*J. Phycol.* **35,** 433–437 (1999)

# SUBMISSION AND RETRIEVAL OF AN ALIGNED SET OF NUCLEIC ACID SEQUENCES[1]

*Susan H. Brawley*[2]

School of Marine Sciences, University of Maine, Orono, Maine 04469

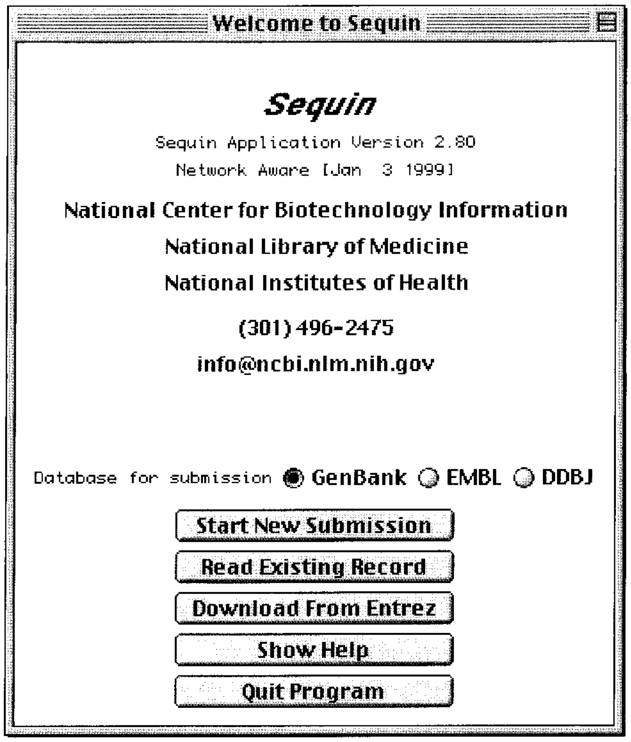
**An important part of many studies published in the *Journal of Phycology* is the construction and analysis of an alignment of related nucleotide sequences. These alignments are valuable to other investigators, especially when the data provide a basis for more than one possible alignment. Publication of alignments, however, is prohibitively expensive, and the data are difficult to retrieve and use in this form. Several electronic archives for these alignments now exist, and information is provided here on the submission and retrieval of alignments at the National Center for Biotechnology Information (NCBI), the European Bioinformatics' Institute (EBI), and TreeBASE. Emphasis is placed on the use of *Sequin* and *Entrez* at the NCBI, because these programs have been modified recently and they provide a fluid and integrated system to submit and retrieve sequence and alignment records and related information. Other useful features of *Entrez* (e.g. free Medline searches) are noted.**

***Key index words:*** **alignments; bioinformatics; *Entrez*; phylogeny; population biology; *Sequin***

The *Journal of Phycology* and most other journals no longer publish the full alignment of nucleic acid sequences from an author's phylogenetic, populational, or other molecular study. A small, key portion of the alignment can still be included as a figure in the manuscript, but several archives are now available for the full alignments. This not only reduces publication cost to journals, but more importantly, makes the information far more usable by the scientific community because of its permanent storage in electronic form. The purpose of this editorial note is to provide guidance to authors for submission of their alignments to such archives and to illustrate how this information and related resources can be retrieved.

Archives for alignments include the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (Bethesda, Maryland) and the European Bioinformatics Institute (EBI, Cambridge, England). The procedures at NCBI (http://www.ncbi.nlm.nih.gov/) are emphasized because submission and retrieval of the alignment and related information are now possible through their programs *Sequin* and *Entrez* (see below). *Sequin* can be used to prepare a single sequence or an alignment of sequences for submission to either GenBank, EMBL, or DDBJ (the DNA Database of Japan). A submitter is requested to specify the database of choice for submission of the information on the first window in *Sequin* (Fig. 1). Data on individual nucleic acid sequences and their accession numbers are shared daily among all three databases (GenBank, EMBL, DDBJ), but in order for the alignment submitted through *Sequin* to be retrieved, it must be submitted to GenBank. The NCBI will not assign an unique number for the alignment, but the alignment will be retrievable by reference to any GenBank accession number within the alignment, as specified below in the discussion of *Entrez*.

