BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 7(4)

EuDBase: An online resource for automated EST analysis pipeline (ESTFrontier) and database for red seaweed *Eucheuma denticulatum*

Zeti Azura Mohamed Hussein[1, 2*], Kok Keong Loke[1, 2], Rabiatul Adawiah Zainal Abidin[1, 2], Roohaida Othman[1, 3]

[1]School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia 43600 UKM Bangi, Selangor, Malaysia; [2]Centre for Bioinformatics Research, Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia; [3]Centre for Gene Analysis and Technology, Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia; Zeti-Azura Mohamed-Hussein – Email: zeti.amh@gmail.com; zeti@ukm.my; *Corresponding author

Received September 21, 2011; Accepted September 22, 2011; Published October 14, 2011

Abstract:
Functional genomics has proven to be an efficient tool in identifying genes involved in various biological functions. However the availability of commercially important seaweed *Eucheuma denticulatum* functional resources is still limited. EuDBase is the first seaweed online repository that provides integrated access to ESTs of *Eucheuma denticulatum* generated from samples collected from Kudat and Semporna in Sabah, Malaysia. The database stored 10,031 ESTs that are clustered and assembled into 2,275 unique transcripts (UT) and 955 singletons. Raw data were automatically processed using ESTFrontier, an in-house automated EST analysis pipeline. Data was collected in MySQL database. Web interface is implemented using PHP and it allows browsing and querying EuDBase through search engine. Data is searchable via BLAST hit, domain search, Gene Ontology or KEGG Pathway. A user-friendly interface allows the identification of sequences either using a simple text query or similarity search. The development of EuDBase is initiated to store, manage and analyze the *E. denticulatum* ESTs and to provide accumulative digital resources for the use of global scientific community. EuDBase is freely available from http://www.inbiosis.ukm.my/eudbase/.

Background:
The species of *Eucheuma* occur throughout the Indo-Pacific region from East Africa to Guam, in China and Japan waters and mostly in algal reef areas of islands in Southeast Asia. In Malaysia, *E. denticulatum* is commonly farmed in Kudat and Semporna in the state of Sabah. *E. denticulatum* is also known as "spinosum" which is a trade name indicating its production of *iota*-carrageenan. *E. denticulatum* has been the focus of many studies due to its unique polysaccharides that constitute its cell wall that are unlike those found in plants. Currently there is no dedicated database available for the expressed sequence tags (ESTs) data of *E. denticulatum* even though the interest in seaweed community has been increased globally due to its economical value. ESTs are significantly important especially for organisms where the genome sequences are not available and they can be used as a basis for structural genomic annotation. Until now only *Ectocarpus siliculosus* (brown algae) has its genome fully sequenced **[1]**. We aim to generate as many ESTs from *E. denticulatum* as possible and to use the encoded information to reveal interesting information on the biosynthetic pathway of *iota*-carrageenan. Bioinformatics analysis has been carried out to facilitate the finding of interesting biological information. To date, we have uploaded 9,057 high quality ESTs to the GenBank EST repository. We present the *E. denticulatum* EST database (EuDBase) consists of EST data, functional annotation and metabolic pathway assignments. The content of EuDBase will continue to increase in parallel with the EST sequencing effort carried out at

 © 2011 Biomedical Informatics

Universiti Kebangsaan Malaysia (UKM). It also provides comparative data for analyses of organism that has no comparable genomic resources. EuDBase also links to ESTFrontier pipeline for comprehensive EST data analyses.

Methodology:

EuDBase and ESTFrontier have been designed for simple and efficient information search and retrieval. EuDBase is composed of two major components i.e. a relational database created using open access MySQL version 5.1.36 and a PHP version 5.3.0 web application that communicates with the database. PHP scripts dynamically execute SQL queries to retrieve data from the database according to user criteria and display them as a standard HTML output. EuDBase database model and its interaction with ESTFrontier are depicted in **Figures 1 & 2**. EuDBase stores raw sequences, assembled sequences as unique transcripts (UTs), translated peptides, BLASTX results, protein signature analysis results from InterProScan, Gene Ontology functional annotations based on BLASTX results and KEGG PATHWAY mapping. Currently, EuDBase contains 9,057 refined EST from 10,031 sequenced ESTs from *E. denticulatum* libraries. ESTFrontier was developed to facilitate EST data processing and functional annotation. Several bioinformatic tools are embedded in the pipeline including Seqclean **[2]**, RepeatMasker, CAP3 **[3]**, ESTScan **[4]**, FrameDP **[5]**, BLASTX **[6]**, InterProScan, InterPro2GO, BLAST2GO **[6]**, AutoFACT **[7]** and KOBAS **[8]**. A comprehensive spreadsheet report in EXCEL format is generated as output files.

