J Mol Evol (1981) 17:334–347

Journal of
Molecular Evolution
© Springer-Verlag 1981

An Evaluation of the Phylogenetic Position of the Dinoflagellate *Crypthecodinium cohnii* Based on 5S rRNA Characterization

Alan G. Hinnebusch[1], Lynn C. Klotz[2], Roger L. Blanken[2], and Alfred R. Loeblich III[3]

[1]Department of Biochemistry, Wing Hall, Cornell University, Ithaca, New York 14853, USA
[2]Department of Biochemical Sciences, Princeton University, Princeton, New Jersey 08544, USA
[3]University of Houston, Marine Science Program, Galveston, Texas 77550, USA

Summary. Partial nucleotide sequences for the 5S and 5.8S rRNAs from the dinoflagellate *Crypthecodinium cohnii* have been determined, using a rapid chemical sequencing method, for the purpose of studying dinoflagellate phylogeny. The 5S RNA sequence shows the most homology (75%) with the 5S sequences of higher animals and the least homology (< 60%) with prokaryotic sequences. In addition, it lacks certain residues which are highly conserved in prokaryotic molecules but are generally missing in eukaryotes. These findings suggest a distant relationship between dinoflagellates and the prokaryotes. Using two different sequence alignments and several different methods for selecting an optimum phylogenetic tree for a collection of 5S sequences including higher plants and animals, fungi, and bacteria in addition to the *C. cohnii* sequence, the dinoflagellate lineage was joined to the tree at the point of the plant-animal divergence, well above the branching point of the fungi. This result is of interest because it implies that the well-documented absence in dinoflagellates of histones and the typical nucleosomal subunit structure of eukaryotic chromatin is the result of secondary loss, and not an indication of an extremely primitive state, as was previously suggested. Computer simulations of 5S RNA evolution have been carried out in order to demonstrate that the above-mentioned phylogenetic placement is not likely to be the result of random sequence convergence.

We have also constructed a phylogeny for 5.8S RNA sequences in which plants, animals, fungi and the dinoflagellates are again represented. While the order of branching on this tree is the same as in the 5S tree for the organisms represented, because it lacks prokaryotes, the 5.8S tree cannot be considered a strong independent confirmation of the 5S result. Moreover, 5.8S RNA appears to have experienced very different rates of evolution in different lineages indicating that it may not be the best indicator of evolutionary relationships.

We have also considered the existing biological data regarding dinoflagellate evolution in relation to our molecular phylogenetic evidence.

Key words: Dinoflagellates – Molecular evolution – Phylogeny – Chromatin – 5S RNA – 5.8S RNA – RNA sequence

Introduction

The dinoflagellates are a group of diverse eukaryotic algae possessing a number of unique cellular properties. These include characteristics of their cell covering, carotenoid pigment composition, the mechanism of mitosis, and the composition and fine structure of the chromosomes (Dodge 1973). The last feature is perhaps the most interesting: the chromosomes remain condensed throughout the cell cycle and fine-structural studies have revealed a fibrillar arrangement which resembles that seen in bacterial nucleoids (Giesbrecht 1962). Characterization of the chromatin of several free-living species has revealed a basic protein to DNA ratio 10–50 times lower than in other eukaryotic chromatin (Rizzo and Nooden 1974) and an absence of the nucleosomal subunit structure (Hamkalo and Rattner 1977; Rizzo and Burghardt 1980; C. Yen and P. M. M. Rae, personal communication) detected in all other nucleated organisms that have been examined (Horgen and Silver 1978). These features, along with the nuclear membrane association of dinoflagellate chromosomes during mitosis (Kubai and Ris 1969; Oakley and Dodge 1974) have led some to suggest that these algae may represent the first

Offprint requests to: A.G. Hinnebusch

0022–2844/81/0017/0334/$02.80

of the extant eukaryotic groups to diverge from the eukaryotic line following its split from the common prokaryotes, prior to the evolution of the histone-based subunit structure of chromatin (the Mesokaryote hypothesis; Dodge 1965; Loeblich 1976).

