CLONING AND CHARACTERIZATION OF GENES INVOLVED IN CARBOHYDRATE METABOLISM IN THE MARINE RED ALGA *GRACILARIA GRACILIS*

by

Arturo O. Lluisma

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
September 1997

0-612-24752-X

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Cloning and Characterization of Genes Involved in Carbohydrate Metabolism in the Marine Red Alga *Gracilaria gracilis*"

by Arturo O. Lluisma

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: September 6, 1997

External Examiner [redacted]

Research Supervisor

Research Supervisor

Examining Committee

DALHOUSIE UNIVERSITY

15 September 1997

AUTHOR: Arturo Obeña Lluisma

TITLE: Cloning and characterization of genes involved in carbohydrate metabolism in the marine red alga *Gracilaria gracilis*

DEPARTMENT: Biology

DEGREE: Ph. D. CONVOCATION: October YEAR: 1997

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABSTRACT

The molecular biology of carbohydrate metabolism in red algae is poorly known. Enzymological studies are few, and no gene for the biosynthesis of sugar nucleotides and polysaccharides has so far been characterized.

To isolate genes involved in carbohydrate metabolism in *Gracilaria gracilis*, genomic libraries were screened with homologous probes prepared either by PCR with degenerate primers, or from cDNAs previously isolated for generating expressed sequence tags (ESTs) from *G. gracilis*. Genes involved in carbohydrate metabolism, photosynthesis, protein synthesis and degradation, amino acid metabolism, and stress response were among those tagged by the ESTs.

Three genes were characterized. These encode galactose-1-phosphate uridylyltransferase (GALT, named *GgGALT1*), a key enzyme for D-galactose metabolism; UDPglucose pyrophosphorylase (UGPase; *GgUGP*), a key enzyme for sugar nucleotide synthesis; and starch branching enzyme (SBE; *GgSBE1*), which helps determine the structure of floridean starch. The three genes are devoid of introns. Each possesses a polyadenylation signal, TAAA, which occurs in all *G. gracilis* genes so far characterized, as well as a potential TATA box. Southern hybridization experiments indicate that the three genes are single-copy, but that other genes related to *GgGALT1* and *GgSBE1* exist. *GgGALT1* and *GgUGP* are each located close to another gene, hinting that occurrence of closely-spaced genes, atypical in eukaryotic genomes, may not be uncommon in the *G. gracilis* genome. The deduced proteins show high sequence similarity with their homologs in other organisms, but intriguing differences, such as nonconservative substitutions at functionally important sites, were observed. The protein encoded by *GgSBE1* lacks an N-terminal portion that could contain a possible target peptide, consistent with the cytosolic localization of floridean starch synthesis. The GgUGP and GgSBE1 proteins are as phylogenetically related to plant as they are to their animal and fungal homologs.

ABBREVIATIONS USED

ADPGlc	adenosine-diphosphate-D-glucose
Gal1P	D-galactose-1-phosphate
GALT	D-galactose-1-phosphate uridylyltransferase
GBE	glycogen branching enzyme
Glc1P	D-glucose-1-phosphate
MW	molecular weight
PCR	Polymerase Chain Reaction
RACE	Rapid Amplification of cDNA ends
SBE	starch branching enzyme
UDPGal	uridine-diphosphate-D-galactose
UDPGlc	uridine-diphosphate-D-glucose
UGPase	uridine-diphosphate-D-glucose pyrophosphorylase

ACKNOWLEDGEMENTS

I am grateful to those who have helped me bring this work to completion. I am especially indebted to Dr. Mark Ragan, my thesis supervisor, who provided invaluable support and guidance. I also thank Dr. Gary Newkirk for the solid administrative support; and the technical officers at the Marine Biology Section, NRC-Institute for Marine Biosciences, Halifax: Jeff Gallant, Janet Munholland, Jane Osborne, Dr. Rama Singh, and especially Colleen Murphy, for providing excellent technical assistance and for their kindness.

Many individuals have also provided help in a variety of ways as I carried out this project for almost half a decade. I thank them all: Becky Field, Dr. Ron McKay, Dr. Michael Reith, Ms. Carolyn Bird, Peter Shacklock, Dr. Yi-hong Zhou, Dr. Christoph Sensen, Dr. James Craigie, Dr. Anna Gawlicka, and others at NRC-IMB.

I also thank those who provided moral support: friends at the Univ. of the Philippines-Marine Science Institute (particularly Dr. Gavino Trono and his group); and most especially my father and mother, and my brother Mario.

I thank the International Development Research Center of Canada for the Ph.D. fellowship; Dr. John van der Meer for the opportunity to carry out this project at NRC-IMB; the P. Lett Fund of the Department of Biology, Dalhousie University for partial financial support; and UP-MSI for the study leave.

CHAPTER I

Principal carbohydrates and their biosynthesis in red algae: an overview

Principal forms of low- and high-MW carbohydrates in the red algae

The red algae (Rhodophyta) synthesize a variety of carbohydrates, but three forms can be regarded as of particular importance, as their biosynthesis accounts for a major portion of carbon flux. These are floridoside, the galactans, and floridean starch.

Floridoside (α-D-galactopyranosyl-(1-2)-glycerol) occurs in most red algae (Kremer and Kirst 1982), including the “primitive” forms *Cyanidium caldarium*, *Cyanidioschyzon merolae*, and *Galdieria sulphuraria* (De Luca and Moretti 1983), as well as some members of the Ceramiales, a group in which floridoside was thought to be absent (Barrow *et al.* 1995). Studies have shown that floridoside is a major photosynthetic product in red algae, and likely functions as a short-term organic-carbon reserve (Kremer 1978, Kirst 1980, Kremer and Kirst 1982, Macler 1986, Wu and Gretz 1993). Its involvement in osmoregulation has also been demonstrated (Reed 1990, Ekman *et al.* 1991). Floridoside is apparently produced only by the red algae (Kremer and Kurst 1982); its occurrence in other organisms has not been observed.

The red algae also synthesize a variety of polysaccharides for cell wall construction, including cellulose. However, the most abundant components of the cell wall in most red algae are galactans, such as agarans (agars) or carrageenans (Craigie 1990). As gel-forming polysaccharides (phycocolloids) with a variety of industrial uses, agars and carrageenans are commercially important (Jensen 1993). Many red algal species, such as members of the genus *Gracilaria* (Armisen 1995), are cultivated in many places of the world as commercial sources of these phycocolloids.

The abundance of D-galactose-containing carbohydrates (floridoside and galactans) in the red algae underscores the importance of D-galactose to red algal metabolism. This is also indicated by the observation that the sugar nucleotide UDPGal, the D-galactosyl donor in the biosynthesis of D-galactose-containing carbohydrates, has been found to be the most abundant sugar nucleotide in the agarophytic red alga *Pterocladia capillacea* (Manley and Burns 1991). The abundance of D-galactose-containing carbohydrates, and the exclusive occurrence of floridoside and certain types of galactans, are among the distinctive characteristics of the red algae.

Like many other organisms, the red algae store carbohydrates as α-1,4 glucans, called floridean starch in red algae (Raven *et al.* 1990). Studies have established that floridean starch differs from plant starch in structure and cellular localization of its synthesis. Whereas plant starch consists of amylose and amylopectin fractions, floridean starch consists mainly of amylopectin-like fractions (Percival and McDowell 1967), although some species apparently produce floridean starch that contains amylose-like fractions (McCracken and

Cain 1981). Interestingly, floridean starch biosynthesis occurs in the cytosol (Pueschel 1990), unlike in plants where starch is biosynthesized and stored in the plastid (Preiss 1991).

Carbohydrate metabolism in the red algae: islands of knowledge

Studies on the biochemistry of carbohydrate metabolism in the red algae are few and far between, but nevertheless have provided enough evidence for constructing the likely network of pathways that yield floridoside, galactans, and floridean starch as products (Fig. 1).

Bean and Hassid (1955) and Kremer and Kirst (1981) elucidated the biosynthetic pathway for floridoside. The pathway involves the condensation of glycerol-3-P (derived from dihydroxyacetone phosphate) and UDPGal, producing floridoside phosphate and UDP; floridoside phosphate is dephosphorylated to yield floridoside (Fig. 1). A key enzyme in this pathway, floridoside phosphate synthase, has been extracted from *Porphyra perforata*, and some of its enzymological properties assayed (Meng and Srivastava 1990).

As floridoside is considered a short-term storage form of organic carbon, its degradation and reutilization of its components are therefore important aspects of red algal carbohydrate metabolism. The biochemistry of floridoside hydrolysis and reutilization of its components (D-galactose and glycerol) is still poorly studied, but recent studies are beginning to provide key information. α-Galactosidase, which hydrolyzes floridoside into D-galactose and glycerol, has

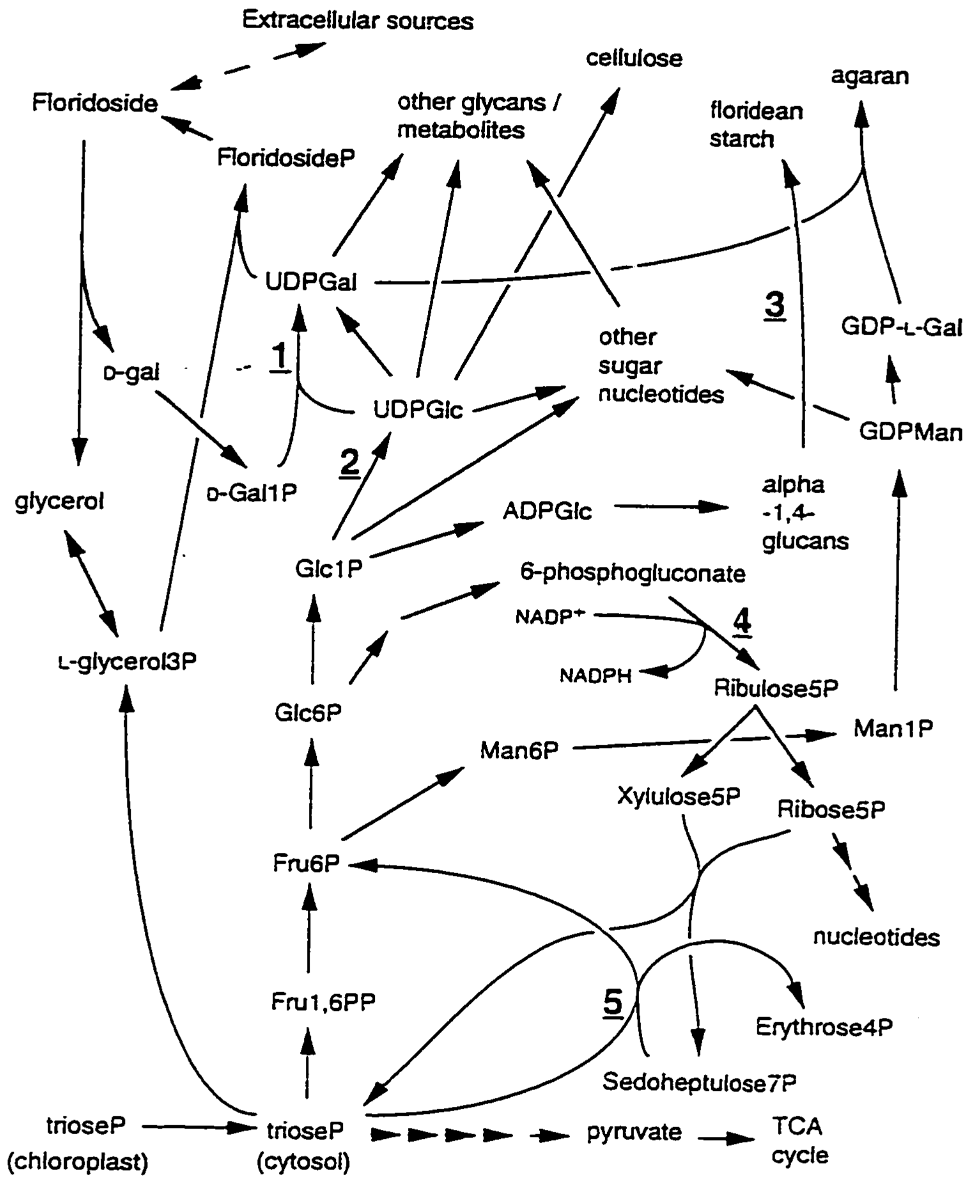

Figure 1. The metabolic pathways leading to the synthesis of agaran, floridean starch, and floridoside in agarophytic red algae (adapted from Manley and Burns 1991). The pentose phosphate pathway is also shown. Not all reactions and substrates/products are shown. The numbers represent the enzymes whose genes have been sequenced and characterized in this thesis (1, 2, and 3), or have been sequenced but have not yet been subjected to sequence analysis (4 and 5). 1=galactose-1-phosphate uridylyltransferase (GALT); 2= UDPglucose pyrophosphorylase (UGPase); 3= starch-branching enzyme (SBE); 4=6-phosphogluconate dehydrogenase (6PGDH); 5=transaldolase (TAL).

been isolated from two species of *Gracilaria, G. tenuistipitata* and *G. sordida*, and some of its enzymological properties determined (Yu and Pedersen 1990). There is also evidence that reutilization of liberated D-galactose proceeds *via* the Leloir pathway; the key enzyme of this pathway, galactose-1-P uridylyltransferase, has been demonstrated to occur in the red alga *Galdieria sulphuraria* (Gross and Schnarrenberger 1995b).

The pathways that lead to the synthesis of cell wall polysaccharides are less studied. In land plants, polysaccharide biosynthesis has been shown to proceed in three main stages (see Brett and Waldron 1996): formation of sugar nucleotides, the glycosyl donors; formation of polysaccharide chains; and post-polymerization modifications (*e.g.*, attachment of certain molecules to polysaccharides by covalent bonds). That similar patterns, and (for some reactions) similar enzymes, are involved in the biosynthesis of polysaccharides in red algae is indicated by the detection and isolation of the sugar donors and enzyme activity, similar to those in plants, in some red algal species. Studies in the 1960s (Su and Hassid 1962) and more recently (Manley and Burns 1991) have shown that sugar nucleotides, such as UDPglucose, ADPglucose, UDPgalactose, GDP-L-galactose, and GDPmannose, occur in the red algae. Manley and Burns (1991) noted, in particular, the relative abundance of UDPGal compared to the other sugar nucleotides. The activities of certain enzymes, particularly UDPglucose pyrophosphorylase, GDPglucose pyrophosphorylase, GDPmannose pyrophosphorylase, and UDPglucose-4-epimerase have also been detected. The data supported the hypothesis that the D- and L-galactose donors for agaran formation are UDPGal and GDP-L-gal, respectively, which are produced by different pathways (Fig. 1; Manley and Burns 1991).

Floridean starch is apparently biosynthesized via the ADPGlc pathway (*sensu* Preiss 1991), the pathway used by plants for starch biosynthesis. Nagashima *et al.* (1971) have shown that the biosynthesis of floridean starch in the red alga *Serraticardia maxima*, although occurring in the cytosol, involves ADPglucose pyrophosphorylase, which produces ADPGlc, and starch synthase (ADPGlc transferase), which preferentially utilizes ADPGlc as the glucosyl donor.

Carbohydrate metabolism in red algae: ocean of questions

Although biochemical studies have provided a general picture of how the red algae construct their carbohydrates, details are wanting. This is especially clear if one wishes to understand how carbohydrate metabolism is regulated at the molecular level in the red algae. The problem of how the activities of each enzyme in the pathways are regulated has remained unaddressed. This will require the study of individual enzymes, at different levels: the gene, the protein, the metabolic pathway, and overall cellular physiology. These studies, however, cannot be done until carbohydrate metabolism genes and proteins from red algae are isolated, something that has not so far been done. Consequently, many areas of investigation remain inaccessible. Specifically, the role of transcriptional and posttranslational control of enzyme activity remains to be elucidated. It is not known, for instance, which enzymes are subject to transcriptional control, which endogenous and/or exogenous signals effect such control, what mechanisms allow those signals to be received, and what are the *cis*-acting elements on the gene that mediate between the signals and transcription.

It is also not known what posttranslational processes control enzyme activities. It is here where knowledge of enzyme properties (kinetic properties, presence of regulatory sites [*e.g.*, phosphorylation or glycosylation sites], allosteric properties, which metabolites serve as allosteric regulators, structure-function relationships, etc.) are required, none of which are presently available.

Aside from the enzymes catalyzing the reactions shown in Figure 1, there are other proteins that are critically involved in polysaccharide biosynthesis that are yet to be isolated from red algae (and from land plants for that matter) and to be studied in detail. These include the sugar nucleotide transporters (biosynthesis of cell wall polysaccharides occurs in the Golgi apparatus, whereas synthesis of sugar nucleotide occurs in the cytosol), as well as the glycosylation primers. Recently, Muñoz *et al.* (1996) have studied a Golgi membrane-localized UDPglucose transporter in pea, which is apparently an antiporter, transporting UDPglucose into the Golgi apparatus in exchange for either UDP or UMP, or both. They also reported that the activity of this transporter could potentially limit polysaccharide biosynthesis. cDNAs for human Golgi membrane UDPGal transporters have already been isolated (Ishida *et al.* 1996, Miura *et al.* 1996);

that UDPgalactose transporters could be rate-limiting in human Golgi galactosylation reactions has also been demonstrated (Toma *et al.* 1996).

The question of what protein primes floridean starch biosynthesis is yet to be answered. As floridean starch biosynthesis occurs in the cytosol, there is a distinct possibility that this protein may be related to glycogenin in animals and fungi. Isolation of this protein will be required to evaluate this possibility. The role of this protein in determining cellular localization as well as rate of floridean starch formation also requires investigation.

The molecular approach to understanding carbohydrate metabolism: lessons from land plant research

The molecular biology of carbohydrate metabolism in land plants has undergone spectacular advances in the last decade. Molecular genetics, in combination with biochemistry, has been routinely used to dissect the properties of enzymes and their genes, and their roles in metabolic pathways. Genetic manipulation of carbohydrate metabolism in plants both for research and for commercial purposes has also become common (for example, see Visser and Jacobsen 1993, Stitt 1995, and Herbers and Sonnewald 1996). Studies of the structure and function of the genes and enzymes form the basis of these advances. Among the interesting findings of these studies are the following (some of which are observed with other non-carbohydrate metabolism-related genes as well):

1. Regulatory *cis*-acting regions of genes may be found upstream and downstream of the protein-coding region (*e.g.*, an AGPase gene: Nakata and Okita 1996).

2. Genes may show unusual gene structure, which could serve regulatory functions (*e.g.*, a single AGPase gene in *Hordeum vulgare* produces two different transcripts by alternative splicing: Thorbjørnsen *et al.* 1996).

3. mRNA and protein levels may be uncoupled, and may indicate posttranslational control (Nakata and Okita 1995).

4. The amount of enzyme activity necessary for normal physiology may be much lower than what is observed (for example, antisense inhibition of UGPase in potato, resulting in 95-96% reduction in enzyme activity, did not adversely affect growth, development, and carbohydrate metabolism, indicating that only 4% of wild-type UGPase activity is sufficient: Zrenner *et al.* 1993).

5. Availability of clones facilitates structure-function relationship studies; cloned DNAs are routinely used in experiments to alter residues (*via* site-directed mutagenesis) of proteins and examining the properties of the altered proteins (*e.g.*, AGPase proteins, Charng *et al.* 1995, Sheng *et al.* 1996).

6. The role of an enzyme's allosteric properties in regulating carbon flux *in vivo* can be tested by replacing that enzyme with a heterologous enzyme that has the same catalytic properties but different allosteric properties (for example, see Stark *et al.* 1995).

7. The inhibition of the activity of an enzyme may affect overall physiology (*e.g.*, inhibition of AGPase in transgenic potatoes alters levels of sugars and expression of storage protein genes: Müller-Röber *et al.* 1992).

The molecular biology of carbohydrate metabolism in red algae: taking the first steps

The general goal of this thesis is to help expand what is known about carbohydrate metabolism in the red algae. This thesis has been based on two premises. First, molecular genetics offers an excellent approach for understanding carbohydrate metabolism in red algae and for paving the way for subsequent biochemical studies. In this approach, one begins by characterizing the genes of the enzymes involved in the pathways, as this represents a concrete step forward to understanding not only the gene but also the enzyme encoded by the gene, and the pathways in which that enzyme participates. Clues about the enzyme's properties are often revealed by analysis of the deduced amino acid sequences, and by comparison of the deduced protein with its homologs in other organisms. These clues could be translated into hypotheses that can be tested by biochemical experiments.

Second, the complete absence of molecular genetic information about carbohydrate metabolism in red algae warrants a broad assault on this problem. Understanding carbohydrate metabolism requires, among other things, understanding the properties of all the enzymes in the pathways, at both the gene and protein levels. It follows that the genes of all the enzymes should be targeted for isolation and characterization.

Thus, a specific goal was aimed at: to clone and characterize some genes involved in carbohydrate metabolism in the red algae.

For this research, the agar-producing marine red alga *Gracilaria gracilis* was chosen as the "model" red alga. *G. gracilis* belongs to a genus whose members are among the most economically important species in the world, being commercial sources of agar (Armisen 1995). Members of the genus *Gracilaria* are also among the most studied in the red algae with respect to physiology and classical genetics.

To date, not a single gene participating in the pathways shown in Figure 1 has been cloned. As the eventual goal is to understand how each enzyme in the pathways contribute to the regulation of carbohydrate metabolism, the gene for each enzyme would be a target for characterization. However, attempts to clone genes from red algae are bound to face serious technical difficulties: heterologous probes cannot be obtained for a number of genes; and available biochemical information and methodologies (*e.g.*, for purifying enzymes to obtain enough pure protein for sequencing) are limited. Thus, at this stage of development of red algal molecular biology and biochemistry, the choice unavoidably has to be partially dictated by the feasibility of cloning.

Of the several approaches to gene cloning that were tried in this research, two yielded positive results. One approach involved the generation of homologous probes *via* PCR reactions that made use of degenerate primers, designed from highly conserved regions of proteins. The second approach involves generation of Expressed Sequence Tags (ESTs, which are partial single-pass sequences of cDNAs), and selecting those that are fragments of genes involved in carbohydrate metabolism. This approach was inspired by the success that has been demonstrated by other EST projects, notably human (Boguski 1995) and *Arabidopsis thaliana* (Delseny *et al.* 1997).

This thesis reports the characterization of three genes from *G. gracilis.* One gene, galactose-1-phosphate uridylyltransferase (GALT, Chapter III), was cloned *via* the EST approach. This enzyme is a key component in D-gal metabolism in red algae, being involved in the reutilization of D-galactose resulting from floridoside hydrolysis (Fig. 1). Two genes, starch branching enzyme (SBE: Chapter IV), and UDPglucose pyrophosphorylase (UGPase: Chapter V), have been cloned using the "PCR-with-degenerate-primers" approach. SBE is a key enzyme that help determine the final structure of floridean starch biosynthesis (Fig. 1). UGPase, as indicated in Figure 1, plays a pivotal role in red algal metabolism, as UDPGlc, its product, is an intermediate of many pathways.

The results of the EST project is reported in Chapter II. One of the obvious benefits of ESTs is that they facilitate the cloning of the "tagged" genes. In addition to the GALT gene, two other carbohydrate metabolism genes, those of 6-phosphogluconate dehydrogenase (6PGDH) and transaldolase (TAL), have also been cloned (but are not included for discussion in this thesis) using ESTs as probes. 6PGDH and TAL are components of the pentose phosphate pathway (Fig.1), which produces NADPH as well as precursor of nucleotides (hence, sugar nucleotides) and aromatic amino acids, among other compounds. Genes involved in cellular processes other than carbohydrate metabolism have also been tagged by the ESTs, some of which have been cloned and sequenced; characterization of these genes is underway and will be reported in the literature in due course.

Chapter II

Expressed Sequence Tags (ESTs) from the marine red alga *Gracilaria gracilis*[1]

Abstract

Expressed sequence tags (ESTs) are partial sequences of cDNAs, and can be used to characterize gene expression in organisms or tissues. We have constructed a 200-sequence EST database from vegetative thalli of *Gracilaria gracilis*, the first ESTs reported from any alga. This database contains recognizable ESTs corresponding to genes of carbohydrate metabolism (seven), amino acid metabolism (three), photosynthesis (five), nucleic acid synthesis, repair and processing (three), protein synthesis (14), protein degradation (six), cellular maintenance and stress response (three), other identifiable protein-coding genes (13), and 146 sequences for which significant matches were not found in existing sequence databases. We have already used this EST database to recover genes of carbohydrate biosynthesis from *G. gracilis*.

[1] This paper was co-authored by A.O. Lluisma and M.A. Ragan, and has been accepted for publication in the Journal of Applied Phycology.

Introduction

Although it has been known since mid-century that DNA is the active principle controlling heredity (Avery *et al.,* 1944) and that genetic information is encoded as the linear sequence of nucleotide bases (Watson and Crick, 1953), it was only after the development of high-throughput automated methods for sequencing DNA in the mid-1980s (Chen, 1994) that it became realistic to try to obtain the complete genetic blueprint of an organism. The first explicit proposal to sequence the entire genome of human was put forward in 1985 (Yager *et al.,* 1994), and by mid-1995 large-scale sequencing was underway. In the same year, the sequence of the 1.83-Mbp genome of the bacterium *Haemophilus influenzae* was released (Fleischmann *et al.,* 1995); at least nine further prokaryote genomes have since been completely sequenced. The first eukaryote genome to be sequenced was of the yeast *Saccharomyces cerevisiae* (12.07 of 13.3 Mbp sequenced, the remainder being repeats; Goffeau *et al.,* 1997); genomes of the nematode *Caenorhabditis elegans* (*ca* 100 Mbp) and the flowering plant *Arabidopsis thaliana* (*ca* 100-150 Mbp) are expected to be completed during 1997 and by 2004 respectively. The *ca* 3000-Mbp human genome is anticipated to be largely complete by about 2003.

For many purposes, however, complete genomic sequencing is neither practical nor cost-effective. Whereas open reading frames (potential genes) tend to be tightly packed together in genomes of prokaryotes, most eukaryote genomes are, so far as is known, constituted predominantly (often more than 90%) of non-coding regions both within (introns) and between genes (intergenic regions). Although the sequences of these non-coding regions may be relevant to some issues of chromosomal structure, genetic regulation and comparative biology, for many central questions *e.g.* of protein structure and cellular metabolism they are not only unnecessary, but indeed greatly complicate the discovery of genes, given that intron/exon boundaries can be difficult to locate in primary sequence data, and that mRNAs can be spliced in alternative ways. Moreover, even an efficient analysis of genomic DNA would provide no information on what genes are actually expressed in a given organism or tissue at any specific time.

These considerations motivated the development of the expressed sequence tag (EST) approach to genomic characterization, first demonstrated for human cDNAs (Adams *et al.,* 1991; Boguski, 1995). In the EST approach, clones from a cDNA library are randomly isolated and partially sequenced, typically from the 5' end and on only one of the two DNA strands. These sequences thus serve as markers, or tags, for genes expressed in the corresponding organism or tissue, and have proven useful in many applications including recovery of full-length cDNA or genomic clones (including those not clonable by classical approaches), discovery of novel genes, recognition of exons, characterization of exon/intron boundaries, delineation of protein families, development of genetic maps, identification of organism- or tissue-specific genes, and investigation of genes of unknown function (Adams *et al.,* 1995; Boguski, 1995; Claverie, 1996; Hillier *et al.* 1996; Rounsley *et al.,* 1996; Delseny *et al.,* 1997; Wolfsberg and Landsman, 1997). Moreover, cDNA libraries (including subtractive libraries) can be prepared from tissues under a wide range of conditions, and the corresponding ESTs can thus be used to identify and characterize not only normal but also developmental states, as well as conditions of stress, pathology and disease.

A publicly accessible EST database (dbEST) is being maintained at the U.S. National Center for Biotechnology Information (NCBI). More than 1 million ESTs derived from 82 organismal species are currently on deposit (dbEST release 053097). Most of these are from human (>719000), mouse (>185000), *A. thaliana* (>31000), *C. elegans* (>30000), and *Oryza sativa* (rice, >12000); at the other end of the spectrum are organisms represented by only one EST. Not a single algal species was currently represented in dbEST (release 053097) prior to submission of the sequences reported herein.

We are interested in the molecular genetics of cell wall biogenesis and carbohydrate biosynthesis in the commercially important agarophyte *Gracilaria gracilis* (formerly *G. verrucosa:* Bird and Kain, 1995). As part of our studies, we generated a small EST database for this red alga. This database has already proven useful in the cloning of certain genes of carbohydrate metabolism from *G. gracilis.*

Materials and methods

A cDNA library prepared from young vegetative thalli of *Gracilaria gracilis* (Stackhouse) Steentoft, Irvine & Farhnam (Steentoft *et al.*, 1994) grown in the laboratory (Zhou and Ragan, 1993) was used as the source of clones. The library, in phage lambda (lambda ZAPII, Statagene, La Jolla CA) was plated out on Luria-Bertani (LB) bacto-agar plates (Sambrook *et al.,* 1989) at low density (<200 pfu per 180 mm-diameter plate), and more than 400 individual plaques were randomly cored out and eluted from the agar plugs at 4°C with 80 µL of SM buffer (50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 8 mM $MgSO_4$, 0.01% gelatin).

The inserts from the clones were amplified by the polymerase chain reaction (PCR). For each clone, 1 µL of eluted phage was combined with 1.5 U *Taq* DNA polymerase (Bio/CAN Scientific, Mississauga, ON), 3.7 pmol each of the vector-specific primers T3 and T7 (Kretz *et al.*, 1993; synthesized in-house), and 200 µM dNTPs, diluted to 50 µL of 1X reaction buffer (BIO/CAN Scientific), and PCR was carried out on a Perkin-Elmer model 9600 thermal cycler (Norwalk, CT) through 35 cycles of denaturation (94°C, 30 s except initial denaturation 90 s), annealing (58°C, 30 s), and extension (72°C, 60 s except final extension 5 min). After PCR amplification, 4 µL of each reaction was subjected to electrophoresis on a 2.5% (w/v) agarose gel to determine the size of the insert.