The alignment database at the European Bioinformatics Institute (http://www.ebi.ac.uk/) is available at an FTP site (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/), and the file ''align.info'' contains information for submitters. The EBI assigns a unique number to the alignment, which should be included in the published article. With this accession number in hand, a reader can view the alignment by ftp (see address above) or by sending an e-mail message to netserv@ebi.ac.uk with GET ALIGN: (the published accession number).DAT in



**Sequin**

Sequin Application Version 2.80

Network Aware [Jan 3 1999]

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

(301) 496-2475

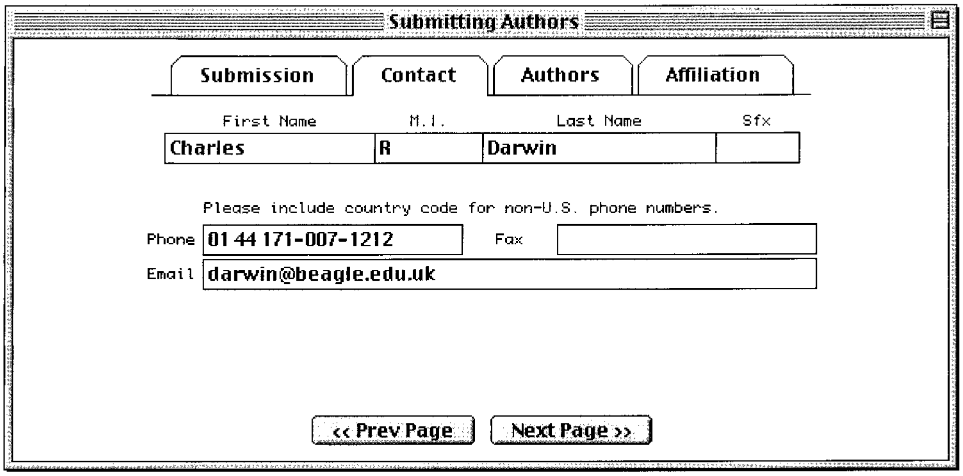
info@ncbi.nlm.nih.gov

Database for submission ● GenBank ○ EMBL ○ DDBJ

FIG. 1. The first window of *Sequin* (NCBI) in the Network-aware mode (reprinted from the *Sequin* Quick Guide).

[1] Editor's Comment.

[2] E-mail jphycol@maine.maine.edu



| First Name | M.I. | Last Name | Sfx |
|---|---|---|---|
| Charles | R | Darwin | |

Please include country code for non-U.S. phone numbers.

Phone 01 44 171-007-1212 Fax

Email darwin@beagle.edu.uk

FIG. 2. *Sequin* is organized as a set of ordered windows with folders that request information from an author. Here, the Contact tab was clicked, and Charles Darwin filled the appropriate information into this folder (reprinted from the *Sequin* Quick Guide).