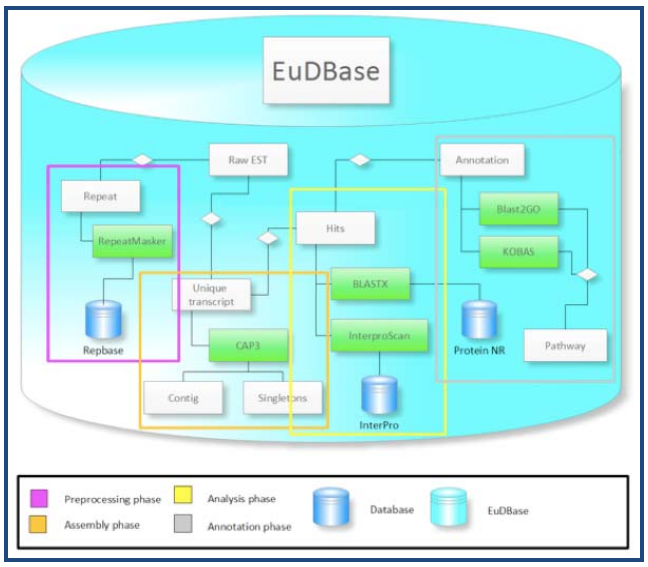

Figure 1: A database model of EuDBase

Utility:

***E. denticulatum* EST statistics in EuDBase:**

StackPack EST assembly pipeline was used to assemble raw EST data resulting to the assembly of 2,275 unique transcripts that consisted of 1,320 consensus sequences and 955 singletons **(Table 1, see supplementary material)**. Sequence similarity search against NCBI nr-database with a cut-off value of 1e-06 showed 961 UTs have significant matched homologues, 145 UTs were categorised as predicted proteins whilst 138 UTs were grouped into hypothetical and unknown proteins **(Table 2, see supplementary material)**. **Table 3 (see supplementary material)** shows the most abundant similarity search of *E. denticulatum* UT data set where 62 ESTs were found to match the RNA-binding proteins. **Table 4 (see supplementary material)** shows the output for domain analysis using InterProScan. We used Blast2GO for the functional annotation in *E. denticulatum* EST and 1935 Gene Ontology terms were assigned on 399 UTs. BLAST2GO used 5 best hits from BLASTX results to annotate each UTs sequence and successfully annotated 823 GO terms under biological process, 578 under molecular function, and 488 under cellular component **(Table 5, see supplementary material)**. We have also performed a pathway mapping of *E. denticulatum* UTs on KEGG pathway to observe their interactions. Using BLAST2GO KEGG pathway mapping functionality along with the complementary support from KOBAS, 57 unique pathways have been mapped with *E. denticulatum* ESTs and 100 UTs were found to map on the pathway of plant hormones biosynthesis and 99 UTs are mapped on the phenylpropanoids biosynthetic pathway. **Table 6 (see supplementary material)** lists 10 most abundant pathways that were mapped with *E. denticulatum* UTs. EuDBase web interface enables users to perform keyword search and browsing against the database. Database users can query the database using keywords together using Boolean operators such as AND, OR and NOT to perform complex queries. EuDBase includes local BLAST server to enable BLAST searching against EuDBase assembled UTs and translated peptides using appropriate BLAST subprograms such as BLASTN, BLASTP and BLASTX **(Figure 4)**.

