DNA RESEARCH 6, 369–373 (1999)

A Large Scale Structural Analysis of cDNAs in a Unicellular Green Alga, *Chlamydomonas reinhardtii*. I. Generation of 3433 Non-redundant Expressed Sequence Tags

Erika ASAMIZU,[1] Yasukazu NAKAMURA,[1] Shusei SATO,[1] Hideya FUKUZAWA,[2] and Satoshi TABATA[1,*]

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan[1] *and Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan*[2]

(Received 23 November 1999)

Abstract

To understand genetic information carried in a unicellular green alga, *Chlamydomonas reinhardtii*, normalized and size-selected cDNA libraries were constructed from cells at photoautotrophic growth, and a total of 11,571 5′-end sequence tags were established. These sequences were grouped into 3433 independent EST species. Similarity search against the public non-redundant protein database indicated that 817 groups showed significant similarity to registered sequences, of which 140 were of previously identified *C. reinhardtii* genes, but the remaining 2616 species were novel sequences. The coverage of full-length protein coding regions was estimated to be over 60%. These cDNA clones and EST sequence information will provide a powerful source for the future genome-wide functional analysis of uncharacterized genes.
Key words: green alga; *Chlamydomonas reinhardtii*; cDNA; EST; normalized library; similarity search

1. Introduction

Chlamydomonas reinhardtii, a single-celled green alga, is a haploid organism that is easily grown and maintained in laboratories. It has served as an important model organism for the study of the flagellar assembly,[1,2] cell wall biogenesis,[3] gametogenesis,[4] phototaxis,[5] and mating processes.[6] Early[7,8] and recent studies using mutants[9,10] also revealed the utility of this organism as a model system to perform genome-wide studies of photosynthetic function.[11] An advantage of using *C. reinhardtii* is that electroporation provides a high frequency of transformation,[12] but systematic marker accumulation and mapping have not been reported.

Although sequencing whole genomes and cDNA are the primary focus of genome analysis in higher eukaryotes, overall information of genes expressed in an organism is obtained by developing cDNA sequencing methods that are more cost effective. In addition, information on the cell types and the environmental conditions affecting expression as well as expression levels can be obtained by systematic sequencing of cDNAs.

With this in mind, we aimed at sequencing cDNA clones from *C. reinhardtii* cells grown under different conditions. As the first part of this project, EST data acquisition from cDNA libraries made from cells at photoautotrophic growth was carried out. In order to obtain gene species in the form containing full coding regions as efficiently, two types of cDNA libraries, normalized and size-selected libraries, were used. In this paper, details of the cDNA library construction and EST data analysis are reported.

2. Materials and Methods

2.1. Strain and culture conditions of cells

C. reinhardtii Dangeard C-9 (mt^-) strain used was obtained from the IAM culture collection at the University of Tokyo and is maintained in the laboratory of H. Fukuzawa at Kyoto University. Cells were photoautotrophically grown in a HSM minimal liquid medium[13] under continuous illumination of 100 $\mu E \cdot m^{-2} \cdot s^{-1}$ from cool-white fluorescent tubes at 28°C with constant bubbling of ordinary air containing 0.04% (v/v) CO_2. When the cultures reached a density of 1×10^6 cells·ml^{-1}, cells were harvested by centrifugation at 2000×g for 10 min at 4°C, washed once with the HSM media, resuspended in TE-buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.5), and immediately frozen in liquid nitrogen.