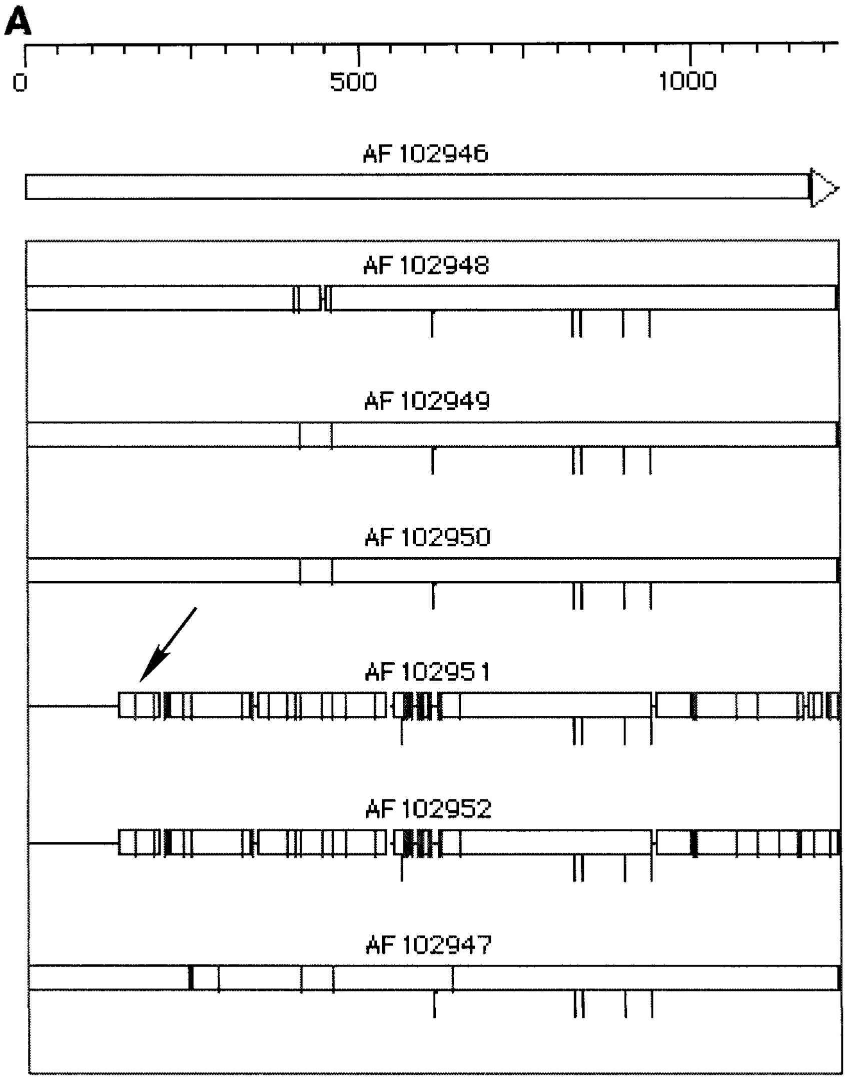


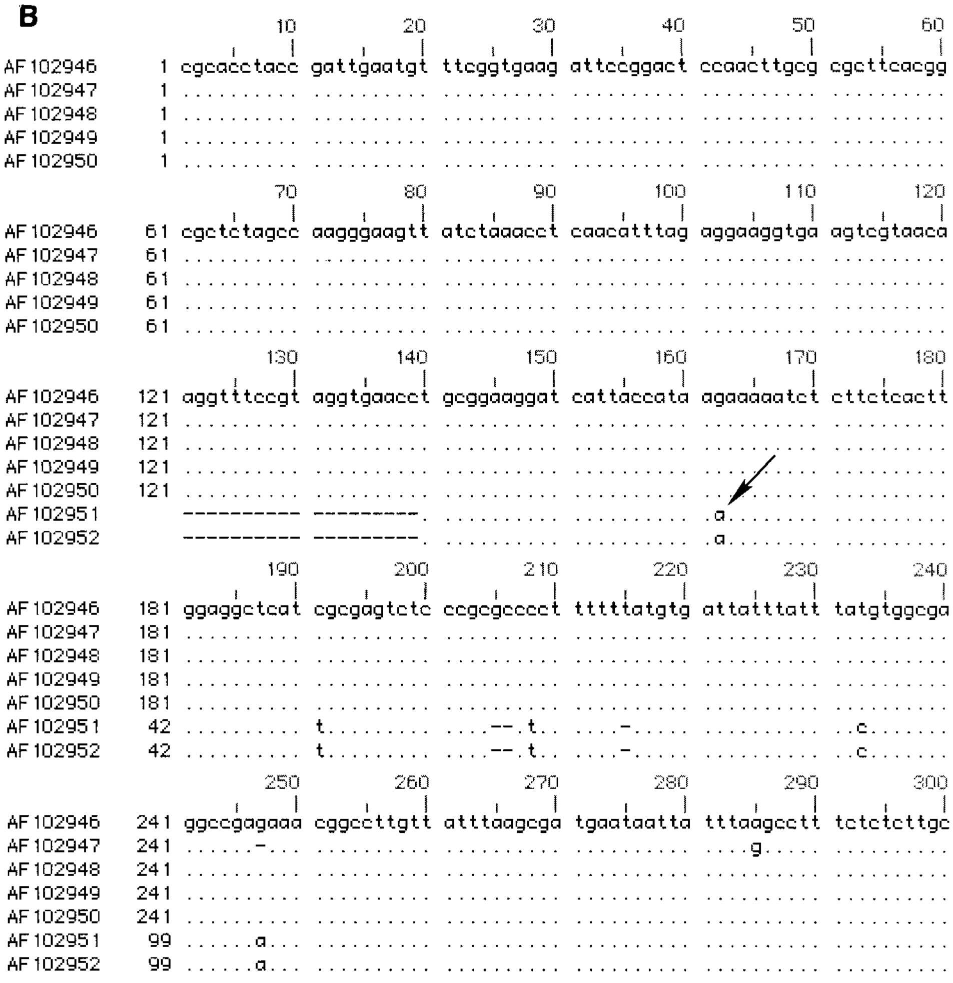
FIG. 3. Alignment view (A) and sequence view (B) of an alignment of rRNA sequences that include ITS1 and ITS2 from *Hesperophycus californicus* (AF102946-51) and *Pelvetiopsis limitata* (AF102951-52) (see Serrão et al. 1999). In the alignment view, the top sequence (AF102946 = GenBank accession number, *Hesperophycus californicus*) is the master sequence and differences in the other sequences compared to the master are indicated by vertical lines across each bar (substitutions), breaks that are connected by a horizontal line (gaps), or lines below the bar (insertions). The nucleotide sequence of the master sequence is printed in sequence view (B). This view takes much more space than the alignment view; thus, only the first 300 positions of the alignment are shown. The arrow (A, B) marks position 162, in which an adenine (A) is substituted for a guanine (G) in AF102951 (*Pelvetiopsis limitata,* California) and AF102952 (*Pelvetiopsis limitata,* Oregon), as shown graphically by a line across the bar in alignment view (A) and by the letter ''a'' in the appropriate position in sequence view.

