ISSN 1672-9145 **Acta Biochimica et Biophysica Sinica** 2006, **38**(11): 812–820 **CN** 31-1940/Q

Sequencing and Analysis of a Genomic Fragment Provide an Insight into the *Dunaliella viridis* Genomic Sequence

Xiao-Ming SUN[1#], Yuan-Ping TANG[1#], Xiang-Zong MENG[1], Wen-Wen ZHANG[1], Shan LI[1], Zhi-Rui DENG[1], Zheng-Kai XU[1,2], and Ren-Tao SONG[1]*

[1] *Shanghai Key Laboratory of Bio-energy Crops, School of Life Sciences, Shanghai University, Shanghai 200444, China;*
[2] *Institute of Plant Physiology & Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China*

Abstract *Dunaliella* is a genus of wall-less unicellular eukaryotic green alga. Its exceptional resistances to salt and various other stresses have made it an ideal model for stress tolerance study. However, very little is known about its genome and genomic sequences. In this study, we sequenced and analyzed a 29,268 bp genomic fragment from *Dunaliella viridis*. The fragment showed low sequence homology to the GenBank database. At the nucleotide level, only a segment with significant sequence homology to 18S rRNA was found. The fragment contained six putative genes, but only one gene showed significant homology at the protein level to GenBank database. The average GC content of this sequence was 51.1%, which was much lower than that of close related green algae *Chlamydomonas* (65.7%). Significant segmental duplications were found within this fragment. The duplicated sequences accounted for about 35.7% of the entire region. Large amounts of simple sequence repeats (microsatellites) were found, with strong bias towards $(AC)_n$ type (76%). Analysis of other *Dunaliella* genomic sequences in the GenBank database (total 25,749 bp) was in agreement with these findings. These sequence features made it difficult to sequence *Dunaliella* genomic sequences. Further investigation should be made to reveal the biological significance of these unique sequence features.

Key words *Dunaliella viridis*; genomic sequence; sequence feature; sequence duplication; simple sequence repeat

Dunaliella is a genus of wall-less unicellular eukaryotic green alga, with important industry applications [1]. Natural β-carotene is mainly produced from this organism [2]. It is one of the most salt-tolerant eukaryotic organisms known to date [3]. It also has exceptional resistance to other stresses, such as high light intensity, dramatic pH change, temperature shocking, etc.. *Dunaliella* is not only an important algae with great value, but also an ideal model system for studying the mechanisms for stress tolerance [4].

As an extremophile organism, *Dunaliella* evolves many unique cellular structures or metabolic features to deal with the extreme environment, including the over accumulation of glycerol and β-carotene under stress conditions [5]. It would be interesting to know if such evolution process might leave markers on the *Dunaliella* genome as well. To date, very few genes have been cloned from *Dunaliella*. Among the clones, most were cDNAs which did not reveal much information about genomic organizations [6,7], and a few were genomic sequences containing only gene coding sequences, providing only limited information about the genic regions in the genome [8]. Therefore, very little is known about the general features of *Dunaliella* genomic sequences, particularly for the non-coding sequence regions between genes (intergenic regions).

In order to facilitate the cloning of *Dunaliella* genes, we constructed a bacterial artificial chromosome (BAC) library from *Dunaliella viridis* genomic DNA (unpublished

Received: July 4, 2006 Accepted: August 20, 2006
This work was partially supported by a grant from the National Natural Science Foundation of China (No. 30471119)
[#] These authors contributed equally to this work
*Corresponding author: Tel, 86-21-66135182; Fax, 86-21-66135163; E-mail, rentaosong@staff.shu.edu.cn

DOI: 10.1111/j.1745-7270.2006.00227.x

data). The library contained about 9200 clones, with an average insert size of about 55 kb. Based on the analysis of this BAC library, the genome size of *Dunaliella viridis* was estimated to be about 100 Mb, which is about the same size of *Chlamydomonas reinhardtii* [9].

In this study, we randomly selected a BAC clone, and sequenced it using a shotgun sequencing approach. This yielded a 29,268 bp genomic sequence of *Dunaliella viridis*, the longest one to date. Using this sequence, we were able to investigate the general features of the *Dunaliella* genomic sequence. Sequence features, such as GC content, significant sequence duplication, abundant and biased simple sequence repeats, etc., were found in this genomic sequence. These sequence features were further verified by other *Dunaliella* genomic sequences available from the database. This study provides an insight into the general sequence features of *Dunaliella* genomic organization, and might facilitate our further study of the molecular mechanism of stress tolerance using *Dunaliella* as a model.