In contrast to these prokaryotic affinities, dinoflagellate nuclei contain large and highly variable quantities of DNA, comparable to higher plant and animal nuclei and atypical of the prokaryotes and most other protists (Cavalier-Smith 1978). One free-living species, *Crypthecodinium cohnii*, has been shown to contain a large fraction (about 40%) of its DNA in repeated sequences which occur, in part, finely interspersed with non-repetitive sequences in a manner closely resembling that of most higher eukaryotic genomes (Hinnebusch et al. 1980). In addition, the other organelles occupying the dinoflagellate cell appear to be no less evolved than in other eukaryotic algae (Kubai and Ris 1969; Dodge 1973) and show definite affinites with other flagellate groups (Taylor 1976). The pigment composition also suggests relatedness with other eukaryotic algae, particularly the chromophytes (Ragan and Chapman 1978). Finally, histochemical studies have detected more typical quantities of basic chromosomal proteins in certain parasitic dinoflagellates than occur in most free-living species (Ris and Kubai 1974; Hollande 1974). In view of these properties, one must consider the possibility that dinoflagellates do not represent the most primitive present-day eukaryotes, but instead are degenerate forms which secondarily lost the otherwise highly conserved features of eukaryotic chromatin structure. Indeed, in recent attempts at constructing a phylogeny of the protists, this group has not been considered the most primitive of the Eukaryota (Sagan 1967; Cavalier-Smith 1975; Taylor 1976; Ragan and Chapman 1978).

We decided to approach the problem of dinoflagellate phylogeny by characterizing the primary structures of the small RNA molecules (5S and 5.8S) found in the cytoplasmic ribosomes of the dinoflagellate, *C. cohnii*. A comparison of these sequences with those known from other organisms can then be used to construct an evolutionary tree for the organisms involved. This approach has been justified by the construction of a vertebrate phylogeny from cytochrome c sequences that is generally consistent with the existing paleontological and biological data (Fitch and Margoliash 1967). In addition, 5S RNA trees have been published which consistently cluster together organisms known to belong to the same phylogenetic groupings (Schwartz and Dayhoff 1978; Hori and Osawa 1979). Molecular phylogenetics is particularly appropriate for the Protista because of the limited and inconclusive nature of the fossil record that exists for this group (Loeblich 1974) and the paucity of reliable morphological phyletic markers among living protists (Taylor 1976). The results of our analysis suggest that the absence of histones and nucleosomes in dinoflagellate chromatin is probably not an indication of a close evolutionary relationship with the Prokaryota.

Methods

Isolation and Sequencing of RNA. Total RNA was extracted from late log-phase axenic cultures of *C. cohnii* growing in MLH (Tuttle and Loeblich 1975). Cells were harvested at 2,500 X g, washed in ten volumes of 0.35 M NaCl, 0.01 M Tris-HCl (pH 8), 1 mM EDTA (SET buffer), and ground in liquid N_2 to a fine powder. After resuspending in SET, the lysate was repeatedly extracted with phenol:chloroform:isoamyl alcohol (1:1:0.02) at room temperature. Cold ethanol was added, DNA was wound out, and the remaining nucleic acids were collected at 10,000 X g. The pellet was rinsed in 95% ethanol, dried in vacuo, and resuspended in 10 mM Tris-HCl (pH 8), 10 mM NaCl, 10 mM $MgCl_2$. Electrophoretically purified DNase I (Worthington), further purified by chromatography on agarose-coupled aminophenylphosphoryl-uridine-2'(3')-phosphate (Maxwell et al. 1977), was added to 100 μg/ml and incubated at 37°C for 20 min with constant agitation. RNA was extracted as above, precipitated with ethanol, and resuspended at 65°C in 8 M urea, 1/2 X TBE (1 X TBE: 50 mM Tris-borate (pH 8.3), 1 mM EDTA) and 0.05% each of bromophenol blue and xylene cyanol.