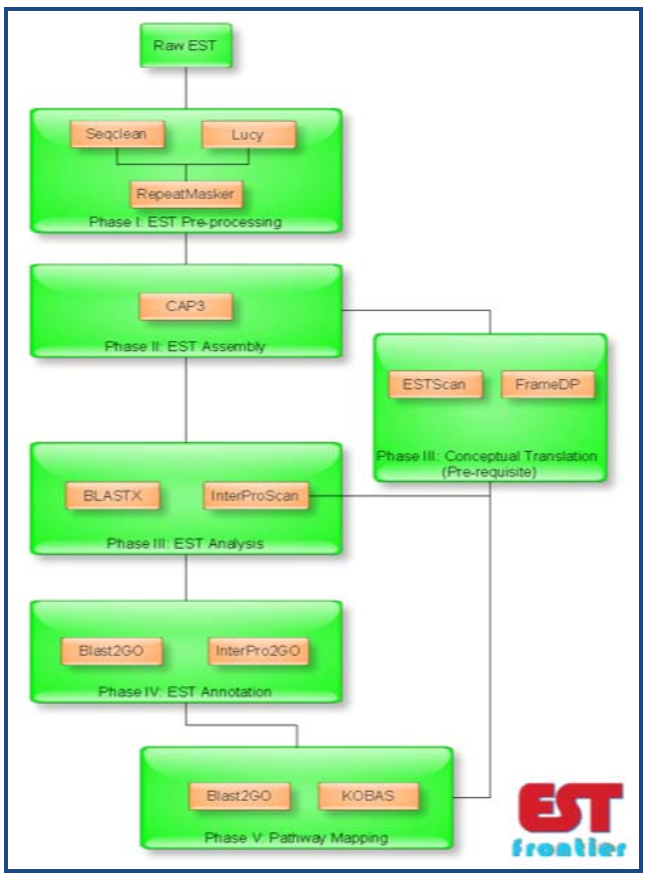

Figure 2: EST analysis pipeline in EuDBase known as ESTFrontier

© 2011 Biomedical Informatics

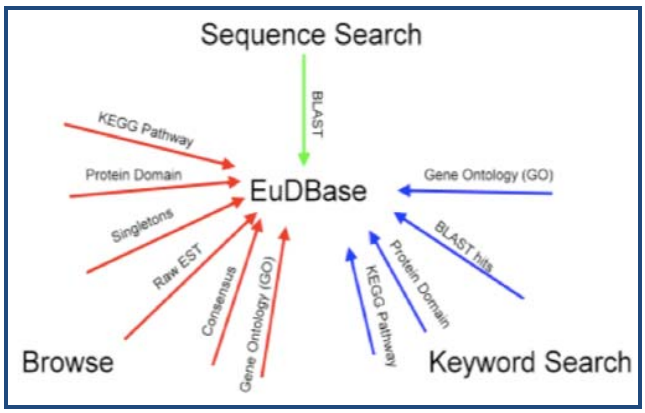

Figure 3: Data mining route in EuDBase. There are three main branches for mining in EuDBase

Figure 4: Snapshots of EuDBase web interface. EuDBase Top page with links to Browse and Search. A) Keyword search results with links toward sequence summary report; B) Consensus sequence summary report; C) Browse EuDBase by raw sequences, singletons, consensus, protein domain, Gene Ontology (GO) and KEGG pathway; D) Raw EST sequence summary report; E) Singleton sequence summary report; F) Consensus sequence summary report.

Future developments:
Eventually EuDBase will incorporate *E. denticulatum* proteomics, transcriptomics and metabolomics data as well as its integration with a genome browser. The server will be periodically upgraded for faster access to accommodate the growing number of data.

Conclusion:
EuDBase is a first online resource for red seaweed that allows for easy data integration and retrieval with the aim of providing a tool to expand the knowledge on *E. denticulatum* functional genomics.

Authors' contributions:
ZAMH formulated the study, directed the work and wrote the manuscript. RAZA worked on the preliminary development of the database. LKK continuously developed, implemented and managed the database and analysis pipeline. RO conceived and directed the molecular biology studies. All authors read and approved the final manuscript.