Materials and Methods

Materials

Bacterial artificial chromosome (BAC) library of *Dunaliella viridis* was constructed in our laboratory (unpublished data). The library was constructed into pCC1BAC vector (Epicenter, Madison, USA) by *Hin*dIII. Totally about 9200 clones were obtained, with an average insert size of 55 kb. The BAC library had about 4-time genome coverage. A clone was randomly picked for this study.

Shotgun library construction

BAC DNA was extracted by the alkaline lysis method. The insert DNA was released from the vector by *Not*I digestion, and purified from a pulse field electrophoresis gel. The purified DNA was physically sheared by a Hydorshear device (GeneMachine, http://genome.nhgri.nih.gov/genemachine/) using a speed code setting of 11, which yielded a DNA shearing fraction of about 2–4 kb. The sheared DNA was end-repaired by T4 DNA polymerase for 30 min at 12 ºC, and the DNA fraction of 2–4 kb was purified using an agarose gel. Recovered DNA fragments were ligated into pUC18, which was prepared by *Sma*I digestion and CIP de-phosphorylation, by T4 DNA ligase, 16 ºC overnight. The ligation products were transformed into *Escherichia coli* (strain DH10B) by electroporation (1 mm curvet, 1800 V/cm) with GenePulser Xcell electroporator (Bio-Rad, Hercules, USA). Transformants were plated on LB plates with ampicillin, X-gal and IPTG. White colonies were picked, and plasmid DNAs were extracted for following restriction digestion and sequencing analysis.

Sequencing analysis and sequence assembly

A total of 480 clones (on five 96-well microtiter plates) were used for sequencing analysis at both directions. Plasmids were extracted in 96-well plates. Sequencing reactions were carried out using universal M13 forward and reverse primers with a DYEnamic ET dye terminator sequencing kit (Pharmacia, Piscataway, USA) according to the manufacturer's instruction. Sequencing products were analyzed on a MegaBACE 4000 capillary DNA sequencer (Pharmacia). Raw data from the DNA sequencer were exported to a Linux cluster, and were base called and assembled by Phred/Phrap programs [10,11]. Sequences generated from this analysis had about 12-times the coverage of the BAC fragment. The resulting contig information was viewed and edited by the Consed program [12,13]. The gaps between neighboring contigs were bridged by PCR fragments. Sequencing of these PCR fragments closed some of the gaps. The assembled contig sequence was digital digestion. The digital digestion results were compared with real restriction enzyme digestion patterns to verify the reliability of the sequence assembly.

A sequence homology search was carried out by the BLAST program against the NCBI GenBank (http://www.ncbi.nlm.nih.gov/BLAST/). The GC content was analyzed by the sliding-window technique (window size=1000 bp, shift size=1 bp) [14]. Gene prediction was carried out using GENSCAN (http://genes.mit.edu/GENSCAN.html) and FGENSH (http://www.softberry.com/berry.html). The organism mode setting was *Arabidopsis* [15]. To improve accuracy, gene prediction results were combined. Only those predicted by both programs were retained for further analysis. Sequence duplication was analyzed by bl2seq (blast on two sequences) [16]. Microsatellites were searched by CENSOR (http://www.girinst.org/censor/index.php), a software tool which screens query sequences against a reference collection of repeats.

Results

Sequencing of *Dunaliella viridis* genomic fragment DQ641395

In order to clone and analyze genes of *Dunaliella viridis*, we had constructed a BAC genomic library of the

Dunaliella viridis (data unpublished). To obtain a general view of the features of the *Dunaliella viridis* genomic sequence, a BAC clone was randomly selected for sequencing analysis.

A shotgun library was constructed from the insert fragment of this BAC clone (see "Materials and Methods"). The insert size of the shotgun library was between 2.5 bp to 3 kb based on checking the insert sizes of ten randomly picked clones (data not shown). A total of 480 clones were sequenced from both directions, and yielded about 12-times the sequence coverage of the genomic fragment. Two ordered contigs were generated from the assembly, with a total sequence length of about 30 kb.

Finishing attempts were carried out to close the gap between two contigs. The gap was estimated to be about 50 bp in size based on the PCR result. However, the gap was flanked by simple sequence repeats, and it was unable to be closed with further sequencing attempts. The reliability of sequence assembly was verified by comparing the digital digestion results of the assembled sequence with restriction digestion of BAC DNA (data not shown).

The resulting assembly comprised two ordered contigs, with a total length of 29,268 bp. A sequencing gap was near the 3′ end of the sequence (from 26,544 bp to 26,593 bp). The sequence had been submitted into GenBank under accession No. DQ641395. This represented the longest genomic fragment from the *Dunaliella* species so far in the GenBank database.

