Generation of 7137 Non-redundant Expressed Sequence Tags from a Legume, *Lotus japonicus*

Erika Asamizu, Yasukazu Nakamura, Shusei Sato, and Satoshi Tabata*

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan

(Received 26 January 2000)

Abstract

For comprehensive analysis of genes expressed in a model legume, *Lotus japonicus*, a total of 22,983 5' end expressed sequence tags (ESTs) were accumulated from normalized and size-selected cDNA libraries constructed from young (2 weeks old) plants. The EST sequences were clustered into 7137 non-redundant groups. Similarity search against public non-redundant protein database indicated that 3302 groups showed similarity to genes of known function, 1143 groups to hypothetical genes, and 2692 were novel sequences. Homologues of 5 nodule-specific genes which have been reported in other legume species were contained in the collected ESTs, suggesting that the EST source generated in this study will become a useful tool for identification of genes related to legume-specific biological processes. The sequence data of individual ESTs are available at the web site: http://www.kazusa.or.jp/en/plant/lotus/EST/.

Key words: Legume; *Lotus japonicus*; cDNA; EST

Elucidation of the molecular mechanisms involved in symbiotic nitrogen fixation in legume plants is gaining in importance from the viewpoint of agricultural resources, and orthodox and molecular genetic analyses have been initiated using a model legume, *Lotus japonicus*. The genome of this plant is relatively small among legume plants, and is estimated to be about 450 Mb. Additional characteristics of this plant is its self-fertile nature and short generation time (2 to 4 months). Agrobacterium-mediated transformation experiments have been reported, and the use of *Ac* transposable element was evaluated toward transposon mutagenesis and tagging. Generation of genetic and molecular markers is essential for map-based cloning, and efforts along this line have been made using two accessions of *Gifu* and Funakura among several accessions of *L. japonicus*. However, higher polymorphism frequency was found between Gifu and Miyakojima accessions recently (unpublished result).

One of approaches for analysis of legume-specific biological processes is the identification of genes involved in these processes by comparative analysis of the expressed gene profiles of *L. japonicus* with those of other plant species. Up to now, more than 45,000 expressed sequence tags (ESTs) have been registered to the public DNA databases for *Arabidopsis thaliana* and *Oryza sativa*, while less than 1500 ESTs have been accumulated for *L. japonicus*. Here, we report large-scale collection of 5' end ESTs from the young plants of *L. japonicus*. Characteristic features of *L. japonicus* ESTs revealed by database search will also be presented.

1. Construction and Qualification of cDNA Libraries

*Lotus japonicus* Miyakojima MG-20 accession was grown on soil in a 20-hr photoperiod of fluorescent illumination (100 μE sec⁻¹ m⁻²) at 22°C for 2 weeks and whole plants were harvested. Preparation of polyadenylated RNA and conversion to cDNA were performed as described previously. Synthesized cDNA was resolved by 1% agarose gel electrophoresis, and fractions ranging from 0.5 to 3 kb and over 3 kb were separately recovered. The recovered fragments were cloned into the *EcoR I-Xho I* sites of pBluescript II SK- plasmid vector (Stratagene, USA) and transformed into the *Escherichia coli* XL1-Blue MRF' strain (Stratagene, USA) by electroporation. The cDNA library containing fragments over 3 kb was named the size-selected library. This library contained 6 x 10⁴ independent clones. Normalization was performed for the library containing 0.5–3 kb fragments as described previously. The normalized library contained 5 x 10⁵ independent clones.

The insert size distribution of clones was analyzed for both the normalized and size-selected libraries (Fig. 1). It was indicated that 73% of clones in the normalized library consisted of species less than 1.5 kb, whereas the...
content of cDNA species longer than 1.5 kb was 58% in the size-selected library. The quality of libraries with respect to the insert size was assessed by comparing the 5' end sequences of redundant isoforms of *Cicer arietinum* mRNA for ribulose 1,5-bisphosphate carboxylase small subunit (182 amino acid residues: aa), *Cucumis sativus* mRNA for glyoxysomal malate dehydrogenase (357 aa), and *Nicotiana tabacum* mRNA for cytokinin binding protein CBP57 (486 aa), and *Pisum sativum* mRNA for LoxG (869 aa). Of 30 sequences analyzed for ribulose 1,5-bisphosphate carboxylase small subunit, 23 (76.7%) contained the translation initiation codon. Similarly, 8 (80.6%) of 10 sequences obtained for glyoxysomal malate dehydrogenase, 25 (80.6%) of 31 sequences for cytokinin binding protein, and 10 (45.5%) of 22 sequences for LoxG contained the translation initiation codon. The result showed that the libraries constructed in this study contains full-length cDNA clones in high proportion for the genes encoding proteins of up to at least 500 aa. All the clones which contained the LoxG translation initiation codon were derived from the size-selected library, indicating the effectiveness of the size selection procedure for the generation of long cDNA species.

2. Features of Generated ESTs

A total of 22,983 cDNA clones including 18,280 clones from the normalized library and 4703 clones from the size-selected library were sequenced from their 5' ends. From the base composition of the sequences of randomly selected 640 ESTs (246,982 bases together), the GC content was estimated to be 49.2%.

To identify the number of independent EST species, clustering of the EST sequences was performed. The end sequences were compared with a dataset of itself using the BLASTN program, and clones that showed over 95% identity for more than 100 bp were included in the same group. As a result, 7137 non-redundant groups were generated. As cDNA clones were not always full length, it is possible that different regions of a single gene were included as non-overlapping ESTs. A more accurate number of independent gene species represented by these EST groups would be obtained by combination with the 3' end sequence information.

