Journal of Applied Phycology **11:** 99–104, 1999.

Occurrence of closely spaced genes in the nuclear genome of the agarophyte *Gracilaria gracilis*

Arturo O. Lluisma[1,2] & Mark A. Ragan[1,3,*]
[1]*Institute for Marine Biosciences, National Research Council of Canada, Halifax, Nova Scotia, B3H 3Z1 Canada*
[2]*Present address: Marine Science Institute, University of the Philippines, 1101 Quezon City, Philippines*
[3]*Canadian Institute for Advanced Research, Program in Evolutionary Biology*

(*Author for correspondence)

Received 1 April 1998; revised 13 November 1998; accepted 16 November 1998

Key words: genome structure, genome organisation, red algae, synteny

Abstract

Little is known about the structure and organisation of nuclear genomes in red algae. In particular, it is not known whether genes are densely or loosely packed, whether gene order is conserved, whether their genes tend to occur in one or multiple copies and whether their nuclear genes tend to be compact or interrupted by numerous introns. Sequencing of cloned genomic DNA from *Gracilaria gracilis* has begun to provide provisional answers to some of these questions. Four pairs of closely spaced genes have been found in *G. gracilis* upon sequencing genomic clones that contain genes for UDPglucose pyrophosphorylase, galactose-1-phosphate uridylyltransferase, the β subunit of tryptophan synthetase, and methionine sulphoxide reductase (a fifth pair of closely spaced genes, encoding polyubiquitin and aconitase, was reported earlier). An open reading frame with significant similarity to another known gene occurs close ($<$1.7 kbp) to each of these genes. In two pairs the intergenic region is less than 400 bp in length, and for these the location of the putative polyadenylation signals indicates that the gene transcripts, encoded on opposite strands, have overlapping (hence complementary) 3′ regions. These somewhat unexpected findings begin to establish a basis for genome-level characterisation of red algae.

Abbreviations: EST – expressed sequence tag; GalT – galactose-1-phosphate uridylyltransferase; MSR – methionine sulphoxide reductase; NMIOR – NAD-dependent myo-inositol oxidoreductase; ORF – open reading frame; PCR – polymerase chain reaction; PTH – peptidyl tRNA hydrolase; SBE – starch branching enzyme; TSβS – tryptophan synthase-β subunit; UGPase – UDPglucose pyrophosphorylase

Introduction

Determining the structure and organisation of genomes is important for a number of reasons. As has been demonstrated for some groups of organisms (*e.g.*, grasses: Barakat et al., 1997; Bennetzen & Freeling, 1997), such characterisation is key not only to understanding how the genomes of extant members of a group are structurally and functionally related, but also to reconstructing the evolution of their genomes from an ancestral form. In addition, genomic characterisation is a prerequisite for comparative mapping, which facilitates the application of information from more readily manipulated 'model' species, to others that are typically of commercial importance (*e.g.*, cereals).

Current understanding of the structure, organisation and complexity of red algal genomes is based largely on studies that employed micro-spectrophotometry, DNA reassociation kinetics, and microscopy. Micro-spectrophotometry has been useful in quantifying the genomic DNA of a number of red algae; DNA reassociation kinetics in estimating the

relative amounts of repetitive and unique DNA, and G+C content (molar fraction of guanine and cytosine in the DNA); and microscopy in estimating genome sizes and number of chromosomes. These techniques have been instrumental in establishing, for example, that in species of *Gracilaria* the chromosome number and nuclear genome size are highly conserved, but the G+C content and relative amounts of unique and repetitive DNA tend to vary (Kapraun et al., 1996a,b; Lopez-Bautista & Kapraun, 1995).