Gene content in DQ641395

A homology search was carried out for DQ641395 using BLASTN against the NCBI GenBank. Only a segment located in 27,562–27,699 bp with significant homology to an 18S ribosomal RNA gene DQ447648 (*e* value $4e^{-66}$) was found. The rest of the sequence showed no significant sequence homology to the GenBank database. The rRNA genes were well known for their sequence conservation among different species. No other homology hits among this sequence suggested that the *Dunaliella* genomic sequence might be highly diverged from other organisms.

Gene prediction was carried out using both the GENSCAN and FGENSH web servers. GENSCAN predicted six putative genes, while FGENSH predicted 12 putative genes (data not shown). The GENSCAN prediction has the tendency to fuse neighboring genes predicted by FGENSH. To improve accuracy, only genes predicted by both programs were reserved for further analysis (**Table 1**). Together with the 18S rRNA non-coding gene that was found by BLASTN analysis, the entire genomic fragment contained seven putative genes yielding a gene density of 4.2 kb/gene.

A homology search was carried out for the predicted gene products using the BLASTP program against the GenBank database. Four genes were found to have homology to the database (**Table 1**). Among them, Gene 4 showed significant homology to retrotransposon nucleocapsid protein from *Cryptococcus neoformans* (AAW44070). Genes 1, 5 and 6 showed limited homology to the database (**Table 1**). Genes 2 and 3 showed no homology to the GenBank database.

Sequence analysis of DQ641395

The overall GC content of DQ641395 was 51.1% [**Fig. 1(A)**]. A 168 kb genomics fragment (AC087762) from the closely related green algae *Chlamydomonas* was also analyzed for GC content [**Fig. 1(B)**]. To our surprise, the average GC content in *Chlamydomonas* was much higher than that of the *Dunaliella viridis*. The fluctuation range of GC content within the *Dunaliella* sequence was 23.9% (from 40.3% to 64.2%). In addition, the 23.9% shift of GC content occurred with a sequence distance as close as 5 kb (from 18,088 bp to 23,492 bp). The data indicated that the sequence composition was very uneven, and a dramatic GC content shift was common in the *Dunaliella* genomic sequences. But such dramatic GC fluctuation was not unique to the *Dunaliella* genomics sequence. Similar

Table 1 **Gene prediction of *Dunaliella viridis* genomic fragment (GenBank accession No. Q641395)**

Gene	Direction	Begin	End	Length	Exon	Total aa	BLASTP result	*e* value
1	+	256	1196	941 bp	2	104	CAE66198	0.78
2	+	3598	5945	2348 bp	4	355	No hit found	NA
3	–	9564	6692	2873 bp	4	389	No hit found	NA
4	+	11,919	13,671	1753 bp	4	350	AAW44070	$5e^{-18}$
5	–	21,032	16,924	4109 bp	9	664	YP_607634	3.7
6	–	23,913	21,488	2426 bp	4	279	XP_426919	2.5

+, same to genomic fragment; –, on the minue chain of genomic fragment. aa, amino acid. NA, not available.