3. Sequence Similarity

Each sequence was translated into its amino acid sequence in six frames and subjected to similarity search against the NCBI-provided non-redundant protein database, nr, using the BLASTX program. Similarity between a deduced amino acid sequence and a known
sequence was judged to be significant when the P-value was less than 1.0-14. The result of analysis indicated that 4445 non-redundant EST groups have significant similarity to deposited sequences and the remaining 2692 groups are novel. Among the 4445 EST groups with significant similarity, 3302 showed similarity to genes with known function and the remaining 1143 to hypothetical genes. Notable AP2 domain containing genes, 21 DREBIA and DREB2B,22 RAV1,23 TINY,24 CONSTANS,25 and SCARECROW26 from A. thaliana were identified in the EST collection.

2. A large number of ESTs showing similarity to pathogen response genes, including hypersensitivity-related genes,15 chitinases,16 lectins,17 and thaumatins18 were identified.

3. One hundred ninety ESTs showing similarity to reported transcription factors and genes containing DNA-binding motifs were obtained. These include BELL1,19 a homeobox-leucine zipper gene,20 AP2 domain containing genes,21 DREBIA and DREB2B,22 RAV1,23 TINY,24 CONSTANS,25 and SCARECROW26 from A. thaliana.

4. Codon Usage

The codon usage of L. japonicus genes was analyzed using predicted translated sequences compiled from 87 randomly selected ESTs, which included 9071 codons (Table 3). As only the 5' end partial sequences were used for analysis, the table does not contain the information of the stop codon usage. The codon usage data shown in Table 3 is inconsistent with those of the codon usage data estimated from 47 L. japonicus genes registered in the databases [http://www.kazusa.or.jp/codon/]. Although our data do not include the 3'-end moieties, a possibility for this inconsistency could be due to the variation of gene species used for codon usage analysis because 27 of the 47 registered genes code for small GTP-binding proteins. It has been reported that the codon preference among genes was highly variable in higher organisms.27

The non-redundant ESTs generated in this study will be a useful resource for comprehensive analysis of gene function and regulation. The EST sequences reported in this paper appear in the GenBank/EMBL/DDJB databases with accession numbers AV406328-AV429310.

Acknowledgements: We thank Dr. Masayoshi Kawaguchi at the University of Tokyo for providing us the seeds of Miyakojima MG-20. We also thank Akiko Watanabe, Miko Yasuda, Kumi Iidesawa, Masako Ishikawa, Mamabu Yamada and Shigemi Sasamoto, for excellent technical assistance. This work was supported by the Kazusa DNA Research Institute Foundation.

References

1. Handegger, K. and Stougaard, J. 1992, Lotus japonicus, an autogamous, diploid legume species for classical and molecular genetics, Plant J., 2, 487-496.
2. Jiang, Q. and Gresshoff, P. M. 1997, Classical and molecular genetics of the model legume Lotus japonicus. Mol. Plant Microbe Interact., 10, 59-68.
3. Handegger, K., Stüller, J., Thyjkaer, T., and Stougaard, J. 1994, Transgenic plants: Agrobacterium-mediated transformation of the diploid legume Lotus japonicus, Cell bi-
16. Melchers, L. S., Apotheker-de Groot, M., van der Knaap, J. A. et al. 1994, A new class of tobacco chitinases homologous to bacterial exo-chitinases displays antifungal activity, Plant J., 5, 469–480.

17. Peumans, W. J. and van Damme, E. J. 1995, The role of lectins in plant defence, Histochem. J., 27, 253–271.

18. Ye, X. Y., Wang, H. X., and Ng, T. B. 1999, First chromatographic isolation of an antifungal thaumatin-like protein from French bean legumes and demonstration of its antifungal activity, Biochem. Biophys. Res. Commun., 263, 130–134.

19. Reiser, L., Modrussan, Z., Margossian, L., Samach, A., Ohad, N., Haughn, G. W., and Fischer, R. L. 1995, The BELL1 gene encodes a homeodomain protein involved in pattern formation in the Arabidopsis ovule primordium, Cell, 83, 735–742.

20. Schena, M. and Davis, R. W. 1994, Structure of homeobox-leucine zipper genes suggests a model for the evolution of gene families. Proc. Natl. Acad. Sci. USA, 91, 8393–8397.

21. Okamura, J. K., Caeter, B., Villarroel, R., Van Montagu, M., and Josefuku, K. D. 1997, The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in Arabidopsis, Proc. Natl. Acad. Sci. USA, 94, 7076–7081.

22. Liu, Q., Kasuga, M., Sakuma, Y. et al. 1998, Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in Arabidopsis, Plant Cell, 10, 1391–1406.

23. Kagaya, Y., Oohmiya, K., and Hattori, T. 1999, RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants, Nucleic Acids Res., 27, 470–478.

24. Wilson, K., Long, D., Swinburne, J., and Coupland, G. 1996, A Dissociation insertion causes a semidominant mutation that increases expression of TINY, an Arabidopsis gene related to APETALA2, Plant Cell, 8, 659–671.

25. Putterill, J., Robson, F., Lee, K., Simon, R., and Coupland, G. 1995, The CONSTANS gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors, Cell, 80, 847–857.

26. Di Laurentio, L., Wysocka-Diller, J., Malamy, J. E. et al. 1996, The SCARECROW gene regulates an asymmetric cell division that is essential for generating the radial organization of the Arabidopsis root, Cell, 86, 423–433.

27. Bernardi, G., Olofsson, B., Filipski, J. et al. 1985, The mosaic genome of warm-blooded vertebrates, Science, 228, 953–958.