The kingdom of Plantae EST Indices: a resource for plant genomics community

Ali Masoudi-Nejad[1,2]
amasoudi@kais.kyoto-u.ac.jp

Ruy Jauregui[2]
ruy@kuicr.kyoto-u.ac.jp

Shuichi Kawashima[2]
shuichi@kuicr.kyoto-u.ac.jp

Susumu Goto[2]
goto@kuicr.kyoto-u.ac.jp

Minoru Kanehisa[2]
kanehisa@kuicr.kyoto-u.ac.jp

Takashi R. Endo[1]
endo@kais.kyoto-u.ac.jp

[1] Laboratory of Plant Genetics, Kyoto University, Kyoto 606-8502, JAPAN
[2] Bioinformatics Center, ICR, Kyoto University. Gokasho, Uji, Kyoto 611-0011, JAPAN

Keywords: Plant ESTs, ESTs clustering, Pathway, gene transcript, ESTs assembling

1 Introduction

In order to understand the function of all genes of an organism, it is now clear that the genome sequence alone may be not enough, especially if the organism shows a high degree of complexity, or like many plants have an extremely large genome (Wheat, 16000 Mb, compare to Arabidopsis, 180 Mb). Expressed Sequence Tag (EST) sequencing is a cost effective way to survey the expressed portion of the genome. The rapidly growing EST-database has become an invaluable tool for gene discovery, gene mapping and genome annotation [1]. The value of these transcripts regardless of how they are used can be enhanced if a set of non-redundant EST-index produced. There are some public databases which have generated such a gene index based on ESTs, including NCBI's UniGene [2], TIGR Gene Indices [3] and PlantGDB [4]. Each one of these database has its own advantage and disadvantages. We have tried a unique and complementary approach to construct an EST-based gene index for the whole kingdom of Plantae ESTs included in NCBI's dbEST, covering about a hundred and eighty different species.

2 Method and Results

One of the main problems with ESTs data is poor quality and contamination (vectors, repeats, and low-complexity regions), which can't be completely avoided. Each one of the public databases mentioned above suffers from one or more of those contaminants; even they have tried to avoid it. Each group have used different vectors and repeats database for decontamination analysis, while none of those reference databases cover all contaminants. To avoid these, we first collected all publicly available vectors and repeat data and constructed our custom non-redundant vector and repeats databases for filtering [Table 1].

Table 1: Vectors and repeats databases used for construction of our custom non-redundant database

Vectors	Repeats
UniVec	TIGR Plant repeats
emvec	TREP Triticae repeats
VectorZ	GIRI Repeat database

EST sequences for all species were downloaded from NCBI's dbEST and trimmed for polyA/T tails and low-complex sequences and sequences more than 100bp were selected. Repeats and vector

sequences were removed using programs Cross_Match and RepeatsMasker using our own custom filter. For clustering and assembling the data, we used CAP3 [5]. CAP3 assembles ESTs from the same gene under more stringent criteria than the other approaches, and was shown superior to TIGR assembler [6], and Phrap in its ability to distinguish gene family members while tolerating sequencing error. Since the efficiency of CAP3 can also be improved or changed by incorporating different option in the clustering algorithm. We used CAP3 to cluster ESTs with various stringency criteria (sequence identity P=80, 90, 92, 95, 97.5). Results showed that P=92 is the best threshold as is stringent enough to separate paralogs while capable of tolerating sequencing error to avoid miss-separation of ESTs from the same gene into two or more clusters. A range of overlap lengths (O) was initially examined but the results shows no significant difference, so O=40 (default) was used for clustering experiments. Finally the data were processed with CAP3 default options and with P=92. The results were slightly different in number of contigs produced by each process. Results of P=92 clustering were chosen for further annotation and analysis. Table 2 shows the results of assembling for few plants.

Table 2. Results of ESTs clustering with CAP3 (default option)

Species	No. of ESTs analyzed*	Contigs	Singletons
Triticum aestivum	**558098**	**31549**	**93823**
Zea mayz	**415211**	**25970**	**36284**
Hordeum vulgare	**381006**	**21380**	**44833**
Glycine max	**334036**	**24884**	**30222**
Arabidopsis thaliana	**322641**	**21879**	**23699**
Oryza sativa	**284007**	**22165**	**39504**
Saccharum officinarum	**246301**	**23810**	**47480**
Sorghum bicolor	**190946**	**19192**	**19215**
Medicago truncatula	**187763**	**16840**	**17695**
Pinus taeda	**173680**	**14353**	**14543**

* ESTs were downloaded from NCBI on 2004-10-6

3 Discussion

Functional annotation of singleton and contigs and mapping to the KEGG pathways will provide a tool for better annotation of the plant genomes and an insight to the network of the genes expressed in the same pathway for better understanding the function of the genes.

References

[1] Adams, M. D., Kerlavage, A. R., Fields, C., and Venter, J. C. (1993) 3,400 new expressed sequenced tags identify diversity of transcripts in human brain. *Nature Genetics*, 4: 256–267

[2] Boguski, M. S. and Schuler, G. D. (1995) ESTablishing a human transcript map. *Nature Genetics*, 10(4), 369–71

[3] John Quackenbush *et al.* (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, 29:159-164

[4] Qunfeng Dong, Shannon D. Schlueter, and Volker Brendel (2004) PlantGDB, plant genome database and analysis tools**.** *Nucleic Acids Res.* 32: D354 - 359.

[5] Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Research*, 6: 829–845

[6] Liang, F., *et al.* (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acids Research*, 28: 3657–3665