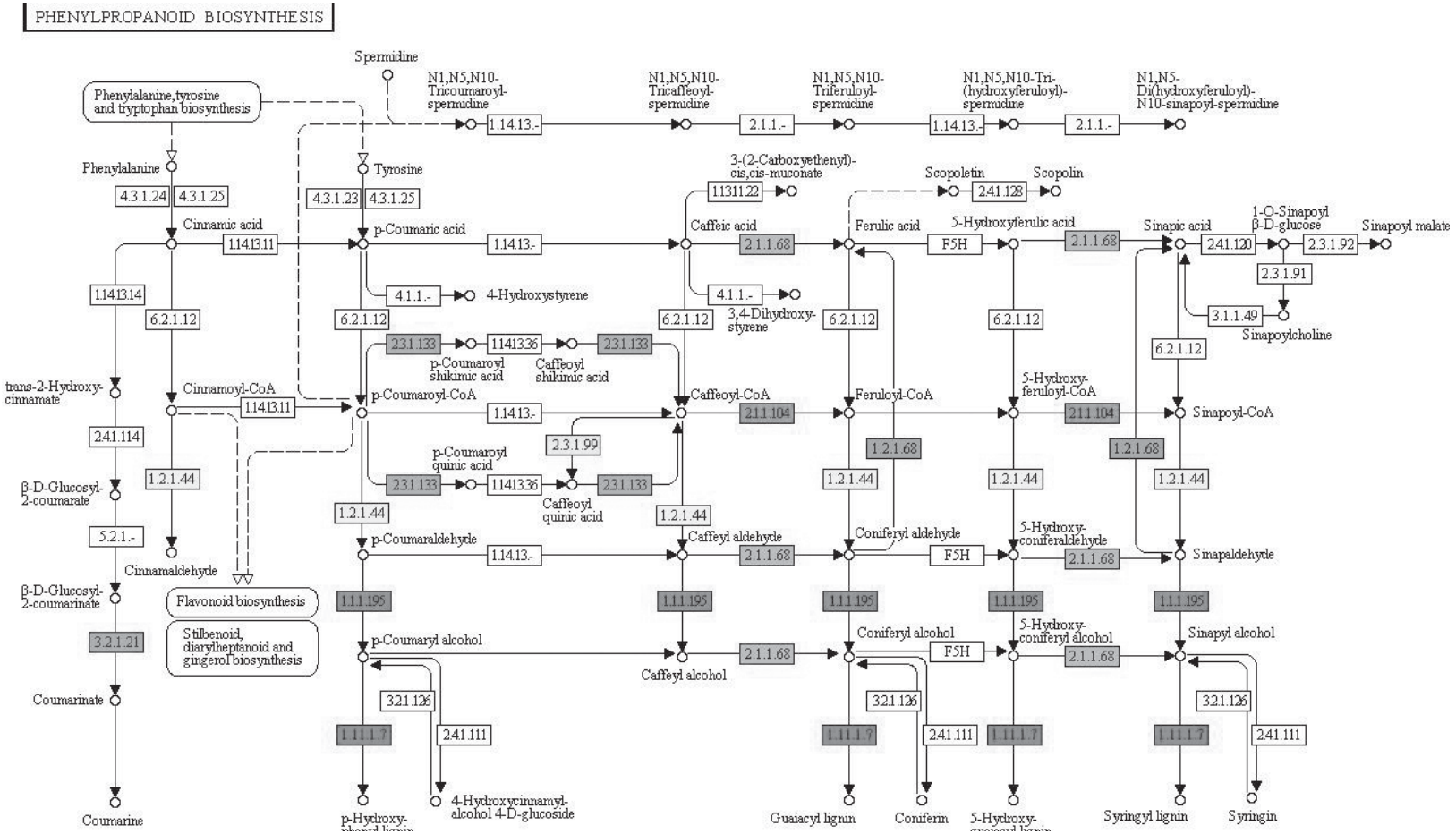


**Fig. 2. Phenylpropanoid Biosynthesis pathway showing stress response enzymes**

deposited in a MySQL relational database. Three different tables were created using SQL for storing sequences, blast hit and functional annotation. The front end of the database was created using HTML\CSS. Web server apache2 was configured and the server side scripting languages PHP/Perl were used for connectivity purpose.