Acknowledgement:
We thank all RO's students for their contribution in generating the EST raw data, Universiti Kebangsaan Malaysia (UKM), the Malaysian Ministry of Higher Education for the Fundamental Research Grant Scheme (UKM-RB-06-FRGS0101-2009) awarded to ZAMH which supported this study and the MOSTI Grant 06-02-02-003-BTK/ER/0016 awarded to RO which supported the laboratory experiments. The work was carried out at the Bioinformatics Lab, Institute of Systems Biology (INBIOSIS), UKM and the laboratory experiments were carried out at the Centre for Gene Analysis and Technology (CGAT), INBIOSIS, UKM. The facilities and financial aid are duly acknowledged.

References:
[1] Cock JM *et al*. *Nature* 2010 **465**: 617 [PMID: 20520714]
[2] Chen YA *et al*. *BMC Genomics*. 2007 **8**: 416 [PMID: 17997864]
[3] Huang X & Madan A. *Genome Res*. 1999 **9**: 868 [PMID: 10508846]
[4] Gouzy J *et al*. *Bioinformatics*. 2008 **25**: 670 [PMID: 19153134]
[5] Altschul SF *et al*. *J Mol Biol*. 1990 **215**: 403 [PMID: 2231712]
[6] Conesa A *et al*. *Bioinformatics* 2005 **21**: 3674 [PMID: 16081474]
[7] Koski LB *et al*. *BMC Bioinformatics*. 2005 **6**: 151 [PMID: 15960857]
[8] Wu J *et al*. *Nucleic Acids Res*. 2006 **34**: W720 [PMID: 16845106]

Edited by P Kangueane
Citation: Mohamed-Hussein *et al*. Bioinformation 7(4): 157-162 (2011)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

BIOINFORMATION

open access

Supplementary material:

Table 1: A summary of *E. denticulatum* EST in EuDBase

Library and EST summary	
Number of raw EST	10031
Mean EST length[a]	422.60
Mean GC percentage	50.25
Number of high quality ESTs	9057
Contig assembly results	
Number of ESTs assembled[b]	9057
Number of consensus	1320
Number of singletons	955
Number of unique transcripts[c]	2275
Contig sizes	
2-10 ESTs	1087
11-20 ESTs	122
21-30 ESTs	20
31-40 ESTs	10
41-50 ESTs	6
>50 ESTs	8

[a]Mean EST length following vector and end clipping; [b]EST assembly parameters were 80% minimum match with 40 minimum base overlap; [c]Unique transcripts are the sum of contigs and singletons.

Table 2: BLASTX analysis results for *E. denticulatum* UTs

Results	Number of hits
	Nonredundant Protein database
Total query UTs	2,390
Significant hits to known proteins	961
Hypothetical protein	110
Predicted protein	145
Unknown protein	28

Table 3: Most abundant EST similarity search of *E. denticulatum* UT data set

UT	Description	No of ESTs	E-value
cn11	RNA binding protein	62	3e-18
cn1085	Ubiquitin	43	6e-18
cn457	Actin	38	2e-146
cn995	Transcription factor	34	1e-28
cn982	Proton gradient regulation 5 (PGR5)	34	5e-24
cn1212	Eukaryotic translation factor	31	3e-44
cn766	mRNA binding post-transcriptional regulator	31	8e-12
cn6	Ferrodoxin oxidoreductase precursor	28	3e-25
cn669	Putative DNA-bindig transcriptional regulator	27	6e-71
cn3	Hypothetical protein SYNW1396	22	6e-19
cn132	Putative inner membrane	21	8e-37
cn1041	Hypothetical protein (predicted similar to gluthione S-transferase)	20	6e-70
cn1078	Glyoxylate carboligase	19	1e-33
cn31	Predicted similar to ubiquitin isoform-1	18	3e-72

Table 4: Gene Ontology of *E. denticulatum* UTs

Library/library group[a]:	EST count
GO term	
All molecular function	578
Molecular_function	8
Binding	83
Carbohydrate binding	3
Nucleic acid binding	2

© 2011 Biomedical Informatics

DNA binding	24
RNA binding	22
Nucleotide binding	59
Protein binding	60
Receptor binding	2
Lipid binding	1
Catalytic	79
Hydrolase activity	60
Motor activity	1
Nuclease activity	2
Enzyme regulator activity	2
Transferase activity	47
Kinase activity	15
Signal transducer activity	3
Receptor activity	4
Structural molecule activity	53
Transcription regulator activity	6
Transcription factor activity	5
Translation factor activity, nucleic acid binding	11
Transporter activity	26

[a]Classified using guidelines of the Gene Ontology Consortium 2001 (http://www.geneontology.org). Indented terms are children of the above parent term. Only mapped GO terms are presented.