Beyond these basic features, however, most structural aspects of red algal genomes remain uninvestigated; in particular, essentially nothing is known about the order and distribution of genes along the chromosomes of red algae. Studies of other eukaryotic genomes have shown that genes may be somewhat clustered in some groups of organisms, *e.g.*, Graminae, but relatively more scattered in others, *e.g.*, mammals (Bernardi, 1995). Groups that diverged relatively recently might be expected to share greater genomic co-linearity. Modern-day grasses, for example, have genomes that in general represent reorganisations of chromosomal segments of an ancestral grass genome (Moore et al., 1995; Bennetzen & Freeling, 1997). Similar studies have not been performed for red algae. Sequencing the entire genome of several representative red algae would be the ideal approach, but such an effort remains to be undertaken.

By sequencing the flanking regions of cloned nuclear genes, it should be possible to determine whether genes tend to be clustered. Several red algal genes have already been cloned, but their flanking regions have received little attention. Heretofore the only report of closely spaced nuclear genes in a red alga was that of Zhou and Ragan (1995), who found that the genes encoding mitochondrial aconitase and polyubiquitin are located about 1.5 kb apart in the genome of *G. gracilis*. The purpose of the present paper is to report that four more pairs of genes are closely spaced in the nuclear genome of *G. gracilis*.

Materials and methods

Isolation and sequencing of genomic clones

DNA was extracted from *G. gracilis* as described previously (Zhou & Ragan, 1993). The genomic library was constructed by partially digesting the DNA with *Sau*3AI, and ligating the genomic fragments to the Lambda-DASH vector (Stratagene, La Jolla, CA).

Table 1. Primers used for amplifying the terminal and intergenic regions from each pair of closely spaced genes

Gene pair	PCR round	F primer	R primer	Expected size of product
GalT-PtRH	1	galTsF-1	galTr-2	∼ 1.4 kb
	2	galTsF1	galTr-2	∼ 1.1 kb
UGPase-helicase	1	ugpsF-1	ugpr-6b	∼ 2.0 kb
	2	ugpsF3b	ugpr-6b	∼ 1.6 kb
NMIOR-TSβS	1	c313f-6	c313r3	∼ 2.0 kb
	2	c313f-5	c313r3	∼ 1.6 kb
MSR-mAT	1	c47f-1	mATr1b	∼ 2.5 kb
	2	c47sF1	mATr1b	∼ 2.0 kb

Clones were isolated from the genomic library by screening with probes for five putative genes from *G. gracilis*: GALT, MSR, SBE, TSβS, and UGPase. The probes used were generated either from PCR products (SBE: Lluisma & Ragan 1998a; UGPase: Lluisma & Ragan, unpublished data) or ESTs (GALT, MSR and TSβS: Lluisma & Ragan, 1997). The isolated clones were sequenced by primer walking from the region of previously known sequence (the region corresponding to the probe) out to the 5′ and 3′ flanking regions. Sequencing was carried out on an ABI 373 (Applied Biosystems) automated sequencer using ABI's Dye Terminator Cycle Sequencing protocol.

Sequence searches

The nucleotide sequences were used as query sequences to search the U.S. National Center for Biotechnology Information (NCBI) nr (non-redundant) peptide sequence database for similar sequences, using the algorithms BLASTX (Gish & Gates, 1993) and BEAUTY (Worley et al., 1995). Searches were carried out using the BCM Search Launcher facility, maintained by the Human Genome Center, Baylor College of Medicine (Smith et al., 1996).

PCR confirmation

Two rounds of PCR reactions were run to verify the close spacing of certain genes in the genome of *G. gracilis* (*i.e.*, to confirm that the observed proximity is not due to the artifactual formation of chimaeras during cloning). Primers specific for each gene in each pair of closely spaced genes were used in the first round, with genomic DNA from *G. gracilis* as template. In the second round, nested primers were used,