RNA was fractionated on 3 mm-thick 10% polyacrylamide, 7 M urea, 1 X TBE slab gels, using a 4% stacking gel (Rubin 1975). The 5S and 5.8S RNAs were identified by coelectrophoresis with yeast 5S and 5.8S standards (gifts from D. Peattie) and were extracted from the gel, end-labeled at their 3' termini, repurified on 0.45 mm 10% polyacrylamide-urea gels and sequenced, all according to Peattie (1979). In the sequencing reactions, twice the amount of labeled RNA was used in the C > U reaction than in the other three base-specific reactions because of a higher background observed on gels for this reaction. In addition, a 10 min reaction time was employed instead of 20–30 min. Thin (0.45 mm) 40 cm-long sequencing gels run at 1.6 kV were used throughout, except for determination of the first few nucleotides from the end, where 1.5 mm-thick, 20 cm-long gels were used and run at 1.0 kV.

Isolation of RNA from Ribosomes. Cytoplasmic ribosomes were isolated according to Werner-Schlenzka et al. (1978), except that they were collected at 40,000 rpm in the Beckman rotor 60 Ti for 90 min. The ribosomal pellet was resuspended in SET, phenol extracted, and ethanol precipitated as above. The RNA was examined on 1.5 mm 10% polyacrylamide-urea gels. Electrophoresis on 3% polyacrylamide-urea gels with known standards verified the occurrence of the large rRNAs in this fraction.

Construction of Evolutionary Trees. C. cohnii 5S and 5.8S RNA sequences were aligned with other known sequences using published alignments for these molecules (Hori and Osawa 1979; Schwartz and Dayhoff 1978; Pavlakis et al. 1979). Alignment was straightforward for 5S RNA, but several alternatives were considered for the collection of 5.8S RNA sequences (see Results). In the calculation of difference matrices from the alignments, a gap in a sequence was counted as only a single difference in comparison with a non-gapped sequence, regardless of the gap length. A gap aligned with another gap was counted as zero sequence difference. Difference matrices were corrected for reverse and duplicate mutations using either the formula of Holmquist (1972) which assumes completely random substitution, or an analogous correction of our own (see Appendix) which assumes an equal probability for the single transition and the two transversions which can occur at any given site (i.e., the transition is twice as likely as either transversion). The length of the sequences assumed in applying these corrections will be indicated as the results are discussed.

The corrected, or in some cases, the uncorrected difference matrix is the starting point in applying the various methods we have employed to select the optimum evolutionary tree. Two of the methods used examine all possible tree topologies and select the one which minimizes a particular property of the tree. For each tree topology (order of branching) considered, the first step is to assign lengths to the branches of the tree. To do so, we have used an unweighted multinomial least-squares analysis which uses only the data in the difference matrix (Blanken et al. submitted). In the first method, we calculate the percent-standard-deviation ($\%s_D$) of the difference matrix calculated from the assigned branch lengths of each topology from the observed difference matrix:

$$\%s_D = [\sum_{i=1}^{N} \sum_{j=1}^{N} \{100\,(d_{ij} - f_{ij})/d_{ij}\}^2 / \{N\,(N-1)/2\}]^{1/2}$$

where N is the number of sequences and d_{ij} and f_{ij} are the ijth elements of the observed and reconstructed difference matrices, respectively. The topology which minimizes this quantity is selected. This method, which will be referred to as the PSD method (Percent-Standard-Deviation), is a modification of the matrix method introduced by Fitch and Margoliash (1967). In the second method, the topology which minimizes the sum of the absolute values of the tree branch lengths, i.e. the most parsimonious tree, is selected. This method will be called the SBL method (Sum-of-Branch-Lengths) and is the same as that introduced by Dayhoff (1978) except for the procedure used to assign branch lengths and the requirement exercised here that every topology be tested. Also, in using both methods, trees with any negative branch lengths having absolute values > 1 are disqualified, since we have found by extensive computer simulations of evolutionary trees that this substantially increases the precision of these methods (Blanken et al. submitted).

The third method we have used is based on a theoretically justified conversion of the difference matrix which is useful for sequences which have evolved at different rates (Klotz et al. 1979). In this approach, each element of the difference matrix is converted by subtracting from it the distances (number of mutations) of the two sequences being compared from a common ancestor of all the sequences involved. In the idealized situation where the exact number of mutations separating a group of sequences is known, this converted difference matrix will always yield the correct topology using simple cluster analysis to calculate the tree (Klotz et al. 1979). We have recently developed two methods of applying this correction (Blanken et al. submitted), only one of which will be used in the present study and is outlined briefly here.