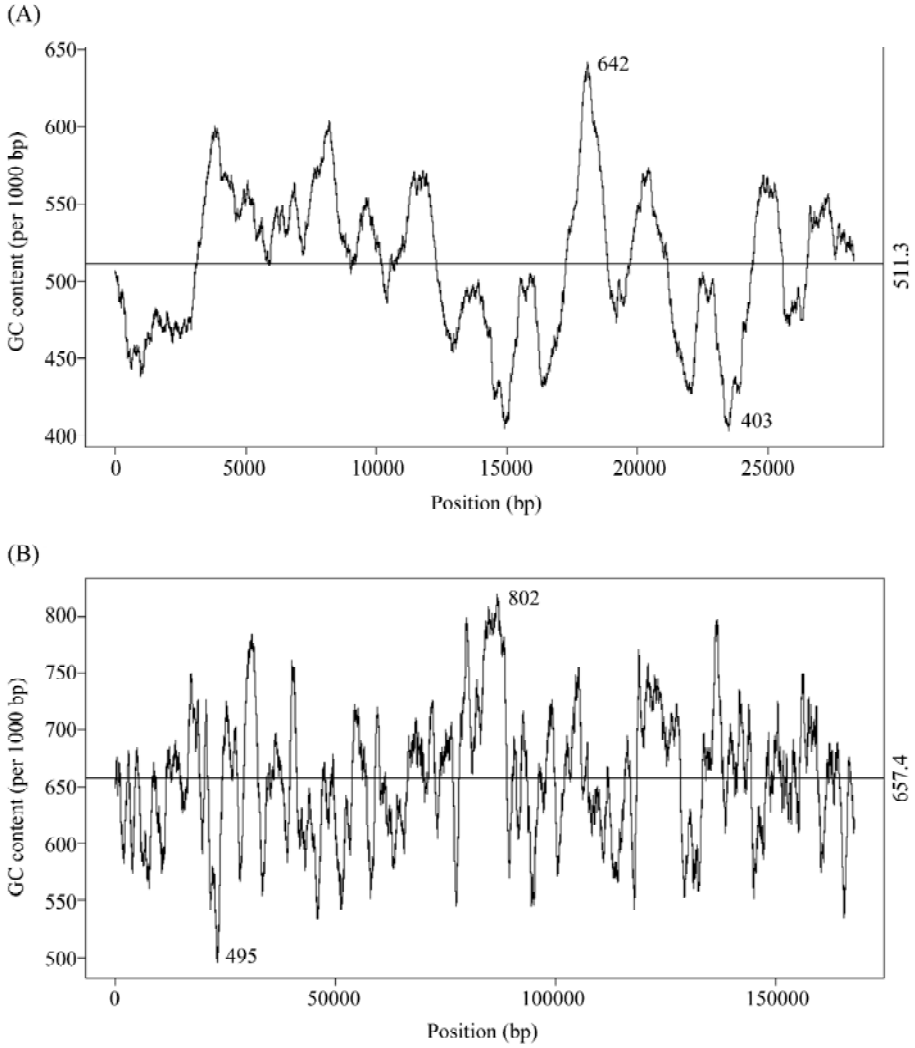

Fig. 1 GC content of *Dunaliella viridis* genomic sequence (GenBank accession No. DQ641395) (A) and *Chlamydomonas* genomic fragment (GenBank accession No. AC087762) (B)

GC fluctuation was also observed in the *Chlamydomonas* genomic sequence.

Table 2 summarizes the GC content in genic and intergenic regions, as well as in exons and introns. Within DQ641395, the genic region and intergenic region each occupied about 50% of the sequence. The GC content of the genic region was 52.5%, which was slightly higher than that of the intergenic region (49.9%). Within the genic region, the total length of exons was 10,060 bp, while that of introns was 4645 bp. This indicated that *Dunaliella* genes had relatively small introns. The GC content for introns was 49.5%, which was very similar to that of the intergenic region. The GC content for exons was higher than that of introns (53.9%).

Table 2 GC content in *Dunaliella viridis* genomic fragment (GenBank accession No. DQ641395)

Sequence type	Length (bp)	GC content
Intergenic	14,513	49.9%
Genic	14,705	52.5%
Exon	10,060	53.9%
Intron	4645	49.5%

Large sequence duplications in DQ641395

As shown in **Fig. 2** and **Table 3**, DQ641395 was

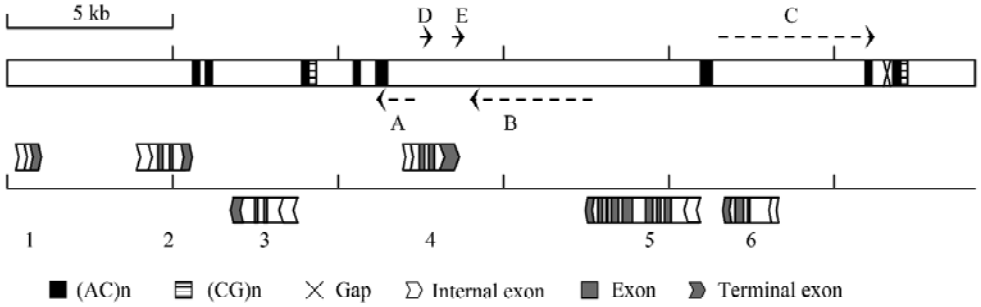

Fig. 2 Map of *Dunaliella viridis* genomic fragment sequence (GenBank accession No. DQ641395)

Dashed arrowhead indicates inverted duplications, and arrowhead indicates tandem duplications in DQ641395. Letter indicates large sequence duplications according to **Table 3**. Number indicates predicted genes according to **Table 1**.

Table 3 Large sequence duplications in the *Dunaliella viridis* genomic fragment (GenBank accession No. DQ641395)

Type	Position	Length (bp)	Identity
Inverted duplication	A (11,363–12,574)	4992	98.3%
	B (14,240–18,019)	4992	
	C (21,383–26,350)	4968	
Tandem duplication	D (12,597–12,846)	250	98.4%
	E (13,988–14,237)	250	

distinguished by a large sequence duplication feature. Except for one tandem duplication, all the duplications were found to be inverted. The longest duplication segment was 4992 bp in length, with 98.3% identity. All the duplicated sequences had greater than 98% sequence identity. The tandem duplication appeared to be inserted into one arm of the long inverted duplication segment (**Fig. 2**). Together, the duplicated sequences counted for 35.7% of the entire sequence.