Table 5: Domain analysis using INTERPRO

Domain	UT Domain Range	E-value	EST Count
RRM (RNA recognition motif) It has a variety of RNA binding preferences and functions, and include heterogeneous nuclear ribonucleoproteins (hnRNPs), proteins implicated in regulation of alternative splicing (SR, U2AF, Sxl), protein components of small nuclear ribonucleoproteins (U1 and U2 snRNPs), and proteins that regulate RNA stability and translation (PABP, La, Hu)	cn111 (69…143)	1.4e-25	62
Kazal_2 This domain is usually indicative of serine protease inhibitors that belong to Merops inhibitor families: I1, I2, I17 and I31. However, kazal-like domains are also seen in the extracellular part of agrins, which are not known to be protease inhibitors.	cn106 (98...127)	3.4e-06	49
Thioredoxin-like Several biological processes regulate the activity of target proteins through changes in the redox state of thiol groups (S2 to SH2), where a hydrogen donor is linked to an intermediary disulphide protein. Such processes include the ferredoxin/thioredoxin system, the NADP/thioredoxin system, and the glutathione/glutaredoxin system.	cn1036 (40…130)	1.3e-08	45
Actin An ubiquitous protein involved in the formation of filaments that are major components of the cytoskeleton.	cn457 (1…262)	3e-182	38
NAC (Nascent polypeptide-associated complex) This is a ribosome-associated entity that binds to the nascent polypeptide after the formation of peptide bond. NAC may prevent binding of ribosome nescnet chains (RNCs) without signal sequence to membranes	cn995 (43…99)	1.1e-19	38
eIF-1A: translation initiation factor it is formerly known as eiF-4C. It is required for maximal rate of protein biosynthesis. It also enhances ribosome dissociation into subunits and stabilizes the binding of the initiator Met-tRNA to 40S ribosomal subunits	cn1212 (46…143)	1.6e-34	31
Fe_bilin_red This family consists of several different but closely related proteins that include phycocyanobilin:ferredoxin oxidoreductase	cn6 (1…71)	3.5e-25	28

© 2011 Biomedical Informatics

EC:1.3.7.5 (PcyA), 15,16-dihydrobiliverdin:ferredoxin oxidoreductase EC:1.3.7.2 (PebA) and phycoerythrobilin:ferredoxin oxidoreductase EC:1.3.7.3 (PebB). Phytobilins are linear tetrapyrrole precursors of the light-harvesting prosthetic groups of the phytochrome photoreceptors of plants and the phycobiliprotein photosynthetic antennae of cyanobacteria, red algae, and cryptomonads.			
CDI (cyclin dependent kinases inhibitors (CDIs) CDI controls the progression of cell cycle and it also involved in cell cycle arrest at G1 phase	cn865 (100…127)	8.9e-07	22
Na_sulph_symp Integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families	cn132 (2…160)	2.9e-56	21
GST (glutathione S-transferase) GSTs in eukaryote participate in the detoxification of reactive electrophillic compounds by catalyzing their conjugation to glutathione.	cn1041 (107…187)	5.8e-08	20

Table 6: Ten most abundant ESTs mapped to KEGG PATHWAY

Pathway_ID	Pathway_name	Total
map04350	TGF-beta signalling pathway	504
map01070	Biosynthesis of plant hormones	247
map02020	Two-component system	59
map00730	Thiamine metabolism	52
map00410	beta-Alanine metabolism	48
map01061	Biosynthesis of phenylpropanoids	40
map00860	Porphyrin and chlorophyll metabolism	38
map00561	Glycerolipid metabolism	35
map00230	Purine metabolism	29
map00630	Glyoxylate and dicarboxylate metabolism	19

© 2011 Biomedical Informatics