It can be shown that in the absence of information about the true ancestral sequence for a group of sequences (which is generally the case) it is possible to use any position on the evolutionary tree as a fixed reference point, in place of the real ancestor, to calculate the converted difference matrix. Thus, the most straightforward approach is simply to use one, or a tight group of present-day sequences as this reference point. The values subtracted from the matrix elements generated from comparisons of the remaining sequences are obtained directly from the matrix row belonging to the sequence chosen as the "ancestor". (When a group of sequences are used as the ancestral sequence, the values in the corresponding matrix rows are first averaged.) Thus, if the vector $\underset{\sim}{X}$ consists of the row of the difference matrix corresponding to the chosen ancestral sequence, then d'_{ij}, the ijth element of the converted difference matrix $\underset{\sim}{D}'$, is given by:

$$d'_{ij} = d_{ij} - x_i - x_j \ ,$$

where x_i and x_j are the ith and jth elements, respectively, of $\underset{\sim}{X}$. The optimum tree is then calculated directly from $\underset{\sim}{D}'$ using simple cluster analysis (Fitch and Margoliash 1967). This method will be referred to below as the PDA method (Present-Day-Ancestor).

Computer Simulations. Simulations of evolutionary trees were conducted to establish confidence estimates for various placements on real trees. The details of these simulations and the rationale behind them are presented in Results, and we limit our discussion here to technical points not dealt with there. The simulations were done using a WANG 2200 computer. Sequences 120 residues in length were simulated according to evolutionary trees presented in Results, beginning with an arbitrarily chosen ancestral sequence in which half of the residues are self-complementary:

$$(AUGC)_{15}(AUGC)_{7.5}(UACG)_{7.5} \ .$$

For a set of X mutations, X random numbers between zero and one were generated by the computer. Those falling between zero and 0.4 designated mutations in the first half of the molecule which were introduced randomly with respect to both position and the nature of the substitution, each being specified by an additional random number. The number of random numbers > 0.4 specified twice the number of mutations which were introduced into the second half of the molecule, as these were introduced in pairs to maintain the self-complementary of this half of the molecule, with only the first member of each pair chosen randomly. In some simulations, substitutions were not random in that transition mutations were made twice as likely as either of the possible transversions at any given site. Each of the nodal sequences in the simulated trees was stored for additional simulations from these points.

In calculating trees from the simulated sequences, the sequences were aligned directly without the use of gaps, and difference matrices were calculated and corrected as described above. The SBL and PSD criteria were used to first select the two optimum four-membered topologies representing the prokaryotic and eukaryotic sides of the tree, and then to join these to produce an 8-membered tree. If the selected tree topology corresponded to that which programmed the simulations, and the branch lengths were reasonably similar, it was used to simulate a ninth sequence from various stored nodal sequences. The optimum position of this ninth sequence was selected using the PSD and SBL methods holding the original eight positions fixed, or by using the PDA method with the simulated prokaryotic sequences as the ancestral sequences.

Results

Partial Nucleotide Sequences of C. cohnii 5S and 5.8S RNA

5S and 5.8S RNAs were isolated by gel electrophoresis from total *C. cohnii* RNA, 3' end-labeled, and sequenced by the method of Peattie (1979). These molecules are found in the 90S cytoplasmic ribosomal pellet; however in this preparation, there are other small RNA species present with very similar electrophoretic mobilities not seen in RNA extracted directly from whole cells (data not shown). We presume that these contaminating species are breakdown products of the larger rRNA, and chose to sequence the molecules isolated from total RNA instead. Figure 1 shows representative autoradiograms of sequencing gels for *C. cohnii* 5S and 5.8S RNAs in which about 100 residues of each can be read. For each of the molecules, every position in the sequence was determined from at least two separate gels. Occasionally, there were uncertainties in the reading from one particular gel (e.g. the 5S positions 85–89 where the identification of the G residues is not convincing), but in all cases, these were resolved either by examining replicate gels, or by running gels of different concentrations.