2.2. Preparation of polyadenylated RNA

Total RNA was extracted by the guanidium thiocyanate/CsCl ultracentrifuge method.[14] Next, 3.9 g of

Communicated by Mituru Takanami
* To whom correspondence should be addressed. Tel. +81-438-52-3933, Fax. +81-438-52-3934, E-mail: tabata@kazusa.or.jp

frozen *C. reinhardtii* cells were ground to powder in liquid nitrogen and treated with 25 ml of 4 M guanidium thiocyanate buffer [4 M guanidium thiocyanate, 0.1 M Tris −HCl (pH 7.5), 10 mM EDTA, 0.5% Sarkosyl, 0.1% β-mercaptoethanol]. After four extractions with 25 ml of phenol/chloroform (pH 8.0), the sample was loaded into four tubes (13PET tube, Hitachi Koki, Japan) containing 3.5 ml of 5.7 M CsCl cushion (5.7 M CsCl, 0.1 M EDTA) and ultracentrifuged at 32,000 rpm for 22 hr at 20°C using himac CP 65β (Hitachi Koki, Japan). The yield of total RNA was 720 μg. Poly(A)$^+$ RNA was purified using Oligotex-dT30 Super (Nippon Roche, Japan).

2.3. *Construction of cDNA libraries*

Synthesis of cDNA was performed using 6.8 μg of Poly(A)$^+$ RNA as a template by cDNA Synthesis kit (Stratagene, USA) following the manufacturer's instructions except for the first strand synthesis step, in which SuperScript II reverse transcriptase (Life Technologies, USA) was used. Poly(A)$^+$ RNA was primed with an oligo(dT)$_{18}$ primer carrying an *Xho*I site, and the first strand synthesis reaction was performed at 52°C. After the second strand systhesis, the termini of cDNAs were converted to blunt-end and *Eco*RI adaptors were ligated to both ends. Then the *Xho*I sites at the 3′ ends of cDNAs were generated by digestion with *Xho*I. 5-Methyl dCTP instead of dCTP was used for synthesis of the first strand to make the inner portions of cDNAs resistant to *Xho*I digestion.

The size selection of cDNAs prior to vector ligation was performed as follows. Synthesized cDNAs were resolved by 1% agarose gel electrophoresis, and fractions ranging from 1 to 3 kb and over 3 kb were separately recovered from the gel using QIAquick Gel Extraction Kit (QIAGEN, Germany). The recovered fragments were cloned into *Eco*RI-*Xho*I sites of pBluescript II SK- plasmid vector (Stratagene, USA) and transformed into *Escherichia coli* XL1-Blue MRF' strain (Stratagene, USA) by electroporation. The cDNA library containing over 3 kb fragments was named size-selected library. The library contained 2×10^4 independent clones.

Normalization was performed for the library containing 1–3 kb fragments as described by Bonaldo et al.[15] At first, a single-stranded library was prepared from 5 μg of plasmid DNA by the combined action of gene II endonuclease of phage F1 (Life Technologies, USA) and exonuclease III (Life Technologies, USA). Then, self-hybridization was performed with an excess of cDNA inserts generated by PCR. For amplification of the inserts, about 5.0 ng of DNA template (single-stranded plasmids) was mixed with 1 μl of 100 μM T7 primer (5′-TAATACGACTCACTATAGGG-3′) and 1 μl of 100 μM SK primer (5′-GCTCTAGAACTAGTGGATC-3′), then PCR was performed by Taq DNA polymerase (TaKaRa, Japan) using a Perkin-Elmer 9600 Thermal Cycler: 7 min ramping up from room temperature to 94°C; 20 cycles of 1 min at 94°C, 2 min at 55°C, 3 min at 72°C, and a final extension for 7 min at 72°C. The PCR product was ethanol-precipitated and dissolved in 1.5 μl of water, and then mixed with 5 μl of single-stranded library DNA (50 ng) in formamide, 0.5 μl (10 μg) of 5′-blocking oligo nucleotide mixture with the following sequences: 5′-GCTCTAGAACTAGTGGATCCCCCGGGCTGCAGGAATTCG-3′ and 5′-AATTCGGCACGAG-3′, and 0.5 μl (10 μg) of 3′-blocking oligo nucleotide mixture with the sequences of 5′-CTCGAGGGGGGGCCCGGTA-3′ and 5′-GTACCCAATTCGCCCTATAGTGAGTCGTATTA-3′. After heating at 80°C for 3 min, 1 μl of 10×buffer [1.2 M NaCl, 0.1 M Tris −HCl (pH 8.0), 50 mM EDTA, and 10% SDS] and 1.5 μl of water was added, and hybridization was performed at 30°C for 24 hr. The remaining single-stranded circles were purified using hydroxyapatite chromatography and converted to double strands using Klenow Fragment (TaKaRa, Japan). The obtained double-stranded DNA was transformed into host *E. coli* by electroporation. The normalized library contained 1×10^6 independent clones.