the subject line of the message. The EBI requests that new sequences (i.e. lacking DDBJ, EMBL, or GenBank accession numbers) in the alignment be submitted separately in order to ensure that accession numbers are assigned to the individual sequences. Data matrices and the resulting phylogenetic trees also can be submitted to TreeBASE (http://www.herbaria.harvard.edu/treebase/) if they pertain to phylogenetic studies, and this site includes helpful information on topics such as the applicability and availability of various software packages for phylogenetic analysis. Any of these means of providing accessibility to an alignment is acceptable. My intention is to discourage authors from including the statement ‘‘alignment available from authors’’ in their Materials and Methods, as this offer does not represent an archival source for the information.

### *Submitting an Alignment with* Sequin

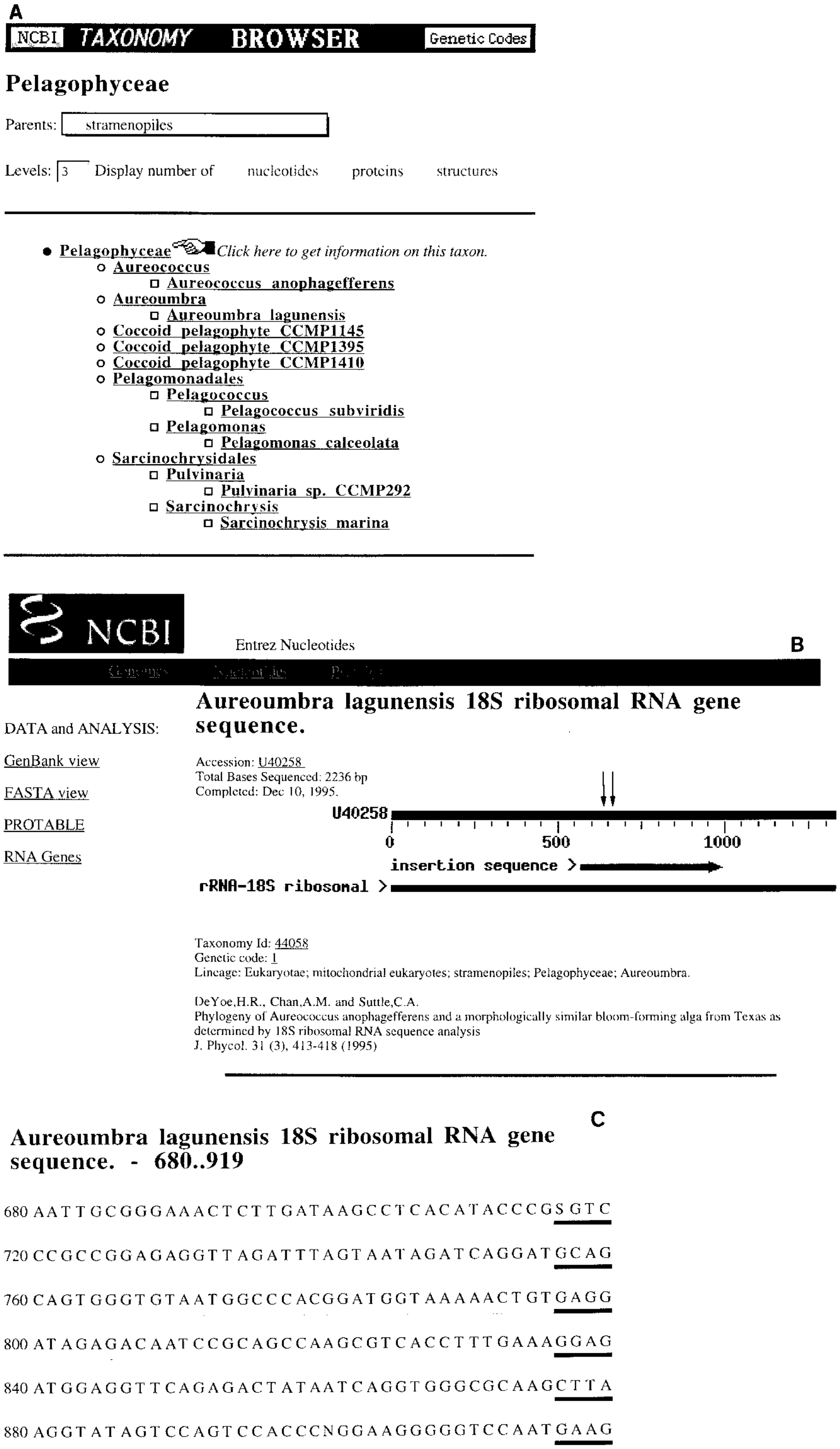
*Sequin* is listed under the heading ‘‘GenBank Sequence Database’’ on the NCBI home page (http://www.ncbi.nlm.nih.gov/). Clicking on *Sequin* brings up several sections that introduce its use, including a ‘‘Quick Guide’’ and instructions on how to download *Sequin* onto the user's computer in a stand-alone mode or in a Net-aware mode (‘‘How to get *Sequin*’’). Authors should move to the section under the Quick Guide labeled ‘‘Before You Begin’’ and review these materials carefully, especially the section ‘‘Sample Data Files’’ under ‘‘Preparing Nucleotide and Amino Acid Data.’’

*Sequin* is arranged as a series of logical windows, each of which prompts the submitter if some of the requested information has not been entered at the appropriate time. Windows are divided into folders and clicking on a folder tab opens a blank form; an author then types the information requested into the form (Fig. 2). The alignment should be prepared before it is imported into *Sequin* under the Nucleotide folder of the ‘‘Organism and Sequences’’ section. *Sequin* handles all of the standard formats in which alignments are prepared (e.g. NEXUS, FASTA, PHYLIP). The position at which annotation (e.g. scientific binomial, strain number, type of sequence, etc.) is attached to each individual sequence within the alignment varies depending on whether FASTA, NEXUS, or PHYLIP is used; authors should view examples in the *Sequin* Quick Guide before beginning the submission. The definition line should give the GenBank, DDBJ, or



```
                      10         20         30         40         50         60
                  |    |     |    |     |    |     |    |     |    |     |    |
AF102946    1 cgcacctacc gattgaatgt ttcggtgaag attccggact ccaacttgcg cgcttcacgg
AF102947    1 .......... .......... .......... .......... .......... ..........
AF102948    1 .......... .......... .......... .......... .......... ..........