PCR products greater than 500 bp were sequenced. Templates were prepared by centrifuging the PCR reactions with 2 mL distilled water through Centricon-100 or -30 columns (Amicon Canada, Oakville, ON). Sequencing was carried out on an Applied Biosystems 373A sequencer (Foster City, CA) following the manufacturer's Dye Deoxy terminator cycle sequencing protocol, and using T3 (one of the PCR primers) as the sequencing primer. As the cDNA library was constructed directionally (Zhou and Ragan, 1993), and the T3 site on the vector is located upstream (5') of the ligated cDNAs, the sequences were predominantly of the "sense" strands.

Data from the sequencer were processed manually. Vector sequences were removed, and ambiguities resolved with reference to original trace data, using Klatte's ABIView software. Sequences were exported and used in both manual and automated querying of the nr (=non-redundant) peptide sequence database at NCBI. Searches were implemented using BLASTX (Gish and States, 1993) under its default options through the BCM Search Launcher, a WWW-accessible molecular database-searching facility maintained by the Human Genome Center, Baylor College of Medicine (BCM; Smith *et al.,* 1996). This facility further processes the results using the BEAUTY algorithm (Worley *et al.*, 1995). For automated searching, we used the BCM Search Launcher Batch Client program (BCM-SLBC) installed locally on a DEC Alpha 2100-5/250 dual-processor workstation under Unix. The BCL-SLBC acts as an interface to the BCM Search Launcher facility.

For Southern hybridizations, DNA was extracted from *G. gracilis* and purified as described by Zhou and Ragan (1993); the DNA (5 μg per reaction) was digested with restriction enzymes, subjected to electrophoresis on a 0.7% agarose gel, and blotted onto Zeta-Probe membranes (Bio-Rad Laboratories, Richmond, CA) using the manufacturer's protocol. Southern hybridization and washings (3 times, 30 min each: 4X, 2X, and 0.5X SSC/0.1% SDS) were performed at 65°C in a Techne hybridization oven (Techne (Cambridge) Ltd, Cambridge, UK) essentially following the protocol described by Sambrook *et al.* (1989); the probes were synthesized from the PCR-amplified inserts and random-prime labelled with α-^{32}P-dCTP. Probes based on the *G. gracilis* aconitase gene (Zhou and Ragan, 1995b) and a eubacterial starch synthase gene (Lluisma, unpublished) were used in positive and negative controls respectively.

Results and discussion

Characterization and reliability of the EST database

Most of the inserts chosen for sequencing ranged between 500 to 1000 bp, while a few were more than 1 kb in length. Single-pass automated sequencing was carried out on more than 200 PCR-amplified inserts, typically yielding reads of from 350 to 550 nucleotides each (after the removal of vector sequences). All ESTs have been deposited in the NCBI database, and can also be accessed via IMB's Webpage (http://www.nrc.ca/imb/***** [remainder of URL to be supplied in proof]), where information on clone availability may be found.

The reliability of an EST database depends to a large extent on the quality of the underlying cDNA library. To test whether any of the ESTs might derive from organisms other than *Gracilaria* (*e.g.* bacteria or epiphytes), we used seven of these ESTs as probes in Southern hybridization experiments with *G. gracilis* genomic DNA; all hybridized strongly (data not shown). Some ESTs showed highest similarities with red algal sequences in the database, and none yielded a suspiciously close match with any of the seven complete genomes (*Saccharomyces cerevisiae* and six prokaryotes) or any other sequence then in the public databases. Thus all available evidence suggests that these ESTs do, in fact, correspond to genes expressed in *G. gracilis*. One EST had significant ($P < 10^{-17}$) similarity with plastidic 50S ribosomal RNA sequences, and was deleted from the EST library.

Eight pairs of overlapping ESTs were identified among these 200, including two pairs with identifiable function: ESTs 84 and 401, tagging the Reiske iron-sulfur protein of the cytochrome b_6f complex, are identical within a 245-bp region of overlap, while ESTs 28 and 225, tagging elongation factor 2, are identical in 132 bp within a 134-bp overlap. The other overlaps ranged from 45 of 45 bp (ESTs 262 and 398) to 363 of 364 bp (ESTs 407 and 417). These results suggest that the cDNA library is not highly redundant.

Identification of EST sequences

Each of the ESTs was used as a query sequence in searching the nr peptide sequence database at NCBI. The search program used was BLASTX, which compares the six-frame conceptual translation of nucleotide sequences (*i.e.*, translations of both strands) with peptide sequences in the database using the BLAST algorithm (Altschul *et al.,* 1990); in the subsequent discussion, probabilities of matches are thus reported at the protein level. Fifty-four of the 200 ESTs showed BLAST scores greater than 100, and are presented in Table 1; this is a conservative criterion, as BLAST scores above 80 generally indicate that a match is significant (Pearson, 1991). Full BLAST results (in html format) for these 54 ESTs can be viewed via IMB's Webpage.

Four of these ESTs tag genes that have previously been reported (as genes or cDNAs) from red algae, encoding polyubiquitin (*Aglaothamnion neglectum*: Apt and Grossman, 1992; *G. gracilis* [as *G. verrucosa*]: Zhou and Ragan, 1995a), actin (*Chondrus crispus*: Bouget *et al.*, 1995), the gamma subunit of R-phycoerythrin (*A. neglectum*: Apt *et al.,* 1993), and a chlorophyll *a/b*-binding protein (S. Tan, A. Ducret, R. Aebersold and E. Gantt, GenBank accession U58680). Many others are first reports for red algae, including tags for genes specifying adenine nucleotide translocase (probably a plastidic isoform), SIR2 (Silent Information Regulator 2), the beta subunit of tryptophan

Table 1. *Gracilaria gracilis* ESTs with significant (BLASTX scores > 100) sequence similarity at the amino acid sequence level to peptide sequences in the NCBI nr (nonredundant) peptide database. For each EST, the maximum BLASTX score in the search result is shown. A WWW version of this table, with hyperlinks to the unedited BLASTX+BEAUTY search results, can be viewed at http://*.*.*/*. All ESTs have been deposited in the NCBI public EST database dbEST; dbEST accession numbers are shown. GenBank accession numbers are AA495495 through AA495694, and GenBank gi numbers 2228816 through 2229015 inclusive. &= ESTs used as probes to isolate genomic clones from a *G. gracilis* genomic library; @= ESTs corresponding to genes already characterized from a red alga (see text).

EST #	putative ID / homolog	BLASTX score (max)	Database match(es)	dbEST Accession no.
Carbohydrate metabolism				
58	galactose-1-phosphate uridylyltransferase &	322	P09580	1140323
93	fructose-bisphosphate aldolase	601	P14540	1140272
100	6-phosphogluconate dehydrogenase &	138	M64328 (1)	1140324
142	transaldolase &	293	P30148, pdb 1ONR	1140273
211	adenine nucleotide translocase	247	Z49227	1140274
230	fructose-bisphosphate aldolase	448	prf 160908	1140275
324	glucose-6-phosphate isomerase, cytosolic	642	P34795	1140276
Amino acid metabolism				
96	S-adenosylmethionine synthetase	342	P18298	1140277
259	S-adenosylhomocysteine hydrolase	409	P50250	1140278
313	tryptophan synthase beta chain 2 &	759	P25269, pir JQ1073	1140279
Photosynthesis				
36	chlorophyll a/b binding protein @	166	U58680	1140280
84	cytochrome b_6-f complex Fe-S subunit	147	Y09612	1140281
140	R-phycoerythrin gamma chain precursor @	417	U72642	1140282
330	protoporphyrin IX Mg chelatase subunit	631	U26916	1140201
401	cytochrome b_6-f complex Fe-S subunit	259	P26292	1140283
DNA/RNA synthesis, repair, and processing				
24	ATP-dependent RNA helicase	698	S47451	1140284
35	DNA repair protein (helicase)	422	L01414, Q00578	1140285
252	polyA (mRNA)-binding protein	145	X89969	1140286
Protein synthesis				
28	elongation factor 2	369	P28996	1140287

Table 1, continued.

39	N-terminal acetyltransferase	216	Q05885	1140288
91	protein disulfide isomerase	174	pir ISMSSS	1140289
131	40S ribosomal protein S9	131	P52810	1140290
137	protein translation factor SUI1 homolog	208	P33278	1140291
187	40S ribosomal protein S7 (S8)	249	P02362	1140292
208	60S ribosomal protein L31	265	P46290	1140293
225	elongation factor 2	713	K03502, M76131	1140294
229	ribosomal protein L7	293	P05737	1140295
275	40S ribosomal protein S18	497	P34788	1140296
283	chaperonin	431	P53451	1140297
289	40S ribosomal protein S12	386	P46405	1140298
371	translation elongation factor EF-3	159	Z73582	1140299
384	eukaryotic peptide chain release factor 1	172	S31445, U40218	1140300
Protein degradation				
184	ubiquitin-conjugating enzyme	536	P46595	1140301
206	26S protease regulatory subunit 4	908	P46466	1140302
219	polyubiquitin @	944	U16852	1140303
232	ubiquitin-protein ligase	263	U58653 (2)	1140304
254	proteasome beta chain precursor	149	P28070, U65636	1140305
333	26S protease regulatory subunit 8	375	X81986	1140306
Cellular maintenance / stress response				
47	methionine sulfoxide reductase &	342	P54150	1140307
183	heat shock 70 KD protein	140	P16394, pir HHUM7B	1140308
379	glutathione S-transferase 1	230	P46436	1140309
Miscellaneous				
42	alpha-aminoacylpeptidase	182	D90731 (3)	1140310
72	adenylyl cyclase-associated protein	105	P40123, P52481	1140311
80	cell division control protein/ER ATPase	595	P46462 (4)	1140312
105	mt-protein/TAT-binding homolog 10)	490	U09358 (5)	1140313
133	phosphatidylinositol 4-kinase alpha	182	U41540	1140314
147	coatomer beta subunit (beta-coat protein)	244	P41810, S54534	1140315
157	actin @	398	P53499	1140316
253	ATP-binding transport protein	169	U64875	1140317
380	SIR2 (Silent Information Regulator 2) &	120	P53685	1140318
382	actin @	225	P53499	1140319
423	Na+/K+-exchanging ATPase alpha subunit	266	P35317	1140320
430	inositol-1,4,5-trisphosphate 5-phosphatase	206	L36818	1140321
440	PELOTA / DOM34 protein	438	U27197	1140322

Table 1, continued.

Notes

(1) Eleven other database entries with this score: M64329, M64330, M64331, M63821, M63823, M63824, M63826, M63827, M63828, M63829, P37756.
(2) Four other database entries with this score: A38564, P22314, Q02053, S12567.
(3) Three other database entries with this score: D90732, P04825, pir DPECN
(4) Four other database entries with this score: P03974, P23787, Q01853, pir A26360.
(5) Two other database entries with this score: P40431, X81068.

synthase, methionine sulfoxide reductase, glutathione transferase, and a DNA repair protein.

Two of these proteins, methionine sulfoxide reductase (MSR) and glutathione transferase, function in the cellular response to stress. Seaweeds are potentially subject to oxidative stress during periods of desiccation or strong solar irradiation, or both, conditions which favor the formation of oxidants (*e.g.* peroxides) which can attack membranes and other biomolecules. Levine *et al.* (1996) have proposed that a cellular mechanism for coping with oxidative stress might involve methionine residues in proteins, which act as endogenous scavengers of oxidants, and MSR, which subsequently reduces methionine sulfoxides (oxidized methionines) back to methionines. The ESTs could be used as probes, *e.g.* in Northern hybridizations, to determine whether the expression of the genes correlates with conditions that subject *G. gracilis* or other red algae to increased oxidative stress.

SIR2 is a component of a protein complex that, at least in the yeast *S. cerevisiae*, helps to silence (transcriptionally inactivate) chromatin domains, and is particularly important in the determination of mating type (Laurenson and Rine, 1992). The SIR2 protein is thought also to participate in other cellular processes including cell-cycle progression, maintenance of chromosomal stability, and DNA recombination (Gottlieb and Esposito, 1989; Brachmann *et al.*, 1995).

Most of the *G. gracilis* ESTs, however, could not be identified by database matching. Some sequence motifs characteristic of certain functions were observed, *e.g.* a DNA-binding motif in EST number 121. In this case, however, we sequenced the flanking regions, and no further similarity was observed; this cDNA might encode a novel DNA-binding protein, *e.g.* a transcription factor. A few *G. gracilis* ESTs match functionally unassigned ESTs or ORFs (open reading frames) from other organisms, while most of these ESTs do not show significant similarity to any sequence in the databases; some of these presumably represent genes specific to *G. gracilis* (or to members of genus *Gracilaria*, family Gracilariaceae, etc.). Analysis of ORFs and ESTs from different organisms have shown that a significant portion (at least 30%) of genes in organisms could be taxon-specific "orphans" (Claverie, 1995; Dujon, 1996; Rounsley *et al.*, 1996; Ragan, unpubl.).

Isolation of G. gracilis *genomic clones using* G. gracilis *ESTs as probes*

We have already used some of these ESTs to isolate genomic clones for some *G. gracilis* genes that putatively code for enzymes of carbohydrate metabolism, including transaldolase, 6-phosphogluconate dehydrogenase, and galactose-1-phosphate uridylyltransferase (Table 1). The former two are enzymes of the pentose phosphate pathway, which produces NADPH as well as biosynthetic precursors for key pathways. Galactose-1-phosphate uridylyltransferase catalyzes the reversible transfer of the uridylyl moiety from UDP-glucose to galactose-1-phosphate, and thereby plays a key role in galactose metabolism. Further characterization of these clones will be described in more detail elsewhere.

Automating the identification of G. gracilis *ESTs*

Because the sequence databases are growing so rapidly (doubling in size about every 18 months), it would potentially be fruitful to re-query the databases on a routine basis, ideally automatically. To this end we have compiled all "anonymous" ESTs in a single file, and installed at IMB a program (the Search Launcher Batch Client program; see Materials and Methods) that automatically compares each of these ESTs against the sequence databases (*e.g.* dbEST) *via* the World Wide Web. New *Gracilaria* or other red algal ESTs can be readily added.

"Data-driven" (as opposed to "problem-driven") approaches are becoming increasingly common not only in gene cloning, but much more broadly throughout biological research, as molecular-sequence data, including ESTs, become increasingly abundant. Just as the human and *A. thaliana* EST databases have proven to be invaluable resources in human biomedicine and plant biology respectively (Boguski, 1995; Hillier *et al.* 1996; Schuler *et al.*, 1996; Delseny *et al.*, 1997), our initial studies suggest that a larger red algal EST database – thousands or tens of thousands of sequences -- would almost certainly be an effective and cost-efficient tool opening up for molecular characterization many hitherto refractory aspects of red algal biology, including the genetics and enzymology of the biosynthesis of cell-wall polysaccharides. Large EST initiatives are typically carried out as multi-laboratory collaborations, and a similar model could be appropriate as well for a red algal EST project.

CHAPTER III

Characterization of a galactose-1-phosphate uridylyltransferase gene from the marine red alga *Gracilaria gracilis*

Introduction

The red algae produce and utilize D-galactose extensively. Galactans, such as the commercially important carrageenans (essentially polymers of D-galactose) and agarans (essentially polymers of D- and L-galactose), are the most abundant component of cell walls in most red algae (reviewed by Craigie 1990). Certain red algal species, including members of the genus *Gracilaria* (Armisen 1995), are cultivated on a large scale in many countries as commercial sources of agars. D-galactose also participates in energy metabolism as well as physiological processes: it is a constituent of floridoside (α-D-galactopyranosyl-(1-2)-glycerol), which is a major photosynthetic product in most red algae (Kirst 1980, Kremer 1978, Kremer and Kirst 1981), functions as a short-term carbon reserve (as sucrose does in higher plants) (Macler 1986), and is involved in osmoregulation (Reed 1990, Ekman *et al.* 1991).

One of the key enzymes in galactose metabolism is galactose-1-phosphate uridylyltransferase (GALT; UDPglucose: α-D-galactose-1-phosphate uridylyltransferase, EC 2.7.7.12), which catalyzes the reversible transfer of the uridine 5' phosphoryl moiety from UDP-glucose to Gal1P to produce UDPGal and Glc1P; UDPGal is a key intermediate of D-galactose metabolism. GALT is a ubiquitous enzyme, and has been described for phyletically diverse organisms. It is a component of the Leloir pathway (see Frey 1996), which allows the entry of free D-galactose into metabolic (specifically, catabolic) pathways. The biochemistry of GALT in the red algae is still poorly characterized. However, at least one function can be ascribed to this enzyme (and to the other Leloir pathway enzymes presumably present in red algae): it allows the utilization and reutilization of D-galactose derived from the hydrolysis of floridoside. α-Galactosidase, which cleaves floridoside to D-galactose and L-glycerol, has been reported for two species of *Gracilaria*, *G. sordida* and *G. tenuistipitata* (Yu and Pedersén 1990); presumably, a D-galactose kinase exists in red algae that phosphorylates D-galactose to Gal1P, which is a substrate of GALT. The floridoside pool can be large (roughly 2-8% of dry weight: Kirst 1980), and is dynamic, increasing and decreasing depending on such factors as salinity (Ekman *et al.* 1991), nitrogen content (Macler 1986), and time of day (Meng and Srivastava 1993). Presumably, reduction of the floridoside pool would require increase in GALT activity to allow D-galactose to reenter the metabolic pathway. This is especially important as GALT is apparently the only enzyme in red algae that can convert Gal1P to UDPGal; Gross and Schnarrenberger (1995a) have shown that whereas GALT activity occurs in the red alga *Galdieria sulphuraria*, UDPGal pyrophosphorylase activity could not be detected (UDPGal pyrophosphorylase is the only other enzyme capable of converting Gal1P to UDPGal).

D-galactose in red algae can also originate from sources other than floridoside. The report that *G. sulphuraria* can utilize exogenous D-galactose as the sole carbon source (Gross and Schnarrenberger 1995a) demonstrates the importance of GALT, and the other Leloir pathway enzymes, in red algae (in particular, those capable of heterotrophy) under heterotrophic conditions. GALT in red alge may also participate in salvaging D-galactose derived from the degradation of certain metabolites. D-galactose is a constituent of a variety of molecules such as glycoproteins and low-MW carbohydrates (*e.g.*, isofloridoside).

Interestingly, the presence of GALT in plants is yet to be confirmed. GALT activity, claimed to be detected in soybean cotyledon (Pazur and Shadaksharaswamy 1961; but see Gross and Schnarrenberger 1995b), was not detected in sugarcane cell cultures (Maretzki and Thom 1978) or cucumber (Gross and Phar 1982). Moreover, neither proteins nor genes of GALT have so far been purified or isolated from plants. In contrast, UDPGal pyrophosphorylase activity has been demonstrated for several plant species (see Dey 1985), raising the possibility that plants mainly use the UDPGal pyrophosphorylase pathway to metabolize D-galactose. Apparently, the red algae and plants have adopted different pathways for utilization of free D-galactose.

The biochemistry and enzymology of GALT from humans and eubacteria, particularly that from *E. coli* (reviewed by Frey *et al.* 1982, Frey 1996), have been well characterized. The enzyme has been shown to be a metalloprotein and to follow the Ping-Pong Bi Bi kinetic mechanism. Some mutations in this enzyme in humans that impair its catalytic activity are known to cause the metabolic disorder galactosemia. The crystal structure of the enzyme from *E. coli* complexed with uridine 5'-diphosphate has been determined at 1.8 Å resolution (Wedekind *et al.* 1995).

In this paper, we characterize a GALT gene from the multicellular agaran-producing marine red alga *G. gracilis* as part of our effort to study the molecular genetics of carbohydrate, particularly galactose, metabolism in red algae.

Materials and methods

Genomic library screening

A *G. gracilis* genomic library constructed previously (Zhou and Ragan 1994) was screened for clones containing GALT genes using standard techniques (Sambrook *et al.* 1989). The probe was prepared from a clone isolated from a *G. gracilis* cDNA library as part of an EST project (Chapter V; Lluisma and Ragan, accepted for publication); the cDNA clone encodes a fragment of a putative GALT gene, identified based on a significant similarity (BLASTX score 322) of the inferred amino acid sequence of the insert with GALT sequences in the NCBI database.

Sequencing

DNA from genomic clones were prepared from phage lysates using the QIAGEN Lambda Mini Kit (QIAGEN Inc., Chatsworth, CA), and used directly for sequencing. Sequences were obtained by walking from the region of known sequence initially determined from the GALT EST (see above). When necessary, portions of the clone were PCR-amplified using *Pfu* polymerase (Stratagene, La Jolla, CA) to produce more templates for sequencing. Both strands were sequenced on an Applied Biosystems 373A sequencer (Applied Biosystems, Foster City, CA) following the manufacturer's Dye Deoxy terminator cycle sequencing protocol. The raw sequence data (the 'traces' or chromatograms) from both strands were compared and edited using the computer programs pregap and gap4, parts of the Staden package (Staden 1996).

Southern analysis

DNA was extracted from *G. gracilis* ("grass" strain), obtained from stocks cultured at the NRC-Institute for Marine Biosciences - Aquaculture Research Station, Sandy Cove, Halifax County, Nova Scotia; this strain, originally collected in Namibia, has been verified as a strain of *G. gracilis* (=*verrucosa*) by analysis of the ssu-rRNA (see Bird *et al.* 1994). Vegetative thalli were first cleaned with filtered seawater, then clean tips were isolated, frozen with liquid N_2, and stored at -70°C for 2 weeks prior to extraction of DNA. DNA was extracted essentially as described (Zhou and Ragan 1993), digested with restriction enzymes, subjected to electrophoresis on a 0.7% agarose gel, blotted on Zeta-Probe GT membranes (Bio-Rad Laboratories, Hercules, CA) following the manufacturer's protocol, and subjected to Southern hybridization with *GALT* probe essentially following standard techniques (Sambrook *et al.* 1989). The probe was labelled with α-^{32}P-dCTP using the Random Primed DNA labelling kit (Boehringer Mannheim, Mannheim, Germany).

RT-PCR

mRNA was extracted from 2 g fresh tissue of laboratory-grown *G. gracilis* using the Invitrogen FastTrack 2.0 kit (Invitrogen Corp., San Diego, CA). Reverse transcription and PCR amplification of the 5' end was performed using the Boehringer Mannheim 5'/3' RACE kit (Boehringer Mannheim, Mannheim, Germany). The gene-specific primer r2 (GTTCCTTAGCTTCCATCTCGATC, positions 652-630 in Fig. 1) was used for reverse transcription. A different primer (r2b: TGTCAAACACGAAAGTGGTATCG, positions 235-213 Fig. 1) was used as the gene-specific primer for the PCR amplification. The PCR-amplified product was purified and used directly for sequencing, with oligonucleotide r2e (CACGAAAGTGGTATCGTACTTGG, see Fig. 1) as primer.

Sequence analysis

GALT amino acid sequences were retrieved from the NCBI protein database and aligned together with the inferred amino acid sequence of GALT from *G. gracilis* using CLUSTAL W (Thompson *et al.* 1994) under its default parameters (pairwise alignments=slow (accurate), gap opening penalty=10, gap extension penalty=0.10 or 0.05) and the BLOSUM series (Henikoff and Henikoff 1992) for scoring.

Results and discussion

Cloning and sequencing of G. gracilis *GALT gene*

Using an EST of a GALT gene from *G. gracilis* (Chapter V; Lluisma and Ragan, accepted for publication) as a probe, we were able to obtain two clones (designated λGalTa and λGalTb), containing identical copies of a GALT gene, from a *G. gracilis* genomic library. In both clones, the GALT gene (*i.e.*, its 5' end) is situated within 600 bp from one end of the insert. One clone (λGalTa) was sequenced starting from the region of known sequence (the portion covered by the EST), and extended in both directions by primer walking. The nucleotide sequence of the protein-coding and flanking regions is shown in Figure 1. The gene encodes a putative GALT protein of 369 amino acids, which shows a high degree of sequence similarity with GALT sequences from other organisms (Fig. 2; see below). Since this gene appears to be one of two GALT genes in *G. gracilis* (see *Southern analysis* section, below), we designate it as *GgGALT1*.

Figure 1. The nucleotide sequence of *GgGALT1*. The numbering scheme assigns +1 to the first position of the start codon (marked with 1). Potential cis-regulatory elements are underlined. The potential TATA box is in bold and underlined. The sequence of the 5' end of cDNA produced by 5'RACE (shown in lowercase letters) is aligned with the genomic sequence, with the site of the sequencing primer r2e underlined. The conceptual translation is shown below the coding region; * marks the stop codon. The putative polyadenylation signal (TGTAAA) downstream of the protein coding region is underlined; the residue immediately upstream of the polyA site is in bold and underlined. The region downstream of *GgGALT1* whose complementary strand contains an ORF potentially encoding a peptidyl-tRNA hydrolase is shown in lowercase letters, with site of the start and stop codons (on the complementary strand) in bold.

```
CATCGATGATCTCTTCCCGCAGTCGAGTGAGCTGCTGGGCCTCACGACGAGCCAGGTAGTTGCGACGCGACT  -473
CCTTTCGCAGGTCATCTACGGACGGTTGTGACGGGGGCTTAGACGGCGTGGACATGGCGACAAGGACGAAGA  -401
CGCGCTGTAGGAACGCGAGAGACAGATGTGAGCGGAGAAATGACTCGGCGTCCGGGAAGAAATTGGAAGCGG  -329
CGACGTTGTCGGACGTGCGTTGCTGACGCGCAACGCCTGTGGCACGCGCAGCAGATTTGTGCAGATTTTGCG  -257
CAGATTTCGGAGCACGACGATACGGCGGCGCGGCGGCAGCCAATCACAGACGCGCGCCACGCATGCCGATTC  -185
CGCTTCGTCGCCACCACGTCATGTTCGCCAATGCACTCAGCGATCAATCGCGTCGCGGACAGCCTGTCCGAC  -113
CCTCAATCGAAACCATTAGTTAAGCGCAGCGGGCGAAGCGGGCTTCGCATCGTCAGCAGACCTCAGCGTCGG   -41
                                                                       g
                                 1
CCTGCGATCGTCTTCTTGTCGTATCGACCCTCCTCAAACAATGTCTGCCAGCTTCGACTACACCGAGCACCC    32
cctgcgatcgtcttcttgtcgtatcgaccctcctcaaacaatgtctgccagcttcgantacaccgagcaccc
                                         M  S  A  S  F  D  Y  T  E  H  P
CCATCGTCGCTACAACCCGCTCTCAGCGCGCTGGATTCTGTGCTCACCCCATCGCGCGAAGCGTCCGTGGCA   104
ccatcgtcgttacaacccgctctcagcgcgctggattctgtgctcaccccatcgcgcgaagcgtccgtggca
 H  R  R  Y  N  P  L  S  A  R  W  I  L  C  S  P  H  R  A  K  R  P  W  Q
GGGCAGCGTTGAAGATCTGCCGCCTGATGAGCGTCCTGAGTACGATCCCAAGGACTACTTGGGGCCGGGAAA   176
gggcagcgttgaagatctaccgcctgatgagcgtcctgagtacgatccgaaggactacttggggccgggaaa
 G  S  V  E  D  L  P  P  D  E  R  P  E  Y  D  P  K  D  Y  L  G  P  G  N
CTTTCGTGTGAACGGCTCCGTGCAGAACCCCAAGTACGATACCACTTTCGTGTTTGACAATGACTTCCAGGC   248
ctttcgtgtgaacggc             <---- primer r2e site
 F  R  V  N  G  S  V  Q  N  P  K  Y  D  T  T  F  V  F  D  N  D  F  Q  A
TTTGCTGGATAACACGCCGCACGGCGAAGTTGGCAGCGTAGAAGACAATGATCTGCTCGTCGCAAAAGCTGT   320
 L  L  D  N  T  P  H  G  E  V  G  S  V  E  D  N  D  L  L  V  A  K  A  V
GCGTGGAAAGTGTCGCGTCGTCTGCTTCTCTCCGAAGCTCAATCTCACAGTCGCCGAGATGACAGTCGAAGA   392
 R  G  K  C  R  V  V  C  F  S  P  K  L  N  L  T  V  A  E  M  T  V  E  E
AATCAAACATGTCGTCGATGCCTGGCTTGAGGAATATGACACCCTCTCCAAGTTGGACTATATCGGCCATGT   464
 I  K  H  V  V  D  A  W  L  E  E  Y  D  T  L  S  K  L  D  Y  I  G  H  V
GCAGATCTTCGAGAACAAGGGGCAGATGATGGGATGCTCAAACCCACATCCTCACGGCCAAATCTGGGCTTC   536
 Q  I  F  E  N  K  G  Q  M  M  G  C  S  N  P  H  P  H  G  Q  I  W  A  S
GGAGTTTGTACCGGAGGAACCTCGCATTGTATTGGAGAACCTCAAGGCGTATCATGAAAAGAAGGGTACCCA   608
 E  F  V  P  E  E  P  R  I  V  L  E  N  L  K  A  Y  H  E  K  K  G  T  H
TATGTTAGAAGACTATGTCAAGATCGAGATGGAAGCTAAGGAACGAATCGTCTGTGAAAATGACACCTTCCT   680
 M  L  E  D  Y  V  K  I  E  M  E  A  K  E  R  I  V  C  E  N  D  T  F  L
TGCCGTCGTGCCCTTCTGGGCAACCTGGCCATTTGAGGTGTTGGTTATGACCAAGAAACGCGTACCATGCCT   752
 A  V  V  P  F  W  A  T  W  P  F  E  V  L  V  M  T  K  K  R  V  P  C  L
CAAGAACTTCAACGAAGCCATGAAGAGCGATCTCGCCGATATCTATCGTCGTGTTTGTGCACGATATGACAA   824
 K  N  F  N  E  A  M  K  S  D  L  A  D  I  Y  R  R  V  C  A  R  Y  D  N
CCTGTTCACCACCCTTTTCCCTTACTCTATGGGTATTCATCAGTCACCTACTACCAATGGTACTGATCCCGC   896
 L  F  T  T  L  F  P  Y  S  M  G  I  H  Q  S  P  T  T  N  G  T  D  P  A
GAAGCATGATTATGCGCACTTTCACATGCACTTCTATCCGCCTCTGCTCCGAAGTGCGACAGTTAGGAAATT   968
 K  H  D  Y  A  H  F  H  M  H  F  Y  P  P  L  L  R  S  A  T  V  R  K  F
CATGGTTGGCTTCGAGATGCTTGGCGAATCACAAAGAGATTTGACCGCAGAGCAAGCCGCGGCGAGACTGCG  1040
 M  V  G  F  E  M  L  G  E  S  Q  R  D  L  T  A  E  Q  A  A  A  R  L  R
TGCCTGCTCTGAAGTGCATTATAACCATGCCCGAAAGAAGGTTGAAGGCAATGGTGACGCTTCTAAGTGAGA  1112
 A  C  S  E  V  H  Y  N  H  A  R  K  K  V  E  G  N  G  D  A  S  K  *
AAGCTGGCGTTGTGCCGCACTTGGGCCCAAGAGTTCTTGAACACCCTGTTCCACAGGCATAGAAACTTTTAA  1184
GCTTAGTCGTGTAAAGACATCCCTTTCGTATTCGTATTCATTGAGCAATGTGATGAGCGATGCTGAGAACGA  1256
CATTGTGCTGGTCTCGTGGCGATAACCATCATGttatcgcttctgtttcgttaacgcgatctggttctgtac  1328
tttttctaaactgggttcctgcacccagaagtcaatggcttccatcacatcccactctacttcttccataat  1400
cttcctatctcctctcgaaaacggcgcgaggacgaaatcagcccactcttgagcaccaagtctcggggaacc  1472
aacgccaacacgcagtctggcataactcattccaccaacgttctgctgaatgctcttcaatccattatgccc  1544
tccagcacttcccttggcgcgcaacctaaggtctccaattggcagggacatatcgtccactactacaagcat  1616
tgcggccacaggggcgtttaaatacttgagagctgcccgcacagctcgaccggatgcattcataaaggttgt  1688
tggcttgagcagcatgatttcttgttgtatagatggaaccgggtaatctctccctgtacactgttttcctt  1760
tctgaactttccagtagcatggcggacggcatactgatctagtaagctgaaacctacattatgcctggtgtt  1832
ttcaaaacgggagcctgggtttcctaagcccacgacaaggagcgtatcaccgcgttctgggggtttcggtcg  1904
accattgggggttttcggtggtggggacggctgtttcgactcactggatgaactgggagtgcctgtagttga  1976
gcatcgcatCACCCTATGCCC
                                                                          1997
```

Figure 1

Figure 2. Alignment of galactose-1-phosphate uridylyltransferase (GALT) sequences. Residues shown in white on a black background are conserved in at least 60% of the sequences; white on grey, 40%. KmarxGALT, *Kluyveromyces marxianus* var *lactis* GALT (NCBI accession number 67063); yeastGALT, *Saccharomyces. cerevisiae* GALT (=GAL7; 625224); mouseGALT, *Mus musculus* GALT (1730188, conflicts with another *M. musculus* GALT sequence occur in this sequence); ratGALT, *Rattus norvegicus* GALT (1083825); humanGALT, *Homo sapiens* GALT (345849); CelegGALT, *Caenorhabditis elegans* GALT (521063); StyphGALT, *Salmonella typhimurium* GALT (120912); EcoliGALT, *Escherichia coli* GALT (120907); HinfGALT, *Haemophilus influenzae* GALT (1169825); FneoGALT, *Filobasidiella* (=*Cryptococcus*) *neoformans* (576777); BfibGALT, *Butyrivibrio fibrisolvens* (1169824); SlivGALT, *Streptomyces lividans* GALT (1169827).