2.4. *Template preparation and sequencing*

Plasmid DNA was prepared by the alkaline lysis method[16] in a 96-well unit. The insert size of each clone was measured by agarose gel electrophoresis after digestion with *Apa*I and *Sma*I. Sequence reaction was performed using a BigDye Terminator Cycle Sequencing Kit (PE Applied Biosystems, USA) and a DYEnamic ET terminator kit (Amersham Pharmacia Biotech, USA), and sequencing was done with an ABI PRISM 377 DNA Sequencer (PE Applied Biosystems, USA) and a MegaBACE 1000 sequencing system (Amersham Pharmacia Biotech, USA).

2.5. *Sequence data analysis*

The vector-derived sequence and ambiguous sequences were removed from the collected EST sequences prior to the computer-aided analyses. Each sequence was translated into amino acid sequences in six frames using the universal codon usage table. Then they were subjected to similarity search against the non-redundant protein database, nr, provided by NCBI using the BLAST algorithm.[17] Grouping of the EST sequences was performed as follows. The end sequences were compared with a data set of itself using the BLASTN program. Clones that showed over 95% identity for more than 100 bp were grouped together.

3. Results and Discussion

3.1. *Features of the generated ESTs*

A total of 11,571 clones, including 7310 clones from the normalized library and 4261 clones from the size-selected

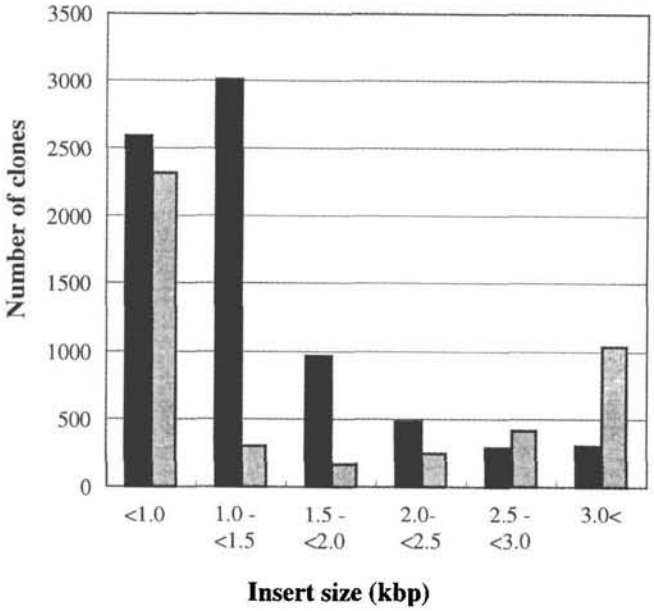

Figure 1. Insert size distribution of cDNA clones from the normalized (solid bars) and the size-selected libraries (gray bars). Insert sizes were measured as described in materials and methods.

library, were sequenced from their 5′ ends. The average length of the EST sequences was 474 bp. The GC content of the ESTs was 63.3%, estimated from the sequences of 50 randomly selected ESTs. To identify the number of independent species, the end sequences were compared with a data set of itself using the BLASTN program. Clones that showed over 95% identity for more than 100 bp were grouped together. As a result, the 5′ end sequences were clustered into 3433 non-redundant groups. However, the real number of independent gene species would be less than 3433, as there is a possibility that different regions of a single gene generate non-redundant ESTs, even though the ESTs were constituted from normalized and size-selected libraries.

