*Journal of Applied Phycology* **9:** 287–293, 1997.



# Expressed sequence tags (ESTs) from the marine red alga *Gracilaria gracilis*

Arturo O. Lluisma[1,2,3] & Mark A. Ragan[1,4*]
[1]*Institute for Marine Biosciences, National Research Council of Canada, 1411 Oxford St, Halifax, NS, Canada B3H 3Z1*
[2]*Department of Biology, Dalhousie University, Halifax, NS, Canada B3H 4J1*
[3]*Present address: Marine Science Institute, University of the Philippines, 1101 Quezon City, Philippines*
[4]*Canadian Institute for Advanced Research, Program in Evolutionary Biology*
(*Author for correspondence. Phone: 902-426-1674; fax 902-426-9413; e-mail: mark.ragan@nrc.ca)

Received 1 May 1997; revised and accepted 31 July 1997

*Key words:* EST, expressed sequence tag, *Gracilaria*, carbohydrate biosynthesis

**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*.

## 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 & 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 *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.

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. protein structure and cellular metabolism they are not only unnecessary, but 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 pro-

vide no information on what genes are expressed in a given organism or tissue at any specific time.

These considerations motivated the development of the *e*xpressed *s*equence *t*ag (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 & 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 one 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. No algal species was represented in dbEST (release 053097) prior to submission of the sequences reported here.

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 & 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 & 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 $\mu$L 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 $\mu$L 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 $\mu$M dNTPs, diluted to 50 $\mu$L 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 $\mu$L of each reaction was electrophoresed 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 & 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 & States, 1993) under its default options through the BCM Search Launcher, a WWW-accessible mole-

cular 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 & Ragan (1993); the DNA (5 $\mu$g per reaction) was digested with restriction enzymes, electrophoresed 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: 4 $\times$, 2 $\times$, and 0.5 $\times$ 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 $\alpha$-$^{32}$P-dCTP. Probes based on the *G. gracilis* aconitase gene (Zhou & 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/*****), 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 at 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 & 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 & 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

*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://www.nrc.ca/imb/home/raganma/esthome.html. 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 dbEST match(es) | 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 10NR | 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 |
| 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 |

*Table 1.* Continued.

| EST# | putative ID/homolog | BLASTX score (max) | Database dbEST match(es) | Accession no. |
|---|---|---|---|---|
| **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 |

(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.

of tryptophan 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 cellular response to oxidative stress involves two major components, the methionine residues in proteins, which act as endogenous scavengers of oxidants, and MSR, which subsequently reduces methionine sulfoxides (oxidized methionines) back to methionines. These ESTs could be used as probes, e.g. in Northern hybridizations, to determine whether these or other culture conditions 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 & 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 & 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; Gaasterland & 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.

## Acknowledgments

We thank Colleen A. Murphy for expert technical assistance; Y.-H. Zhou for access to the cDNA library; and P. F. Shacklock for assistance with the *Gracilaria* cultures. A.O.L. thanks the International Development Research Centre (Canada) and the Patrick F. Lett Endowment for financial support. M.A.R. thanks the Canadian Institute for Advanced Research, Program in Evolutionary Biology for fellowship support. Issued as NRCC no. 39762.

## 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: 3–174.

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

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

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

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.

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

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.

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

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.

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.

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: 3–10.

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

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

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.

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 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.

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.

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

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

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

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

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

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.

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.

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

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

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

Wolfsberg TG, Landsman D (1997) A comparison of expressed sequence tags (ESTs) to human genomic sequences. Nucl. 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.

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

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 (1995a) Characterization of the polyubiquitin gene in the marine red alga *Gracilaria verrucosa*. Biochim. biophys. Acta 1261: 215–222.

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

**World Wide Web URLs in text**

ABIView software: http://www.paranoia.com/đhk/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/home/raganma/esthome.html
Human genome database:http://www.ornl.gov/TechResources/Human_Genome/project/ launcher.html
Microbial genome database: http://www.tigr.org/tdb/mdb/mdb.html
Yeast genome database: http://www.mips.biochem.mpg.de/mips/YEAST and http://genome-www.stanford.edu/Saccharomyces/