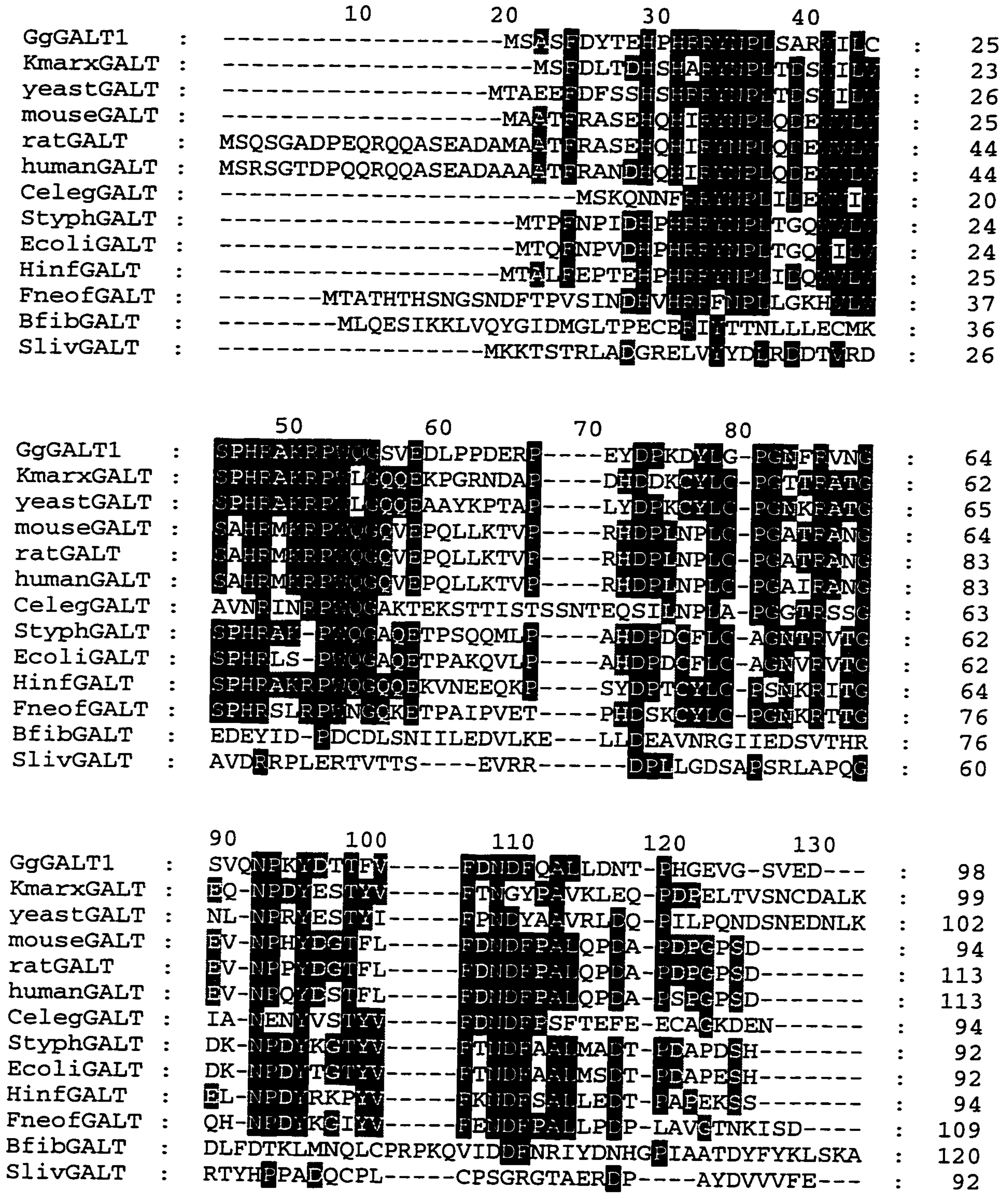
10 20 30 40
50 60 70 80
90 100 110 120 130
GgGALT1
KmarxGALT
yeastGALT
mouseGALT
ratGALT
humanGALT
CelegGALT
StyphGALT
EcoliGALT
HinfGALT
FneofGALT
BfibGALT
SlivGALT
25 23 26 25 44 44 20 24 24 25 37 36 26
64 62 65 64 83 83 63 62 62 64 76 76 60
98 99 102 94 113 113 94 92 92 94 109 120 92

Figure 2

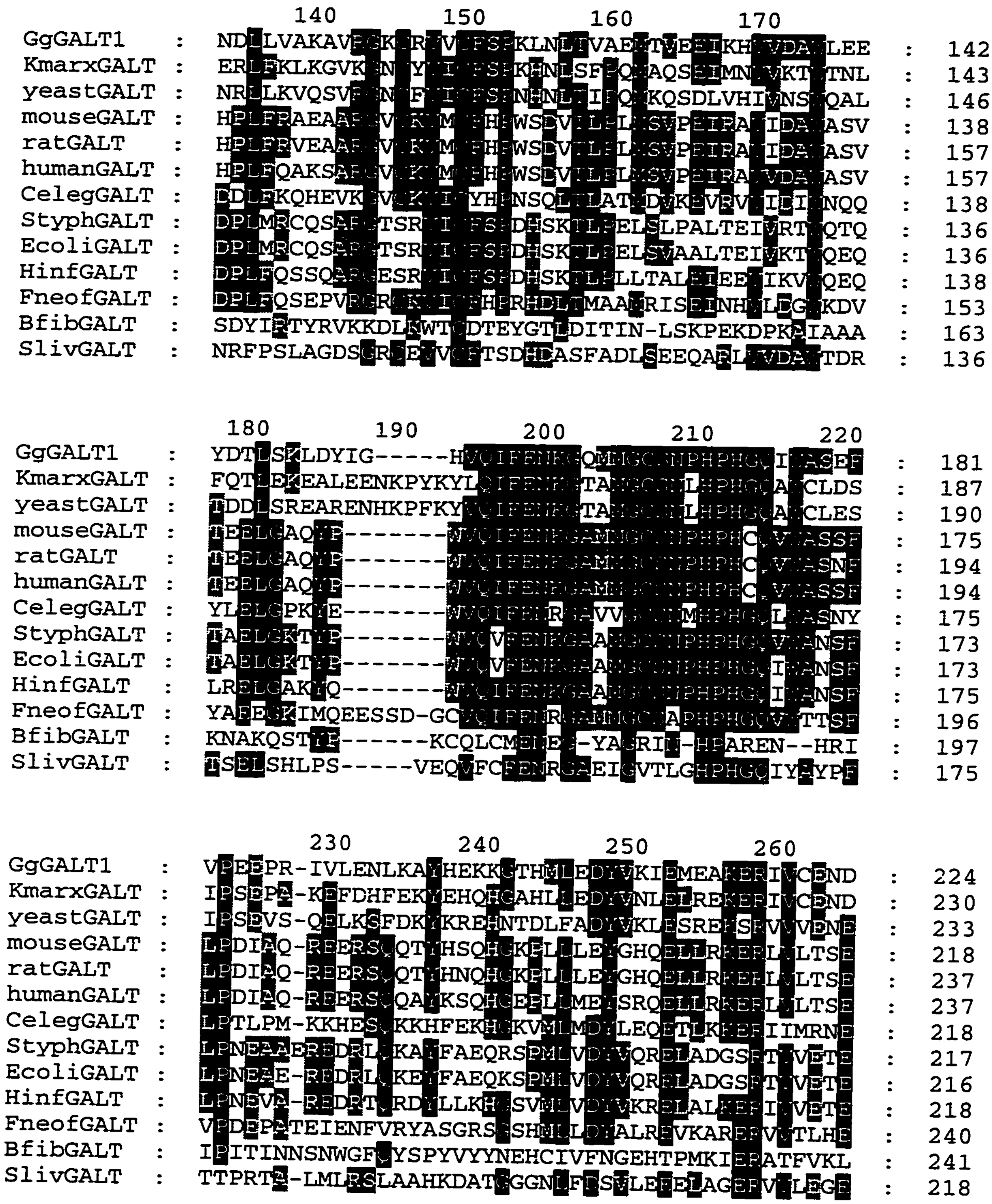

Figure 2, continued

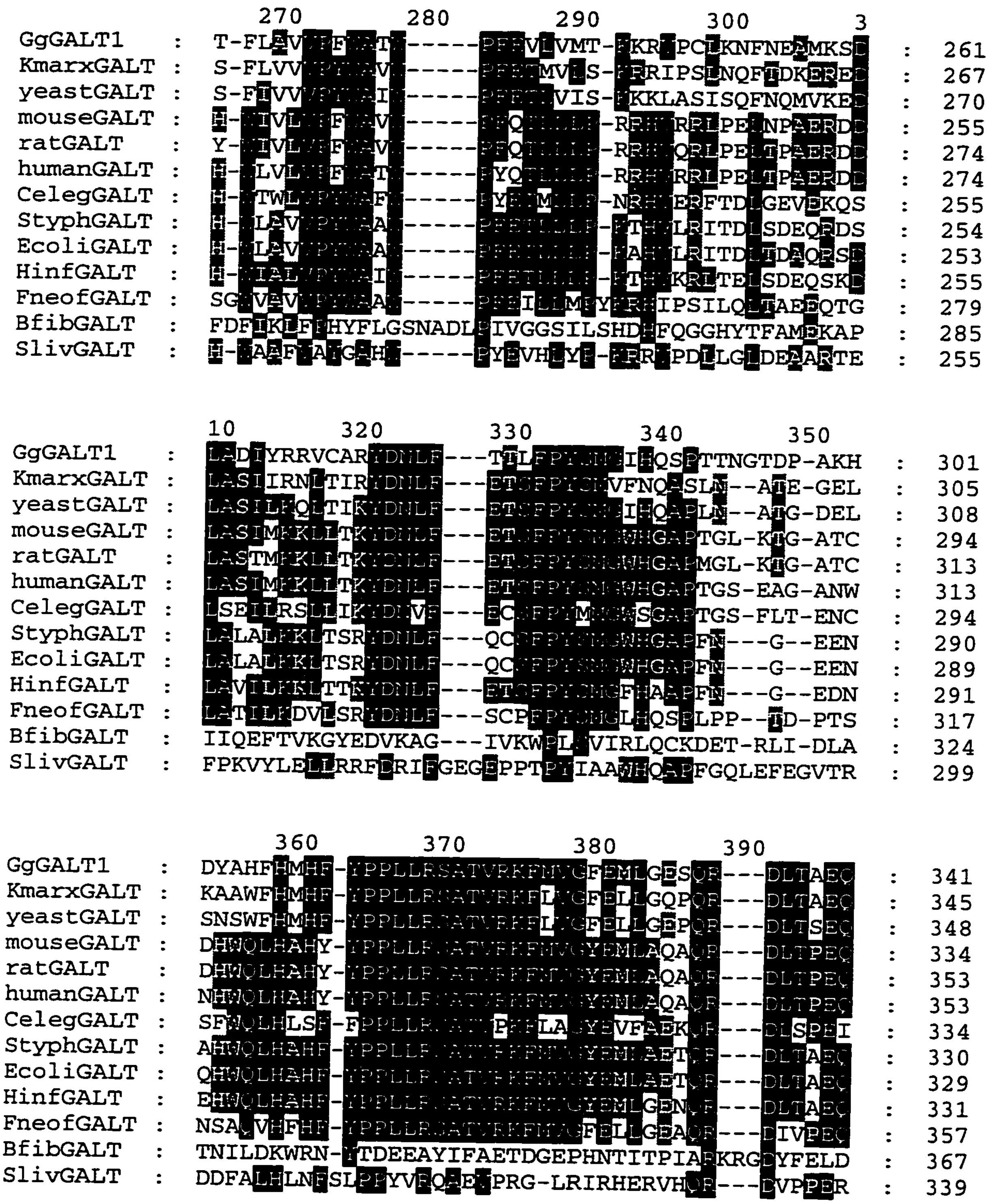

Figure 2, continued

```
               400        410        420
GgGALT1    :  AAARLPACS-EVHYNHARKKVEGNGDASK  :  369
KmarxGALT  :  AADELKALSGEVHYLAKLEQEESKT----  :  370
yeastGALT  :  AAEKLPNLDGQIHYLQRL-----------  :  366
mouseGALT  :  AAERLPALP-EVHYCLAQKDKETAAIA--  :  360
ratGALT    :  AAERLPVLP-EVHYCLTQKDKETAATA--  :  379
humanGALT  :  AAERLPALP-EVHYHLGQNDRETATIA--  :  379
CelegGALT  :  AAKTLSEID-EEHYSKKLQ----------  :  352
StyphGALT  :  AAERLPAVS-DIHFRESGV----------  :  348
EcoliGALT  :  AAERLPAVS-DIHFRESGV----------  :  347
HinfGALT   :  AAERLPALS-EVHYKERTK----------  :  349
FneofGALT  :  AAVELPESLPHKRATLSNDKPYNP-----  :  381
BfibGALT   :  PLESTCRHA-SLALAVVLQRRDW------  :  389
SlivGALT   :  AAERLPEVA-DVHERE-------------  :  354
```

Figure 2, continued

The size of the putative GgGALT1 protein (369 amino acids) falls within the range of sizes (ca 350 to 390 amino acids) of GALT proteins from other organisms (Fig. 2). A GALT enzyme purified from the primitive red alga *G. sulphuraria* is apparently a homodimer, and has an apparent subunit size of 42 kD, roughly equivalent to 378 amino acids (Gross and Schnarrenberger 1995b).

Features of the gene

To determine the potential transcription initiation site of *GgGALT1*, we obtained the sequence of the 5' portion of its transcript by 5' RACE; the sequence is shown in Figure 1. The results suggest that the transcription initiation site is at (or very close to) position -41 in Figure 1. Whether the sequence around this site is similar to corresponding sequences in other red algal genes is not clear, as the consensus sequence for transcription initiation site in red algal (particularly *G. gracilis*) genes remains to be clearly established. Preliminary data indicate that it may not be similar to its counterparts in plants, animals and yeast (Zhou and Ragan 1994).

A potential TATA box, ATTAGTTAA, is present less than 60 bp upstream of the putative transcription initiation site (Fig. 1). We obtained the sequence using lambda clones as templates and verified the G residue that appears to interrupt the putative TATA box. Whether or not this putative TATA box is functional *in vivo* remains to be verified by biochemical experiments. Four CAAT motifs and three ACGT elements are also present upstream of the putative TATA box (Fig. 1).

The translation initiation codon appears to be situated (at least) 41 bp downstream of the transcription initiation site; this is the first AUG downstream of the putative 5' end of the transcript (the next AUG is situated more than 120 bp downstream of this codon). This codon appears to be in the right context (AAACAAUG, the start codon is underlined) as site of translation initiation. The residue at position -2, C, is the same as the canonical residue at the corresponding position in transcripts of red algal genes; the canonical sequence at the translation initiation sites in red algal transcripts appears to be YYCRCYAUG (Zhou and Ragan 1996), while that of plants is UAAACAAUG (Joshi 1987). Although the codon and its 5'-flanking sequence on the *GgGALT1* transcript actually resembles perfectly the plant canonical sequence, it does not depart considerably from red algal consensus sequence. The residues at positions -1 (A) and -3 (A), while different from the apparent red algal canonical sequence, have been found in some red algal genes (Zhou and Ragan 1996).

The *GgGALT1* gene appears to be devoid of introns. Comparison of the amino acid sequence deduced from the *GgGALT1* ORF with GALT sequences in the database suggests that no intron exists in the *GgGALT1* gene. In addition, comparison of the cDNA (obtained by 5' RACE) and genomic sequences indicates that no intron exists in the 5' portion of the *GgGALT1* gene (Fig. 1). GALT genes or ESTs have not yet been characterized from plants. GALT genes have already been cloned from *S. cerevisiae* (Tajima *et al.* 1985), rat (Heidenreich et al 1993), and human (Leslie *et al.* 1992). The yeast gene is intronless, whereas rat and human GALT genes each contain ten introns. The lack of intron in *GgGALT1* is consistent with the observation that red algal genes are intron-poor; most genes reported from red algae so far either have a single intron (for example, genes for GapA, GapC, and mitochondrial aconitase in *G. gracilis* [as *G. verrucosa*: Zhou and Ragan 1994, 1995a,b] and a β-tubulin in *Chondrus crispus* [Liaud *et al.* 1995]) or no intron at all (for example, genes for triose phosphate isomerase in *G. gracilis* [Zhou and Ragan 1995b], and GapC in *C. crispus* [Liaud *et al.* 1993]) although the corresponding genes in other organisms contain multiple introns.

Inspection of the 3' untranslated region of the cDNA and the corresponding region in the genomic sequence reveals the presence of the putative polyA signal UGUAAA, situated 41 nucleotides downstream of the termination codon and 20 bp upstream of the polyA site; the polyA site was determined from the comparison of the sequence of the cDNA obtained by EST sequencing (Chapter V; Lluisma and Ragan, accepted for publication) and that of *GgGALT1*. No other potential polyA signal can be found. It has become evident that UAAA is part of a highly conserved polyadenylation signal in red algae. Initial analysis of 3' untranslated regions of red algal genes by Zhou and Ragan (1996), and inspection of the sequences of other red algal genes, *e.g.*, genes encoding polyubiquitin and γ-subunit of R-phycoerythrin in *Aglaothamnion neglectum* (Apt and Grossman 1992, Apt *et al.* 1993), mitochondrial aconitase in *G. gracilis* (as *G. verrucosa*: Zhou and Ragan 1995a), and β-tubulin in *C. crispus* (Liaud *et al.* 1995), indicate that, in a majority of red algal genes, the UAAA motif is found close to and upstream of polyadenylation sites, and thus likely serves as part of a polyA signal (or positioning element). Whereas in animals AAUAAA is a highly conserved polyA positioning element, in plants and yeast AAUAAA-like elements are degenerate and may even be absent (see reviews by Guo and Sherman 1996, Hunt 1994, Wu *et al.* 1995).

The polyA sites in red algae have been observed to occur immediately 5' to an A residue, similar to those in vertebrates and plants (Zhou and Ragan 1996). The polyA site on the *GgGALT1* transcript (Fig. 1) conforms to this rule (Fig. 1), although we have not verified whether the actual cleavage site is immediately 5' or 3' of A-1214.

Southern analysis

Southern analysis (Fig. 3) revealed that two related GALT genes exist in *G. gracilis*, as indicated by two bands present on each lane after stringent washing (0.2X SSC/0.1%SDS, 30 min at 65°C, performed twice). It appears that the two genes may not be identical copies, as the intensity of hybridization seems to differ. The presence of two GALT genes in *G. gracilis*, and possibly in

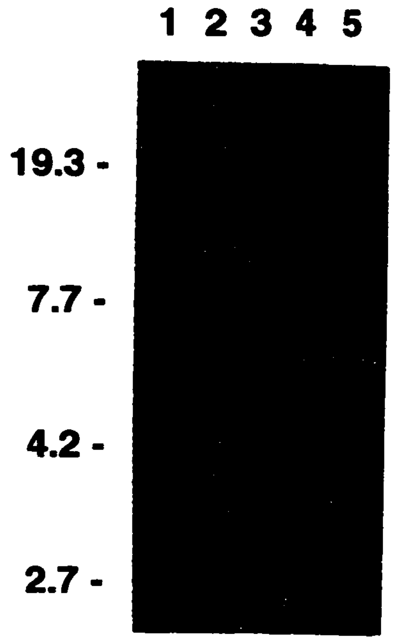

Figure 3. Determination of *GgGALT1* copy number by Southern hybridization analysis. Genomic DNA (5 μg per reaction) from *G. gracilis* was digested with *Eco*R I (lane 1), *Hind* III (lane 2), *Sal* I (lane 3), *Xba* I (lane 4), and *Xho* I (lane 5). The probe was prepared from a *GgGALT1* fragment (position 571 to position 1182, Fig. 1; no sites for the five restriction enzymes used in digesting the genomic DNA occur within this region). Final washing was with 0.2X SSC/0.1% SDS at 65°C for 30 min, performed twice. The numbers on the left of the figure indicate the size of the markers, in kb.

other red algae, may be an indication of the importance of GALT to red algal physiology. Obviously, it would be interesting to know whether the two genes have different physiological roles. Sequencing of the other gene and comparing their expression patterns would be the first steps to answering this question, and the *GgGALT1* probe clearly can be used to isolate the second gene. It remains to be seen if other red algae also have two GALT genes; only a single isoform was detected by isolation of a GALT enzyme from the red alga *G. sulphuraria* (Gross and Schnarrenberger 1995b).

Features of the inferred protein

GALT protein sequences have already been inferred (from sequenced genes) for a number of organisms. An alignment of some sequences with the inferred GgGALT1 is shown in Figure 2. Based on the alignment, the GgGALT1 protein is roughly as similar in sequence to GALTs in fungi (49% sequence identity with *K. marxianus*, 49% with *S. cereviasiae*) and animals (48% with mouse, 47% with human) as it is to eubacterial GALTs (*E. coli*, 45%; *H. influenzae*, 50%). Phylogenetic analysis of the sequences (data not shown) suggests, however, that the GgGALT1 protein sequence is more related to fungal than to animal or eubacterial GALTs.

Highly conserved residues are highlighted in the alignment, a number of which have already been shown to be important to the enzyme's structure and function. The GgGALT1 residues corresponding to the consensus sequence at the active site, GCSNPHPHGQ (Wedekind *et al.* 1996), are a perfect match (positions 205-214 in the alignment). H212 is the nucleophilic catalyst that is transiently nucleotidylated (Field *et al.* 1989), and Q214, whose mutation to R is the predominant cause of galactosemia among the Caucasian population (Elsas *et al.* 1994), helps in stabilizing the nucleotidyl intermediate (Wedekind *et al.* 1996). Three residues involved in binding iron, namely, H358, H359, H338, and E229 (numbers refer to the position in the alignment in Figure 2) (Wedekind *et al.* 1995), which are conserved in many GALTs, are conserved in GgGALT1: E229 (an acidic residue) is replaced by V (a hydrophobic residue). The binding of iron and/or zinc is required by the *E. coli* GALT for catalytic activity (Ruzicka *et al.* 1995).

Interestingly, three of the four residues (C76, C79, H154, and H210) involved in binding zinc (Wedekind *et al.* 1995), conserved in fungal and bacterial but not in mammalian GALTs, are also non-conservatively substituted in the GgGALT1 protein (C76D, C79G, H154L); the substitution pattern between GgGALT1 and mammalian GALTs is apparently different, however. Wedekind *et al.* (1995) have already noted that nonconservation of zinc-binding residues in mammalian GALTs may indicate that animal GALTs lack a metal binding site analogous to that in yeast and bacterial GALTs. Using the same argument, GgGALT1 may also lack an analogous zinc-binding site. Zinc-binding is important for ensuring that the active site is in the right conformation (Wedekind *et al.* 1995); H210, one of the H residues that bind zinc, is part of the active site.

It will be interesting to determine (*e.g.*, by crystallography or structural modelling) what alternative mechanisms are employed by red algal and animal GALTs to ensure the right conformation of the active site.

A gene occurs downstream of GgGALT1

Sequencing the region downstream of *GgGALT1* revealed an ORF of 232 codons encoded on the opposite strand (Fig. 1); the stop codon for this ORF is located 179 bp downstream of *GgGALT1* stop codon. The deduced amino acid sequence of this ORF shows a BLASTX score of 225 against a peptidyl tRNA hydrolase (PTH) from *Bacillus subtilis*, with 43% identity over 178 amino acids (of 188) of the *B. subtilis* PTH sequence.

The 179-bp separation between *GgGALT1* and the *pth*-like ORF is the smallest yet observed between any two red algal nuclear genes; approximately 1.5 kbp separates the *G. gracilis* genes encoding polyubiquitin and mitochondrial-type aconitase (Zhou and Ragan 1995a), while 376 bp separate the *G. gracilis* genes for UDPglucose pyrophosphorylase (UGPase) and a DNA helicase (Chapter V). Such proximity is most remarkable for the nuclear genome of a eukaryote, where (to the extent this has been characterized) genes are typically many kilobases apart. More data will be required to determine whether this proximity is characteristic of *Gracilaria*, or red algal, nuclear genomes.

Conclusion

We have cloned a GALT gene from the red alga *G. gracilis*, apparently one of two related genes in the genome of this alga. It remains to be seen if the second gene is expressed *(i.e.*, functional), and, if functional, how the two genes may differ in expression and catalytic properties. The functional significance of the structural features observed in *GgGALT1* gene and its product also remains to be verified by biochemical experiments. Specifically, the hypothesis that the GgGALT1 enzyme does not require zinc for catalytic activity, as indicated by the structural analysis, needs to be investigated. The observation that another gene, potentially encoding a peptidyl tRNA hydrolase, is located just downstream of *GgGALT1* indicates that occurrence of closely-spaced genes may not be unusual in the *G. gracilis* genome.

CHAPTER IV

A starch-branching enzyme gene from the marine red alga *Gracilaria gracilis*

Introduction

The red algae utilize α-1,4-glucans, known as floridean starch, as reserve polysaccharides (Raven *et al.* 1990). A number of characteristics distinguish the biosynthesis and structure of floridean starch from those of starch (glycogen) in other organisms, particularly green plants. Floridean starch in red algae is synthesized not in plastids but in the cytosol (Pueschel 1990). The properties of floridean starch, such as average unit chain length, average internal chain length, affinity for iodine, and limiting viscosity number, are also intermediate between animal glycogen and plant amylopectin, although they are somewhat more similar to those of amylopectin (Percival and McDowell 1967, Aspinall 1970). The floridean starch granule, compared to that of plant starch, shows greater variability in shape; in addition to the more common spherical shape, ovoid, obovoid, cylindrical, and pyriform shapes have also been observed (Sheath *et al.* 1981). In addition, whereas plant starch consists of both amylose and amylopectin, starch in most red algae consists solely of amylopectin-like material (*i.e.*, floridean starch) (Craigie 1974, Percival and McDowell 1967), although amylose fractions were observed in floridean starch in some species of red algae (McCracken and Cain 1981). These differences indicate that the physiology, molecular biology and evolutionary history of floridean starch biosynthesis may differ in major respects from those of starch or glycogen in other organisms.

One of the key enzymes in starch biosynthesis is the starch branching enzyme (SBE; 1,4-α-D-glucan:1,4-α-D-glucan 6-α-D-(1,4-α-D-glucano)-transferase, EC 2.4.1.18), responsible for the formation of amylopectin from α-1,4-D-glucan chains. Early enzymological studies of plant SBEs led to the recognition of two types of SBEs according to their substrate preference: the Q-enzymes, which can only use amylose as substrate and produce only moderately branched amylopectins, and the amylopectin-branching enzymes, which can act on both amylose and amylopectin, although the distinction may not always be clear-cut (Manners 1985). Recent cloning and characterization of SBE genes from plants has led to the recognition of two families of plant SBEs (Burton *et al.* 1995, Commission on Plant Gene Nomenclature 1994), the *Sbe1* family (or family B), and the *Sbe2* family (or family A); isoforms not clearly belonging to either family may exist. These families apparently became established before the divergence of the monocots and dicots, as isoforms of both families occur in the two plant groups. Members of these two families differ in enzymological properties (Martin and Smith 1995, Guan *et al.* 1997).