3.2. Size distribution of clones

The insert size distribution of clones was analyzed for both the normalized and size-selected libraries. As shown in Fig. 1, most of the clones in the normalized library contained inserts of less than 1.5 kb species. On the other hand, clones with inserts of longer than 3 kb were relatively abundant in the size-selected library, although the content of clones with inserts less than 1 kb inserts was also high. The high content of clones with shorter inserts is likely to be due to contamination by the shorter fragments at the size-selection step, because it was possible to lower the content by running electrophoretic separation of fragments at a relatively low voltage (unpublished observation).

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 previously reported *C. reinhardtii* genes and to known genes from other organisms are indicated.

Similarity	Number of Groups	Number of Clones
Genes from *C. reinhardtii*	140	1622
Genes from other organisms	677	6116
No similarity	2616	3833
Total	3433	11571

3.3. Genes identified by database search

The EST sequences were compared with the public non-redundant protein sequence database with the BLASTX program, and P < 1.0e-14 was taken as the level of significance. As shown in Table 1, 140 EST groups among the 3433 groups showed similarity to previously identified *C. reinhardtii* genes and 677 groups to genes from a wide variety of organisms. These 817 homologous groups were comprised from a total of 7738 clones. The remaining 2616 novel groups were comprised from 3833 clones, suggesting the possibility that genes expressed at low levels are included in the novel groups.

Coverage of full-length protein-coding regions was analyzed using the alignment between the EST sequence and the corresponding known protein sequence. By alignment of 100 randomly chosen clones from the normalized library, 64% were shown to contain the N-terminus. Clones with inserts of longer than 3 kb from the size-selected library were also analyzed, and 62% were found to contain the N-terminus. This result implies that genes encoding large proteins are likely to be represented in the size-selected library. The clone with the longest insert (7.2 kb), which was shown to contain the full coding region, had similarity to a mouse transcriptional repressor gene.[18]

Genes whose functions could be predicted from the similarity were classified into 16 categories based on their biological roles.[19] The number of EST groups classified in each category are summarized in Table 2. The search results including the names of proteins encoded by the *C. reinhardtii* genes and other organisms and all the EST clones with their accession numbers are provided through the Internet at http://www.kazusa.or.jp/en/plant/chlamy/EST/.

Notable genes in these categories are as follows:

1. A homologue of the gene for *Zea mays* chroloplast SecA[20] protein was obtained. Because the protein translocation system across the internal thylakoid membranes of chloroplasts is not completely undertood,[21] *C. reinhardtii* may be an appropriate model system to analyze this process. This gene was

Table 2. Classification of 817 genes matched with non-redundant ESTs into 16 functional categories.

Functional categories	Number of non-redundant ESTs
Translation	179
Photosynthesis and respiration	85
Energy metabolism	65
Transport and binding proteins	52
Amino acid biosynthesis	48
Regulatory functions	46
Transcription	40
Cellular processes	35
Biosynthesis of cofactors, prosthetic groups, and carriers	23
Fatty acid, phospholipid and sterol metabolism	23
Central intermediary metabolism	20
Cell envelope	15
Purines, pyrimidines, nucleosides, and nucleotides	11
DNA replication, restriction, modification, recombination, and repair	6
Other categories	38
Hypothetical	131
	817

classified into the category of cellular processes.

2. The largest number of genes previously identified in *C. reinhardtii* were included in the category of photosynthesis and respiration. These include genes for light harvesting complex I protein precursor, oxygen-evolving enhancer protein 2,[22] ribulose-1,5-bisphosphate carboxylase/oxygenase activase, cytochrome b6f 4.6 kDa subunit PetM,[23] photosystem I 20 kDa protein, chlorophyll a/b-binding protein, ferredoxin,[24] ubiquinol-cytochrome c oxidoreductase,[25] and apoplastocyanin (PC6-2) precursor.[26]

3. Homologues of genes encoding key proteins for morphogenesis in higher plants were obtained. These include *Arabidopsis NPH* gene[27] and *APETALA2* gene.[28] They were included in the class of regulatory functions.