These sequence duplications were involved in three predicted genes (**Fig. 2**). Gene 4 involved part of the tandem duplication, while Gene 5 and Gene 6 shared part of the invert duplication. Despite Gene 5 and Gene 6 sharing highly duplicated sequences, they encoded completely different genes. Therefore sequence duplication followed by sequence divergence formed three different genes. Such sequence duplication and divergence mechanisms provide an efficient way for increasing gene numbers and diverging gene functions during evolution.

Repetitive sequences in DQ641395

Another apparent feature of the sequence was the abundant simple sequence repeats which were identified as two different classes of repetitive sequences (**Table 4**). One

Table 4 Repetitive sequences in the *Dunaliella viridis* genomic fragment sequence (GenBank accession No. DQ641395)

Repeat class	Number	Length (bp)
Interspersed repeat[a]	7	673
DNA transposon	2	147
P	1	72
hAT	1	75
LTR retrotransposon	2	159
Gypsy	1	75
Non-LTR retrotransposon	3	367
L1	1	45
Simple repeat[b]	18	1549
Total	25	2222

L1, long interspersed repetitive element 1; LTR, long terminal repeat.

[a]Details of interspersed repeats

From	To	Type	Direction	Length (bp)
13,294	13,368	ATGP3I	+	75
5756	5830	hAT-2n1_DR	–	75
8520	8564	L1MD1_5	–	45
18,630	18,701	P-1_CR	–	72
8565	8739	R7Ag2	–	175
3921	4067	RTAg4	–	147
4280	4363	SZ-48LTR	–	84

[b]Details of simple repeats

From	To	Type	Direction	Length (bp)
5656	5737	$(AC)_n$	+	82
5952	6063	$(AC)_n$	–	112
8867	8932	$(AC)_n$	–	66
10,501	10595	$(AC)_n$	+	95
11,357	11560	$(AC)_n$	+	204
21,048	21375	$(AC)_n$	–	328
26,166	26350	$(AC)_n$	–	185
27,230	27335	$(AC)_n$	–	106
26,460	26491	$(CAA)_n$	+	32
8933	8990	$(CG)_n$	+	58
27,336	27354	$(CG)_n$	+	19
5199	5236	$(GCA)_n$	–	38
6903	6940	$(GCA)_n$	–	38
7061	7087	$(GCA)_n$	–	27
8402	8483	$(GCA)_n$	–	82
20,894	20934	$(GCA)_n$	–	41
15,265	15282	$(GGGA)_n$	–	18
24,122	24139	$(GGGA)_n$	+	18

class was interspersed repeats, the other was simple repeats. Interspersed repeats included sequence segments derived from transposons or retrotransposons (including long terminal repeats and non-LTR retrotransposons). Two repeats were from DNA transposons, five repeats were from retrotransposons. The interspersed repeats counted for about 1/3 of all repetitive sequences

The remaining 2/3 repetitive sequences were contributed by simple sequence repeats. Simple sequence repeats were also called microstatellites, formed by a tandem array of repeated sequence units of two to a few nucleotides. A total of 18 simple sequence repeats were identified, with a sequence length of 1549 bp (5.3% of the entire sequence). In terms of number and length, $(AC)_n$ was the most abundant. The longest continuous repeat was 164 times (328 bp). There was a strong bias toward $(AC)_n$, which counted for 76.0% of all the microsatellites found in this sequence. Some types of microsatellites were not detected at all within the 29,268 bp sequence.

The analysis of other *Dunaliella* genomic sequences from database

Due to very limited genomic sequences available in the GenBank database, we could only collect three sequences with a length of longer than 5 kb (NR: AY567972, 14650 bp; DCA1: AF541981, 6029 bp; DCA: AY232669, 5070 bp) for similar sequence analysis. All these sequences were genic sequences containing a single gene.

The GC content of these three sequences was consistent with what was found in DQ641395 (**Fig. 3**). The average GC content for each genomic sequence was about 53%, similar to that of DQ641395 in this study. Similar GC fluctuation was found in all three sequences. Particularly in the NR sequence, a 15% fluctuation was found within 1 kb range. The other two sequences had 12.2% and 9.7% GC fluctuation within 4 kb and 3 kb range respectively.

Similar microstatellites feature was also found in these sequences (**Table 5** and **Fig. 4**), with exactly the same bias toward to the $(AC)_n$ microstatellites. It was interesting to notice that no interspersed repeats were found in all three genes, suggesting that these repeats might have uneven distributions in the genome.