Biochemical studies in the late 1960's by Fredrick (1968, 1971) showed that three SBE isozymes occur in the red alga *Rhodymenia pertusa*, two of which were found to be Q-type enzymes, while the third was found to be a "dual action" enzyme, capable of branching both amylose and amylopectin. Biochemical work on red algal SBEs has been in the doldrums since the 1970s. The biochemistry and enzymology of this enzyme in red algae has remained poorly known, and much work remains to be done to understand its role in floridean starch biosynthesis, especially at the molecular level. The evolutionary relationship of red algal SBEs with those of plants has also remained to be clarified. In this paper, we report the cloning and characterization of an SBE gene from the marine red alga *G. gracilis* as a first step to elucidate the molecular biology of floridean starch biosynthesis. We confirm the presence of other potential SBE genes in the genome of *G. gracilis*, and briefly touch on the evolutionary implications of our findings.

Materials and Methods

Construction of a G. gracilis *genomic library*

DNA was extracted from *G. gracilis* ("grass" strain) as previously described (Chapter III), partially digested with *Sau*3A I, and ligated to the Lambda-DASH II vector (Stratagene, La Jolla, CA) using the manufacturer's protocol. The recombinant phage was packaged using the Gigapack III Gold Packaging Extract (Stratagene, La Jolla, CA). The library was constructed with *E. coli* XL1-Blue MRA (P2) strain (Stratagene, La Jolla, CA) as host.

Isolation of genomic clones

The strategy we used to clone an SBE gene from *G. gracilis* relied on the use of PCR to generate a homologous probe. Degenerate PCR primers were designed based on the conserved residues of SBE as revealed by multiple alignments of amino acid sequences of SBEs obtained from the NCBI protein database. Although several primers were tested, only one pair gave a desired product. PCR reactions using the "forward" primer Fb (TA[TC]GCNGA[GA][AT][GC][GCT]CA[TC]GA[TC]CA, which corresponds to the conserved protein region YAESHDQ), and the "reverse" primer Rb ([TC]TCNGG[GA]TGNCC[GA]AA[TC]TC[GA]TT, which corresponds to the conserved region NEFGHPE), with *G. gracilis* genomic DNA as template, yielded a product of about 220 bp; subsequent sequencing confirmed that the product is a fragment of a probable SBE gene. This PCR product was used as probe to screen the genomic library of *G. gracilis* using standard protocols (Sambrook *et al.* 1989). Two clones were isolated and one clone was sequenced.

Sequencing

Lambda DNA was isolated and purified using the protocol described by Sambrook *et al.* (1989) and directly used for sequencing; both strands were sequenced. For the the initial sequencing, primers were designed based on the known sequence of the previously sequenced fragment; sequencing in both directions proceeded by "primer walking". When necessary, fragments of the gene were PCR-amplified using *Pfu* polymerase (Stratagene, La Jolla, CA) to generate more template for sequencing. Sequencing was carried out on an Applied Biosystems 373A sequencer (Applied Biosystems, Foster City, CA) following the manufacturer's Dye Deoxy terminator cycle-sequencing protocol. The raw sequence data was edited using the computer programs pregap and gap4 of the Staden package (Staden 1996).

Southern blotting

The DNA and procedures used were as described in Chapter III. A fragment of the SBE clone (from position 759 to position 1504, Fig 1) was PCR-amplified and used as template for synthesis of a probe. The probe was synthesized and labelled with α-^{32}P-dCTP using the Random Primed DNA labelling kit (Boehringer Mannheim, Mannheim, Germany).

Mapping of the 5' and 3' ends of the SBE mRNA

mRNA extraction and 5' mapping of the mRNA using 5' RACE were as described (Chapter III). To map the 3' end, we reverse-transcribed mRNA from *G. gracilis* using the Pharmacia T-Primed First-Strand Ready-To-Go kit (Pharmacia Biotech, Uppsala, Sweden). Aliquots of the reaction were then used in PCR, using the primer f2 (Fig. 1) as the gene-specific primer, and an anchor primer, ATTCGCGGCCGCAGGAATT; the primers were synthesized in-house on an Expedite Nucleic Acid Synthesis System (Millipore, Bedford, MA). The PCR product was desalted by centrifugation thru Centricon-100 concentrators (Amicon, Beverly, MA) with 2 ml distilled water, and sequenced using primer f2.

Sequence analysis

Amino acid sequences of branching enzymes (BEs, which include SBEs and glycogen branching enzymes, or GBEs) were obtained from the NCBI protein database and aligned using CLUSTAL W (Thompson *et al.* 1994) under its default parameters: pairwise alignments=slow (accurate), gap opening penalty=10, gap extension penalty=0.10 or 0.05, and the BLOSUM series (Henikoff and Henikoff 1992) for scoring. Phylogenetic trees were constructed using two methods: the Neighbor-Joining method (Saitou and Nei 1987), as implemented in Treecon (version 1.15) (Van de Peer and De Wachter 1994); and protein parsimony as implemented in the PHYLIP package (Felsenstein 1989). The N- and C-terminal portions of the alignment (positions 1-215 and 950-end, Fig. 2) were excluded from the analyses, as the quality of the alignment and placement of gaps in this region appear unreliable. Trees were inferred either including or excluding insertion-deletion (indel) regions.

Results and discussion

Cloning and sequencing of an SBE gene from G. gracilis

Degenerate PCR primers were used to amplify a portion of an SBE gene from genomic DNA of *G. gracilis.* A PCR product was confirmed by sequencing to encode a portion of a probable SBE gene and was used as a probe to screen a *G. gracilis* genomic library, resulting in the recovery of two clones. The sequence obtained from one clone (λ25) is shown in Figure 1. It contains an ORF of 767 codons. The inferred amino acid sequence aligns well with other SBE sequences (Fig. 2), and is highly similar with those of green plants (40-45% sequence identity), human (46%) and *Saccharomyces cerevisiae* (45%), indicating that the ORF encodes an SBE. Since there are apparently two or more SBE genes in *G. gracilis* based on the results of the Southern analysis (see below), we designate this gene as *GgSBE1*.

More than 1.4 kb of the 5' flanking region was sequenced. Most ORFs found in this region contain fewer than 100 codons except for one which is at least 240 codons. However, when the entire sequence of the 5' flanking region was used to search the NCBI database (using BLASTX), no sequences with significant similarity were found (BLASTX scores < 80).

Figure 1. The nucleotide sequence of *GgSBE1*. The numbering scheme assigns +1 to the first position of the start codon (marked with 1). Potential *cis*-regulatory elements are underlined, including two pyrimidine-rich regions. The potential TATA box is in bold and underlined. The sequence of the 5' and 3' ends of cDNAs (shown in uppercase letters) is aligned with the genomic sequence, with the site of the sequencing primers r3c (5' RACE) and f2 (3' RACE) underlined. The conceptual translation is shown below the coding region; * marks the stop codon. The putative polyadenylation signal (CGTAAA) downstream of the protein coding region is underlined. The last 4 residues (GATC, positions 2616-2619) comprise the *Sau*3A I site where the vector and the genomic insert were ligated.

tttgcggtgcgttagcctcgtttgccaagttataccaagatgcagtggcgccatatcagcgagctgggaagctacgtc -1406
tagtagattctgagaacgaagataaggcgtcgcctgtggaagggattcggttcatctccaccttcggccatgctccag -1328
gccattgtgcgattgaagtaagaaaagacggggaaagaatggtttcatcggcgatgcatgggtatcaacggtaagct -1250
agaccgagttttgatcgagagaattgtgatgcgcggtgtcggtgagtgggactaacttgttgatgtggcatgttttg -1172
gttcacgttgatgcaggcggatcaattgcaaaacccggagtggggcttcttcacagaaagcaaccgcactgaggctta -1094
cctttcacgggaaggactcttgacagatgtggccagaacggaggctcttgtggttgcgtatcacgagcatttcccggg -1016
gctgggatatgtgcaggagcgaggacctttctttgaatgggtgccgaatgcaccacagctgctggggttcggcgtgaa -938
gagaaagtgctgagaagtgctctcactggtgttttcgaagctgatatggtagataccttcagagggtgaaggttctgc -860
tattctcgtatcgttgggcttatgctttccgaggatgtttcacttggcgcgtggtgcattttgctttgcatgacaaga -782
actctaggctgcgtgtatccattcaccgtgaatcagccttggtgataaagaagcagcgaccgcaatgcacttcggtga -704
ttataagcgcgccacttgtaccttcgtagttgtaattctagaaggacaaactgttctgaaccttcgtagttgtaattc -626
tagaaggacaaactgttttgaacctgcagtgtggttgaagccactaaacatacagccaatagcactctgtcgatggtg -548
tgccacgcagccgtcgtcgtcgtccgtaagtccatttgactttgcagatactgccaggaggcgctttgattcacatcg -470
tgtgatggacatgaagataccacgccacggctcaccgcgaccggacgctcaggggtctctcgagacgcgacagcagca -392
gacgccacaagctacgactgaaaaagatgctgagatcgggcgtcgttacgttggcagtcggtggtggctgccacgccg -314
cgagtgcggcgcttcgattggcgaaaaagcatactgtcagcgttcaacgggatatccgtggcccaacagctcgagctc -236
gcagccacaaaccgtcggctgcaaaacggcgcgtgcaatcggcgcgcggtgtgccgatttctcgagcctgcgcacag -158
ccgctcgttattcacctc**attatta**cgctgccgcttgagaaagtgccccctccccccctttatgggctcaaatccgcgc -80
ttctgaagtgccctctccatacgtagcgcccgcttcttctcgctcctactttctcggtgctgtttctcgcgtctccac -2
CTTCTTCTCGCTCCTACTTTCTCGGTGCTGTTTCTCGCGTCTCCAC

1
catgggatcagaggacccccactacgtcgcctggaaggacaacaaagacggaacggcttgcatccgcgatgaccgcta 77
CATGGGATCAGAGGACCCCCACTACGTCGCCTGGAAGGACAACAAAGACGGAACGGgTTGtATcCGCGATGACCGCTA
M G S E D P H Y V A W K D N K D G T A C I R D D R Y
cttagaacccttcgcggacgccttgcgttaccgctattcaaagtactccgagattctgtccgccatagagtccagcga 155
CTTAGAACCCTTCGCGGACGCCTTGCGTTACCGcTATTCAAAGTACTCCGAGATTCTGTCCGCCATAGAGTCCAGCGA
L E P F A D A L R Y R Y S K Y S E I L S A I E S S E
gaagtcgctggagaacttttccaggggctatgaaaccatgggcattcatcgcgtcgaaggcggagtccgctacagaga 233
GAAGTCGcTGGAGAACTTTT <---- primer r3c site
K S L E N F S R G Y E T M G I H R V E G G V R Y R E
atgggcacccagcgcccgtgagatgtacttcttcggtgagttcaataactgggatcgtaatgcactccccatggaacg 311
W A P S A R E M Y F F G E F N N W D R N A L P M E R
cgatgagttcggaatctggtcctgcttcatcccggaggcagagccgggcgtctctccaatcaagcatggctcaaaggt 389
D E F G I W S C F I P E A E P G V S P I K H G S K V
caaggccgccgtcgtgcctcatcaagggccctggcttgatcgtaatcccgcctgggccaccttctgcgtgcaagatac 467
K A A V V P H Q G P W L D R N P A W A T F C V Q D T
caaaactttcctgtacgatacagtgttttgggatcctccggagaaattcaagtggactgcaccggatcacgtgaagtg 545
K T F L Y D T V F W D P P E K F K W T A P D H V K C
tccggactcgcttcgcatctatgaatgccatgtaggaatggggtccaatgatctcaaggttggctcctaccgcgagtt 623
P D S L R I Y E C H V G M G S N D L K V G S Y R E F
cgcggataatgttttaccacgtattaaggaaacaggctatactgccttgcagattatggccattatggaacatgccta 701
A D N V L P R I K E T G Y T A L Q I M A I M E H A Y
ctatgccagctttggctatcacgtcaccaatttctttgcgattagttccagatgcggcataccagaggatctgaagta 779
Y A S F G Y H V T N F F A I S S R C G I P E D L K Y
cctcatcgacaaggcccaccagcttggattgtatgtctttatggatgtcgtccactcacatgcttcgagcaattctat 857
L I D K A H Q L G L Y V F M D V V H S H A S S N S M
ggatggtatcaacaactttgatggcactgaccatcaatatttccacgaaggtgagcgtggacgccattctctgtggga 935
D G I N N F D G T D H Q Y F H E G E R G R H S L W D
ttctaggcttttcaactatggacattgggaagtgctccgtttcctcttatccaatctaaggtggtatatggaggagta 1013
S R L F N Y G H W E V L R F L L S N L R W Y M E E Y
tcactttgacggctttcgttttgatggcgtgacatccatgctatacttgcattccggaattggcgtgcagttcaccgg 1091
H F D G F R F D G V T S M L Y L H S G I G V Q F T G
taactacagtgaatacttcggcttccaagttgatgtcgatgcatgcgtgtatatgatgctcgctaataagcttgtaca 1169
N Y S E Y F G F Q V D V D A C V Y M M L A N K L V H
cgatctttaccccgatgtcgcagttactatcgcagaagatgtgagtggtatgccgactctatgcgtgcctgtagaccg 1247
D L Y P D V A V T I A E D V S G M P T L C V P V D R
gggagggcttgggtttgactatcgccttgccatggccattcctgatatgtggatagaggtgctagaaaagaaagga 1325
G G L G F D Y R L A M A I P D M W I E V L E K E K D
tgaaaactggaacatgggcaacattgtcttcacactaacaaaccgtcgttggaacgagaagtctattgggtattgtga 1403
E N W N M G N I V F T L T N R R W N E K S I G Y C E
gagtcatgaccaagcactagtcggggacaaaacaattgcgttctggttgatggatgctgcaatgtacactgacatgag 1481
S H D Q A L V G D K T I A F W L M D A A M Y T D M S
ttgtaatggttatccctcgcctgcggtcgagcgaggcatcgctcttcacaaaatgattcgactgctcaccatgtgctt 1559
C N G Y P S P A V E R G I A L H K M I R L L T M C L

Figure 1

gtctggtgagggataccttacattcatgggaaatgaatttggccaccccgaatgggttgatttcccacgtgaaggaaa 1637
S G E G Y L T F M G N E F G H P E W V D F P R E G N
cggcaattcttatcagcatgcgcgtcgccgatgggatctttgtgacaatgagtcgcttcgttacaagcatttgtacga 1715
G N S Y Q H A R R R W D L C D N E S L R Y K H L Y E
gtttgagaaaatcatacatgcgctggataacgcacatcccttctgtagattccaagcacatcagtatatagtgctgca 1793
F E K I I H A L D N A H P F C R F Q A H Q Y I V L Q
acatgaaggagacaagttaattgtggtggaaaagggtgatcggttgttgtttgtcttcaacttccaccattctcagtc 1871
H E G D K L I V V E K G D R L L F V F N F H H S Q S
gtacagtgattatcgaattggaacctactggggaggccggtacaaactggttcttgattcagatggaatgaacactgg 1949
Y S D Y R I G T Y W G G R Y K L V L D S D G M N T G
tggacatggtcgggtgcattgggatgtggtgcatactacgaggacggaacagtggcacaatcgaccatactacttgca 2027
primer f2 ----> CTACGAGGACGGAACAGTGGCACAATCGACCATACTACTTGCA
G H G R V H W D V V H T T R T E Q W H N R P Y Y L Q
ggtatacatcccagcgaggacatgtcaggtgtatcactgctttgaaacttgggaagaagagaaagagaaggggagaa 2105
GGTATACATCCCAGCGAGGACATGTCAGGTGTATCACTGCTTTGAAACTTGGGAAGAAGAGAAAGAGAAGGGGAGAA
V Y I P A R T C Q V Y H C F E T W E E E K E K G E K
ggacaaacagactgccaaaaagtccgaaggagttgttgatgacacgaaagagaaagcaggggtgacgaaaaagctaa 2183
GGACAAACAGACTGCCAAAAAGTCCGAAGGAGTTGTTGATGACACGAAAGAGAAAGCAGGGGGTGACGAAAAAGCTAA
D K Q T A K K S E G V V D D T K E K A G G D E K A N
cactagggttgaggaggtctcagcggttgcagccaagattgacgaggcagtgcatttgaacggttcacaaaaggcctc 2261
CACTAGGGTTGAGGAGGTCTCAGCGGTTGCAGCCAAGATTGACGAGGCAGTGCATTTGAACGGTTCACAAAAGGCCTC
T R V E E V S A V A A K I D E A V H L N G S Q K A S
agcaaagcgacaggaaaaggttgtagctggaaaagcgtaaacaggtctgcttgttgtggaacgggctgtcaatcgaga 2339
AGCAAAGCGACAGGAAAAGGTTGTAGCTGGAAAAGCGTAAACAGGTCTGCTTGTTGTGGAACGGGCTGTCAATCGAGA
A K R Q E K V V A G K A *
actggacccaagcactgagtgctgtgtccatgcatggcgcctttaacctggtcccaacgtgtatccagctgggacggg 2417
ACTGGACCCAAGCACTGAGTGCTGTGTCCATGCATGGCGCCTTTAACCTGGTCCCAACGTGTATCCAGCTGGGACGGG

gggaggtgcgtgtgtttaatttgtccttc**gtaaa**gcttcgggttgtgttttgtctagtttgttctacgatgcgtctgc 2495
GGGAGGTGCGTGTGTTTAATTTGTCCTT**CGTAAA**GCTTCGGGTTGTGTTTTGaCTAAAAAAAAAAAAAAAAA

gattctgaaagctacctagatggctagaacttctgttgccattgtgttcaccatcggagctttaaaactacaccaaa 2573
gaaaagtattcccaggaagaacttgtgcttttgactttcattgatc 2619

Figure 1, continued

Figure 2. Alignment of starch-branching enzyme (sbe) and glycogen-branching enzyme (gbe) sequences. Residues shown in white on a black background are conserved in at least 60% of the sequences; white on grey, 40%. GgraSBE, sequence deduced from *GgSBE1*, this paper; maize IIB, *Zea mays* sbeIIB (NCBI accession no. 1169911); ricesbe3, *Oryza sativa* sbe-3 (436052); TaessbeII, *Triticum aestivum* sbeII (1885344); peasbeI, *Pisum sativum* sbeI (1345570); AthasbeII, *Arabidopsis thaliana* sbe class II (726490); yeastgbe, *Saccharomyces cerevisiae* gbe (1076979); human-gbe, *Homo sapiens* gbe (1082408); ricesbeI, *O. sativa* sbe1 (421991); ricesbeQ, *O. sativa* sbeQ (399544; conflicts with another rice sbeQ sequence was noted); maizesbeI, *Z. mays* sbeI (600872); TaessbeI, *T. aestivum* sbe (1935006); StubsbeQ, *Solanum tuberosum* sbe-Q (1169912); peasbeII, *P. sativum* sbeII (1345571); Syntisgbe, *Synechocystis* sp. strain PCC6803 gbe (1707936); Syncusgbe, *Synechococcus* sp. strain PCC7942 (121297); Scoelgbe, *Streptomyces coelicolor* gbe (2127516); Mtubegbe, *Mycobacterium tuberculosis* gbe (1707934); Ecoli-gbe, *Escherichia coli* gbe (66573); Atumegbe, *Agrobacterium tumefaciens* gbe (1707933); Bsubtgbe, *Bacillus subtilis* gbe (1084216).

```
                          10        20        30        40        50
Ggrasbe1   : --------------------------------------------------- :  -
maizeIIB   : -----MAFRVSGAVLG---GAVRAPRLTGG---GEGSLVFRHTGLFLTR-- : 38
ricesbe3   : --MAAPASAVPGSAAGLRAGAVRFPVPAGARSWRAAAELPTSRSLLSGRRF : 49
TaessbeII  : ----MATFAVSGATLG-----VARPAGAGG----G--LLPRSGSERRGGVD : 36
peasbeI    : -----MVYTISGIRFPVLPSLHKSTLRCDRRASSHSFFLKNNSSSFSRT-- : 44
AthasbeII  : ----RARVRFPHLPSIKKKNSSLHSFNEDLRRSNAVSFSLRKDSRSSGK-- : 45
yeastgbe   : --------------------------------------------------- :  -
human-gbe  : --------------------------------------------------- :  -
ricesbe1   : --------------------------------------------------- :  -
ricesbeQ   : --------------------------------------------------- :  -
maizesbeI  : --------------------------------------------------- :  -
TaessbeI   : --------------------------------------------------- :  -
StubsbeQ   : ----------------------------------------MEINFKVLSKP : 11
peasbeII   : --------------------------------------------------- :  -
Syntisgbe  : --------------------------------------------------- :  -
Syncusgbe  : --------------------------------------------------- :  -
Scoelgbe   : MTPRPSSSGPDPRKTTGKKPAGKTPTGKKPAKAAKKKAPRRTTASANASAT : 51
Mtubegbe   : ------------------------------------------MSRSEK--- :  6
Ecoli-gbe  : --------------------------------------------------- :  -
Atumegbe   : ---------------------------------------MKKPLNSAEE--- : 10
Bsubtgbe   : --------------------------------------------------- :  -

                     60        70        80        90       100
Ggrasbe1   : --------------------------------------------------- :  -
maizeIIB   : --GARVGCSGTHGAMRAAAAARKAVMVPEGEND--GLASRADSAQFQS--- : 82
ricesbe3   : PGAVRVGGSGGRVAVRAAGAS-GEVMIPEGESD--GMPVSAGSDDLQLPAL : 97
TaessbeII  : LPSLLLRKKDSSRAVLSRAASPGKVLVPDGESD--DLASPAQPEELQIPED : 85
peasbeI    : --SLYAKFSRDSETKSSTIAESDKVLIPEDQDNSVSLADQLENPDITSEDA : 93
AthasbeII  : --VFARKPSYDSDSSSLATTASEKLRGHQSDSSS-SASDQVQSRDTVSDDT : 93
yeastgbe   : --------------------------------------------------- :  -
human-gbe  : --------------------------------------------------- :  -
ricesbe1   : ----MLCLTSSSSSAPAPLLP-----SLADRPSP-GIAGGGGNVRLSVVSS : 41
ricesbeQ   : ----MLCLTSSSSSAPPPLLP-----SLADRPSP-GIAGGGGNVRLSVVSS : 41
maizesbeI  : ----MLCLVSPSS-SPTPLPPPRRSRSHADRAAPPGIAGGG-NVRLSVLSV : 45
TaessbeI   : ----MLCLTAPSCSPSLPPRP---SRPAADRPGP-GISGGG-NVRLSAVPA : 42
StubsbeQ   : IRGSFPSFSPKVSSGASRNK------ICFPSQHSTGLKFGSQERSWDISST : 56
peasbeII   : -------ATTTTTTHNSKNK------QYLAKQKPVELTLGYQNPNGCKVCS : 38
Syntisgbe  : ---------MTYTINADQVHQ----IVHNLHHDPFEVLGCHPLGDHGKVNQ : 38
Syncusgbe  : --MTGTTPLPSSSLSVEQVNR----IASNQEQNPFDILGPHPYEHEGQAG- : 44
Scoelgbe   : TSVSGAEVAVSPAPDAADRER----LLAGTHHDPHAVLGAHRVPGGVAFR- : 97
Mtubegbe   : ------LTGEHLAPEPAEMAR----LVAGTHHNPHGILGAHEYDDHTVIR- : 46
Ecoli-gbe  : ---------MSDRIDRDVINA----LIAGHFADPFSVLGMHKTTAGLEVR- : 37
Atumegbe   : --------KKTGDITKAEIEA----IKSGLHSNPFQIIPLHETPEGFSAR- : 48
Bsubtgbe   : --------------------------------------------------- :  -
```

Figure 2

```
                 110        120        130        140        150
Ggrasbe1    : -------------------------------------------------- :    -
maizeIIB    : ---DELEVPDISEETTCGAG---------------------VADAQALNRV :  109
ricesbe3    : D--DELSTEVGAEVEIESSG---------------------ASDVEGVKRV :  125
TaessbeII   : I--EEQTAEVNMTGGTAEKLESSEPTQGI-----------VETITDGVTKG :  123
peasbeI     : QNLEDLTMKDGNKYNIDESTSSYREVGDEKGSVTSSSLVDVNTDTQAKKTS :  144
AthasbeII   : QVLGNVDVQKTEEAQETETLDQTSALSTSG----------SISYKEDFAKM :  134
yeastgbe    : -------------------------------------------------- :    -
human-gbe   : ------------------------------------------MAAPMTPA :    8
ricesbe1    : -------PRRSWPGKVKTNF----------------------SVPATARK :   62
ricesbeQ    : -------PRRSWPGKVKTNF----------------------SVPATARK :   62
maizesbeI   : Q----CKARRSGVRKVKSKF----------------------ATAATVQE :   69
TaessbeI    : P----SSLRWSWPRKAKSKF----------------------SVPVSAPR :   66
StubsbeQ    : P-----KSRVRKDERMKHSS----------------------AISAVLTD :   79
peasbeII    : F-----GSKGSIYQKVSSGF----------------------KGVSVMTD :   61
Syntisgbe   : WV-IRAYLPTAEAVTVLLP---------------------TDRREVIMTT :   66
Syncusgbe   : WV-IRAYLPEAQEAAVICP---------------------ALRREFAMHP :   72
Scoelgbe    : -----VFRPYALAVTVLS-----------------------GELRVGLHD :  119
Mtubegbe    : -----AFRPHAVEVVALV-----------------------GKDRFSLQH :   68
Ecoli-gbe   : -----ALLPDATDVWVIEPK---------------------TGRKLAKLEC :   62
Atumegbe    : -----CFIPGAEEVSVLT-L---------------------DGNFVGELKQ :   72
Bsubtgbe    : -------------------------------------------------- :    -

                160        170        180        190        200
Ggrasbe1    : ----MGSEDPHYVAWKDNKDGTACIRDDRYWEPFADALRYFYSKYSEILSA :   47
maizeIIB    : ------------RVVPPPSDGQKIFQIDPMLQGYKYHLEYRYSLYRRIRSD :  148
ricesbe3    : VE--ELAAEQKPRVVPPTGDGQKIFQMDSMLNGYKYHLEYRYSLYRRLRSD :  174
TaessbeII   : VK--ELVVGEKPRVVPKPGDGQKIYEIDPTLKDFRSHLDYRYSEYRRIRAA :  172
peasbeI     : VHSDKKVKVDKPKIIPPPGTGQKIYEIDPLLQAHRQHLDFRYGQYKRIREE :  195
AthasbeII   : SHSVDQEVGQ--RKIPPPGDGKRIYDIDPMLNSHRNHLDYRYGQYRKLREE :  183
yeastgbe    : -------------MYNIPDNVKGAVEFDPWLKPFADVLSERRYLADKWLYD :   38
human-gbe   : ARPEDYEAALNAALADVPELAR-LLEIDPYLKPYAVDFQRRYKQFSQILKN :   58
ricesbe1    : NK------TMVTVVEEV--DHLPIYDLDPKLEEFKDHFNYRIKRYLDQKCL :  105
ricesbeQ    : NK------TMVTVVEEV--DHLPIYDLDPKLEEFKDHFNYRIKRYLDQKCL :  105
maizesbeI   : DK------TMATAKGDV--DHLPIYDLDPKLEIFKDHFRYRMKRFLEQKGS :  112
TaessbeI    : DY------TMATAEDGV--GDLPIYDLDPKFAGFKEHFSYRMKKYLDQKHS :  109
StubsbeQ    : DNS-----TMAPLEEDVKTENIGLLNLDPTLEPYLDHFRHRMKRYVDQKML :  125
peasbeII    : DKS-----TMPSVEEDF--ENIGILNVDSSLEPFKDHFKYRLKRYLHQKKL :  105
Syntisgbe   : VHH--------PNFFECVLELEE-PKNYQLRITENGHERVIYDPYGFKTPK :  108
Syncusgbe   : VHH--------PHFFETWVPEET-LEIYQLRITEGERERIIYDPYAFRSPL :  114
Scoelgbe    : DG---------DGFFSGLVPLKD-VPAHRLLVAYEGRE-QEVEDPYRFLPT :  159
Mtubegbe    : LD---------SGLFAVALPFVD-LIDYRLQVTYEGCEPHTVADAYRFLPT :  109
Ecoli-gbe   : LDS--------RGFFSGVIPRRKNFFRYQLAVVWHGQQ-NLIDDPYRFGPL :  104
Atumegbe    : IDP--------DGFFEGRIDLSK-RQPVRYRACRDDAE-WAVTDPYSFGPV :  113
Bsubtgbe    : ----------------------------------------------MAAAS :    5
```