Table 2. Sequences of oligonucleotide primers (5′→3′)

galTsF-1	GAGAACCTCAAGGCGTATCATG
galTsF1	CACTTTCACATGCACTTCTATCC
galTr-2	GGGTGATGCGATGCTCAACTAC
ugpsF-1	AAGCTGGCTGTGCTCAAGCTC
ugpsF3b	AAGACACGTGCCGACATCAAG
ugpr-6b	AGGTACAATGCTGTGCTAGGAG
c313f-6	TCCAACAGGACGCCTGCTTTTG
c313F-5	CTGTGAGAAGCCCATTTCGAATG
c313r3	GGGGAGCGGCGCGTTTTGAAC
c47f-1	CTCAAGTCACCGGTCGACACAC
c47f1	GCACGTCGGTGGTTCCGATG
mATr1b	CACCATCTCAAGGGATATTCCG

with the product from the first round of PCR as template. The names and sequences of the primers are shown in Tables 1 and 2, respectively. The PCR reactions contain the following, per 100 μL reaction: 200 μM dNTP, 1X reaction buffer, 2.5 U Taq, 450 ng total DNA from *G. gracilis* (first-round PCR only), and 100 ng of each primer, except that the product of the first round (0.5 μL) was used as template for the second round. The PCR reaction buffer and *Taq* polymerase were obtained from Pharmacia. PCR reactions were performed using a Perkin Elmer 9600 thermal cycler, with the reaction parameters set as follows: 94 °C for 2 min; 30 cycles of denaturation (94 °C, 1 min), annealing (62 °C or 55 °C, 1 min) and extension (72 °C, 6 min); and final extension at 72 °C for 5 min.

Results

Proximity of genes in G. gracilis

We have isolated genomic clones containing putative homologs of five genes in *G. gracilis* (encoding GalT, MSR, SBE, TSβS, and UGPase), and sequenced at least 500 bp outward from the 5′ and 3′ ends of these protein-coding regions. Full sequences of the genes encoding GalT, SBE and UGPase, including their flanking regions, have been deposited in GenBank under accession numbers AF036247, AF042842 and AF100788 respectively. Partial sequences of putative genes encoding MSR and TSβS were reported previously (as ESTs: Lluisma & Ragan, 1997) and have been deposited in GenBank under accession numbers gi1140307 and gi1140279 respectively. BLAST searches were used to compare translations of the flanking regions with sequences in the GenBank peptide database.

In the clone containing a gene (GalT) encoding galactose-1-phosphate uridylyltransferase (GALT) there occurs an ORF less than 1 kb downstream from GalT (Table 3). The BLAST results (score >100; Figure 1) suggest strongly that this downstream ORF encodes a peptidyl tRNA hydrolase (PTH). Sequence analysis, and the 3′ RACE data (Lluisma & Ragan 1998b), indicate that GALT and the PTH ORFs, which are encoded on opposite strands, may produce transcripts with overlapping and complementary 3′ ends; base-paring of the 3′ regions could theoretically interfere with the translation from the transcripts. Whether this occurs *in vivo*, and if so its functional significance, remain to be investigated.

We also found that a putative DNA helicase is adjacent to the gene for UDPglucose pyrophosphorylase (UGPase) in the nuclear genome of *G. gracilis* (Table 3). Although only a fragment of the DNA helicase gene has been sequenced (unpublished data), it nonetheless shows significant similarity (BLAST scores >100) with a number of DNA helicase entries in the GenBank. This putative helicase ORF has two interesting features. First, a potential 96-bp phase-0 GT/AG spliceosomal intron occurs near the 3′ end of this ORF; introns appear to be infrequent in red algal nuclear genes, and this is the first one localised in the 3′ end of a gene. Second, similar to the GalT-PTH pair, the putative UGPase and DNA helicase genes probably produce overlapping transcripts (unpublished data); again, possible physiological implications remain to be investigated.

