Generation of 10,154 Expressed Sequence Tags from a Leafy Gametophyte of a Marine Red Alga, *Porphyra yezoensis*

Itoshi NIKAIADO,1 Erika ASAMIZU,2 Maiko NAKAJIMA,3 Yasukazu NAKAMURA,2 Naotsune SAGA,1,3 and Satoshi TABATA2,*

Graduate School of Marine Science and Technology, Tokai University, Shimizu, Shizuoka 424-8610, Japan,1 Kazusa DNA Research Institute, Kisarazu, Chiba 292-0812, Japan,2 and Center for Advanced Technology, Tokai University, Shimizu, Shizuoka 424-8610, Japan 3

(Received 16 March 2000; revised 19 May 2000)

Abstract

A total of 10,154 5'-end expressed sequence tags (EST) were established from the normalized and size-selected cDNA libraries of a marine red alga, *Porphyra yezoensis*. Among the ESTs, 2140 were unique species, and the remaining 8014 were grouped into 1127 species. Database search of the 3267 non-redundant ESTs by BLAST algorithm showed that the sequences of 1080 species (33.1%) have similarity to those of registered genes from various organisms including higher plants, mammals, yeasts, and cyanobacteria, while 2187 (66.9%) are novel. Codon usage analysis in the coding regions of 101 non-redundant EST groups showing significant similarity to known genes indicated the higher GC contents at the third position of codons (79.4%) than the first (62.2%) and the second position (45.0%), suggesting that the genome has been exposed to high GC pressure during evolution. The sequence data of individual ESTs are available at the web site http://www.kazusa.or.jp/en/plant/porphyra/EST/.

Key words: *Porphyra yezoensis; Rhodophyta; red alga; EST; codon usage*

*Porphyra* is a genus of marine red algae, and a total of 133 species have been reported from all over the world including several species of economic importance. *Porphyra* has a unique dimorphic life cycle consisting of two generations, a leafy gametophyte and a filamentous sporophyte, of completely different morphologies. The developmental process of its life cycle has been intensively investigated using one of the *Porphyra* species, *Porphyra yezoensis*.2–5 A laboratory culture system developed for this species has made it possible to reproduce a complete life cycle within a few months. Because of this system, *Porphyra* has recently been recognized as a model plant for fundamental and applied studies in marine biological sciences.6 Moreover, the small genome size estimated to be 2.7–5.3 x 10⁸ bp consisting of 2–7 chromosomes,7,8 which is the same order of magnitude as those of *Arabidopsis thaliana* and *Oryza sativa*, offers many approaches to study the genetics and genomics of *Porphyra*. Genetic analysis techniques are currently being developed for including a transformation system and AFLP mapping.9–13

The collection of expressed sequence tags (ESTs) has been performed to compile gene constituents of various genomes in a cost effective way (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). A large number of ESTs have been accumulated for a wide variety of model plants including *O. sativa*,14 *A. thaliana*,15 liverwort16 and *Chlamydomonas reinhardtii*.17 They are mostly from Viridiplantae (green plants), and only a limited number of ESTs are available for plants in other classes. With a final objective of understanding the genetic system in red algae, we initiated a large scale EST analysis of *P. yezoensis*. As the first part of this project, 10,154 ESTs have been established from normalized and size-selected cDNA libraries from a mixture of gametophytic and sporophytic tissues of *P. yezoensis* Ueda (strain TU-1). In this paper, details of EST collection and the features of the obtained ESTs revealed by the computer analyses are reported.

1. Construction and Qualification of cDNA Libraries

Leafy gametophytes of *Porphyra yezoensis* Ueda (strain TU-1) were grown in weekly renewed ESL medium containing 3.5% Sealife powder (Marinetech Co., Ltd., Japan) and 1% (v/v) ESS₂ stock solution (pH 8.0) in a 10 hr photoperiod of fluorescent illumination (50 μmol·m⁻²·s⁻¹) at 15°C with constant air...
bubbling for 4 weeks. Development of sporophytes was observed which constituted less than 5% of the harvested thalli. Total RNA was extracted by the modified SDS/phenol method. Briefly, 5.0 g of frozen thalli of Porphyra yezoensis were ground to powder in liquid nitrogen and treated with 10 ml of 1 M Tris-HCl (pH 9.0)/1.0% SDS and 10 ml of phenol (pH 9.0). After extraction with 10 ml of phenol/chloroform (pH 9.0) for four times, the sample was mixed with the equal volume of 4 M LiCl and kept at 4°C for overnight. Total RNA was precipitated by centrifugation at 8000×g for 30 min. Poly(A)+ RNA was purified from the total RNA as described previously. Eight micrograms of Poly(A)+ RNA was used as a template for synthesis of cDNA. Size-selection and cloning of cDNA was performed as described. Normalization was performed for the library containing 0.5–3 kb fragments as described. The normalized library contained 2×10^5 independent clones. The cDNA library containing fragments of over 3 kb was named the size-selected library. This library contained 1×10^4 independent clones.