Figure 2, continued

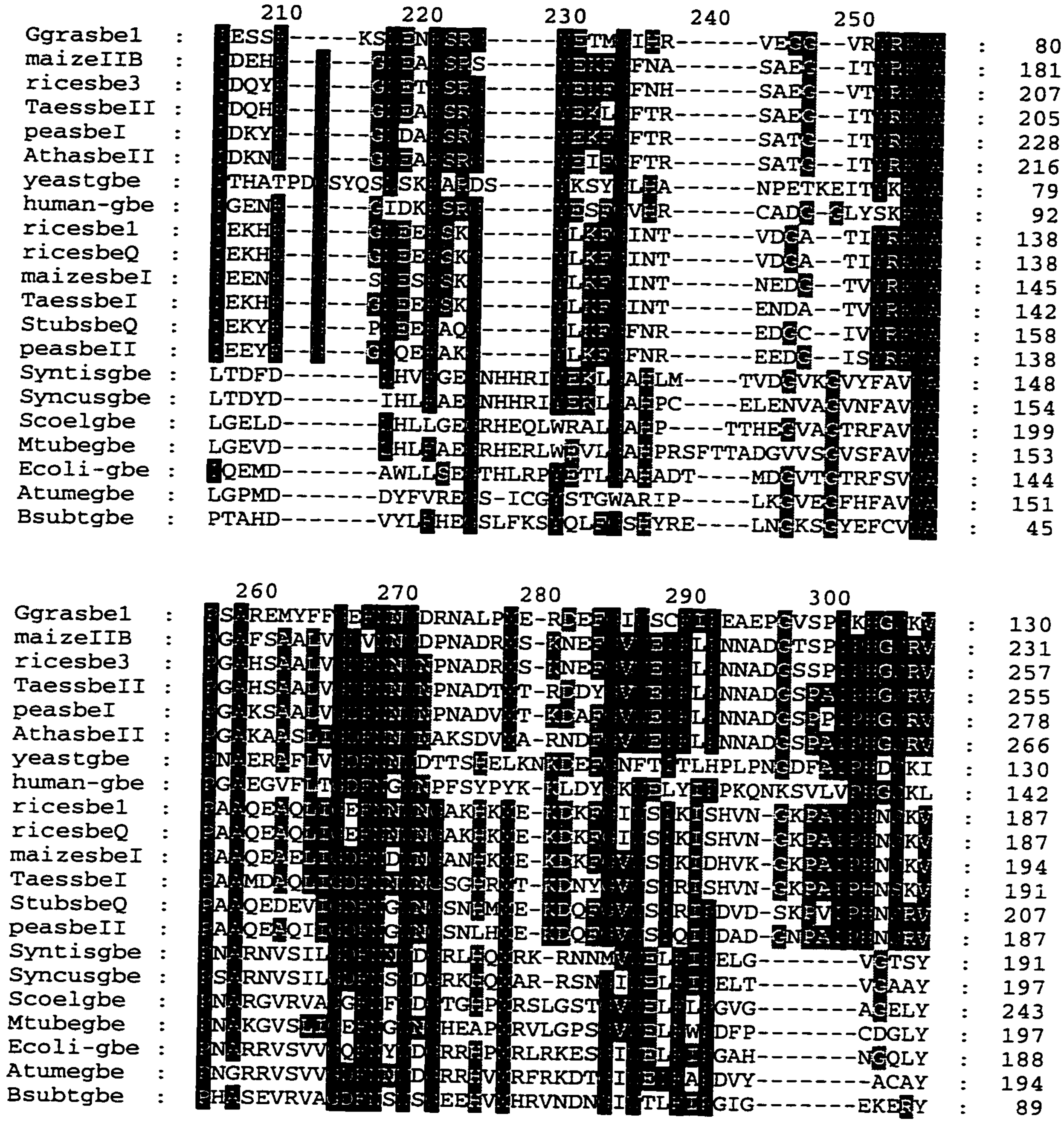

Figure 2, continued

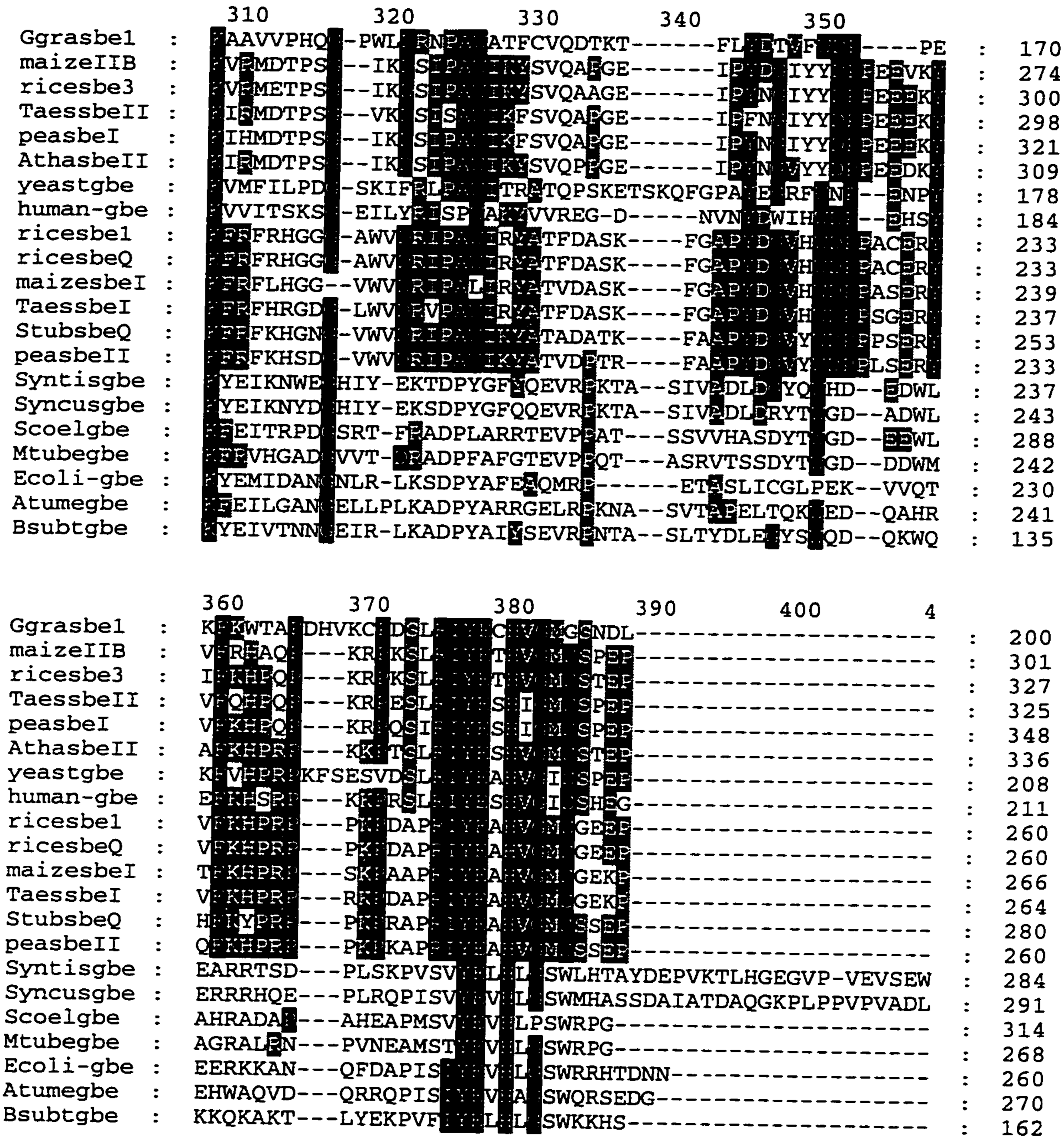

Figure 2, continued

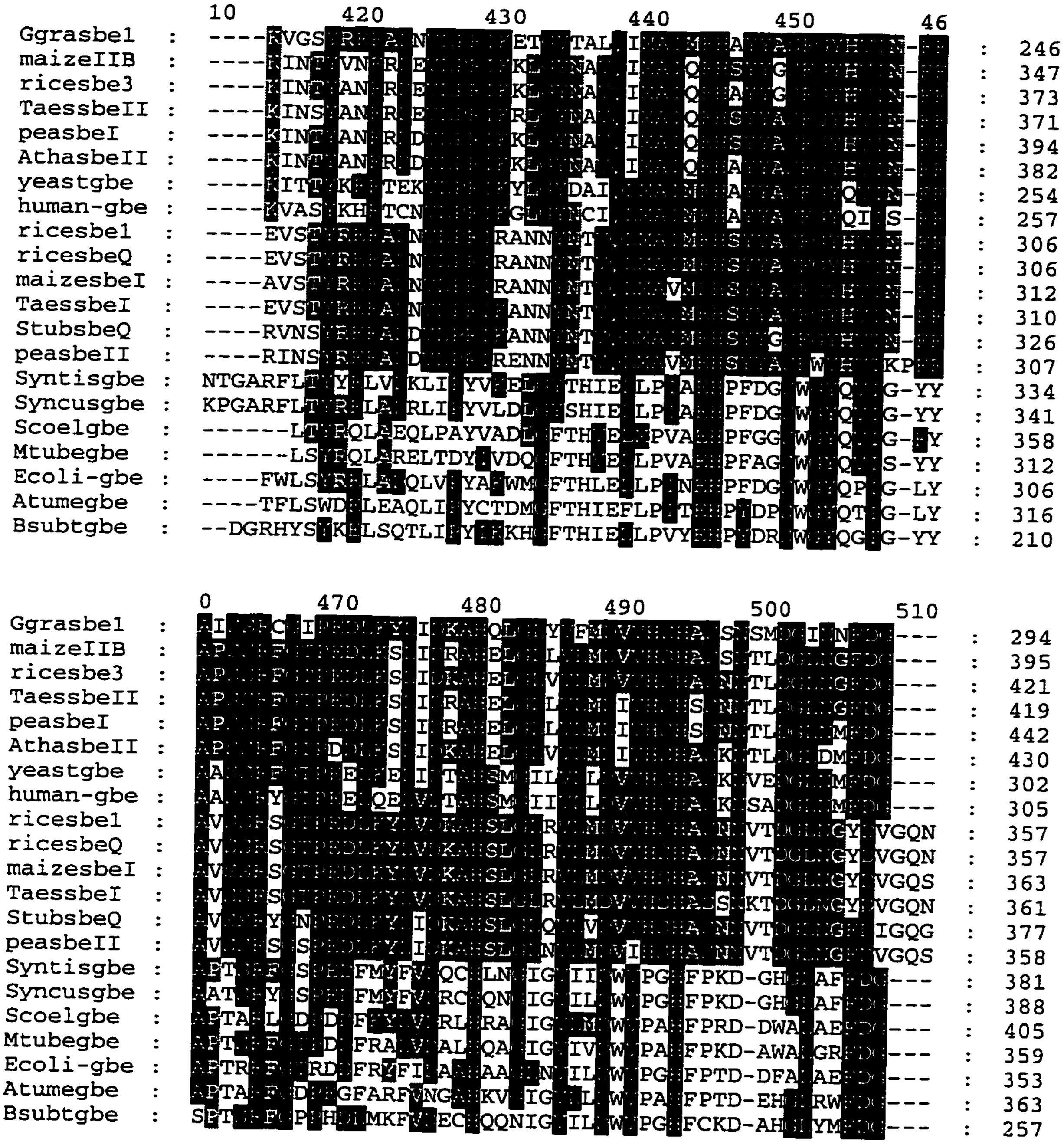

Figure 2, continued

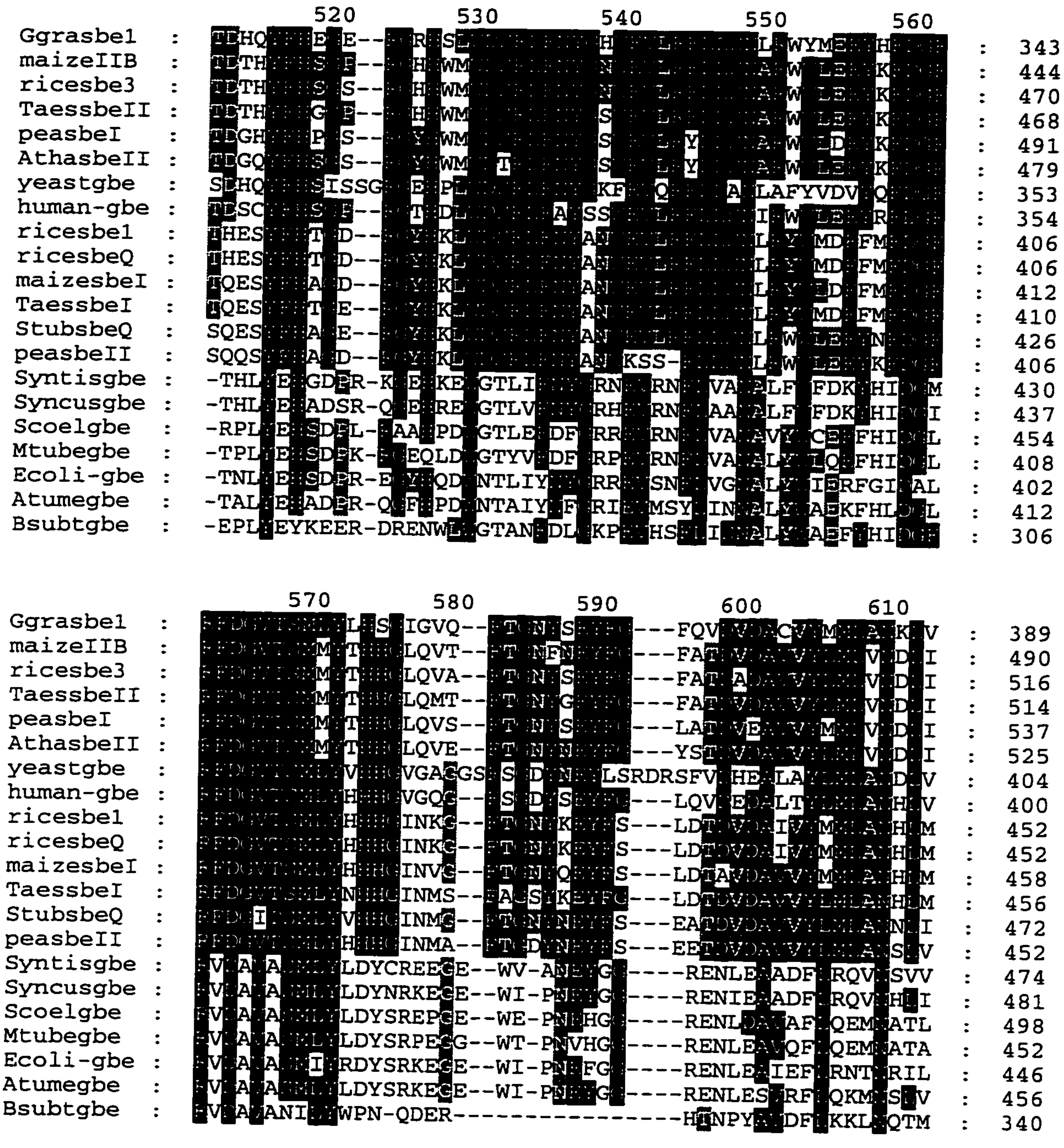

Figure 2, continued

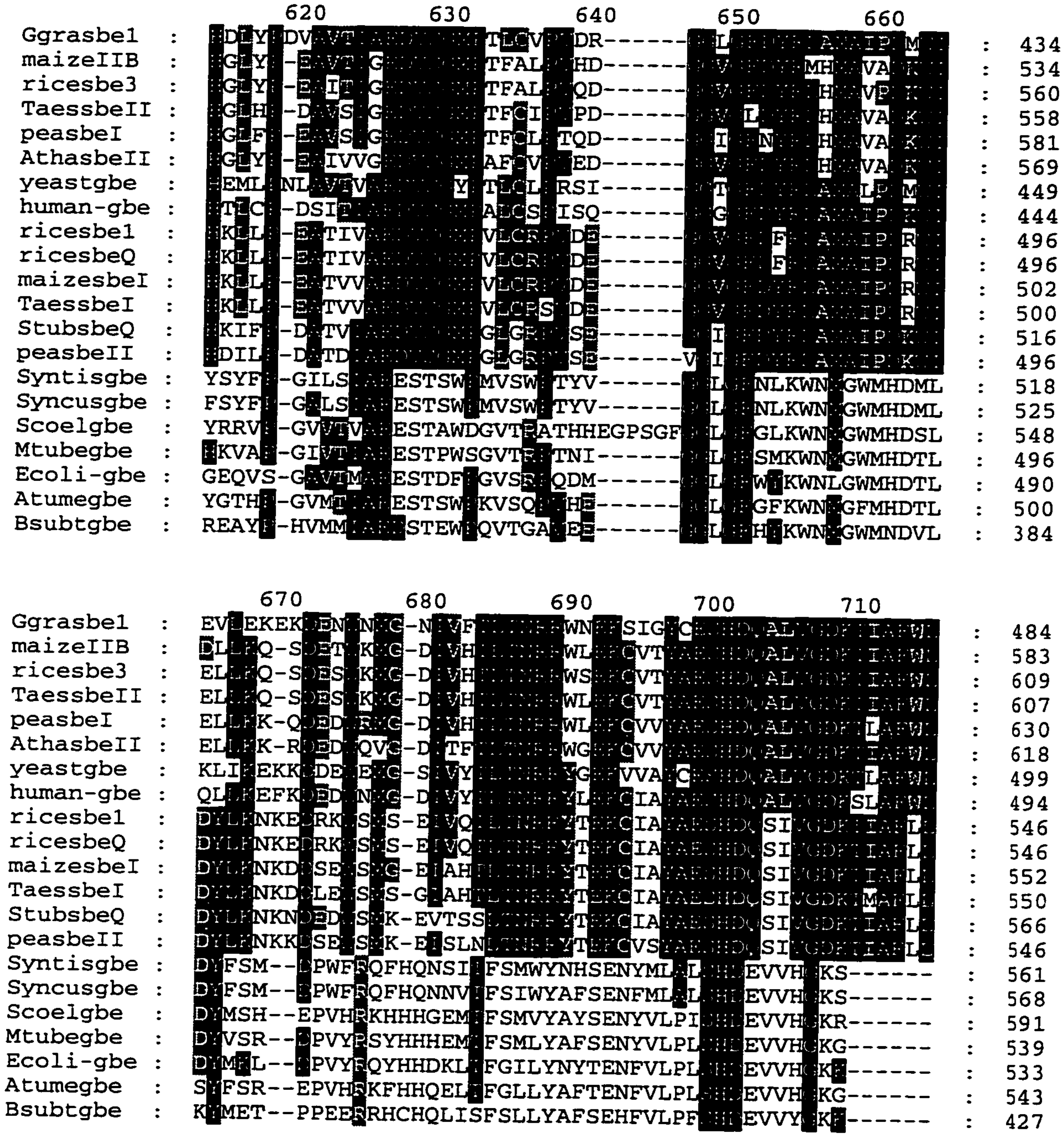

Figure 2, continued

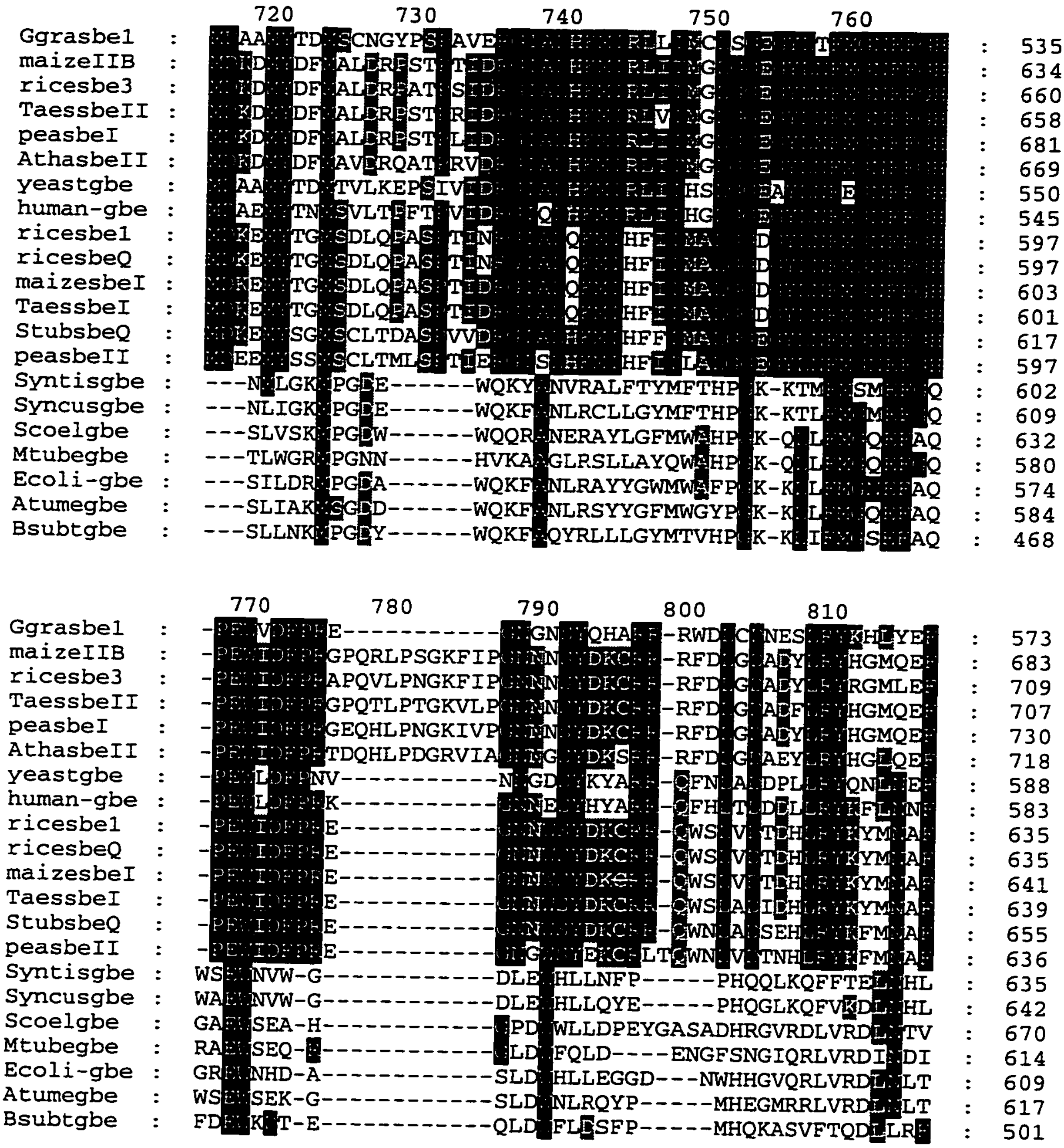

Figure 2, continued

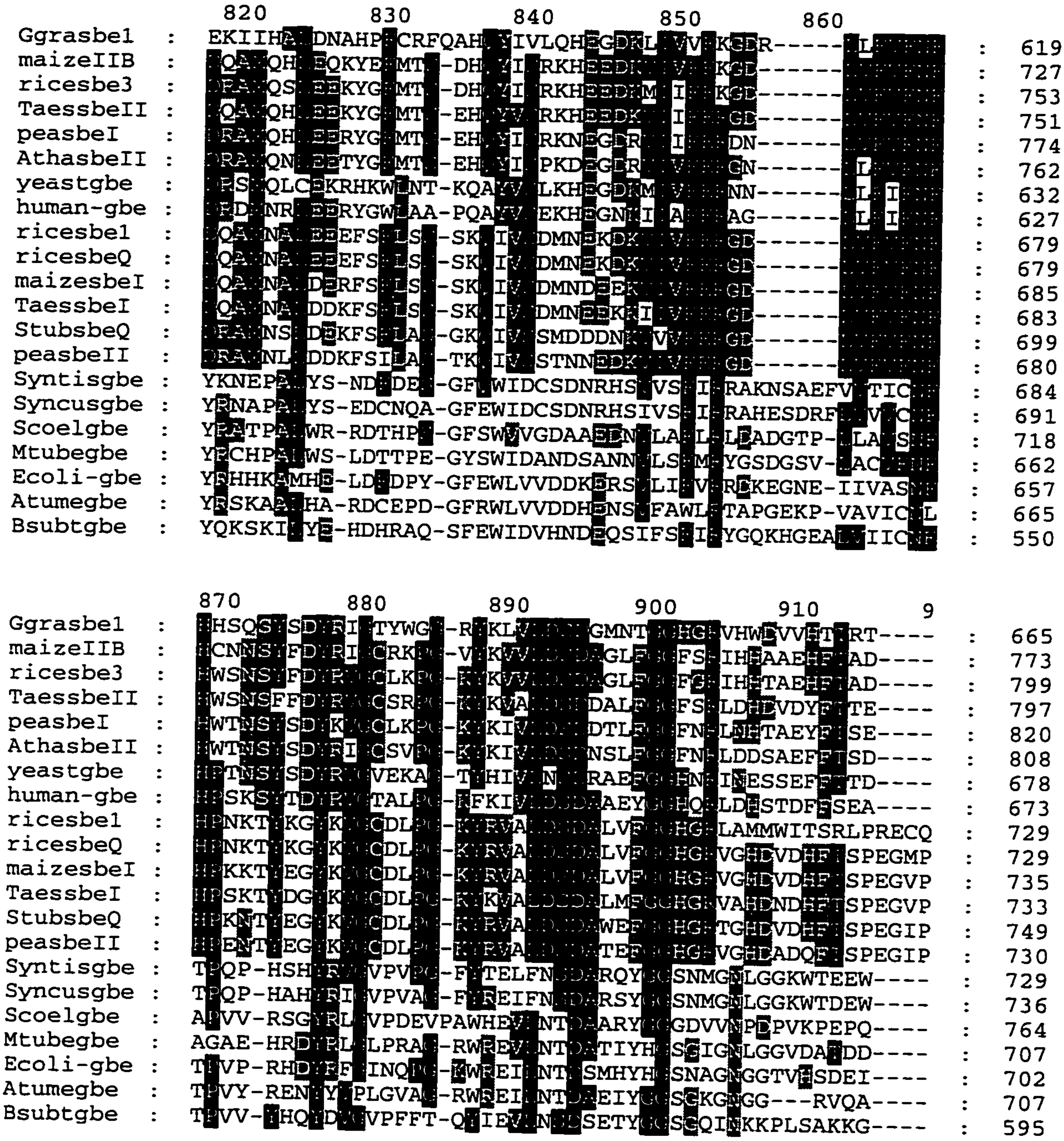

Figure 2, continued

```
              20         930       940       950       960       97
Ggrasbe1   :  ----EQWHNF-DYYLQVSIDARSCQVVHCFETWEEEKEKGEKDKQTAKKSE  :  711
maizeIIB   :  ----CSHDNF-DYAFSVYTDSRTCVVYAPVE-------------------  :  799
ricesbe3   :  ----CSHDNF-DYAFSVISDSRTCVVYAPAE-------------------  :  825
TaessbeII  :  ----HPHDNF-DRAFSVYTDSRTAVVYALTE-------------------  :  823
peasbeI    :  ----GWYDDF-DRAFLVYADSRTAVVYALADGVESEPIELSDGVESEPIEL  :  866
AthasbeII  :  ----GRHDDF-DCAFMVYADCRTAVVYAAVDDDDDDERSSLVPIGLLPEDV  :  854
yeastgbe   :  ----LEWNNF-KNFLQVYIDSRVALVLALKE-------------------  :  704
human-gbe  :  ----FEHNGF-DYALLVYIDSRVALILQNVDLPN----------------  :  702
ricesbe1   :  EYQKQISTTA-LTHSKSFPRFVPVWAVYRVDEDREELRRGGAVASGKIV-T  :  778
ricesbeQ   :  GVPETNFNNF-DNAFKVLSDPRTCVAVYRVDEDREELRRGGAVASGKIV-T  :  778
maizesbeI  :  GVPETNFNNF-DNAFKVLSDPRTCVAVYRVDEAGAGRRLHAKAETGKTSPA  :  785
TaessbeI   :  GVPETNFNNF-DNAFKILSDSRTCVAVYRVEEKAEKPKDEGAASWGKTA-L  :  782
StubsbeQ   :  GVPETNFNGFQIPAKCCLLREHVWLITELMNACQKLKITRQTFVVSYYQQP  :  800
peasbeII   :  GIPETNFNNF-DNAFKVLSDEHTCVVVYRVDERQEESNNPNLGSVEETFAA  :  780
Syntisgbe  :  -----SFHEQ-DYALDLCLPFLSVLVLKLSQNAEENTVPAEEASNIA----  :  770
Syncusgbe  :  -----SCHNF-DYALDLCLPFLTTLVLELASGPES----LSEAANSPL---  :  774
Scoelgbe   :  -----GWHGF-DAAIRLTLPFLATVWLRPA-------------------  :  788
Mtubegbe   :  -----PWHGF-DAAAVLVLPFTSALWLTPA-------------------  :  731
Ecoli-gbe  :  -----ASHGF-QHALSLTLPFLATIWLVREAE-----------------  :  728
Atumegbe   :  -----VDAGG-EIGAMLVLPFLATIMLEPEN------------------  :  732
Bsubtgbe   :  -----ALHHK-DCYITMTIPFYGISILRAVKKRGEIKR-------------  :  627

              0         980       990       1000      1010      1020
Ggrasbe1   :  GVVDDTKEKAGGDEKANTRVEEVSAVAAKIDEAVHLNGSQKASAKRQEKVV  :  762
maizeIIB   :  ---------------------------------------------------  :  -
ricesbe3   :  ---------------------------------------------------  :  -
TaessbeII  :  ---------------------------------------------------  :  -
peasbeI    :  SVGVESEPIELSVEEAESEPIERSVEEVESETTQQSVEVESETTQQSVEVE  :  917
AthasbeII  :  ---------------------------------------------------  :  -
yeastgbe   :  ---------------------------------------------------  :  -
human-gbe  :  ---------------------------------------------------  :  -
ricesbe1   :  EYIDVEAT------SGETISGGWKG-----------SEKDDCGKKGMKFVF  :  812
ricesbeQ   :  EYIDVEAT------SGETISGGWKG-----------SEKDDCGKKGMKFVF  :  812
maizesbeI  :  ESIDVKAS------RASSKED--K-------------EATAGGKKGWKFAR  :  815
TaessbeI   :  GYIDVEATGVKDAADGEATSGSEKA-----------STGGDSSKKGINFVF  :  822
StubsbeQ   :  ISRRVTRNLKIRYLQISVTLTNACQKLKFTRQTFLVSYYQQPILRRVTRKL  :  851
peasbeII   :  ADTDVAR---IPDVSMESEDSNLDR---------IEDNSEDAVDAGILKVE  :  819
Syntisgbe  :  ---------------------------------------------------  :  -
Syncusgbe  :  ---------------------------------------------------  :  -
Scoelgbe   :  ---------------------------------------------------  :  -
Mtubegbe   :  ---------------------------------------------------  :  -
Ecoli-gbe  :  ---------------------------------------------------  :  -
Atumegbe   :  ---------------------------------------------------  :  -
Bsubtgbe   :  ---------------------------------------------------  :  -
```

Figure 2, continued

		1030		
Ggrasbe1	:	AGKA--------	:	766
maizeIIB	:	------------	:	-
ricesbe3	:	------------	:	-
TaessbeII	:	------------	:	-
peasbeI	:	SETTQ-------	:	922
AthasbeII	:	------------	:	-
yeastgbe	:	------------	:	-
human-gbe	:	------------	:	-
ricesbe1	:	RSSDEDCK----	:	820
ricesbeQ	:	RSSDEDCK----	:	820
maizesbeI	:	QPSDQDTK----	:	823
TaessbeI	:	LSPDKDNK----	:	830
StubsbeQ	:	KDSLSTNISTDA	:	863
peasbeII	:	REVVGDN-----	:	826
Syntisgbe	:	------------	:	-
Syncusgbe	:	------------	:	-
Scoelgbe	:	------------	:	-
Mtubegbe	:	------------	:	-
Ecoli-gbe	:	------------	:	-
Atumegbe	:	------------	:	-
Bsubtgbe	:	------------	:	-

Figure 2, continued

Features of the gene

We mapped the 5' end of the *GgSBE1* transcript by 5' RACE. The sequence of the 5' end-fragment of the cDNA is aligned with the genomic sequence in Figure 1. Our results indicate that the 5' end is at or close to position -47. A potential TATA box is located 85 bp further upstream from this site. Upstream of the putative TATA box are found three CAAT elements, one in an inverted orientation. Potential GATA boxes and ACGT elements were likewise identified. Pyrimidine-rich regions (indicated in Fig. 1) were found downstream of the TATA box. The first AUG codon downstream of the putative 5' end of the transcript appears to be in the right context for translation initiation; the sequence TCCACC**AUG** (with the translation initiation codon in bold) conforms to the apparent canonical sequence found at the translation initiation site in red algal genes (YYCRCY**AUG**: Zhou and Ragan 1996), and in fact is identical to the sequence in the corresponding site in the polyubiquitin gene from *G. gracilis* (Zhou and Ragan 1995b).