A clone containing a potential gene for the β subunit of tryptophan synthase (TSβS) was isolated by hybridisation with the corresponding cDNA, reported earlier as part of our EST survey of *G. gracilis* (Lluisma & Ragan, 1997). Sequencing the region of the clone where TSβS is localised revealed that within 700 bp upstream of the gene can be found the 3′ end of an ORF whose conceptual translation is highly similar to sequences of hypothetical proteins in the GenBank peptide database that, in turn, are potential homologs of NAD-dependent myo-inositol oxidoreductase (NMIOR) of *B. subtilis* (Table 3). The *G. gracilis* ORF also matches *B. subtilis* NMIOR itself, although somewhat less strongly. Pairwise alignment reveals that the *G. gracilis* ORF is truncated in its 5′ region (data not shown).

Table 3. Results of BLAST searches revealing the potential identitites of ORFs found close to the putative GalT, MSR, TSβS, and UGPase genes in *G. gracilis*. ORF sequences have been deposited in GenBank with accession numbers AF036247, AF121271, AF121272 and AF100788. Relative orientation refers to the 5′→3′ direction of the coding strands; *e.g.*, A → ← B means that sequences A and B are encoded on opposite strands and transcribed in opposite directions. Intergenic region refers to the number of nucleotides separating the putative protein-coding regions of the gene and the nearby ORF

Cloned gene	Relative orientation	Intergenic region	GenBank entry with highest similarity to ORF		Potential homolog of ORF
			Accession no.	BLAST score	
GalT	GalT→ ←ORF	179 bp	gi|586021	225	PTH[1]
MSR	MSR→ ←ORF	<1.4 kb*	gi|2739362[a]	201	Transporter[2]
TSβS	ORF→ TSβS→	<670 bp**	gi|2635242[b]	251	NMIOR(?)[3]
UGPase	UGP→	376 bp	gi|2495146	205	DNA helicase[4]
	←ORF		gi|2495145	205	

* The intergenic region has not been fully sequenced; however, the tail-to-tail orientation of MSR and the ORF, and the alignment of the ORF with matching sequences, allows an estimate of a maximum length for the intergenic region.

** It is not possible to identify the start codon of TSβS, because 5'RACE experiments have not been done, and because we may expect a signal peptide not readily alignable with those of plant TSβS sequences to be encoded in this region. Thus at present we can determine only a maximum length for the intergenic region.

[1] Other PTH entries matching the ORF with BLAST scores >100: gi 2499989; gi 2507267; gi 2499987; gi 1346894; gi 1172722; gi 1346896; gi 2499986; gi 2499988; gi 1093598; gi 131550.

[2] Other entries matching the ORF with BLAST scores >100; mostly encoding membrane-bound transporter; some are sequences that encode mitochondrial ADP/ATP carrier protein: gi 2132987; gi 2393737; gi 1523933; gi 2132389; gi 2463664; gi 113465; gi 100424; gi 113457; gi 1729671; gi 2191150; gi 2804436; gi 2132884.

[3] Although the ORF was most similar to gi 2635242, the match between the ORF and NMIORs from *B. subtilis* and *S. griseus* also appear significant (BLAST score >100).

[4] Other DNA helicase entries matching the ORF with BLAST scores >100: gi 1709995; gi 101069; gi 5022; gi 2058510; gi 2134009; gi 119540; gi 296645; gi 131812; gi 2408082.

[a] This entry is indicated as sequence from *Arabidopsis thaliana* that is similar to peroxisomal calcium-dependent solute carrier.

[b] This entry was labelled 'similar to opine catabolism"; when used as a query sequence, it showed highest similarity with NMIORs from *Bacillus subtilis* (36% sequence identity) and *Streptomyces griseus* (30%).

Another pair of closely spaced genes from *G. gracilis* discovered *via* gene sequencing putatively encode methionine sulphoxide reductase (MSR) and a potential membrane-bound transporter, possibly a mitochondrial ATP/ADP transporter (Table 3). The MSR gene, like that for TSβS, was cloned following identification using a partial cDNA from the EST survey (Lluisma & Ragan, 1997) as a probe. The transporter-encoding ORF is located <1.4 kb downstream of the MSR coding region.