Figure 1 shows the distribution of the insert length of the clones which were subjected to sequence analysis. The average lengths of the inserts for the normalized and size-selected libraries were 0.97 kb and 2.09 kb, respectively. The quality of libraries with respect to coverage of 5'-termini was assessed by comparison of the 5'-end sequences to known protein sequences. Among the 106 clones randomly chosen from the normalized library and the 105 clones from the size-selected library, 74 (69.8%) and 69 (65.7%) were found to contain the translation initiation codon indicating that roughly two-thirds of the cDNA clones in both libraries are full-length.

### Table 1. The result of similarity search against the public non-redundant protein database. The numbers of EST groups and clones that showed similarity to known genes from various organisms are indicated.

| Similarity                          | Number of groups | Number of clones |
|-------------------------------------|------------------|-----------------|
| Genes of known function             | 934              | 3799            |
| Hypothetical genes                  | 146              | 915             |
| No similarity                       | 2187             | 5440            |
| Total                               | 3267             | 10154           |

a) showed similarity to genes of known function, b) showed similarity to hypothetical genes that have no definition of function, c) showed no similarity.

### 2. Features of Generated ESTs

Single-pass sequencing from the 5' end of the cDNA was performed for 10,154 clones from both normalized and size-selected libraries according to the procedure described previously. The vector-derived sequence and ambiguous sequences were removed from the collected EST sequences prior to the computer analyses. The average length of ESTs thus obtained was 470 bp and the average GC content was 65.2%.

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, a total of 3267 non-redundant
Table 2. Classification of 934 EST groups based on the annotations of their homologous protein entries in the public databases.

| Functional categories                  | Number of non-redundant groups |
|----------------------------------------|-------------------------------|
| Energy metabolism                      | 130                           |
| Protein synthesis                      | 112                           |
| Protein fate                           | 104                           |
| Cellular structure, organization and biogenesis | 77                           |
| Transport and binding proteins         | 72                            |
| Regulatory functions                   | 55                            |
| Signal transduction                    | 37                            |
| Cellular processes                     | 36                            |
| Amino-acid biosynthesis                | 32                            |
| DNA metabolism                         | 21                            |
| Fatty acid and phospholipid metabolism | 21                            |
| Purines, pyrimidines, nucleosides, and nucleotides | 16                            |
| Central intermediary metabolism        | 14                            |
| General transcription                  | 13                            |
| Biosynthesis of cofactors, prosthetic groups, and carriers | 7                            |
| Growth and development                 | 7                             |
| Pathogen responses                     | 7                             |
| Secondary metabolism                   | 5                             |
| Environmental response                 | 4                             |
| Other categories                       | 2                             |
| Unclassified                           | 162                           |
| **Total**                              | **934**                       |

EST groups were obtained, of which 1127 were generated by clustering of 8014 ESTs while 2140 were unique. It should be noted however that the number of EST species generated does not indicate the number of the corresponding genes because non-redundant ESTs may originate from different regions of a single gene.

3. Database Search

Each sequence from the 3267 non-redundant EST species was translated into amino acid sequences 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$^{-14}$. As a result, 1080 (33.1%) of the 3267 clustered ESTs showed sequence similarity to genes registered in the public databases, and 2187 (66.9%) were novel sequences (Table 1). Similar features have been reported in the EST analysis of a green alga, *Chlamydomonas reinhardtii*, where 2616 (76.2%) out of 3433 non-redundant EST species were novel.17 By contrast, only 2692 (37.7%) out of 7137 non-redundant EST species in a legume, *Lotus japonicus*, were new sequences.22 The reason for this feature could be that fewer sequences from red and green algae have been registered in the public DNA databases compared to those from higher plants. Another possibility is that, because the Rhodophyta group including *P. yezoensis* and the Viridiplantae group (green plant) are evolutionary distant, genes specific to each group and/or with highly divergent sequences may be present. The details of search results are provided through World Wide Web at http://www.kazusa.or.jp/en/plant/porphyra/EST/.