AF102949    1 .......... .......... .......... .......... .......... ..........
AF102950    1 .......... .......... .......... .......... .......... ..........
                      70         80         90        100        110        120
                  |    |     |    |     |    |     |    |     |    |     |    |
AF102946   61 cgctctagcc aagggaagtt atctaaacct caacatttag aggaaggtga agtcgtaaca
AF102947   61 .......... .......... .......... .......... .......... ..........
AF102948   61 .......... .......... .......... .......... .......... ..........
AF102949   61 .......... .......... .......... .......... .......... ..........
AF102950   61 .......... .......... .......... .......... .......... ..........
                     130        140        150        160        170        180
                  |    |     |    |     |    |     |    |     |    |     |    |
AF102946  121 aggtttccgt aggtgaacct gcggaaggat cattaccata agaaaaatct cttctcactt
AF102947  121 .......... .......... .......... .......... .......... ..........
AF102948  121 .......... .......... .......... .......... .......... ..........
AF102949  121 .......... .......... .......... .......... .......... ..........
AF102950  121 .......... .......... .......... .......... .......... ..........
AF102951      ---------- ---------. .......... .......... .a........ ..........
AF102952      ---------- ---------. .......... .......... .a........ ..........
                     190        200        210        220        230        240
                  |    |     |    |     |    |     |    |     |    |     |    |
AF102946  181 ggaggctcat cgcgagtctc ccgcgcccct ttttatgtg attatttatt tatgtggcga
AF102947  181 .......... .......... .......... .......... .......... ..........
AF102948  181 .......... .......... .......... .......... .......... ..........
AF102949  181 .......... .......... .......... .......... .......... ..........
AF102950  181 .......... .......... .......... .......... .......... ..........
AF102951   42 .......... t......... ....--.t.. ....-..... .......... ..c.......
AF102952   42 .......... t......... ....--.t.. ....-..... .......... ..c.......
                     250        260        270        280        290        300
                  |    |     |    |     |    |     |    |     |    |     |    |
AF102946  241 ggccgagaaa cggccttgtt atttaagcga tgaataatta tttaagcctt tctctcttgc
AF102947  241 ......-... .......... .......... .......... ....g..... ..........
AF102948  241 .......... .......... .......... .......... .......... ..........
AF102949  241 .......... .......... .......... .......... .......... ..........
AF102950  241 .......... .......... .......... .......... .......... ..........
AF102951   99 ......a... .......... .......... .......... .......... ..........
AF102952   99 ......a... .......... .......... .......... .......... ..........
```