With only one of the five selected genes, that encoding starch branching enzyme (SBE; Lluisma & Ragan, 1998a), was no ORF detected in the proximal flanking regions. More than 1 kb (1483 bp) was sequenced upstream of the start codon, but only a short (318 bp) 3′ flanking region could be sequenced due to the proximity of the 3′ end of the gene to the insert-vector junction. BLAST searches with these regions yielded no significant matches.

We investigated the possibility that the close spacing of these four pairs of genes might be due to cloning artefacts. Two rounds of PCR reactions using primers designed for each gene within each pair were conducted, using genomic DNA from G. *gracilis* as template.

Figure 1 shows the results of the first and second round of PCR. In the first round, with genomic DNA from *G. gracilis* as template, products of the expected size (see Table 1) were obtained for each gene pair. These amplified fragments were used as template for the subsequent round, using nested primers (Table 2). The second-round PCR (Figure 1) shows amplification of smaller fragments of the expected size (Table 1), confirming that the PCR products of the first round

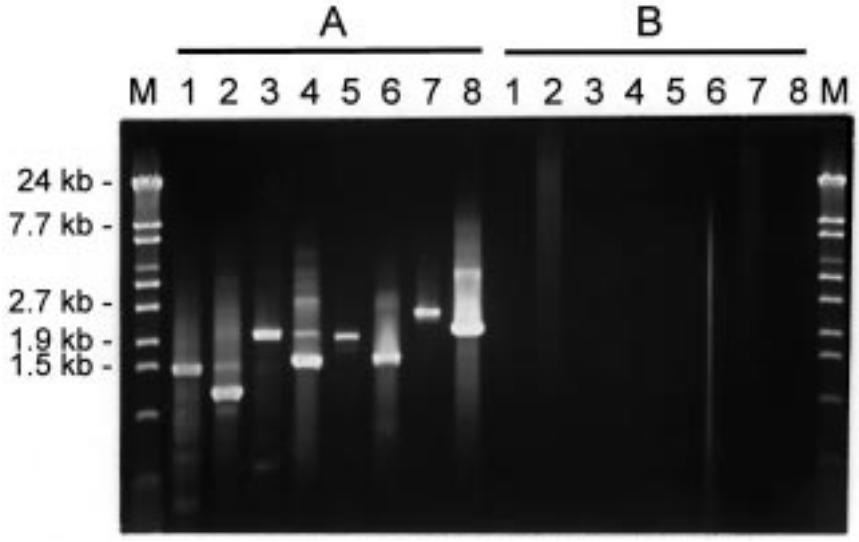

Figure 1. PCR amplification of intergenic regions from four pairs of closely spaced genes in *Gracilaria gracilis*, using the primers shown in Table 2. The PCR amplified products are from the pairs GalT-PTH (lanes 1 and 2), UGPase-DNA helicase (lanes 3 and 4), NMIOR-TSβS (lanes 5 and 6), and MSR-transporter (lanes 7 and 8). First round PCR: lanes 1, 3, 5, and 7. Second round PCR: lanes 2, 4, 6 and 8. Panel A, PCR reactions with genomic DNA template; Panel B, PCR reactions with no DNA template (negative controls).

were the desired genomic fragments. Negative controls (without DNA) yielded no amplification products (Table 1). These results confirm that the close spacing of genes indicated by sequencing of genomic clones reflects their actual proximity in the genome of *G. gracilis*.

Discussion

Characterization of red algal nuclear genomes has so far focused mainly on the most general properties, *e.g.*, genome size, number and size of chromosomes, G+C content, and relative sizes of kinetic components. Recombination studies have helped to determine linkage groups of loci whose alleles encode distinguishable phenotypic characteristics (for example, see van der Meer, 1990). Finer-scale resolution of genome structure in red algae has received little attention. One fundamental characteristic is the relative spacing of genes. Although extensive genomic sequencing would provide a global view, sequencing the flanking regions of cloned genes could be expected to provide at least a first indication of whether genes tend to be clustered or scattered in red algal genomes. The results we report herein confirm this expectation. An additional example, the 1.5-kb spacing between nuclear genes encoding polyubiquitin and mitochondrial aconitase in *G. gracilis* (as *G. verrucosa*; Zhou & Ragan, 1995), was reported previously. To the extent that this initial sample is representative, gene clustering may not be uncommon in the nuclear genome of *G. gracilis*.