Since the unreacted, full-length strands obscure the reading of the 5' ultimate residues, we attempted to sequence 5' end-labeled molecules prepared with γ-^{32}P-ATP and T4 polynucleotide kinase (Maxam and Gilbert 1980). However, the sequence ladders we obtained appeared to be mixtures of two or more sequences, probably due to heterogeneity at the 5' ends (data not shown). We estimate that two to three additional residues that we have not identified could exist at the 5' ends of these molecules.

The sequence determinations we have made are shown in Fig. 2 arranged in secondary structure models similar to those proposed by other workers for homologous 5S (Fox and Woese 1975) and 5.8S (Nazar et al. 1975) RNAs. The dinoflagellate sequences conform well to these proposed models, and represent further demonstrations of compensatory mutations which maintain complementarity in the helical regions predicted by them, despite extensive sequence divergence from the other known sequences for *C. cohnii* 5.8S RNA (see below). At least in the case of 5S RNA, the proposed model is also consistent with much of the available data regarding base-pairing in the molecule free in solution (Vigne and Jordan 1977; Noller and Garret 1979), although it cannot be considered to have been rigorously proven.

The arm of the *C. cohnii* 5S structure which ends in the loop labeled L has been arranged to display more helicity than occurs in the Fox and Woese (1975) model proposed for prokaryotic 5S RNA and analogous structures suggested for eukaryotic sequences (Vigne and Jordan 1977; Hori and Osawa 1979). In fact, the *C. cohnii* model contains both of the helical regions in this portion of the structure suggested by the latter two groups of workers for eukaryotic 5S RNA, and could even include three additional base pairs involving loop L residues (1 GU and 2 GC base pairs), in the manner suggested by Nishikawa and Takemura (1974) for *Torulopsis utilis* 5S RNA. While the characteristics of this region of the proposed 5S RNA structures are not absolutely diagnostic of the origin of a 5S sequence, in possessing a purine-rich loop at the end of this stem-loop structure which is at least four residues in length, *C. cohnii* 5S RNA appears to be more eukaryotic-like: the great majority of prokaryotic 5S molecules, when arranged in this fashion, display a three-residue, exclusively pyrimidine loop at the end of a duplex stem (Fox and Woese 1975; Hori and Osawa 1979).

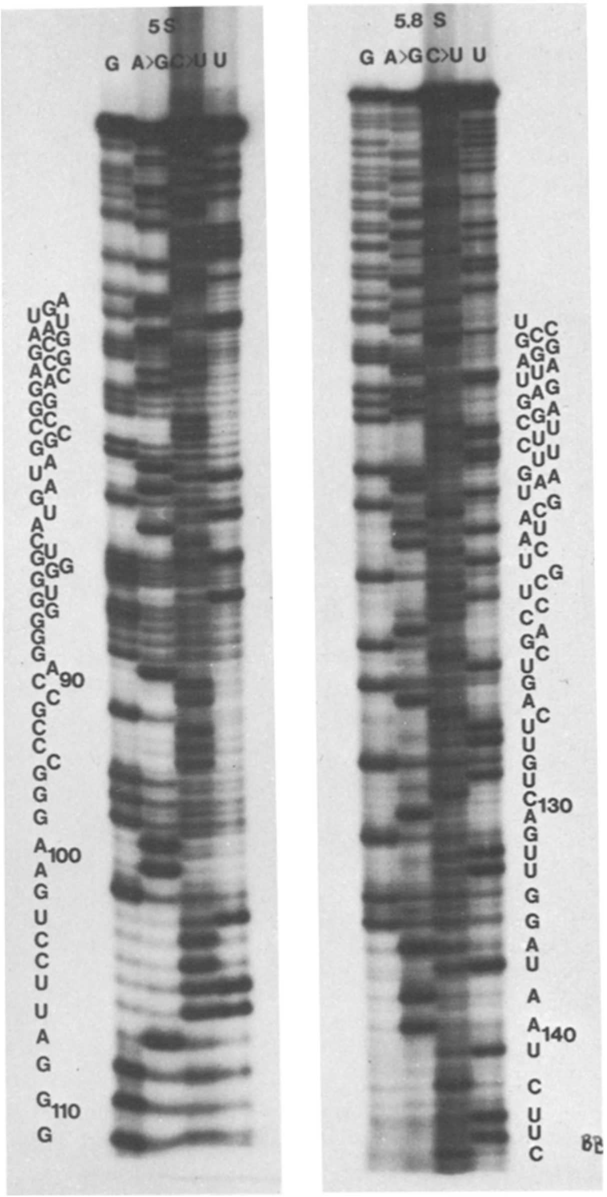

Fig. 1. Sequence ladders for a portion of *C. cohnii* 5S and 5.8S RNAs. The numbering begins at the last 5' residues read from our gels, which may be a few residues from the true 5' ends. Note that it does not correspond to the numbering in Fig. 3