Comparison of the cDNA and genomic sequences (Fig. 1), and comparison of the inferred GgSBE1 amino acid sequence with other SBE sequences (Fig. 2) indicate that *GgSBE1* is devoid of introns; the sequence similarity between the GgSBE1 protein, deduced from the genomic sequence, with other SBEs extends throughout the entire length of GgSBE1. The yeast glycogen branching enzyme is also intronless (Thon *et al.* 1992), although the *sbe1* gene from rice contains 13 introns (Kawasaki *et al.* 1993). The absence of intron in *GgSBE1* is, however, not surprising; known red algal nuclear genes are either intronless (for example, genes for triose phosphate isomerase in *G. gracilis* [as *G. verrucosa*: Zhou and Ragan 1995c], and GapC in *C. crispus* [Liaud *et al.* 1993]) or, more commonly, interrupted only by a single intron (for example, genes for GapA, GapC, and mitochondrial aconitase in *G. gracilis* [as *G. verrucosa*: Zhou and Ragan 1994, 1995a,d], and GapA and β-tubulin genes in *C. crispus* [Liaud *et al.* 1993, 1995]).

The 3' end of the transcript was sequenced using the 3' RACE technique. Comparison of its sequence with that of the genome (Fig. 1) reveals that the polyA (cleavage) site occurs 171 bp downstream of the termination codon; a potential polyA signal, CGUAAA, occurs 21 bp upstream of the polyA site. All cDNAs from *G. gracilis* that have been characterized so far contain the TAAA motif, occurring close to and upstream of the polyA site. In plants, AAUAAA-like elements, analogous to the highly conserved AAUAAA motif in animals, are degenerate or may even be absent (Hunt 1994, Wu *et al.* 1995). The short region between the polyA site and the putative polyA signal is rich (86%) in G and U residues; GU-rich regions have been shown to be important for efficient polyadenylation in animals and plants, and are located upstream (in plants) or downstream (in animals) of the AAUAAA-like motifs (see Wu *et al.* 1995).

Southern analysis

To determine the copy number of *GgSBE1*, I performed Southern analyses with a *GgSBE1* fragment as a probe. The results, shown in Figure 3A, indicate that there are at least two SBE-related sequences in the *G. gracilis* genome, as indicated by multiple bands observed per lane after moderately stringent washing conditions (final wash with 0.5X SSC/0.1%SDS, 65°C, 30 min, performed twice). Other *G. gracilis* genes (*e.g.*, *GgUGP*, Chapter V) used as controls in the Southern experiments showed single bands at this stringency. However, a further, more stringent washing (0.1X SSC/0.1% SDS, 65°C for 30 min, performed twice) of the same blot yielded only a single band per lane (Fig. 3B), suggesting that *GgSBE1* is single-copy in the *G. gracilis* genome This finding is consistent with the observation that three distinct SBE isozymes occur in another red alga, *R. pertussa* (Fredrick 1968, 1970).

Multiple isoforms of SBEs also exist in plants (Burton *et al.* 1995, Martin and Smith 1995). Interestingly, pea SBEI (a member of the Sbe2 family) and pea SBE II (a member of the Sbe1 family) genes do not cross-hybridize in Southern blotting experiments, even at low stringency (2X SSC, 55°C; Burton *et al.* 1995). Whether or not this indicates that SBEs in plants are more divergent than are SBEs in red algae remains to be investigated.

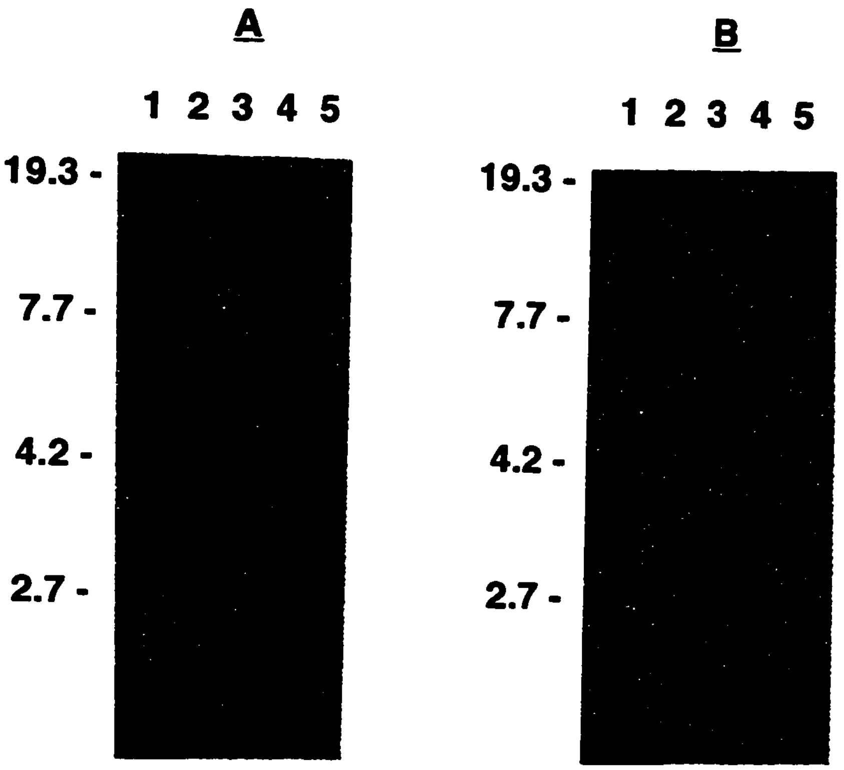

Figure 3. Determination of *GgSBE1* copy number by Southern hybridization analysis. Genomic DNA (5 μg per reaction) from *G. gracilis* was digested with *Eco*R I (lane 1), *Hin*d III (lane 2), *Sal* I (lane 3), *Xba* I (lane 4), and *Xho* I (lane 5). The probe was prepared from a *GgSBE1* fragment (position 759 to position 1504, Fig. 1), which contains no sites for *Eco*R I, *Sal* I, *Xba* I, or *Xho* I, and one site for *Hin*d III (positions 1159-1164). Final washing was with 0.5X SSC/0.1% SDS at 65°C for 30 min, performed twice (**A**). The same blot was rewashed twice with 0.1X SSC/0.1% SDS at 65°C for 30 min (**B**). The numbers on the left of each panel indicate the size of the markers, in kb.

Features of the GgSBE1 protein: structural and phylogenetic analysis

Roughly 50% of the residues in GgSBE1 are identical with those at corresponding sites in plant SBEs and yeast and animal GBEs (as aligned in Fig. 2). Some residues that are highly conserved not only among SBEs and GBEs but also among members of the α-amylase family (to which the SBEs and GBEs belong; see, for example, Romeo *et al.* 1988, Baba *et al.* 1991, Jespersen *et al.* 1993) are also conserved in GgSBE1. These include H493, D564, E625, H700 and D701 (numbering according to the alignment in Fig. 2), residues involved in substrate binding or in catalysis (see Svensson 1994, Kuriki *et al.* 1996).

It has been proposed that group-specific differences in the nature of residues on certain β-strand → α-helix connecting loops, and the lengths of these loops, help determine substrate specificity of members of the α-amylase family (Jespersen *et al.* 1993, Svensson 1994). Burton *et al.* (1995) have noted that members of the Sbe2 family in plants contain an insertion of 11 amino acids in the loop connecting β-strand 8 and α-helix 8 that is absent in members of the Sbe2 family, and that this difference could possibly account for the difference in the average lengths of branches they transfer. This 11-amino acid insertion is also absent in GgSBE1 and in human and yeast GBEs.

The GgSBE1 protein, like the yeast and human GBEs, lacks a region at its N-terminus that would correspond to those in green plant SBEs. This difference can be partly explained by the fact that plant, but not animal or yeast, BEs possess transit peptides; since starch synthesis is cytosolic in red algae (see Pueschel 1990), red algal SBEs would not be expected to contain a transit peptide. However, the mature peptides of Sbe2 isoforms in green plants as well as the eubacterial GBEs have an N-terminal extension absent in BEs from red algae, animals, and fungi.

To examine relationships among BEs, we aligned 21 SBEs and GBEs, removed the ambiguously alignable regions, and inferred distance trees by the neighbor-joining and protein parsimony methods (see Methods). Both methods produced trees with essentially similar topologies. Essentially the same topology was also recovered whether the regions common to only a few of the sequences (*e.g.* shared insertions) were kept in the alignment, or removed. The tree produced by the Neighbor Joining method is shown in Figure 4.

The BEs separate cleanly into two groups, each present in 100% of the bootstrap replicates: BEs from eukaryotes (*S. cerevisiae*, human, *G. gracilis* and green plants: Fig. 4, top 14 sequences), and GBEs from eubacteria, including cyanobacteria (bottom seven sequences). GgSBE1 is resolved on one of the deeper branches within the eukaryote group, although bootstrap support within this region of the tree is modest. SBE isoforms in green plants divide solidly into two subgroups, corresponding to Sbe1 and Sbe2 families, each supported in 100% of the bootstrap replicates. SBEs from both monocots and dicots are present in each subgroup, strongly suggesting that these two isoforms became

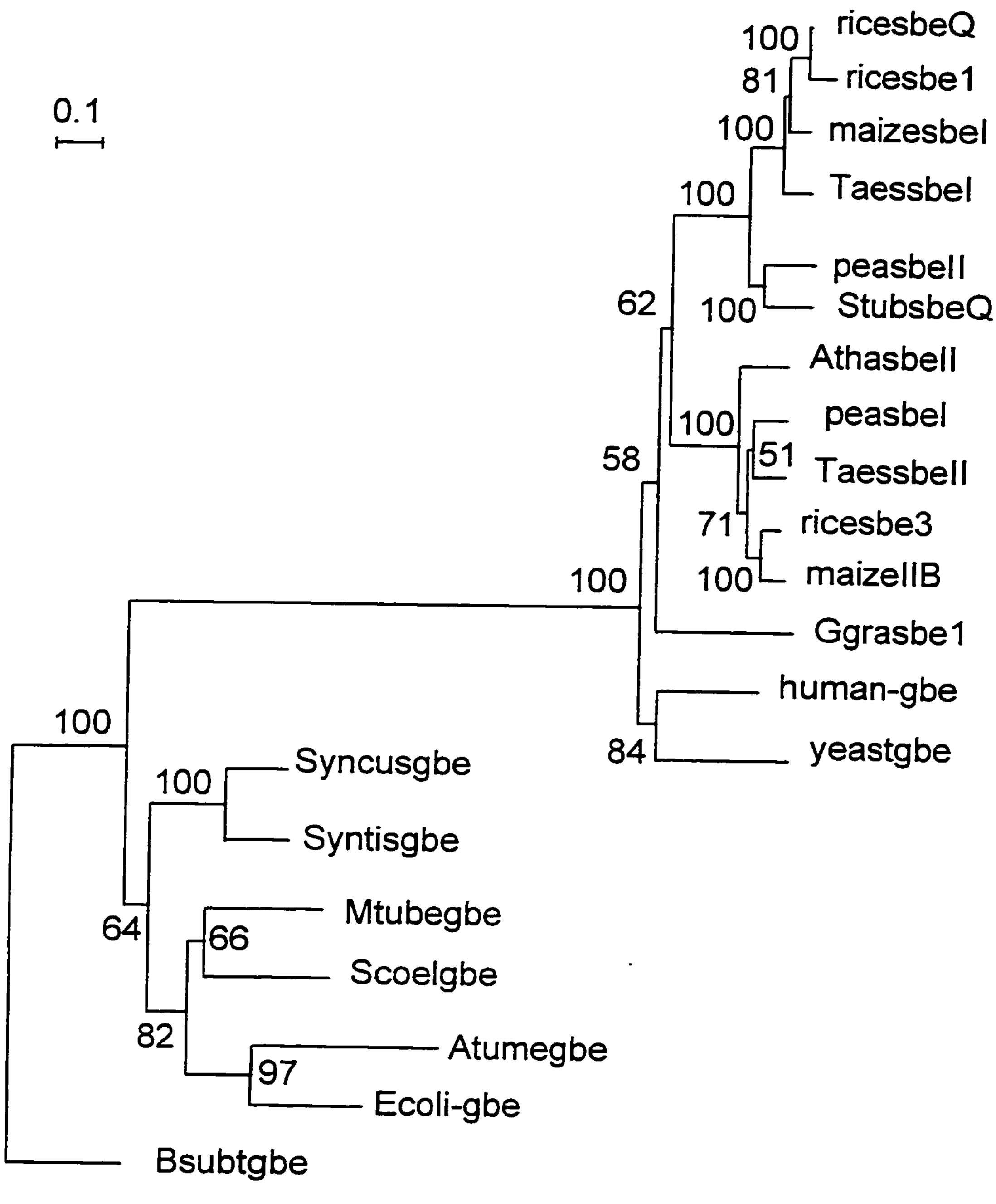

Figure 4. A phylogenetic tree of selected starch- and glycogen-branching enzymes constructed using the neighbor joining method. Distances were calculated using the formula of Kimura (1983); insertion-deletion regions ("indels") were excluded from the analysis. Numbers represent bootstrap values out of 100. The scale (top, left) indicates the branch length corresponding to 0.1 substitution per site. See Figure 2 for full names of the sequences.

established prior to the divergence of monocot and dicot lineages. Complete interpretation of these results (*e.g.* whether GgSBE1 specifically resembles one or the other of the green plant SBE groups) must await the identification and sequencing of other SBE-like sequences in *G. gracilis*.

The grouping of eukaryote BEs (Fig. 4) bears further comment. Both red algae (Nagashima *et al.* 1971, Sheath *et al.* 1979) and green plants (Preiss 1991) use the ADP-glucose pathway (*sensu* Preiss 1991) for synthesis of α-1,4-glucans. As red algae and green plants appear closely related phylogenetically (Ragan and Gutell 1995), it seems a reasonable extrapolation that their ADP-glucose pyrophosphorylases (AGPases) and starch synthases should each be homologous. Green plant AGPases and starch synthases bear significant sequence similarity to the corresponding enzymes in eubacteria, particularly cyanobacteria (Charng *et al.* 1992, Ainsworth *et al.* 1993), consistent with plants having acquired this biosynthetic capability through endosymbiosis with a cyanobacterium-like organism leading to establishment of the plastid. By contrast, animals and fungi utilize the UDP-glucose pathway (*i.e.,* UGPase and glycogen synthase) for α-1,4-glucan biosynthesis. Glycogen synthase uses UDP-glucose as substrate, and is thought to be only remotely related to bacterial (hence plant) starch synthases (Browner *et al.* 1989; Farkas *et al.* 1990). Nonetheless the BEs of animals and fungi are clearly very similar to those of green plants and red algae, at least in contrast with eubacterial (including cyanobacterial) GBEs (Fig. 4).

One might thus hypothesize that in red algae, the endosymbiotically acquired AGPase and starch synthases were combined with the ancestrally eukaryotic BE to construct the pathway of floridean starch biosynthesis in the cytosol. On the other hand, in green plants, where starch is biosynthesized within the plastid, not only must the AGPase and starch synthase genes have retained or acquired sequences encoding plastid-localization signals (*e.g.* transit peptides), but sequences coding such signals must have been acquired by the ancestrally eukaryotic BE gene(s). Zhou and Ragan (1994) and Ragan and Gutell (1995) have suggested that the apparently common position of an intron in the plastid transit peptide coding region of the plastid-localized *GapA* gene of green plants and *G. gracilis* [as *G. verrucosa*] indicates that the red algal – green plant lineage acquired its plastid before the separation of the red algal and green plant lineages. By contrast, the SBE gene appears to have acquired a plastid transit peptide coding region in green plants but not in *G. gracilis* (or, by extension, in red algae).

Conclusion

We have cloned an SBE gene from the red alga *G. gracilis*. The gene is apparently one of at least two SBE genes in this alga. Presumably, the products of these genes differ in function (as might be indicated by differences in expression or catalytic activity), as is the case in green plants. It would thus be important to isolate and sequence the other gene(s) so that comparative structural and functional analysis of the SBEs in *G. gracilis* can be carried out. This would be necessary to ascertain the function of each isoform in floridean starch biosynthesis. The cloning and characterization of *GgSBE1* would also facilitate more intensive biochemical studies of this gene and its product.

CHAPTER V

The UDPglucose pyrophosphorylase gene from the marine red alga *Gracilaria gracilis*

Introduction

UDP-glucose (UDPGlc) plays a key role in carbohydrate metabolism in the red algae. Its main biosynthetic function is as precursor of UDP-galactose (UDPGal), the D-galactosyl donor in the biosynthesis of galactans and of floridoside (α-D-galactopyranosyl-(1-2)-glycerol) (Su and Hassid 1962, Kremer and Kirst 1981, Manley and Burns 1991). Galactans, such as agarans and carrageenans, are the most abundant component of the cell walls in most red algae (Craigie 1990), whereas floridoside is a key photosynthetic product that functions as a short-term low-MW carbohydrate reserve (Kremer 1978, Macler 1986). There is some evidence that UDPGlc might also serve as a minor glucosyl donor in the biosynthesis of floridean starch (Nagashima *et al.* 1971).

The key enzyme for the biosynthesis of UDPGlc is UDPglucose pyrophosphorylase (UGPase; EC 2.7.7.9), which uses UTP and Glc1P as substrates. Plant, animal, and eubacteral UGPases are well characterized at the protein and gene levels (*e.g.*, Turnquist and Hansen 1973, Kleczkowsi 1994). That this enzyme occurs in red algae was first demonstrated in the early 1960s when UDPGlc, *inter alia*, was isolated from the red alga *Porphyra perforata* (Su and Hassid 1962). More recently, Manley and Burns (1991) demonstrated

UGPase activity in the red alga *Pterocladia capillacea*. However, characterization of red algal UGPases at the protein and gene levels has not yet been accomplished.

Herein we report the cloning of the UGPase gene from the agarophytic marine red alga *Gracilaria gracilis*, and describe some features of the gene and the deduced protein.

Materials and Methods

Screening of genomic libraries and sequencing of clones

Construction of a genomic library of *G. gracilis* ("grass" strain) was as described (Chapter IV). A homologous UGPase probe was produced by a PCR-mediated approach. Degenerate primers were designed based on highly conserved regions of UGPases, as revealed by a multiple alignment of UGPase sequences obtained from the NCBI protein database. PCR with one pair of primers (oligo-A, GGNGGN[TC]T[GT]GGNAC[GT][AT]C[GT]ATGGG), corresponding to the highly conserved amino acid sequence GGLGTSMG, and oligo-B, [GA]T[TC]NCC[GA]TG[GT]CCNGG[GT]GG[GA]TACCA, corresponding to the highly conserved amino acid sequence WYPPGHGD) yielded a product that was confirmed by subsequent sequencing to encode part of a potential UGPase gene. Using this PCR product as a probe, we screened the genomic library of *G. gracilis* and isolated two clones (λ27 and λ29); in each, the putative

UGPase-coding region was truncated close to its 3'-end, as revealed by sequencing using vector-specific primers. One clone (λ27) was further sequenced on both strands; primers for initial sequencing were designed based on the known sequence (from the initial PCR product). To obtain the sequence of the entire gene and of the 3'-flanking region, additional clones were obtained from another *G. gracilis* genomic library, constructed by Zhou and Ragan (1994); the sequence of the 3' portion of the UGPase gene and its 3'-flanking region was obtained from one clone (λUGPb). Lambda DNA purification, library screening, and sequencing methodology were as described (Chapter III).

3' RACE

mRNA was extracted from laboratory-grown *G. gracilis* using the Invitrogen FastTrack 2.0 kit (Invitrogen Corp., San Diego, CA). The 3' end of the UGPase transcript was reverse-transcribed using the Pharmacia T-primed Ready-To-Go kit (Pharmacia Biotech, Uppsala, Sweden), and PCR-amplified using primer f2b (positions 746-767, Fig. 1) as the gene-specific primer, together with an anchor primer (ATTCGCGGCCGCAGGAATT). As two PCR products were found on the gel, these were separately reamplified by PCR using the same primers. The PCR products were desalted by centrifugation thru Centricon-100 concentrators (Amicon, Beverly, MA) with 2 ml distilled water, and used directly for sequencing.

Southern analysis

DNA extraction and Southern blotting were as described (Chapter III). The probe was synthesized and labelled with α-^{32}P-dCTP using the Random Primed DNA labelling kit (Boehringer Mannheim, Mannheim, Germany), and a portion of the cloned DNA fragment (positions -33 to 424 in Fig. 1; amplified by PCR) as template.

Sequence analysis

UGPase amino acid sequences were obtained from the NCBI database and aligned, together with the UGPase sequence from *G. gracilis*, using CLUSTAL W (Thompson *et al.* 1994) under its default parameters: pairwise alignments=slow (accurate), gap opening penalty=10, gap extension penalty=0.10 or 0.05, and the BLOSUM series (Henikoff and Henikoff 1992) for scoring. The same alignment was used for construction of phylogenetic trees, except that the first 66 positions were deleted from the data set, as alignment in this region is ambiguous. Tree construction was carried out using the Neighbor Joining algorithm (Saitou and Nei 1987) as implemented in Treecon (Van de Peer and DeWachter 1994).

Results and discussion

Cloning of the G. gracilis *UGPase gene*

Using a homologous probe to screen a *G. gracilis* genomic library, I recovered two clones, each of which was found to contain a UGPase-coding region truncated near the 3'-end; one clone (λ27) was sequenced. To obtain the sequence of the rest of the UGPase gene, another clone (λUGPb) was isolated from another *G. gracilis* genomic library (Zhou and Ragan 1994). As UGPase is encoded by a single gene in *G. gracilis* (see below), we are confident that the clone λUGPb contains the same gene; the sequence at the region of overlap (339 bp) was 100% identical (data not shown), confirming their identity. The final sequence, reconstructed from the two clones (λ27 and λUGPb), is shown in Figure 1. It contains an ORF of 496 codons. Comparison of the deduced amino acid sequence with other UGPase sequences (Fig. 2) shows significant sequence identity between the deduced protein with the sequences from *Solanum tuberosum* (50%), human (51%) and *Saccharomyces cerevisiae* (48%), confirming that the gene we isolated encodes a UGPase. We designate this gene as *GgUGP*.

The 5' flanking region has proved recalcitrant to sequencing. I was also unable to obtain the 5'-end of the transcript by the 5'-RACE technique, using methodology that has yielded the 5' ends of other *G. gracilis* genes

Figure 1. The nucleotide sequence of *GgUGP*. The numbering scheme assigns +1 to the first position of the putative start codon (marked with 1). The putative TATA box is in bold and underlined. The conceptual translation is shown below the coding region; * marks the stop codon. The 3'-RACE revealed two polyadenylation sites; the putative polyadenylation signal (ATTAAA) of the first site, which overlaps with the termination codon, and that of the second site (AATAAA), are underlined.The sequence of the 3'-end cDNAs is aligned with the genomic sequence and is shown in uppercase letters, with the site of the sequencing primers f4b and f5 underlined; the site of primer f2b (used in the RTPCR) is underlined (position 746-767). The portion of the sequence representing the 3' end of a putative DNA helicase gene (encoded on the complementary strand) is italicized and underlined.

```
tgcaccccccctccgctccgctttgttcgcttcttgcaccccccgttccaccgcttcacccgtggaaccttctgctgctat   -60
                                                                 1
cccgttcgcactgctgttcgtttatttaagccaccatgatgccaacggaaaaggagccatgaatcgcgactccaggtctc     22
                                                                 M  N  R  D  S  R  S
tgcaggacttcaagggcgtcatggacaagtccgccgcctccaccgtcgccgagaagctcactgtcatgaaccagatggccg   103
L  Q  D  F  K  G  V  M  D  K  S  A  A  S  T  V  A  E  K  L  T  V  M  N  Q  M  A
ccaatgagctcgagaagatgaccgattctgagaccaccggcttcgtcgagttgtacggccgctacatgagcgaacgttcca   184
A  N  E  L  E  K  M  T  D  S  E  T  T  G  F  V  E  L  Y  G  R  Y  M  S  E  R  S
aaaaggccgaaatcaagtgggatctcatcgaacagcccagtgaaaacatgctgcaaaagtacgataccttgccaaagccgg   265
K  K  A  E  I  K  W  D  L  I  E  Q  P  S  E  N  M  L  Q  K  Y  D  T  L  P  K  P
ccaccgacgaagaactcgcttcgctcctttccaagctggctgtgctcaagctcaacggcggtcttggtacctccatgggct   346
A  T  D  E  E  L  A  S  L  L  S  K  L  A  V  L  K  L  N  G  G  L  G  T  S  M  G
gcaaaggacccaagagcgtcatcgaagtgcgtgatgacaccaccttcttggacctcattgttcagcagattggtcagctca   427
C  K  G  P  K  S  V  I  E  V  R  D  D  T  T  F  L  D  L  I  V  Q  Q  I  G  Q  L
acaagaaccatcccacggccaacgtcccccctgcttctcatgaactctttcaacaccgactctgagaccgcaaagatcattc   508
N  K  N  H  P  T  A  N  V  P  L  L  L  M  N  S  F  N  T  D  S  E  T  A  K  I  I
gcaagtaccaggataccagtgttaccctcaccaccttccagcaatctcgttaccccaggatcgtcaaggagtctctcgaac   589
R  K  Y  Q  D  T  S  V  T  L  T  T  F  Q  Q  S  R  Y  P  R  I  V  K  E  S  L  E
cgatgccgctcacacacgaccactatgcccatgaggactggtaccctccaggtcacggagatttcttccaatctatttaca   670
P  M  P  L  T  H  D  H  Y  A  H  E  D  W  Y  P  P  G  H  G  D  F  F  Q  S  I  Y
actcgggattggttgatacccttcttgcgcagggtaaggagtacatctttgtctcaaacgttgacaatcttggcgccactg   751
N  S  G  L  V  D  T  L  L  A  Q  G  K  E  Y  I  F  V  S  N  V  D  N  L  G  A  T
tcgatctcaacattctcaagaatgtcgttgaccgcgaagtcgagtactgcatggagctcaccgacaagacacgtgccgaca   832
V  D  L  N  I  L  K  N  V  V  D  R  E  V  E  Y  C  M  E  L  T  D  K  T  R  A  D
tcaagggtggtaccattatttcatacgacggaaaggtgtctctgctcgaggtcgctcaggtccctgccaagtacgttgaag   913
I  K  G  G  T  I  I  S  Y  D  G  K  V  S  L  L  E  V  A  Q  V  P  A  K  Y  V  E
agttcaagtctatttccaagtttaaggtgttcaacaccaataacatttgggtatcgttacgagcaatcaagcgcgtcatgc   994
E  F  K  S  I  S  K  F  K  V  F  N  T  N  N  I  W  V  S  L  R  A  I  K  R  V  M
aatccggcgagatgaagcttgatatcattgtgaacaacaaggaagtcaagggtaccaaggtcatccagctcgaaagcgcta  1075
Q  S  G  E  M  K  L  D  I  I  V  N  N  K  E  V  K  G  T  K  V  I  Q  L  E  S  A
ttggagcagccattggctacttcaacaatgcgtgcggtgtcaacgttcctcgttcccgtttccttcctgtcaagtcaacct  1156
I  G  A  A  I  G  Y  F  N  N  A  C  G  V  N  V  P  R  S  R  F  L  P  V  K  S  T
ccgatctcatgctcattcagagcaacatgtacaaccttaaatctggctctctggttatgaaccctgcccgccaatttacta  1237
S  D  L  M  L  I  Q  S  N  M  Y  N  L  K  S  G  S  L  V  M  N  P  A  R  Q  F  T
caactccagtgattaagcttggaaaggagtttaagaaggtcgctcaatatcttgaacgtctgggtagcatccctgacattt  1318
                                    sf4b ---->                          CTGACATTT
T  T  P  V  I  K  L  G  K  E  F  K  K  V  A  Q  Y  L  E  R  L  G  S  I  P  D  I

tggagcttgaccatctcactgtctcaggtgatgtctactttggggctaacactactctgaaaggaaccgttatcgtggtag  1399
TGGAGCTTGACCATCTCACTGTCTCAGGTGATGTCTACTTTGGGGCTAACACTACTCTGAAAGGAACCGTTATCGTGGTAG
L  E  L  D  H  L  T  V  S  G  D  V  Y  F  G  A  N  T  T  L  K  G  T  V  I  V  V

caaaccctgggaacaccatcatgattccagaaggctcagttctcgagaacaaggtcgttcttggttctctccatgtgattc  1480
CAAACCCTGGGAACACCATCATGATTCCAGAAGGCTCAGTTCTCGAGAACAAGGTCGTTCTTGGTTCTCTCCATGTGATTC
A  N  P  G  N  T  I  M  I  P  E  G  S  V  L  E  N  K  V  V  L  G  S  L  H  V  I

cgcattaaacatgtcctggtaggctacggtcagtgcgcgtactatacattaagcatcttcatttggtaggcccgctatcgc  1561
CGCATTAAACATGTCCTGGTAGGCTACGGTCAGTGCGCGTACTAAAAAANAAAAAAAAAAA
P  H  *

agcatgtgcaggccatcctgaagcaatcggtgagcccatcagttagaatgttggtccatgccgtggttgaatcacgaatgg  1642
sf5 -->        ATCCTGAAGCAATCGGTGAGCCCATCAGTTAGAATGTTGGTCCATGCCGTGGTTGAATCACGAATGG

ataccgacattttcggcatgaaggaccattcatgacctcgtcctgcgaccgtactacgagcgagaaaggagagattgtctg  1723
ATACCGACATTTTCGGCATGAAGGACCATTCATGACCTCGTCCTGCGACCGTACTACGAGCGAGAAAGGAGAGATTGTCTG

cgaacaatttactgctacgacacgacgcgtagtgagttagcgaattagtttgtcgaagggcacctcaatctgatacgcgct  1804
CGAACAATTTACTGCTACGACACGACGCGTAGTGAGTTAGCGAATTAGTTTGTCGAAGGGCACCTCAATCTGATACGCGCT

aatacacgaattcaataaacaatgcaagaactatagggtgtttcggttttgaacttaatcttactctggtgcggtccggct  1885
AATACACGAATTCAATAAACAATGCAAGAACTATAGGGTGTTTCGGTTTTGAAAAAAAAAAAAAAAAAAAAN

tccgatgaattcatagtttttccggtggcttctcctccactcgagacttagctaccgatggttgcgccatgtcaagcagaaa  1966
agtacgtgcttcagcaattgcggtaccgagatcgagatccaaggagtcaacagacaggaactgggcaatccatttcggaag  2047
cttgcttcttagtttggcacgcgtaaaccgtttgtcggcgaatataacgattccgtaatcgttcttgttccgaattacacg  2128
ccctgcgcactgagccgcttggcgcattgcctattgtccgttttggtgacgattcagaagaagaactgttagatttctcct  2209
agcacagcattgtaccttggcaaaagactgcagaacaaaacttacatcgaaaacaaggaagtcgtcctcccgtatttggta  2290
cttttctgccagaaaccgcaggcgagcctttaacactctggattcggtgtattggaaa                           2348
```

Figure 1

Figure 2. Alignment of UDP-glucose pyrophosphorylase (UGPase) sequences. Residues shown in white on a black background are conserved in at least 60% of the sequences; white on grey, 40%. G. gracilis, *Gracilaria gracilis* UGPase (GgUGP), this paper; human, *Homo sapiens* UGPase (NCBI accession number 731050); bovine, *Bos taurus* UGPase (731049); barley, *Hordeum vulgare* UGPase (2117937); yeast, *Saccharomyces cerevisiae* UGPase (UGP1, 1585157); potato, *Solanum tuberosum* UGPase(322794); pig, *Sus scrofa* UGPase (1752677); slime mold, *Dictyostelium discoideum* UGPase (136738); C. elegans, *Caenorhabditis elegans* UGPase (1326259).