The discovery of closely spaced genes in a red alga raises further questions, *e.g.*, whether the majority of genes in the *G. gracilis* genome tend to occur in clusters, and the extent to which corresponding genes in other red algae show the same (or similar) patterns. The results presented in this paper demonstrate that sequencing the flanking regions of cloned genes constitutes a valid and efficacious approach to surveying patterns of gene spacing in red algal nuclear DNA. Further characterisation of the structure of red algal genomes may be expected to yield practical benefits as well as fundamental insights into red algal biology.

Acknowledgements

We thank the staff of NRC-Institute for Marine Biosciences, Halifax, especially Ms Colleen Murphy, for excellent technical assistance. Issued as NRCC 42275.

References

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. Proc. natl Acad. Sci. U.S.A. 94: 6857–6861.

Bennetzen JL, Freeling M (1997) The unified grass genome: synergy in synteny. Genome Res. 7: 301–306.

Bernardi G (1995) The human genome: organization and evolutionary history. Ann. Rev. Genet. 29: 445–476.

Gish W, Gates DJ (1993) Identification of protein coding regions by database similarity search. Nature Genet. 3: 266–272.

Kapraun DF, Lopez-Bautista J, Bird KT (1996a) DNA base composition heterogeneity in some agarophytes (Gracilariales, Rhodophyta) from Mexico and the Philippines. J. appl. Phycol. 8: 229–237.

Kapraun DF, Lopez-Bautista J, Trono G, Bird KT (1996b) Quantification and characterization of nuclear genomes in commercial red seaweeds (Gracilariales) from the Philippines. J. appl. Phycol. 8: 125–130.

Lluisma AO, Ragan MA (1997) Expressed Sequence Tags (ESTs) from the marine red alga *Gracilaria gracilis*. J. appl. Phycol. 9: 287–293.

Lluisma AO, Ragan MA (1998a) Cloning and characterization of a nuclear gene encoding a starch-branching enzyme from the marine red alga *Gracilaria gracilis.* Curr. Genet. 34: 105–111.

Lluisma AO, Ragan MA (1998b) Characterization of galactose-1-phosphate uridylyltransferase gene from the marine red alga *Gracilaria gracilis*. Curr. Genet. 34: 112–119.

Lopez-Bautista J, Kapraun DF (1995) Agar analysis, nuclear genome quantification and characterization of four agarophytes (*Gracilaria*) from the Mexican Gulf Coast. J. appl. Phycol. 7: 351–357.

Moore G, Devos KM, Wang Z, Gale MD (1995) Grasses, line up and form a circle. Curr. Biol. 5: 737–739.

Smith RF, Wiese BA, Wojzynski MK, Davidson DB, Worley KC (1996) BCM Search Launcher – an integrated interface to molecular biology database research and analysis services available on the World Wide Web. Genome Res. 6: 454–462.

Worley KC, Wiese BA, Smith RF (1995) BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity results. Genome Res. 5: 173–184.

van der Meer J (1990) Genetics. In Cole KM, Sheath RG (eds), Biology of the Red Algae. Cambridge University Press, New York: 103–121.

Zhou Y-H, Ragan MA (1993) cDNA cloning and characterization of the nuclear gene encoding chloroplast glyceraldehyde-3-phosphate dehydrogenease from the marine red alga *Gracilaria verrucosa.* Curr. Genet. 23: 483–489.

Zhou Y-H, Ragan MA (1995) Characterization of the nuclear gene encoding mitochondrial aconitase in the marine red algal *Gracilaria verrucosa.* Plant mol. Biol. 28: 635–646.