4. As a characteristic of *C. reinhardtii*, chlamyopsin[29] and flagellar radial spoke protein genes[30] were obtained.

Among the 3433 non-redundant ESTs, only 817 (23%) showed similarity to known genes. One reason could be that the number of sequences from algae species registered in the public database is still small. According to information on the number of entries by taxonomies provided by NCBI (http://www.ncbi.nlm.nih.gov:80/Taxonomy/tax.html), 61,102 among the 63,622 entries for Viridiplantae (green plant) entries in the non-redundant protein database are of higher plants, and only 2414 are of green algae. For the prediction of gene function, further experimental approaches are obviously required. One promising approach would be the examination of transcripsional characteristics by an array monitoring system. Generation of a macro-array covering 3433 EST groups is in progress.

The EST sequences reported in this paper appear in the GenBank/EMBL/DDBJ databases with accession numbers AV386458-AV398028.

Acknowledgments: We thank A. Watanabe and M. Yamada for excellent technical assistance. This work was supported by the Kazusa DNA Research Institute Foundation.

References

1. Aselson, C. M. and Lefebvre P. A. 1998, Genetic analysis of flagellar length control in *Chlamydomonas reinhardtii*: A new long-flagella locus and extragenic supressor mutations, *Genetics*, **148**, 693–702.
2. Diener, D. R., Curry, A. M., Johnson, K. A. et al. 1990, Rescue of a paralyzed flagella mutant of *Chlamydomonas* by transformation, *Proc. Natl. Acad. Sci. USA*, **87**, 5739–5743.
3. Woessner J. P. and Goodenough U. W. 1992, Zygote and vegetative cell wall proteins in *Chlamydomonas reinhardtii* share a common epitope, $(\text{Ser Pro})_x$, *Plant Sciences*, **83**, 65–76.
4. Gloeckner, G. 1997, Cloning and characterization of *LRG5*, a gene involved in blue light signaling in *Chlamydomonas* gametogenesis, *Plant J.*, **42**, 1264–1268.
5. Hegemann, M. 1997, Vision in microalgae, *Planta*, **203**, 265–274.
6. Ferris, P. J. and Goodenough, U. W. 1994, A mating type-linked gene cluster expressed in *Chlamydomonas* participates in the uniparental inheritance of the chloroplast genome, *Cell*, **76**, 1135–1145.
7. Moll, B. and Levine, R. P. 1970, Characterization of photosynthetic mutant strain of *Chlamydomonas reinhardtii* deficient in phosphoribulokinase activity, *Plant Physiol.*, **27**, 1–7.
8. Levine, R. P. and Goodenough, U. W. 1970, The genetics of photosynthesis and of the chloroplast in *Chlamydomonas reinhardtii*, *Ann. Rev. Genetics*, **4**, 397–408.
9. Spreitzer, R. J. and Mets, L. J. 1980, Non-mendelian mutation affecting ribulose-1,5-bisphosphate carboxylase/oxygenase, *Nature*, **285**, 114–115.
10. Erickson, J. M., Rahire, M., and Rochaix, J. D. 1984, *Chlamydomonas reinhardtii* gene for the 32,000 mol. wt. protein of photosystem II contains four large introns and is located entirely within the chloroplast inverter repeat, *EMBO J.*, **3**, 2753–2762.
11. Rochaix, J. D. 1995, *Chlamydomonas reinhardtii* as the photosynthetic yeast, *Annu. Rev. Genet.*, **29**, 209–230.
12. Shimogawara, K., Fujiwara, S., Grossman, A. R., and Usuda, H. 1998, High efficiency transformation of