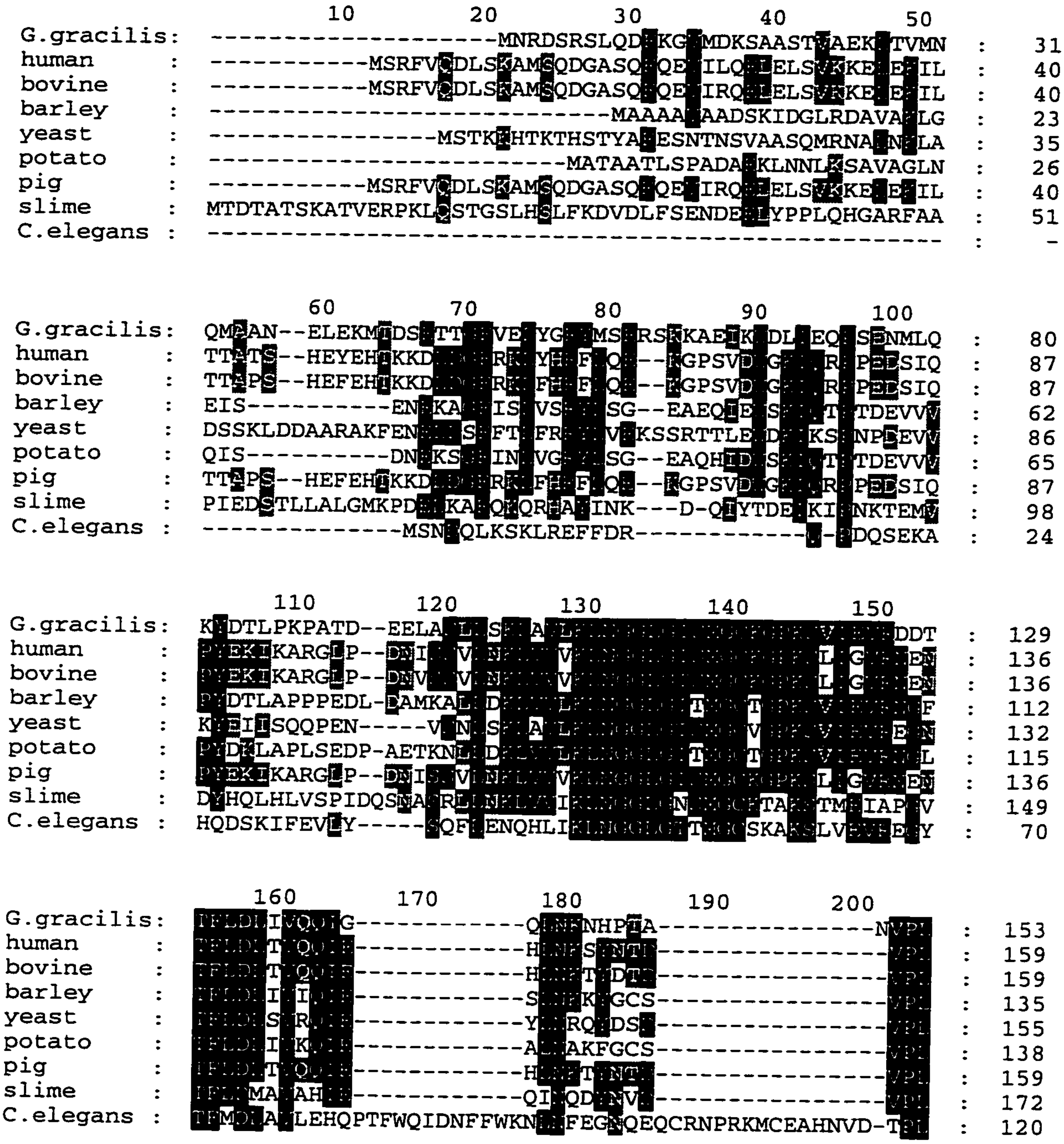
G.gracilis
human
bovine
barley
yeast
potato
pig
slime
C.elegans

Figure 2

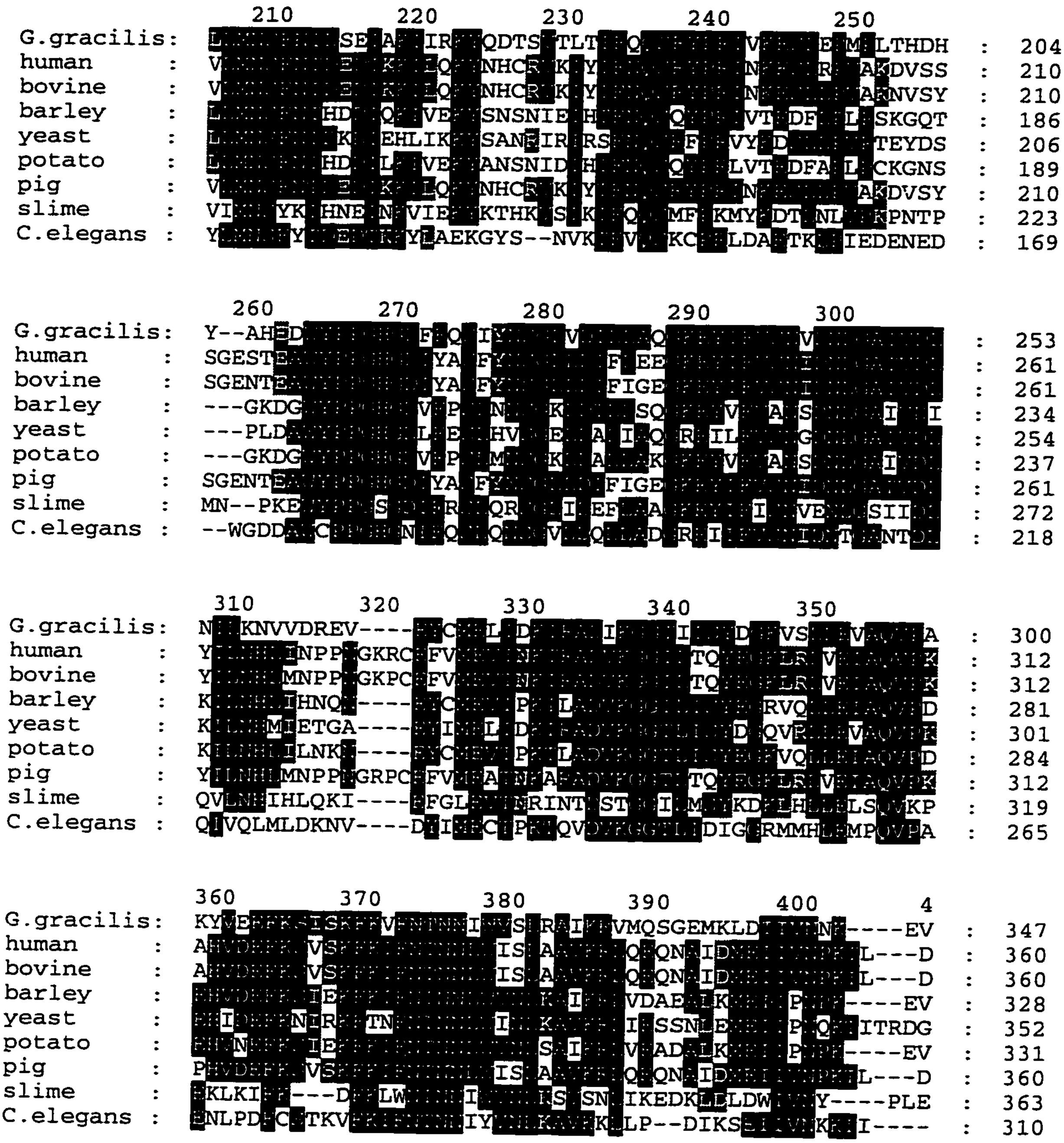

Figure 2, continued

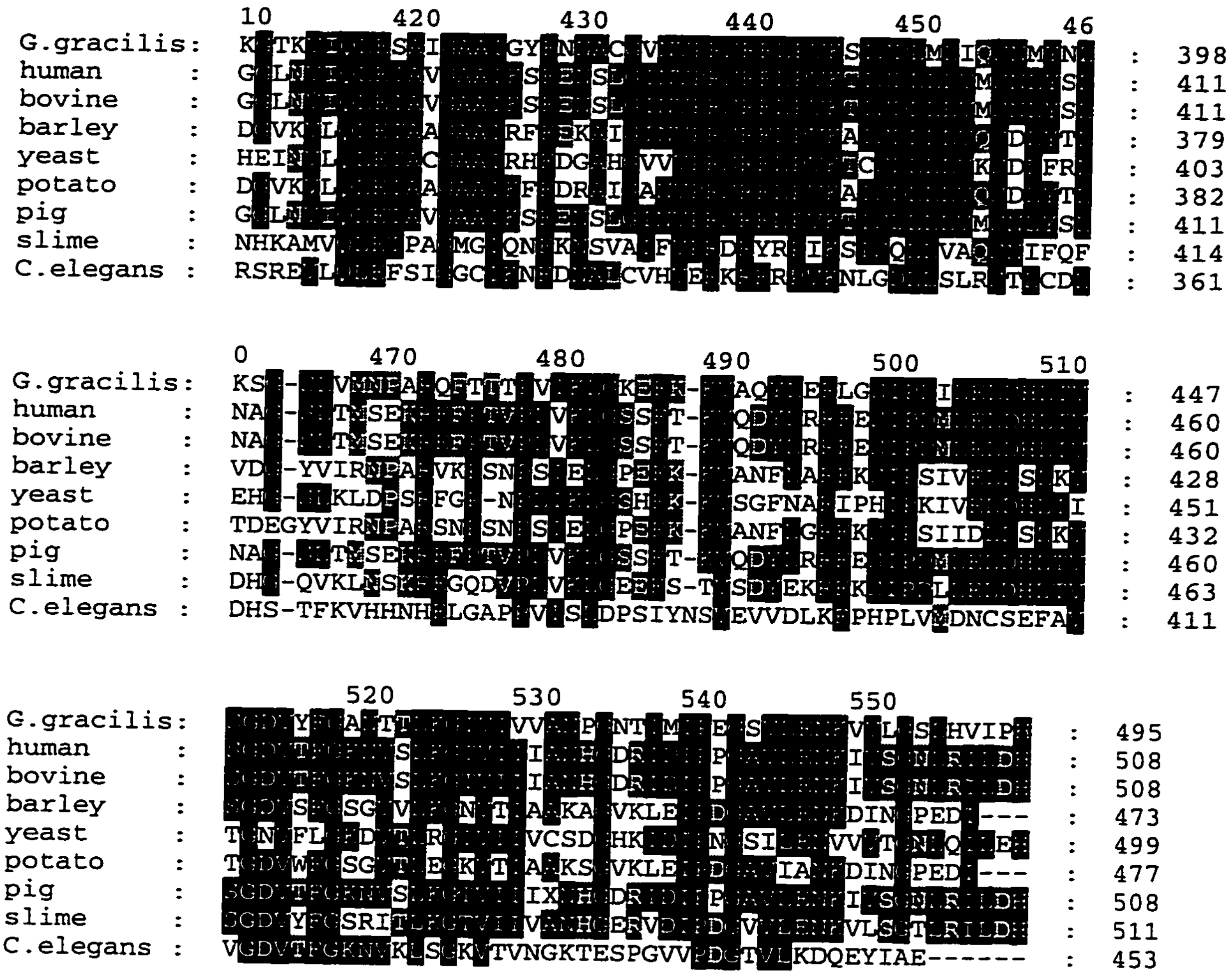

Figure 2, continued

(*GgGALT1*, Chapter III; *GgSBE1*, Chapter IV).

Features of the gene

A number of ATG codons in-frame with the UGPase-encoding ORF are present just downstream of the only potential TATA box (Fig. 1). In the absence of data on the location of the 5' end of the transcript, it is not possible to definitively identify the start codon. The first ATG codon downstream of the putative TATA box is in the right context as a translation initiation site, but is probably too close (6 bp). Another ATG codon, 30 bp downstream of the putative TATA box (Fig. 1), also appears to be in the right context for translation initiation; its 5'-flanking sequence (GCC**ATG**, the start codon in bold) conforms to the canonical sequence, RCY**ATG**, at translation initiation sites in red algal genes so far characterized (Zhou and Ragan 1996). We thus provisionally designate this codon as the translation start site of the UGPase gene.

The amino acid sequence deduced from the genomic sequence of *GgUGP* shows significant sequence similarity with other UGPase sequences throughout the whole length of the deduced protein (Fig. 2); no region in the deduced sequence appears to be an "insertion" relative to the other sequences, suggesting that *GgUGP* is devoid of introns. Other intronless genes have already been reported in red algae, *e.g.*, genes for triose phosphate isomerase in *G. gracilis* (Zhou and Ragan 1995c), and GapC in *C. crispus* (Liaud *et al.* 1993). In contrast, the UGPase gene from potato is interrupted by 19 introns (Borovkov *et al.* 1997), while that of *Dictyostelium discoideum* by three introns (Ragheb and Dottin 1987).

Reverse transcription and PCR amplification of the 3' end of *GgUGP* transcripts were carried out to determine the location of polyA sites and polyA signals. Two PCR products differing in size were obtained and sequenced using primer sf4b; the sequence of the smaller fragment is shown in Fig. 1. The longer transcript was sequenced with another primer sf5 (Fig. 1). Alignment of the sequences with the genomic sequence (Fig. 1) confirms that there are two types of UGPase transcripts in *G. gracilis*, differing in size by 332 bp due to alternative polyadenylation sites. Each site apparently has its own polyA signal, AUUAAA for the shorter transcript, found 33 bp upstream of the polyA site, and AAUAAA for the larger one, 32 bp upstream of the other polyA site. The polyA signal for the shorter transcript overlaps with the termination codon, a case observed with another *G. gracilis* gene, polyubiquitin (Zhou and Ragan 1995b). Interestingly, the polyubiquitin gene also has an alternative polyA site with its own polyA signal, and, just like the UGPase transcript, the polyA signal for the longer polyubiquitin transcript is also AAUAAA. The presence of UAAA signal in both polyadenylation sites in *GgUGP* and in the polyubiquitin gene probably indicates the importance of this highly conserved polyadenylation signal in *G. gracilis*. The UAAA motif has been observed in all gene trancripts so far characterized from *G. gracilis* (see Zhou and Ragan 1996; see also Chapter III and Chapter IV).

The relevance of alternative polyadenylation sites in the *GgUGP* transcript to the physiology of *G. gracilis* remains to be evaluated.

UGPase is single-copy in G. gracilis

The result of the Southern analysis is shown in Figure 3. The final wash was of moderate stringency (0.5X SSC/0.1% SDS, 65°C, 30 min, performed twice); at this stringency, multiple copies of other *G. gracilis* genes (*GgGALT1*, Chapter III; *GgSBE1*, Chapter IV) were clearly detected. However, in the case of UGPase, only one band was found per lane (Fig. 3), indicating that the UGPase gene is single-copy; in lane 3, a small band (< 1 kb) is observed, but this is explained by the presence of a *Xho* I restriction site, CTCGAG (positions 112-117 in Fig. 1), within the region spanned by the probe. UGPase from potato (Borovkov *et al.* 1996) has also been shown to be single-copy, although allelic isoforms have been described from this plant (Sowokinos *et al.* 1997).

The GgUGP protein and its phylogenetic relationship with other UGPases

As shown in Figure 2, GgUGP shares a high level of sequence identity (around 50%) with other eukaryotic UGPases, implying that results of structural studies on other UGPases may also apply to GgUGP, especially if highly conserved residues are involved. Affinity-labelling studies of the potato (Kazuta *et al.* 1991) and bovine (Konishi *et al.* 1993) UGPases have identified lysine residues (K336, K402, K444, K486, and K488, numbering according to the

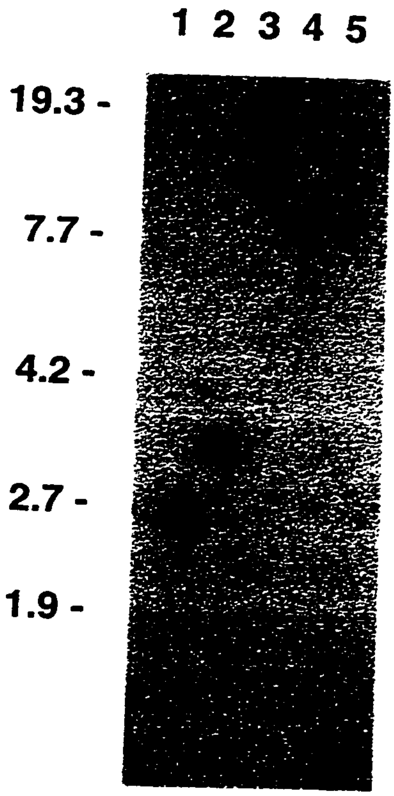

Figure 3. Determination of *GgUGP* copy number by Southern hybridization analysis. Genomic DNA (5 μg per reaction) from *G. gracilis* was digested with *Eco*R I (lane 1), *Hind* III (lane 2), *Sal* I (lane 3), *Xba* I (lane 4), and *Xho* I (lane 5). The probe was prepared from a *GgUGP* fragment (positions -33 to 424, Fig. 1; this region does not contain restriction sites for the enzymes used in digesting the genomic DNA (*Eco*R I, *Hind* III, *Sal* I, and *Xba* I), except for *Xho* I, which has a site, CTCGAG, [positions 112-117] within this fragment). Final washing was with 0.5X SSC/0.1% SDS at 65°C for 30 min, performed twice. The numbers on the left of the figure indicate the size of the markers, in kb.

alignment in Fig. 2) that are at, or close to, the substrate-binding site, three of which (K444, K336, and K402) have been confirmed by site-directed mutagenesis studies to be important to enzyme activity (Katsube *et al.* 1991); K444 is particularly critical to enzyme activity, while K336 and K402 are believed to be involved in binding pyrophosphate or α-D-glucose-1-phosphate. These residues are highly conserved among the UGPases, including GgUGP. There are, however, highly conserved residues whose mutation does not adversely affect enzyme activity, at least in the human liver UGPase (Chang *et al.* 1996); examples of such mutations include C140S, H311R, W263S, R437H, R471Q, and R495H (numbering according to the alignment in Fig. 2). The highly conserved H311 is substituted in GgUGP with N.

Trees were constructed to determine the phylogenetic relationship of GgUGP with other UGPases. To determine the effects of insertion-deletion regions ("indels") in the alignment on the tree topology, analyses that include and exclude indels were performed. The results are shown in Figure 4. Inclusion/exclusion of indels caused only a small difference on tree topology. The plant UGPases (from a monocot and from a dicot) formed a stable group with 100% bootstrap support, as did the animal UGPases. GgUGP and the UGPases from *S. cerevisiae* and *C. elegans* branched off between animal and plant UGPases, although the order of branching is uncertain, as indicated by low bootstrap support. The placement of *C. elegans* UGPase is problematic, as its sequence lacks strong similarity (< 32%) with those of other animals. The

Figure 4. Phylogenetic trees of selected UGPases. Distances were calculated using the formula of Kimura (1983); insertion-deletion regions ("indels") were either included in (**A**) or excluded from (**B**) the analysis. Numbers represent bootstrap values out of 100. Full names of the sequences are given in Figure 2. The scale bars (top, left in each figure) indicate the branch length corresponding to 0.1 (A) or 0.02 (B) substitution per site.

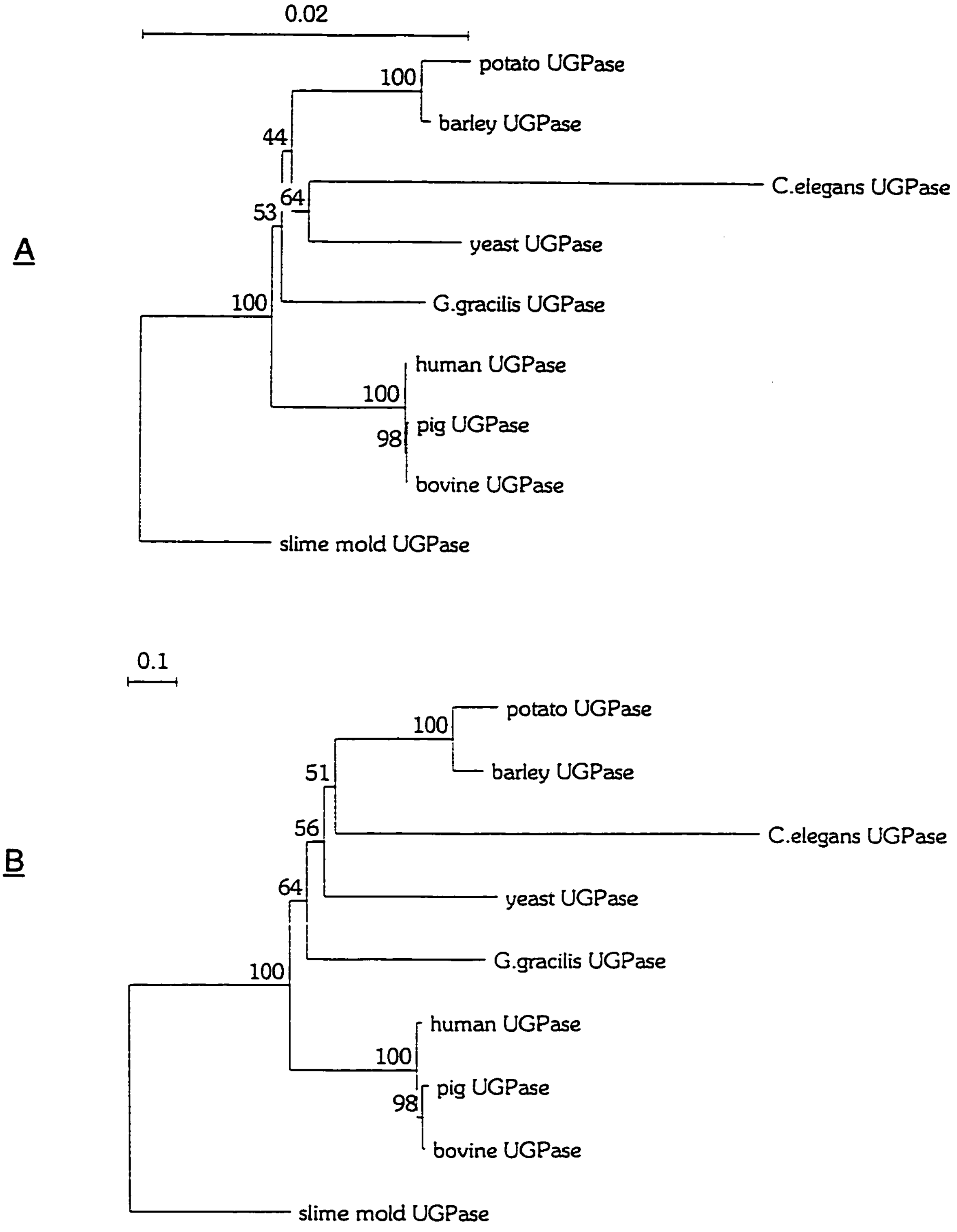
A
0.02
potato UGPase
barley UGPase
C.elegans UGPase
yeast UGPase
G.gracilis UGPase
human UGPase
pig UGPase
bovine UGPase
slime mold UGPase
100
44
64
53
100
100
98
B
0.1
potato UGPase
barley UGPase
C.elegans UGPase
yeast UGPase
G.gracilis UGPase
human UGPase
pig UGPase
bovine UGPase
slime mold UGPase
100
51
56
64
100
100
98

Figure 4

trees apparently support the general conclusion that GgUGP is as related to plant as it is to fungal and animal UGPases.

An intron-containing gene occurs just downstream from GgUGP

We sequenced the region downstream from *GgUGP* and used the sequence to search the NCBI database (using BLASTX). The search yielded interesting results.

The conceptual translation of the sequence has significant similarity with C-terminal portions of DNA helicases (Fig. 5A), indicating that this region (italicized in Fig. 1) is likely the 3' portion of a potential DNA helicase gene (encoded on the complementary strand relative to the *GgUGP*). The stop codon is 376 bp downstream of the stop codon of *GgUGP*. The proximity of these two genes is yet another indication that, as noted in Chapter III, proximity of genes in the *G. gracilis* genome may not be uncommon; the proximity of polyubiquitin and mACN genes (Zhou and Ragan 1995a) and GALT and PTH genes (Chapter III) has already been observed.

A potential 96-bp, phase-0 intron was also found in this sequence. As indicated by the alignment (Fig. 5A), the otherwise strong sequence similarity is interrupted by the presence in the *G. gracilis* sequence of a 32 amino acid insertion that contains an in-frame stop signal. Inspection of the nucleotide sequence (Fig. 5B) reveals that the amino acid insertion is "encoded" by a likely

```
G. gracilis     ......FQYTESRVLKARLRFLAEKYQIREDDFLVFDVSFVLQSFAKVQCCARRNLTVLLL
S. pombe        ...624YQYTESRVLKARLEFLRDTYQIREADFLTFD------------------------
M. musculus     ...625YVYTQSRILKARLEYLRDQFQIRENDFLTFD------------------------
D. discoideum   ...621---TESKVLRARLEFLRDRYQIRENEFLTFD------------------------

G. gracilis     NRHQNGQ*AMRQAAQCAGRVIRNKNDYGIVIFADKRFTRAKLRSKLPKWIAQFLSVDSLDL
S. pombe        --------AMRHAAQCLGRVLRGKDDHGIMVLADKRYGRSDKRTKLPKWIQQYITEGATNL
M. musculus     --------AMRHAAQCVGRAIRGKTDYGLMVFADKRFARADKRGKLPRWIQEHLTDSNLNL
D. discoideum   --------AMRTASQCVGRVIRGKSDYGIMIFADKRYNRLDKRNKLPQWILQFCQPQHLNL

G. gracilis     DLGTAIAEARTFLLDMAQPSVAKSR--VEEKPPENYEFIGSRTAPE*--------------
S. pombe        STDMSLALAKKFLRTMAQPFTASDQEGISWWSLDDLLIHQKKALKSAAIEQSKHEDEMDID
M. musculus     TVDEGVQVAKYFLRQMAQPFHREDQLGLSLLSLEQLQSEETLQRIEQIAQQL*--------
D. discoideum   STDMAISLSKTFLREMGQPFSREEQLGKSLWSLEHVEKQSTSKPPQQQNSAINSTITTSTT

G. gracilis     ---------------
S. pombe        VVET*----------
M. musculus     ---------------
D. discoideum   TTTTTSTISETHLT*
```

Figure 5A. Alignment of the peptide fragment deduced from the portion of a putative DNA helicase gene found downstream of *GgUGP* (italicized portion in Figure 1) with the C-terminal portions of DNA helicase homologs from *Mus musculus* (NCBI accession number 2114484), *Dictyostelium discoideum* (RepD, 2058510), and *Schizosaccharomyces pombe* (RAD15, 5022); numbers at the start of sequences indicate positions in the original sequence of the first residue in the fragment. - indicates gaps inserted to achieve alignment; * indicates a position encoded by a stop codon. Alignment was produced using CLUSTAL W (Thompson *et al.* 1994).

tttccaatacaccgaatccagagtgttaaaggctcgcctgcggtttctggcagaaaagtaccaatacgggaggacga
F Q Y T E S R V L K A R L R F L A E K Y Q I R E D D

cttccttgttttcgat***gt***_aagtttgttctgcagtcttttgccaaggtacaatgctgtgctaggagaatcta_***ac***_agt_
F L V F D

tcttcttctgaatcgtcaccaaaacggacaat***ag***gcaatgcgccaagcggctcagtgcgcagggcgtgtaattcggaa
A M R Q A A Q C A G R V I R N

caagaacgattacggaatcgttatattcgccgacaaacggtttacgcgtgccaaactaagaagcaagcttccgaaatg
K N D Y G I V I F A D K R F T R A K L R S K L P K W

gattgcccagttcctgtctgttgactccttggatctcgatctcggtaccgcaattgctgaagcacgtacttttctgct
I A Q F L S V D S L D L D L G T A I A E A R T F L L

tgacatggcgcaaccatcggtagctaagtctcgagtggaggagaagccaccggaaaactatgaattcatcggaagccg
D M A Q P S V A K S R V E E K P P E N Y E F I G S R

gaccgcaccagagtaa
T A P E *

Figure 5B. The nucleotide sequence of the 3'-portion of a putative gene encoding a DNA helicase homolog in *G. gracilis*; this sequence is complementary to the italicized portion in Figure 1 (positions 2348-1865). The conceptual translation is shown below the nucleotide sequence. The portion likely to be an intron is italicized; the dinucleotides conserved at the 5' and 3' ends (GT and AG, respectively) and at the branch site (AC) in red algal spliceosomal introns (Zhou and Ragan 1996) are in bold and underlined.

intron whose sequences at the potential 5' (AT:GTAAGT; : indicates the presumptive exon-intron junction) and 3' (TAG:G) splice sites and at the potential branch site (CTAAC) conform to the canonical sequences for spliceosomal introns in red algal genes (Zhou and Ragan 1996). This intron is apparently close to the 3' end of the gene; all other red algal genes characterized to date are devoid of introns at the 3' region.