None of these sequences was observed to have significant sequence duplication. In the analysis of DQ641395, large sequence duplications were found between different genes. While these genomic sequences obtained from the database all contained only a single gene. This might explain why no obvious sequence duplication was found in these three *Dunaliella* genomic sequences.

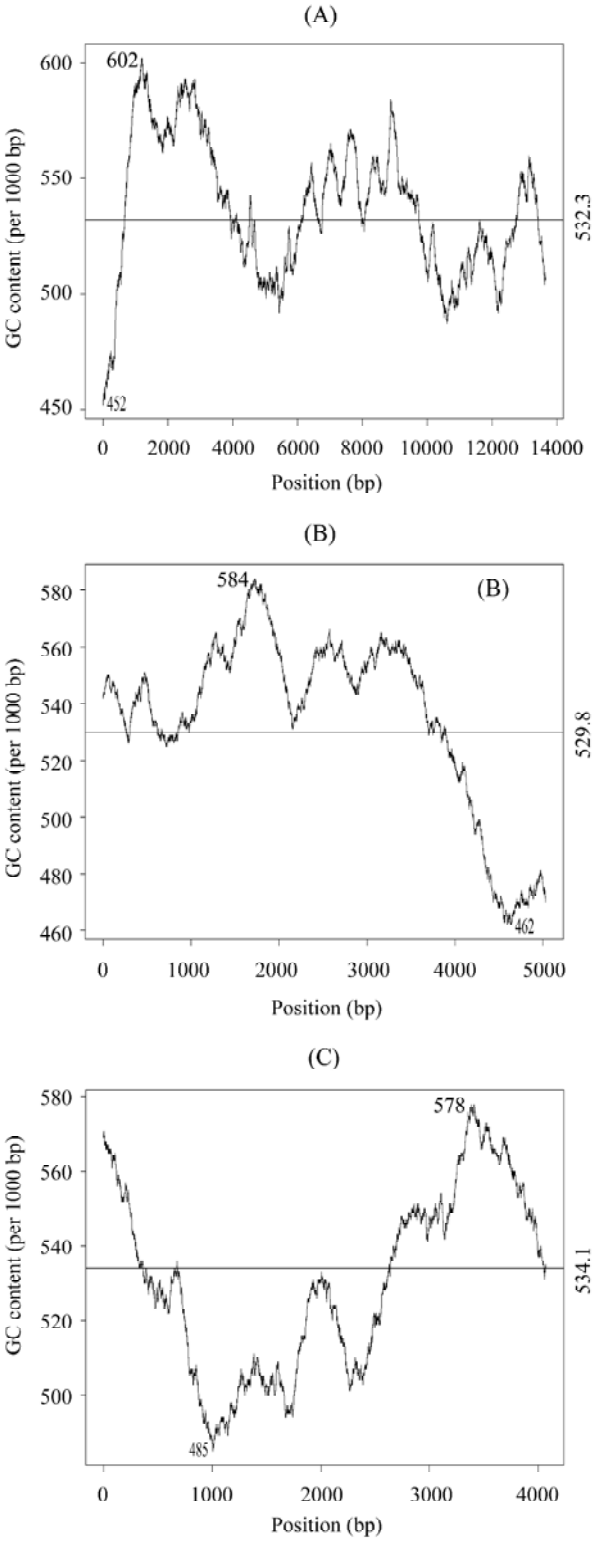

Fig. 3 GC content of nitrate reductase gene from *Dunaliella viridis* (NR) (A), duplicated carbonic anhydrase gene from *Dunaliella salina* strain UTEX 1644 (DCA1) (B) and carbonic anhydrase gene from *Dunaliella salina* (DCA) (C) genomic fragment sequences

Table 5 Repetitive sequences in NR, DCA1 and DCA

Type	Length	Repetitive sequences				
		From	To	Type	Direction	Length (bp)
NR	14,650 bp	2097	2127	$(AC)_n$	–	31
		2566	2822	$(AC)_n$	+	257
		3362	3402	$(AC)_n$	–	41
		3681	3720	$(AC)_n$	+	40
		6943	7158	$(AC)_n$	+	216
		8577	8767	$(AC)_n$	–	191
		12,072	12,225	$(AC)_n$	–	154
		13,459	13,542	$(AC)_n$	–	84
		349	374	$(GAAAA)_n$	–	26
DCA1	6029 bp	1	225	$(AC)_n$	–	225
DCA	5070 bp	65	202	$(AC)_n$	–	138
		368	386	$(GGAA)_n$	+	19

NR, nitrate reductase gene from *Dunaliella viridis* (GenBank accession No. AY567972); DCA1, duplicated carbonic anhydrase gene from *Dunaliella salina* strain UTEX 1644 (GenBank accession No. AF541981); DCA, carbonic anhydrase gene from *Dunaliella salina* (GenBank accession No. AY232669). +, same to genomic fragment; –, on the minue chain of genomic fragment.