**Results and Discussion**

From the 5451 EST sequences retrieved from ESTtik database (Table 1.) processing was performed using SeqClean tool resulting 5424 good quality EST sequences which were used for extended analysis. By the contig assembly using CAP3 tool 2505 non redundant EST (421 contigs and 2084 singletons) sequences were obtained. The primary sequence analysis shows total GC content of non redundant EST collection as 44.16% and the sequence length ranges from 102 residues to 1280 residues. Primary sequence analysis was done by locally developed Perl script (DSA.pl). All the contigs and singletons sequences and their primary sequence analysis results found in the EST collection is available in the

**Table 1. Drought induced EST library information**

| Library name | Genotype | Description | No. of sequence |
|---|---|---|---|
| DROUGHTLS_KZ0ACAF | Scavina6 | Leaves submitted to drought stresses | 2766 |
| DROUGHTRS_KZ0ACAE | Scavina6 | Roots submitted to drought stresses | 2685 |

following URL: http://220.227.88.254/cacao/sequence.html.

Blastx similarity search was done against the non-redundant database, from which significant E-value with <e-10 was used for annotation. Results obtained were stored in MySQL database and the user interface is available in the link: http://220.227.88.254/cacao/table_blast.html. Gene Ontology classification (GO), HMMER search against Pfam database, Interproscan and Enzyme search were done using Blast2go tool. This functional annotation shows most of the EST are responsible for stress response, signal transduction and transcription factors (http://220.227.88.254/cacao/table_function.html). Totally 392 enzymes corresponding to 102 metabolic pathways (http://220.227.88.254/cacao/metabolic_pathway.html) were identified. Figure 2 showing important enzymes in Phenylpropanoid biosynthetic pathway enzymes including Caffeate O-Methyltransferase (E.C: 2.1.1.68), Acyltransferase (E.C: 2.3.1.133), Coffeoyl-CoA O-Methyltransferase (E.C: 2.1.1.104), Cinnamoyl-CoA reductase (E.C: 1.2.1.44) and Cinnamyl alchohol dehydrogenase (E.C:1.1.1.195) involved in stress response were identified.

**Conclusion**

We have developed a Cocoa stress EST database (URL: http://220.227.88.254/cacao/) which includes the sequence, structural, functional and metabolic pathway information about the drought induced EST sequences of cocoa. This database provides the information about the functional genes which involved in stress response. This information is further useful for the reconstruction of stress response pathways in cocoa.

**Acknowledgement**

This work was supported by a grant from Department of Information Technology (DIT), Government of India.

## REFERENCES

1. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17): 3389-3402.

2. Bowers, J.H., Bailey, B.A., Hebbar, P.R., Sanogo, S. and Lumsden, R.D. 2001. The impact of plant disease on world chocolate production. *Plant Health Progress* 2001.

3. Chanliau, S. and Cross, E.1996. Influence du traitement post-recolte et de la torrefaction sur ly development de l'arome cacao. *12th Alliance's Inter Cacao conf*, Salvadr de Bahia (Brazil): 959-964.

4. Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J. and Cox, A. 2003. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res* 13(4): 662-672.

5. Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.

6. Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res* 9(9): 868-77.

7. RepeatMasker (http://www.repeatmasker.org).

8. The Gene Index project (http://compbio.dfci.harvard.edu/tgi/software/).

9. Wood, G.A.R. and Lass, R.A. 2001. *Cacao,* 4th edition. Blackwell, Oxford. 620 p.

10. Xavier Argout. 2008. Towards the understanding of the cacao 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.