Conclusion

We have cloned the UGPase gene, apparently single-copy, from the red alga *G. gracilis*. The characterization of this gene paves the way for more intensive studies of the biochemistry and molecular biology of UGPase in this alga. The gene produces two transcripts that differ in length due to the presence of alternative polyadenylation sites; the possibility that this constitutes a mechanism for regulating gene expression remains to be investigated. The observation that another gene, potentially encoding a DNA helicase, is located just downstream of *GgUGP* indicates that occurrence of closely-spaced genes may not be unusual in the *G. gracilis* genome.

Chapter VI

Conclusions

Currently, little is known about molecular biology of carbohydrate metabolism in red algae. This work was intended as a contribution to the effort of elucidating this aspect of red algal biology. The results of this work are the following: the generation of ESTs, and the cloning and characterization of three genes, encoding galactose-1-P uridylyltransferase (the gene was named *GgGALT1*), a starch branching enzyme (*GgSBE1*), and UDPglucose pyrophosphorylase (*GgUGP*), from the red alga *G. gracilis*.

One of the important outcomes of this work is the demonstration of the utility of ESTs for cloning carbohydrate metabolism-related genes. The availability of ESTs facilitated the production of homologous probes used to screen genomic libraries. Three genes, encoding galactose-1-P uridylyltransferase, transaldolase, and 6-phosphogluconate dehydrogenase, were among those successfully isolated from a genomic library using ESTs as probes.

This work has also demonstrated the utility of ESTs for cloning genes involved in a variety of other cellular processes. Genes for tryptophan synthase β subunit, methionine sulfoxide reductase, and Silent Information Regulator 2 have been cloned using ESTs as probes. As indicated in Table1 (Chapter II) a variety of genes can be targeted for cloning using the ESTs; the further characterization of these genes would facilitate investigations into various areas of red algal molecular biology.

This work has also identified the major features in the primary structure of the cloned genes and their putative products. Biochemical experiments will be necessary to ascertain the functional significance of these features. In addition, the characterization of the three genes has yielded information relevant to our understanding of the structure and organization of red algal genes. In particular, the results of this work support the following generalizations:

1. Red algal genes are intron-poor. The three genes described in this thesis (*GgGALT1*, *GgSBE1*, and *GgUGP*) all lack apparent introns. However, an instance of an intron occurring close to the 3' end of a gene (found downstream of *GgUGP*) was also observed.

2. An invariant UAAA sequence is an important component of the polyA signal in red algal genes. In the transcripts of each of the three genes characterized, a UAAA sequence was found close to and upstream of the polyadenylation sites. The UAAA (as TAAA) sequence was also observed downstream of the protein-coding region of putative genes whose mRNAs remain to be sequenced (*e.g.*, a gene encoding peptidyl tRNA hydrolase [PTH] downstream of *GgGALT1*, and a gene encoding DNA helicase downstream of *GgUGP*). In addition, each of the

two polyA sites found in the *GgUGP* gene appears to have its own UAAA-containing polyadenylation signal.

3. Occurrence of closely-spaced genes may not be uncommon in the genome of *G. gracilis*. Two of the characterized genes, *GgGALT1* and *GgUGP1*, are each located close to another gene (potentially encoding PTH and a DNA helicase, respectively). That two *G. gracilis* (as *G. verrucosa*) genes, polyubiquitin and mACN, are located close to each other has already been reported (Zhou and Ragan 1995a). Interestingly, the 3' ends of the transcripts of *GgGALT1* and *GgUGP1* apparently overlap those of PTH and DNA helicase genes, respectively; because these neighboring genes are encoded on opposite strands, the overlapping regions are thus complementary and can potentially anneal by base-pairing. Whether this has functional significance (*e.g.*, as a mechanism for regulating the expression of the genes) remains to be investigated; however, the physiological roles of *GgGALT1* and *GgUGP1* appear to be unrelated to those of their downstream neighbors.

Although these generalizations are consistent with what is known about red algal genes so far characterized, more red algal genes need to be studied to establish the validity of these generalizations.

Prospects

Studies have shown that the control of flux through metabolic pathways is distributed among the enzymes in the pathways (for example, see Srere 1994 and references therein). The distributive nature of flux control, which has also been demonstrated for the starch biosynthetic pathway in land plants (Smith *et al.* 1995), thus makes it desirable to include for analysis as many enzymes as possible when studying metabolic pathways.

To study enzymes intensively, cloning and characterization of the genes encoding the enzymes would be an advantage, if not a necessity. Currently, however, only a handful of red algal genes has been cloned. Thus, at this stage, cloning and characterizing more genes, perhaps using the EST approach, deserves high priority if we are to advance our understanding of the molecular biology and biochemistry of carbohydrate metabolism in red algae. On the other hand, progress can also be made by carrying out more intensive studies of those genes that have already been sequenced. Their expression as a function of environmental (such as light and the concentration of inorganic nitrogen and phosphorus) and other variables can now be investigated, for example, by using cloned DNA as probes. Enzymological properties of the enzymes can also be studied, perhaps using expression systems to produce recombinant enzymes and biochemical techniques such as site-directed mutagenesis. Clearly, a variety of studies is awaiting to be done.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and Human Genome Project. Science 252: 1651-1656

Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CA, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, *et al.* [86 co-authors] (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377 (Suppl.): 3-174

Ainsworth C, Clark J, Balsdon J (1993) Expression, organisation and structure of the genes encoding the waxy protein (granule-bound starch synthase) in wheat. Plant Mol Biol 22: 67-82

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410

Apt KE, Grossman AR (1992) A polyubiquitin cDNA from a red alga. Plant Physiol 99: 1732-1733

Apt KE, Hoffman NE, Grossman AR (1993) The γ subunit of R-phycoerythrin and its possible mode of transport into the plastid of red algae. J Biol Chem 268: 16208-16215

Armisen R (1995) World-wide use and importance of *Gracilaria*. J Appl Phycol 7: 231-243

Aspinall GO (1970) Polysaccharides. Pergamon Press, Oxford

Avery OT, MacLeod CM, McCarty M (1944) Studies on the chemical nature of the substance inducing transformation of *Pneumonococcus* types. J Exp Med 79: 137-158

Baba T, Kimura K, Mizuno K, Etoh H, Ishida Y, Shida O, Arai Y (1991) Sequence conservation of the catalytic regions of amylolytic enzymes in maize branching enzyme - I. Bichem Biophys Res Commun 181: 87-94

Barrow KD, Karsten U, King RJ, West JA (1995) Floridoside in the genus *Laurencia* (Rhodomelaceae: Ceramiales) - a chemosystematic study. Phycologia 34: 279-283

Bean B, Hassid WZ (1955) Assimilation of $^{14}CO_2$ by a photosynthesizing red alga, *Iridophycus flaccidum* . J Biol Chem 212: 411-425

Bird CJ, Kain JM (1995) Recommended names of included species of Gracilariaceae. J Appl Phycol 7: 335-338

Bird CJ, Ragan MA, Critchley AT, Rice EL, Gutell RR (1994) Molecular relationships among the Gracilariaceae (Rhodophyta): further observations on some undetermined species. Eur J Phycol 29: 195-202

Boguski MS (1995) The turning point in genome research. Trends Biochem Sci 20: 295-296

Borovkov AY, McClean PE, Secor GA (1997) Organization and transcription of the gene encoding potato UDP-glucose pyrophosphorylase. Gene 186: 293-297

Bouget FY, Kerbourc'h C, Liaud MF, Loiseaux de Goër S, Quatrano RS, Cerff R, Kloareg B (1995) Structural features and phylogeny of the actin gene of *Chondrus crispus* (Gigartinales, Rhodophyta). Curr Genet 28: 164-172

Brachmann CB, Sherman JM, Devine SE, Cameron EE, Pillus L, Boeke JD (1995) The SIR2 gene family, conserved from bacteria to humans, functions in silencing, cell cycle progression, and chromosome stability. Genes Dev 9: 2888-2902

Browner MF, Nakano K, Bang AG, Fletterick RJ (1989) Human muscle glycogen synthase cDNA sequence: a negatively charged protein with an asymmetric charge distribution. Proc Natl Acad Sci 86: 1443-1447

Burton RA, Bewley JD, Smith AM, Bhattacharyya MK, Tatge H, Ring S, Bull V, Hamilton WDO, Martin C (1995) Starch branching enzymes belonging to distinct enzyme families are differentially expressed during pea embryo development. Plant J 7: 3-15

Chang H-Y, Peng H-L, Chao YC, Duggleby RG (1996) The importance of conserved residues in human liver UDPglucose pyrophosphorylase. Eur J Biochem 236: 723-728

Charng Y-Y, Kakefusa G, Iglesias AA, Buikema WJ, Preiss J (1992) Molecular cloning and expression of the gene encoding ADP-glucose pyrophosphorylase from the cyanobacterium *Anabaena* sp. strain PCC 7120. Plant Mol Biol 20: 37-47

Charng Y-Y, Sheng J, Preiss J (1995) Mutagenesis of an amino acid residue in the activator-binding site of cyanobacterial ADP-glucose pyrophosphorylase causes alteration in activator specificity. Arch Biochem Biophys 318: 476-480

Chen EY (1994) The efficiency of automated DNA sequencing. In: Adams MD, Fields C, Venter JC (eds) Automated DNA sequencing and analysis. Academic Press, London, pp 3-10

Claverie J-M (1995) Exploring the vast territory of uncharted ESTs. In Browne MJ, Thurlby PL (eds), Genomes, Molecular Biology and Drug Discovery. Academic Press, London: 55-71

Commission on Plant Gene Nomenclature (1994) Nomenclature of sequenced plant genes. Plant Mol Biol Reptr 12: S2-S109.

Craigie JS (1974) Storage products. In: Stewart WDP (ed) Algal Physiology and Biochemistry. Blackwell Scientific Publications Ltd, Oxford, pp 206-235

Craigie JS (1990) Cell walls. In: Cole KM, Sheath RG (eds) Biology of the Red Algae. Cambridge University Press, Cambridge. pp 221-257.

De Luca P, Moretti A (1983) Floridosides in *Cyanidium caldarium*, *Cyanidioschyzon merolae* and *Galdieria sulphuraria* (Rhodophyta, Cyanidiophyceae). J Phycol 19: 368-369

Delseny M, Cooke R, Raynal M, Grellet F (1997) The *Arabidopsis thaliana* cDNA sequencing projects. FEBS Letters 403: 221-224

Dey PM (1985) D-galactose-containing oligosaccharides. In: Dey PM, Dixon RA (eds) Biochemistry of Storage Carbohydrates in Green Plants, Academic Press, London. pp 53-129

Dujon B (1996) The yeast genome project: what did we learn? Trends Genet 12: 263-270

Ekman P, Yu S, Pedersen M (1991) Effects of altered salinity, darkness, and algal nutrient status on floridoside and starch content, α-galactosidase activity and agar yield of cultivated *Gracilaria sordida*. Br Phycol J 26: 123-131

Elsas LJ, Dembure PP, Langley S, Paulk EM, Hjelm LN, Fridovich-Keil J (1994) A common mutation associated with the Duarte galactosemia allelle. Am J Hum Genet 54: 1030-1036

Farkas I, Hardy TA, DePaoli-Roach AA, Roach PJ (1990) Isolation of the *GSY1* gene encoding yeast glycogen synthase and evidence for the existence of a second gene. J Biol Chem 265: 20879-20886

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166

Field TL, Reznikoff WS, Frey PA (1989) Galactose1-phosphate uridylyltransferase: identification of histidine-164 and histidine 166 as critical residues by site-directed mutagenesis. Biochemistry 28: 2094-2099.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, *et al.* [40 co-authors] (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512

Fredrick JF (1968) Glucosyltransferase isozymes in algae-II. Properties of branching enzymes. Phytochemistry 7: 931-936

Fredrick JF (1971) Polyglucan branching isoenzymes of algae. Physiol Plant 24: 55-58

Frey PA (1996) The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. FASEB J 10: 461-470

Frey PA, Wong L-J, Sheu K-F, Yang S-L (1982) Galactose-1-phosphate uridyltransferase: detection, isolation, and characterization of the uridyl enzyme. Methods Enzymol 87: 20-36

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

Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, *et al.* [633 co-authors] (1997) The yeast genome directory. Nature (Lond.) 387 (suppl.): 5-105

Gottlieb S, Esposito RE (1989) A new role for a yeast transcriptional silencer gene, SIR2, in regulation of recombination in ribosomal DNA. Cell 56: 771-776

Gross KC, Phar DM (1982) A potential pathway for galactose metabolism in *Cucumis sativus* L., a stachyose transporting species. Plant Physiol 69: 117-121

Gross W, Schnarrenberger C (1995a) Heterotrophic growth of two strains of the acido-thermophilic red alga *Galdieria sulphuraria*. Plant Cell Physiol 36: 633-638

Gross W, Schnarrenberger C (1995b) Purification and characterization of a galactose-1-phosphate: UDP-glucose uridyltransferase from the red alga *Galdieria sulphuraria*. Eur J Biochem 234: 258-263

Guan H, Li P, Imparl-Radosevich J, Preiss J, Keeling P (1997) Comparing the properties of *Escherichia coli* branching enzyme and maize branching enzyme. Arch Biochem Biophys 342: 92-98

Guo Z, Sherman F (1996) 3'-end-forming signals of yeast mRNA. TIBS 21: 477-481

Heidenreich RA, Mallee J, Segal S (1993) Rat galactose-1-phosphate uridyltransferase coding sequence, transcription start site and genomic organization. DNA Seq 3: 311-318

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89: 10915-10919

Herbers K, Sonnewald U (1996) Manipulating metabolic partitioning in transgenic plants. Trends Biotechnol 14: 198-205

Hillier LaD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, *et al.* [33 co-authors] (1996) Generation and analysis of 280,000 human expressed sequence tags. Genome Res 6: 807-828

Hunt AG (1994) Messenger RNA 3' end formation in plants. Annu Rev Plant Physiol Plant Mol Biol 45: 47-60

Ishida N, Miura N, Yoshioka S, Kawakita M (1996) Molecular cloning and characterization of a novel isoform of the human UDP-galactose transporter, and of related complementary DNAs belonging to the nucleotide-sugar transporter gene family. J Biochem 120: 1074-1078

Jensen A (1993) Present and future needs for algae and algal products. Hydrobiologia 260/261: 15-23

Jespersen HM, MacGregor EA, Henrissat B, Sierks MR, Svensson B (1993) Starch- and glycogen-debranching and branching enzymes: prediction of structural features of the catalytic $(\beta/\alpha)_8$-barrel domain and evolutionary relationship to other amylolytic enzymes. J Protein Chem 12: 791-805

Katsube T, Kazuta Y, Tanizawa K, Fukui T (1991) Expression in *Escherichia coli* of UDP-glucose pyrophosphorylase cDNA from potato tuber and functional assessment of the five lysyl residues located at the substrate-binding site. Biochemistry 30: 8546-8551

Kawasaki T, Mizuno K, Baba T, Shimada H (1993) Molecular analysis of the gene encoding a rice starch branching enzyme. Mol Gen Genet 237: 10-16

Kazuta Y, Omura Y, Tagaya M, Nakano K, Fukui T (1991) Identification of lysyl residues located at the substrate-binding site in UDP-glucose pyrophosphorylase from potato tuber: affinity labelling with uridine di- and triphosphopyridoxals. Biochemistry 30: 8541-8545

Kimura M (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge

Kirst GO (1980) Low MW carbohydrates and ions in rhodophyceae: quantitative measurement of floridoside and digeneaside. Phytochem 19: 1107-1110

Kleczkowski LA (1994) Glucose activation and metabolism through UDP-glucose pyrophosphorylase in plants. Phytochemistry 37: 1507-1515

Kremer BP (1978) Patterns of photoassimilatory products in Pacific Rhodophyceae. Can J Bot 56: 1655-1659

Kremer BP, Kirst GO (1981) Biosynthesis of 2-*O*-D-glycerol-α-D-galactopyranoside (floridoside) in marine Rhodophyceae. Plant Sci Lett 23: 349-357

Kremer BP, Kirst GO (1982) Biosynthesis of photosynthates and taxonomy of algae. Z Naturforsch 37c: 761-771

Kretz K, Callen W, Hedden V, Kaderli M (1993) Improved primers for the Bluescript phagemid vector. [Stratagene] Strategies 6: 15-16

Kuriki T, Guan H, Sivak M, Preiss J (1996) Analysis of the active center of branching enzyme II from maize endosperm. J Protein Chem 15: 305-313

Laurenson P, Rine J (1992) Silencers, silencing, and heritable transcriptional states. Microbiol Rev 56: 543-560

Leslie ND, Immerman EB, Flach JE, Florez M, Fridovich-Keil JL, Elsas LJ (1992) The human galactose-1-phosphate uridyltransferase gene. Genomics 14: 474-480

Levine RL, Mosoni L, Berlett BS, Stadtman ER (1996) Methionine residues as endogenous antioxidants in proteins. Proc Natl Acad Sci USA 93: 15036-15040

Liaud M-F, Valentin C, Brandt U, Bouget F-Y, Kloareg B, Cerff R (1993) The GAPDH gene system of the red alga *Chondrus crispus*: promotor structures, intron/exon organization, genomic complexity and differential expression of genes. Plant Mol Biol 23: 981-994

Liaud M-F, Brandt U, Cerff R (1995) The marine red alga *Chondrus crispus* has a highly divergent β-tubulin gene with a characteristic 5' intron: functional and evolutionary implications. Plant Mol Biol 28: 313-325

Macler BA (1986) Regulation of carbon flow by nitrogen and light in the red alga, *Gelidium coulteri*. Plant Physiol 82: 136-141

Manley SL, Burns DJ (1991) Formation of nucleoside diphosphate monosaccharides (NDP-sugars) by the agarophyte *Pterocladia capillacea* (Rhodophyceae). J Phycol 27: 702-709

Manners DJ (1985) Starch. In: Dey PM, Dixon RA (eds) Biochemistry of Storage Carbohydrates in Green Plants, Academic Press, London. pp. 149-203

Maretzki A, Thom M (1978) Characteristics of a galactose-adapted sugarcane cell line grown in suspension culture. Plant Physiol 61: 544-548

Martin C, Smith AM (1995) Starch biosynthesis. Plant Cell 7: 971-985

McCracken DA, Cain JR (1981) Amylose in floridean starch. New Phytol 88: 67-71

Meng J, Srivastava LM (1990) Extraction, assay and some properties of floridoside phosphate synthase from *Porphyra perforata* (Rhodophyta). J Phycol 26: 683-688

Meng J, Srivastava LM (1993) Variations in floridoside content and floridoside phosphate synthase activity in *Porphyra perforata* (Rhodophyta). J Phycol. 29: 82-84

Miura N, Ishida N, Hoshino M, Yamauchi M, Hara T, Ayusawa D, Kawakita M (1996) Human UDP-galactose translocator: molecular cloning of a complementary DNA that complements the genetic defect of a mutant cell line deficient in UDP-galactose translocator. J Biochem 120: 236-241

Muñoz P, Norambuena L, Orellana A (1996) Evidence for a UDP-glucose transporter in Golgi apparatus-derived vesicles from pea and its possible role in polysaccharide biosynthesis. Plant Physiol 112: 1585-1594

Müller-Röber B, Sonnewald U, Willmitzer L (1992) Inhibition of the ADP-glucose pyrophosphorylase in transgenic potatoes leads to sugar-storing tubers and influences tuber formation and expression of tuber storage protein genes. EMBO J 11: 1229-1238

Nagashima H, Nakamura S, Nisizawa K, Hori T (1971) Enzymic synthesis of floridean starch in a red alga, *Serraticardia maxima*. Plant Cell Physiol 12: 243-253

Nakata PA, Okita TW (1995) Differential regulation of ADP-glucose pyrophosphorylase in the sink and source tissues of potato. Plant Physiol 108: 361-368

Nakata PA, Okita TW (1996) *Cis*-elements important for the expression of the ADP-glucose pyrophosphorylase small-subunit are located both upstream and downstream from its structural gene. Mol Gen Genet 250: 581-592

Pazur JH, Shadaksharaswamy M (1961) Galactose-1-phosphate uridyl transferase activity in soybean extracts. Biochem Biophys Res Commun 5: 130-134

Pearson WR (1991) Identifying distantly related protein sequences. Curr Opin Struct Biol 1: 321-326

Percival E, McDowell R (1967) Chemistry and enzymology of Marine Algal Polysaccharides. Academic Press, London.

Preiss J (1991) Biology and molecular biology of starch synthesis and its regulation. In: Miflin BJ, Miflin HF (eds) Oxford Surveys of Plant Molecular and Cell Biology, vol 7. Oxford University Press, Oxford. pp 59-114

Pueschel CM (1990) Cell Structure. In: Cole CM, Sheath RG (eds) Biology of the Red Algae. Cambridge University Press, Cambridge, pp. 7-41

Ragan MA, Gutell RR (1995) Are red algae plants? Bot J Linn Soc 118: 81-105

Ragheb JA, Dottin RP (1987) Structure and sequence of a UDP glucose pyrophosphorylase gene of *Dictyostelium discoideum*. Nucleic Acids Res 15: 3891-3906

Raven JA, Johnston AM, MacFarlane JJ (1990) Carbon metabolism. In: Cole KM, Sheath RG (eds) Biology of the Red Algae. Cambridge University Press, Cambridge, pp 171-202

Romeo T, Kumar A, Preiss J (1988) Analysis of the *Escheria coli* glycogen gene cluster suggests that catabolic enzymes are encoded among biosynthetic genes. Gene 70: 363-376

Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC, Kerlavage AR (1996) The construction of *Arabidopsis* expressed sequence tag assemblies. Plant Physiol 112: 1177-1183

Ruzicka FJ, Wedekind JE, Kim J, Rayment I, Frey PA (1995) Galactose-1-phosphate uridylyltransferase from *Escherichia coli*, a zinc and iron metalloenzyme. Biochemistry 34: 5610-5617

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular Cloning. A Laboratory Manual. Second Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, vol. 3: A.1 and A.4.

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tomé P, *et al.* [104 co-authors] (1996) A gene map of the human genome. Science 274: 540-546

Sheath RG, Hellebust JA, Sawa T (1979) Floridean starch metabolism of *Porphyridium purpureum* (Rhodophyta) I. Changes during ageing of batch culture. Phycologia 18: 149-163

Sheath RG, Hellebust JA, Sawa T (1981) Ultrastructure of the floridean starch granule. Phycologia 20: 292-297

Sheng J, Charng Y-Y, Preiss J (1996) Site-directed mutagenesis of lysine[382], the activator-binding site, of ADP-glucose pyrophosphorylase from *Anabaena* PCC 7120. Biochemistry 35: 3115-3121

Smith AM, Denyer K, Martin CR (1995) What controls the amount and structure of starch in storage organs? Plant Physiol 107: 673-677.

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

Sowokinos JR, Thomas C, Burrel MM (1997) Pyrophosphorylases in potato V. Allelic polymorphism of UDP-glucose pyrophosphorylase in potato cultivars and its association with tuber resistance to sweetening in the cold. Plant Physiol 113: 511-517

Srere P (1994) Complexities of metabolic regulation. Trends Biochem Sci 19: 519-520

Staden R (1996) The Staden sequence analysis package. Mol Biotechnol 5: 233-241

Stark DM, Timmerman KP, Barry GF, Preiss J, Kishore GM, (1992) Regulation of the amount of starch in plant tissues by ADP glucose pyrophosphorylase in potatoes. Science 258: 287-292

Steentoft M, Irvine LM & Farnham WF (1994) Two terete species of *Gracilaria* and *Gracilariopsis* (Gracilariales, Rhodophyta) in Britain. Phycologia 34: 113-127

Stitt M (1995) Regulation of metabolism in transgenic plants. Annu Rev Plant Physiol Plant Mol Biol 46: 341-368

Su J-C, Hassid WZ (1962) Carbohydrates and nucleotides in the red alga *Porphyra perforata* II. separation and identification of nucleotides. Biochemistry 1: 474-480

Svensson B (1994) Protein engineering in the α-amylase family: catalytic mechanism, substrate specificity, and stability. Plant Mol Biol 25: 141-157.

Tajima M, Nogi Y, Fukusawa T (1985) Primary structure of the *Saccharomyces cerevisiae* GAL7 gene. Yeast 1: 67-77

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680

Thon VJ, Vigneron-Lesens C, Marianne-Pepin T, Montreuil J, Decq A, Rachez C, Ball AG, Cannon JF (1992) Coordinate regulation of glycogen metabolism in the yeast *Saccharomyces cerevisiae*. Induction of glycogen branching enzyme. J Biol Chem 267: 15224-15228

Thorbjørnsen T, Villand P, Kleczkowski LA, Olsen O-A (1996) A single gene encodes two different transcripts for the ADP-glucose pyrophosphorylase small subunit from barley (*Hordeum vulgare*). Biochem J 313: 149-154

Toma L, Pinhal MAS, Dietrich CP, Nader HB, Hirschberg CB (1996) Transport of UDP-galactose into the Golgi lumen regulates the biosynthesis of proteoglycans. J Biol Chem 271: 3897-3901

Turnquist RL, Hansen RG (1973) Uridine diphosphoryl glucose pyrophosophorylase. The Enzymes, 3rd ed. vol 8, pp. 51-71

Van de Peer Y, De Wachter R (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Applic Biosci 10: 569-570

Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. Nature (Lond.) 171: 964-967

Wedekind JE, Frey PA, Rayment I (1995) Three-dimensional structure of galactose-1-phosphate uridylyltransferase from *Escherichia coli* at 1.8 Å resolution. Biochemistry 34: 11049-11061

Wedekind JE, Frey PA, Rayment I (1996) The structure of nucleotidylated histidine-166 of galactose1-phosphate uridylyltransferase provides insight into phosphoryl group transfer. Biochemistry 35: 11560-11569.

White O, Dunning T, Sutton G, Adams M, Venter JC, Fields C (1993) A quality control algorithm for DNA sequencing projects. Nucleic Acids Res 21: 3829-3838

Wolfsberg TG, Landsman D (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucleic Acids Res 25: 1626-1632

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

Wu L, Ueda T, Messing J 1995) The formation of mRNA 3' ends in plants. Plant J 8: 323-329

Wu Y, Gretz MR (1993) Stable isotope labelling method for studies of saccharide metabolism in Agardhiella subulata. Hydrobiologia 260/261: 595-600

Yager TD, Zewert TE, Hood LE (1994) The Human Genome Project. Acc Chem Res 27: 94-100

Yu S, Pedersen M (1990) The α-galactosidase of *Gracilaria tenuistipitata* and *G. sordida* (Gracilariales, Rhodophyta) Phycologia 29: 454-460

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

Zhou Y-H, Ragan MA (1994) Cloning and characterization of the nuclear gene encoding plastid glyceraldehyde-3-phosphate dehydrogenase from the marine red alga *Gracilaria verrucosa*. Curr Genet 26: 79-86

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

Zhou Y-H, Ragan MA (1995b) Characterization of the polyubiquitin gene in the marine red alga *Gracilaria verrucosa*. Biochim Biophys Acta 1261: 215-222

Zhou Y-H, Ragan MA (1995c) Cloning and characterization of the nuclear gene and cDNAs for triosephosphate isomerase of the marine red alga *Gracilaria verrucosa*. Curr Genet 28: 317-323.

Zhou Y-H, Ragan MA (1995d) The nuclear gene and cDNAs encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase from the marine red alga *Gracilaria verrucosa*: cloning, characterization and phylogenetic analysis. Curr Genet 28: 324-332

Zhou Y-H, Ragan MA (1996) Nuclear-encoded protein-encoding genes of the agarophyte *Gracilaria verrucosa* (Gracilariales, Rhodophyta). Hydrobiologia 326/327: 429-436

Zrenner R, Willmitzer L, Sonnewald U (1993) Analysis of the expression of potato uridinediphosphate-glucose pyrophosphorylase and its inhibition by antisense RNA. Planta 190: 247-252

World Wide Web URLs referred to in text:

ABIView software: http://www.paranoia.com/~dhk/abiview.html

Arabidopsis thaliana genome database: http://genome-www.stanford.edu/Arabidopsis/AGI/ and http://www.mips.biochem.mpg.de/mips/ATHALIANA

BCM Search Launcher: http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/

BLASTX at NCBI: http://www.ncbi.nlm.nih.gov

Caenorhabditis elegans genome database: http://eatworms.swmed.edu/genome.shtml

dbEST database: http://www.ncbi.nlm.nih.gov/dbEST/

Gracilaria ESTs at IMB: http://www.nrc.ca/imb/***** [remainder of URL to be supplied in proof].

Human genome database:http://www.ornl.gov/TechResources/Human_Genome/project/launcher.html

Microbial genome database: http://www.tigr.org/tdb/mdb/mdb/html project.html

Yeast genome database: http://www.mips.biochem.mpg.de/mips/YEAST and http://genome-www.stanford.edu/Saccharomyces/

IMAGE EVALUATION TEST TARGET (QA-3)

150mm

6"

APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989