FIG. 3. *Continued.*



TAXONOMY BROWSER

**Pelagophyceae**

Parents: stramenopiles

Levels: 3 Display number of nucleotides proteins structures

- Pelagophyceae Click here to get information on this taxon.
  - ○ Aureococcus
    - □ Aureococcus anophagefferens
  - ○ Aureoumbra
    - □ Aureoumbra lagunensis
  - ○ Coccoid pelagophyte CCMP1145
  - ○ Coccoid pelagophyte CCMP1395
  - ○ Coccoid pelagophyte CCMP1410
  - ○ Pelagomonadales
    - □ Pelagococcus
      - □ Pelagococcus subviridis
    - □ Pelagomonas
      - □ Pelagomonas calceolata
  - ○ Sarcinochrysidales
    - □ Pulvinaria
      - □ Pulvinaria sp. CCMP292
    - □ Sarcinochrysis
      - □ Sarcinochrysis marina

Entrez Nucleotides

**Aureoumbra lagunensis 18S ribosomal RNA gene sequence.**

DATA and ANALYSIS:

GenBank view

FASTA view

PROTABLE

RNA Genes

Accession: U40258
Total Bases Sequenced: 2236 bp
Completed: Dec 10, 1995.

Taxonomy Id: 44058
Genetic code: 1
Lineage: Eukaryotae; mitochondrial eukaryotes; stramenopiles; Pelagophyceae; Aureoumbra.

DeYoe,H.R., Chan,A.M. and Suttle,C.A.
Phylogeny of Aureococcus anophagefferens and a morphologically similar bloom-forming alga from Texas as determined by 18S ribosomal RNA sequence analysis
J. Phycol. 31 (3), 413-418 (1995)

**Aureoumbra lagunensis 18S ribosomal RNA gene sequence. - 680..919**

```
680 A A T T G C G G G A A A C T C T T G A T A A G C C T C A C A T A C C C G S G T C

720 C C G C C G G A G A G G T T A G A T T T A G T A A T A G A T C A G G A T G C A G

760 C A G T G G G T G T A A T G G C C C A C G G A T G G T A A A A A C T G T G A G G

800 A T A G A G A C A A T C C G C A G C C A A G C G T C A C C T T T G A A A G G A G

840 A T G G A G G T T C A G A G A C T A T A A T C A G G T G G G C G C A A G C T T A

880 A G G T A T A G T C C A G T C C A C C C N G G A A G G G G G T C C A A T G A A G
```

FIG. 4. A window in the Taxonomy section of *Entrez* (A). Clicking on *Aureoumbra lagunensis* opens the window shown in (B). Clicking on (B) in the position indicated by the double arrows opens the window shown in (C), the actual sequence of the 18S rRNA gene between positions 680 and 919 of the submission. *Entrez* provides rapid and well-integrated access to many types of information.