Chlamydomonas reinhardtii by electroporation, *Genetics*, **148**, 1821–1828.
13. Sueoka, N. 1960, Mitotic replication of deoxyribonucleic acid in *Chlamydomonas reinhardtii*, *Proc. Natl. Acad. Sci. USA*, **46**, 83–91.
14. Sambrook, J., Fritsch, E. F., and Maniatis T. 1989, Molecular cloning: A laboratory manual, 2nd edn. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, pp F.4-F.5.
15. Bonaldo, M. F., Lennon, G., and Soares, M. B. 1996, Normalization and subtraction: two approaches to facilitate gene discovery, *Genome Res.*, **6**, 791–806.
16. Stowers, L., Herrnstadt, C., Grothe, A. et al. 1992, Rapid isolation of plasmid DNA, *Am. Biotechnol. Lab.*, **10**, 48.
17. 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.
18. Ayer, D. E., Lawrence, Q. A., and Eisenman, R. N. 1995, Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3, *Cell*, **10**, 767–776.
19. Kaneko, T., Sato, S., Kotani, H. et al. 1996, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, **3**, 109–136.
20. Voelker, R., Mendel-Hartvig, J., and Barkan, A. 1997, Transposon-disruption of a maize nuclear gene, tha1, encoding a chloroplast SecA homologue: *in vivo* role of cp-SecA in thylakoid protein targeting, *Genetics*, **145**, 467–478.
21. Smith, T. A. and Kohorn, B. D. 1994, Mutations in a signal sequence for the thylakoid membrane identify multiple protein transport pathways and nuclear suppressors, *J. Cell Biol.*, **126**, 365–374.
22. Mayfield, S. P., Rahire, M., Frank, G., Zuber, H., and Rochaix, J. D. 1987, Expression of the nuclear gene encoding oxygen-evolving enhancer protein 2 is required for high levels of photosynthetic oxygen evolution in *Chlamydomonas reinhardtii*, *Proc. Natl. Acad. Sci. USA*, **84**, 749–753.
23. de Vitry, C., Breyton, C., Pierre, Y., and Popot, J. L. 1996, The 4-kDa nuclear-encoded PetM polypeptide of the chloroplast cytochrome b6f complex. Nucleic acid and protein sequences, targeting signals, transmembrane topology, *J. Biol. Chem.*, **271**, 10667–10671.
24. Stein, M., Jacquot, J. P., and Miginiac-Maslow, M. 1993, A cDNA clone encoding *Chlamydomonas reinhardtii* preferredoxin, *Plant Physiol.*, **102**, 1349–1350.
25. Atteia, A. and Franzen, L. G. 1996, Identification, cDNA sequence and deduced amino acid sequence of the mitochondrial Rieske iron-sulfur protein from the green alga *Chlamydomonas reinhardtii*. Implications for protein targeting and subunit interaction, *Eur. J. Biochem.*, **237**, 792–799.
26. Merchant, S., Hill, K., Kim, J. H., Thompson, J., Zaitlin, D., and Bogorad, L. 1990, Isolation and characterization of a complementary DNA clone for an algal preapoplastocyanin, *J. Biol. Chem.*, **265**, 12372–12379.
27. Motchoulski, A. and Liscum, E. 1999, *Arabidopsis* NPH3: A NPH1 photoreceptor-interacting protein essential for phototropism, *Science*, **286**, 961–964.
28. Okamuro, J. K., Caster, B., Villarroel, R., Van Montagu, M., and Jofuku, 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.
29. Deininger, W., Kroger, P., Hegemann, U., Lottspeich, F., and Hegemann, P. 1995, Chlamyrhodopsin represents a new type of sensory photoreceptor, *EMBO J.*, **14**, 5849–5858.
30. Curry, A. M., Williams, B. D., and Rosenbaum, J. L. 1992, Sequence analysis reveals homology between two proteins of the flagellar radial spoke, *Mol. Cell Biol.*, **12**, 3967–3977.