In Fig. 3, the dinoflagellate sequences are shown aligned with several other homologous sequences (see legend for complete species names). The 5S RNA alignment (Fig. 3A) is adapted from Hori and Osawa (1979), and is based on the secondary structure model of the type shown for *C. cohnii* 5S RNA in Fig. 2A. Based on this alignment, *C. cohnii* 5S RNA shows the most homology (about 75%) with the animal sequences and the least homology (< 60%) with the prokaryote

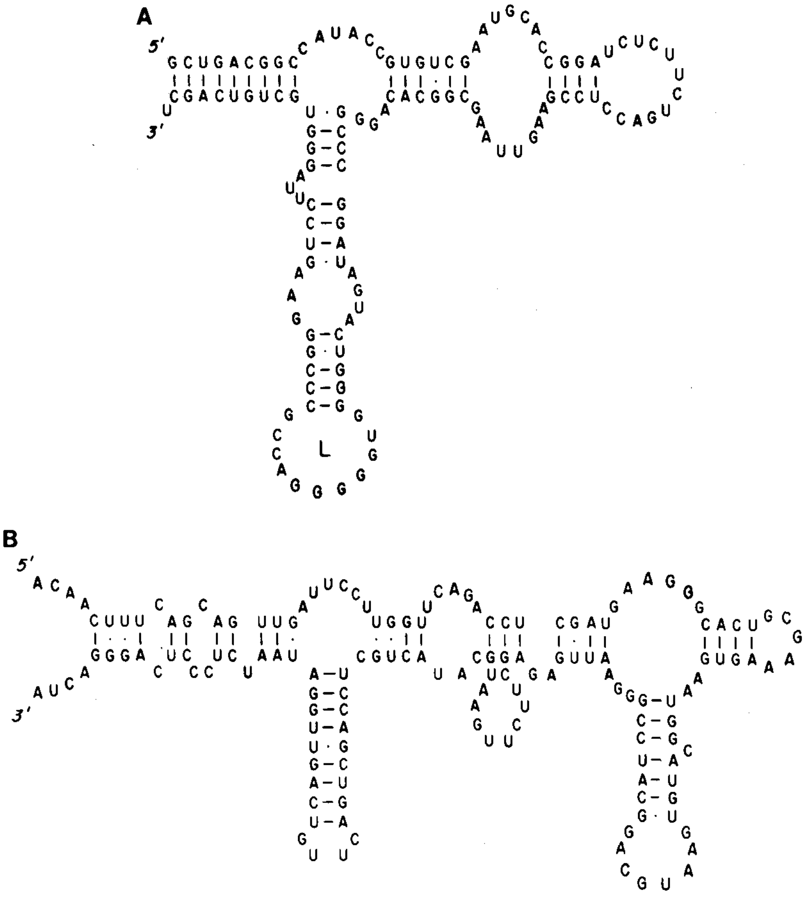

Fig. 2 A and B. Secondary structure models of *C. cohnii* 5S and 5.8S RNAs. Note that these sequences may be incomplete in any of the following ways: (1) there could be a few additional residues at the 5' ends; (2) there could occur pseudouridines which would register only as gaps in the sequence ladders, although no such gaps were apparent; (3) there could be sugar methylations which go undetected with the sequencing method used here