Genes which showed sequence similarity to those registered in the public DNA databases were classified according to the biological roles or biochemical functions23 (Table 2). Notable features found so far are as follows:

1. 41.9% of them were similar to protein sequences from higher plant species largely obtained from *Arabidopsis thaliana*. However, significant portions showed similarity to genes of divergent species: 18.0% showed similarity to proteins from mammals, 8.2% to yeast, 7.1% to cyanobacteria (*Synechocystis* sp., *Synechococcus* sp., and *Anabaena* sp.), 5.4% to green alga (*Volvox* sp., *Chlamydomonas* sp., and *Chlorella* sp.), and 5.4% to red alga (*Gracilaria* sp., *Porphyra purpurea*, *Cyanidium caldarium*, and *Porphyridium cruentum*).

2. A homologue of phytohormone-induced protein in higher plants have not been found in the *P. yezoensis* EST collection. It is notable that a homologue of Dwarf1-encoding gene was obtained, which was reported to be involved in brassinosteroid biosynthesis.24

3. Several homologues of genes in higher plants associated with growth and development were obtained. These include *AGO1*,25 *CONSTANS*,26 and a senescence-associated gene.27

4. Features of Protein-Coding Regions

To characterize the features of coding regions of *P. yezoensis* genes, consensus sequences were generated by assembly of EST sequences. The sequences of 101 non-redundant EST species which showed relatively high similarity (score >200 and P<1.0$^{-15}$) to known gene sequences were translated into amino acid sequences. Sixty-one out of 101 EST sequences contained a putative stop codon followed by the 3′ untranslated region. The GC content of the predicted coding region (61.7%) was slightly higher than that of the predicted 3′ untranslated region (59.3%). The codon usage was analyzed using sequences of predicted coding regions of 101 assembled non-redundant EST sequences containing 21,739 codons. As shown in Table 3, no strong bias in frequency of codon usage was observed. One notable feature is that the codons with G/C at the third position are frequently utilized: The GC content of the third position (79.4%) was much higher than that at the first (62.2%) and second (45.0%) positions (Table 3). This suggests that
the genome of \textit{P. yezoensis} has been exposed to high GC pressure during evolution. These characteristics of base preference may be useful for prediction of protein-coding regions in the \textit{Porphyra yezoensis} genome. The EST sequences reported in this paper appear in the GenBank/EMBL/DDBJ databases with accession numbers AV429311-AV439464.

Acknowledgments: We thank M. Yamada for excellent technical assistance. This work was supported by the Kazusa DNA Research Institute Foundation, High-technology Research Center Program from the Promotion and Mutual Aid Corporation for Private Schools, and Research and Development Program from Tokai University.

References

1. Yoshida, T., Notoyama, M., Kikuchi, N., and Miyata, M. 1997, Catalogue of species of \textit{Porphyra} in the world, with special reference to the type locality and bibliography, \textit{Nat. Hist. Res., Special Issue}, 3, 5-18.
2. Kitade, Y., Taguchi, G., Shin, J.-A., and Saga, N. 1998, \textit{Porphyra} monospore system (Bangiales, Rhodophyta): A model for the developmental biology of marine plants, \textit{Phycol. Res.}, 46, 17-20.
3. Yamazaki, A., Nakanishi, K., and Saga, N. 1998, Axenic tissue culture and morphogenesis in \textit{Porphyra yezoensis} (Bangiales, Rhodophyta), \textit{J. Phycol.}, 34, 1082-1087.
4. Kitade, Y., Fukuda, S., and Saga, N. 1999, Preliminary study on early development of monospore of \textit{Porphyra yezoensis} (Bangiales, Rhodophyta), \textit{Fish Genet. Breed. Sci.}, 28, 27-34.
5. Kitade, Y., Yamazaki, S., Watanabe, T., and Saga, N. 1999, Structural features of a gene encoding the vacuolar \(H^+\)-ATPase c subunit from a marine red alga, \textit{Porphyra yezoensis}, \textit{DNA Res.}, 6, 307-312.
6. Kuwano, K., Aruga, Y., and Saga, N. 1996, Cryopreservation of clonal gametophytic thalli of \textit{Porphyra} (Rhodophyta), \textit{Plant Sci.}, 116, 117-124.
7. Kapraun, D. F., Hinson, T. K., and Lemus, A. J. 1991, Karyology and cytophotometric estimation of Inter- and Intraspecific nuclear DNA variation in four species of \textit{Porphyra} (Rhodophyta), \textit{Phycologia}, 30, 458-466.
8. Le Gall, Y., Brown, S., Marie, D., Meijjad, M., and Kloareg, B. 1993, Quantification of nuclear DNA and GC content in marine macroalgae by flow cytometry of isolated nuclei, \textit{Protoplasma}, 173, 123-132.
9. Miura, A. 1985, Genetic analysis of the variant color types of \textit{Porphyra yezoensis} (Rhodophyta, Bangiaceae). In: Hara, H. (ed) Origin and Evolution of Diversity in Plants and Plant Communities, Academia Scientific Book Inc., Tokyo, pp. 270-284.
10. Ohme, M. and Miura, A. 1988, Tetrad analysis in conchospore germings of \textit{Porphyra yezoensis} (Rhodophyta, Bangiales), \textit{Plant Sci.}, 57, 135-140.
11. Reith, M. and Munholland, J. 1993, A high-resolution gene map of chloroplast genome of the red alga \textit{Porphyra purpurea}, \textit{Plant Cell}, 5, 465-475.
12. Liu, Q. Y., van der Meer, J. P., and Reith, M. E. 1994, Isolation and characterization of phase-specific complementary DNAs from sporophytes and gametophytes of Porphyra purpurea (Rhodophyta) using subtracted complementary DNA libraries, J. Phycol., 30, 513-520.