EMBL accession number for any sequence that is included in an alignment but which already has an accession number. The GenBank staff assigns new numbers to unaccessioned sequences in the alignment and communicates these to the author by e-mail.

It is important to arrange the sequences within the alignment in an order of biological interest to take full advantage of analytical displays of the alignment in *Sequin.* For example, the first sequence in the alignment is the "master sequence." Thus, in a phylogenetic study, the master sequence might be the outgroup or the basal sequence within the ingroup, with other sequences arranged with respect to phylogenetic distance. In other studies, the master sequence might be a key isolate based on geography, etc. When sequences are arranged in such an order, their degree of similarity can be assessed quickly using an alignment view (i.e. using the Display Format pop-up menu and choosing Graphic). Differences between individual sequences in the alignment with respect to the master sequence are shown by red lines that indicate insertions, gaps, and substitutions. This is illustrated on the cover of this issue of the *Journal* and in Figure 3 as a comparison between the alignment view (Fig. 3A) and sequence view (Fig 3B). The sequence view is obtained in the Display Format menu by choosing Sequence.

Following entry of the requested information, *Sequin* carries out a validation procedure (a final review that the information is complete), then provides the e-mail address to which the finished record should be posted. Again, information on single-sequence accessions is exchanged among GenBank, DDBJ, and EMBL, but not information on alignments, so the alignment should be e-mailed to NCBI in order to make its retrieval possible. It must include at least one new sequence to which a GenBank number will be assigned.

### *Using* Entrez *to View an Alignment*

*Entrez* has been revised to accommodate searches for submitted alignments. At present, *Entrez* opens from the NCBI home page with the heading: Search WWW *Entrez* at NCBI with the following choices: Nucleotides, Proteins, 3Dstructures, Genomes, Taxonomy, Literature-PubMed. By the time this issue of the *Journal* is published, a new heading, Populations, will have been added to the list. Clicking on Populations permits a reader to access any alignment that has been submitted to GenBank by using a search strategy based on the published accession number for any GenBank sequence included within the alignment or by any other term the sequence would be indexed under. Please indicate in the Materials and Methods of your article in the *Journal of Phycology* that such an alignment is available at NCBI or elsewhere.

Using many of the other main menus in *Entrez* leads a viewer to information of phycological interest. For example, the complete genome of *Synechocystis* PCC6803 in both alignment and sequence views can be found by clicking into the Genomes section of *Entrez*. Clicking into Literature-PubMed and doing a literature search on cryptochrome (a blue-light photoreceptor) in Medline retrieved the references to 55 articles on 27 February 1999. Clicking on Taxonomy leads a viewer through a list of taxa organized as shown in Figure 4, which leads to an alignment or sequence view of nucleotide sequences for any taxon of interest (e.g. *Aureoumbra lagunensis*) for which information has been submitted to DDBJ, GenBank, or EMBL. In *Entrez*, one can move quickly between a relevant citation, the alignment and sequence views of a nucleotide submission, or taxonomic information associated with the sequence by clicking on the appropriate underlined item or on the alignment view of the sequence itself.

If revisions to a submission of an alignment are required (e.g. to change a scientific name or add a sequence), the original submitter can download the record from *Entrez* in the Net-aware mode of *Sequin* to make the changes that are required. Please contact the staff at NCBI for help if any difficulties in submission or revision of an alignment at NCBI are experienced (e-mail: info@ncbi.nlm.nih.gov; telephone: 301-496-2475).

I thank Drs. Jonathan Kans, Colombe Chappey, Ilene Mizrachi, and James Ostell of NCBI for their help and interest, both with respect to preparation of this article and for implementing procedures to archive alignments at NCBI. Very helpful information on TreeBASE was provided by Dr. W. Piel and on submission of alignments at EBI by Dr. Mary Ann Tuli. Drs. Joyce Longcore and Wytze Stam also provided valuable comments on an earlier version of this note.

Serrão et al. 1999. *J. Phycol.* 35:382-394.