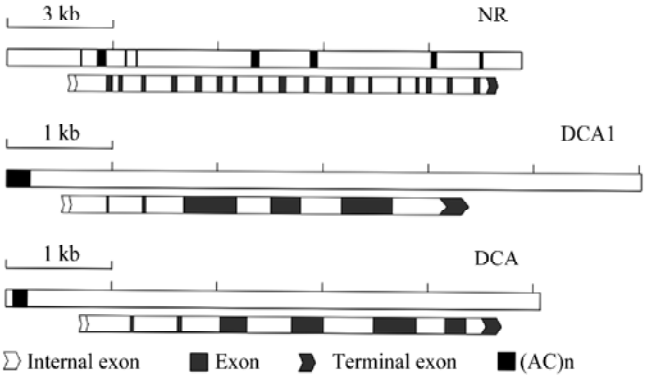

Fig. 4 Map of nitrate reductase gene from *Dunaliella viridis* (NR), duplicated carbonic anhydrase gene from *Dunaliella salina* strain UTEX 1644 (DCA1), and carbonic anhydrase gene from *Dunaliella salina* (DCA) genomic fragment sequences

Discussion

Dunaliella is an extremophile unicellular eukaryotic green alga with exceptional tolerance to salt as well as various other environmental stresses. Previous studies mainly focused on physiology or biochemistry to tackle the possible mechanism of its extraordinary stress tolerance [17–20]. Only a few genes were cloned from this organism so far [8], and very little was known about its sequence organization in the genome. In this study, we sequenced a 29,268 bp genomic fragment of *Dunaliella viridis* (DQ641395), the longest one so far for the *Dunaliella* species, and have found some sequence features related to the *Dunaliella* genome.

First of all, the sequence had relatively poor homology to the sequences available in the GenBank database. Only an 18S rRNA sequence was found to have significant homology to the database at the nucleotide level. 18S rRNA was well known for its sequence conservation cross species. The rest of the sequences in the entire 29,268 bp showed no significant homology to the database. Besides this 18S rRNA, there were six predicted genes in this sequence. At the amino acid level, only one gene showed significant homology to the GenBank database. The rest of the genes had either poor homology or no homology to the database. Therefore, the poor sequence homology to the database suggested the *Dunaliella* genomic sequence was highly diverged and very different to other organisms.

The second sequence feature was its sequence composition. The average GC content of 29,268 bp sequence and the other three *Dunaliella* genomic sequences (NR, DCA1 and DCA) was around 51%–53%. This number was much lower than that of another closely related green algae, *Chlamydomonas reinhardtii* (65.7%). The GC distribution in the *Dunaliella* genomic sequence was not even. Fluctuation of GC content was apparent, even within very a small sequence range. For DQ641395, the fluctuation of 23.9% GC content occurred within only a 5-kb sequence range. The NR and two DCA genomic sequences all showed similar GC fluctuation. But such features were not unique to that of *Dunaliella*. Similar GC

fluctuation was also observed in the *Chlamydomonas* genomic sequence. Similar to other organisms, in the *Dunaliella* genomic sequence, the GC content of genic regions was higher than that of intergenic regions, and the GC content of exons was higher than that of introns.

The third sequence feature was abundant and biased microsatellites. In 29,268 bp sequence or other 25,749 bp sequences from the database, abundant microsatellites were found. In all the cases, strong bias towards the $(AC)_n$ type was observed, indicated that $(AC)_n$ could be the most abundant di-nucleotide simple sequence repeats in the *Dunaliella* genome. We also observed a strong bias of $(GCA)_n$ type microsatellites in the 29,268 bp sequence, but such observation could not be confirmed by other sequences from the database, mainly due to limited data availability. In the 29,268 bp sequence, other interspersed repetitive sequences were also found, though they were not apparent in the sequences from the database. Therefore, more genomic sequences would be needed for evaluation of the abundance of interspersed repeats in the *Dunaliella* genome.

Significant sequence duplications were also found in the 29,268 bp sequence. But such sequence duplication was not observed in the sequences obtained from the database. This might be due to very limited sequences available from the database (total only 25,749 bp in size). The duplication segments in the 29,268 bp sequence contained genes. Therefore, such sequence duplication happened in between genes. The sequences obtained from the database all only contained a single gene or part of a gene. If the sequence duplication would occur between genes, then we would not expect to observe them within a single gene. Would sequence duplication be a common feature in *Dunaliella* genome? More genomic sequencing would be needed to answer this question.