13. Hamada, J., Shin, J.-A., and Miura, A. 1994, Mapping of centromere distances of light green and light red genes in Porphyra yezoensis (Rhodophyta, Bangiales), Jpn. J. Phycol., 42, 401-406.

14. Sasaki, T., Song, J., Koga-Ban, Y. et al. 1994, Toward cataloguing all rice gene: large-scale sequencing of randomly chosen rice cDNA from a callus cDNA library, Plant J., 6, 615-624.

15. Höfte, H., Desprez, T., Arnaud, J. et al. 1993, An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNA from Arabidopsis thaliana, Plant J., 4, 1051-1061.

16. Nagai, J., Yamamoto, K., Sakaida, M., Yoda, H., Fukuzawa, H., and Ohyama, K. 1999, Expressed sequence tags from immature female sexual organ of a liverwort, Marchantia polymorpha, DNA Res., 6, 1-11.

17. Asamizu, E., Nakamura, Y., Sato, S., Fukuzawa, H., and Tabata, S. 1999, A large scale structural analysis of cDNAs in a unicellular green alga, Chlamydomonas reinhardtii. I. Generation of 3433 non-redundant expressed sequence tags, DNA Res., 6, 369-373.

18. Kitade, Y., Yamazaki, S., and Saga, N. 1996, A method for extraction of high molecular weight DNA from the macroalga Porphyra yezoensis (Rhodophyta), J. Phycol., 32, 496-498.

19. Naito, S., Hirai, M. Y., Chino, M., and Komeda, Y. 1994, Expression of a soybean (Glycine max [L.] Merr.) seed storage protein gene in transgenic Arabidopsis thaliana and its response to nutritional stress and to abscisic acid mutations, Plant Physiol., 104, 497-503.

20. Bonaldo, M. F., Lemmon, G., and Soares, M. B. 1996, Normalization and subtraction: two approaches to facilitate gene discovery, Genome Res., 6, 791-806.

21. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool, J. Mol. Biol., 215, 403-410.

22. Asamizu, E., Nakamura, Y., Sato, S., and Tabata, S. 2000, Generation of 7,137 Non-redundant expressed sequence tags from a legume, Lotus japonicus, DNA Res., 7, 127-130.

23. Lin, X., Kaul, S., Rounsley, S., et al. 1999, Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana, Nature, 402, 761-768.

24. Choe, S., Dilkes, B. P., Gregory, B. D. et al. 1999, The Arabidopsis dwarf1 mutant is defective in the conversion of 24-methylenecholesterol to campesterol in brassinosteroid biosynthesis, Plant Physiol., 119, 897-907.

25. Bohmert, K., Camus, I., Bellini, C., Bouchez, D., Caboche, M., and Benning, C. 1998, AGO1 defines a novel locus of Arabidopsis controlling leaf development, EMBO J., 17, 171-180.

26. 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.

27. Panavas, T., Pikula, A., Reid, P. D., Rubinstein, B., and Walker, E. L. 1999, Identification of senescence-associated genes from daylily petals, Plant Mol. Biol., 40, 237-248.