All these unique sequence features caused difficulties to sequence the *Dunaliella* genomic sequences efficiently. Highly fluctuated GC content would cause problems during sequencing, increasing the sequencing failure rate. A long stretch of microsatellite was a major cause of the sequencing gap (the gap due to sequencing failure). Sequence duplication would complicate sequence assembly. All these sequence features of lone terminal repeats could also cause instability of cloned genomic sequences in a regular cloning vector. To counteract these difficulties, we made the *Dunaliella* genomic library using the BAC vector, which would provide excellent sequence stability [21]. We sequenced the genomic sequence by a high coverage shotgun sequencing approach [22]. The newer capillary sequencer model MegaBACE4500 was used and produced longer and higher quality of sequences. Although it is still difficult, sequencing long genomic sequences from *Dunaliella* now is feasible.

Acknowledgements

We would like to thank Bin-Bin HUANG and Liang-Liang ZhOU for their help with sequencing analysis.

References

1 Ben-Amotz A, Katz A, Avron M. Accumulation of β-carotene in halotolerant algae: Purification and characterization of β-carotene-rich globules from *Dunaliella bardawil* (Chlorophyceae). J Phycol 1982, 18: 529–537
2 Pulz O, Scheibenbogen K, Grob W. Biotechnology with cyanobacteria and microalgae. In: Rehm HJ, Reed G eds. A Multi-Volume Comprehensive Treatise Biotechnology. Weinheim: Wiley-VCH Verlag GmbH 2001
3 Ben-Amotz A, Avron M. The biotechnology of cultivating the halotolerant alga *Dunaliella* for industrial products. Trends Biotech 1990, 8: 121–126
4 Cowan AK, Rose PD, Horne LG. *Dunaliella salina*: A model system for studying the response of plant cells to stress. J Exp Bot 1992, 43: 1535–1547
5 Ben-Amotz A, Shaish A, Avron M. Mode of action of the massively accumulated beta-carotene of *Dunaliella bardawil* in protecting the alga against damage by excess irradiation. Plant Physiol 1989, 91: 1040–1043
6 Fisher M, Gokhman I, Pick U, Zamir A. A salt-resistant plasma membrane carbonic anhydrase is induced by salt in *Dunaliella salina*. J Biol Chem 1996, 271: 17718–17723
7 Fisher M, Gokhman I, Pick U, Zamir A. A structurally novel transferrin-like protein accumulates in the plasma membrane of the unicellular green alga *Dunaliella salina* grown in high salinities. J Biol Chem 1997, 272: 1565–1570
8 Li Q, Gao X, Sun Y, Zhang Q, Song R, Xu Z. Isolation and characterization of a sodium-dependent phosphate transporter gene in *Dunaliella viridis*. Biochem Biophys Res Commun 2006, 340: 95–104
9 Harris EH. Chlamydomonas as a model organism. Annu Rev Plant Physiol Plant Mol Biol 2001, 52: 363–406
10 Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998, 8: 186–194
11 Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998, 8: 175–185
12 Gordon D. Viewing and editing assembled sequences using consed. In: Baxevanis AD, Davison DB eds. Current Protocols in Bioinformatics. New York: John & Wiley Co. 2004
13 Gordon D, Abajian C, Green P. Consed: A graphical tool for sequence finishing. Genome Res 1998, 8: 195–202
14 Fujimori S, Washio T, Tomita M. GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics 2005, 6: 26
15 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol 1997, 268: 78–94
16 Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 1999, 174: 247–250
17 Semenenko VE, Abdullaev AA. Parametric control of β-carotene

biosynthesis in *Dunaliella salina* cells under conditions of intensive cultivation. J Sov Plant Physiol 1980, 27: 22–30

18 Loeblich LA. Photosynthesis and pigments influenced by light intensity and salinity in the halophile *Dunaliella salina* (Chlorophyta). J Mar Biol Ass UK 1982, 62: 493–508

19 Curtain CC, West SM, Schlipalius L. Manufacture of β-carotene from the salt lake alga *D. salina*: The scientific and technical background. J Biotechnol 1987, 1: 51–57

20 Massyuk NP. Morphology, taxonomy, ecology and geographic distribution of the genus *Dunaliella* teod. and prospects for its potential Utilization. Naukova Dumka (Kiev) 1973, 244

21 Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 1992, 89: 8794–8797

22 Bouck J, Miller W, Gorrell JH, Muzny D, Gibbs RA. Analysis of the quality and utility of random shotgun sequencing at low redundancies. Genome Res 1998, 8: 1074–1084

Edited by
Qi-Yun LI