sequences. In addition, the dinoflagellate 5S RNA lacks two sequences found to be highly conserved in prokaryotes, but not in eukaryotes: the sequence 5'CC*GAAC*3' at positions 46–51, and the longer sequence at positions 78–87. The italicized part of the former is believed to interact with the "GTψC loop" of prokaryotic tRNA (Erdmann 1976), the latter with the 23S rRNA (Herr and Noller 1975). It has been suggested that the eukaryotic equivalent of the tRNA binding site, 5'GA$\overset{A}{U}$C3', is involved in binding only initiator tRNA (Erdmann 1976), which has 5'GIψC3' in place of 5'GTψC3' in loop IV (Larue et al. 1979). If so, the novel occurrence of the sequence 5'GACC3' in *C. cohnii* 5S RNA represents the third of the three possible tetranucleotides capable of base-pairing with the initiator tRNA sequence, based on the pairing potential of inosine (Crick 1966). Finally, the dinoflagellate molecule contains a sequence between positions 101–118 that is complementary to 14 nucleotides in yeast 18S rRna and is closely related to a sequence found in all eukaryotic 5S RNA molecules at this position. This complementarity is probably involved in the association of the ribosomal subunits in protein synthesis (Azad 1979). An analogous complementarity exists for prokaryotic 5S RNA and 16S rRNA at the same position in these molecules but involving different sequences.

The alignment of the 5.8S RNA sequences in Fig. 3B is adapted from Pavlakis et al. (1979) to include the recently reported broad bean sequence (Tanaka et al. 1980) and the *C. cohnii* sequence (see below for discussion of this alignment). In this case, the *C. cohnii* sequence has diverged from all other known sequences to about the same extent (45%). Despite this low degree of homology, there is little doubt that this is in fact a 5.8S ribosomal RNA, as indicated by the large number of conserved residues shared by all of the 5.8S molecules (56 out of about 156, Fig. 3B). In particular, three long conserved regions occur at positions 34–51, 69–80, and 102–111. As already noted by Pavlakis et al. (1979), the 3', 30% of the molecule has undergone much more evolution than the rest of the molecule, especially between positions 118–145. The broad bean and dinoflagellate sequences provide further evidence of the variation permitted in this part of the molecule.

Since the sequencing method employed here cannot give a direct indication of the occurrence of modified residues, it is possible that there are a few pseudouridines in the dinoflagellate molecules which we have been unable to detect as noticeable gaps in the sequence ladders (Peattie 1979). Only certain fungal 5S RNA molecules contain pseudouridine (at one position). However, nine out of the ten 5.8S RNA molecules that

A

```
                  00000000011111111112222222222333333333344444444445555555555666666666677777777778
                  12345678901234567890123456789012345678901234567890123456789012345678901234567890

X. laevis         -GCCU-ACGGCC-ACACC-ACCCUG-AAAGUGC-CCGAUCUCGU-CUGAUC-UCGGAAGCCAAGCAGGGUCGGGCCUGGU
D. melanogaster   -GCCA-ACGACC-AUACC-ACGCUG-AAUACAU-CGGUUCUCGU-CCGAUC-ACCGAAAUUAAGCAGCGUCGGGCGCGGU
Tomato            -GGAU-GCGAUA-CCAUC-AGCACU-AACGCAC-CGGAUC-CAU-CAGAAC-UCCGAAGUUAAGCGUGCUUGGGCGAGAG
Sunflower         --GGU-GCGAUA-CCAUC-AGCACU-AAUGCAC-CGGAUC-CAU-CACAAC-UCCGCAGUUAAGCGUGCUUGGGCGAGAG
S. cerevisiae     -GGUU-GCGGCC-AUAUC-UACCAG-AAAGCAC-CGUUUCCCGU-CCGAUCAACUGψAGUUAAGCUGGUAAGAGCCUGAC
P. membranaefac.  -GGUU-GCGGCC-AUAUC-UAGCAG-AAAGCAC-CGUUUCCCGU-CCGAUCAACUGψAGUUAAGCUGCUAAGAGCCUGAC
C. cohnii         -GCUG-ACGGCC-AUACC-GUGUCG-AAUGCAC-CGGAUCUCUU-CUGACC-UCCGAAGUUAAGCGGCACAGGGCCCGGA
E. coli           UGCCUGGCGGCC-GUAGC-GCGGUG-GUCCCAC-CUGACCCCAUGCCGAAC-UCAGAAGUGAAACGCCGUAGCGCC-GAU
P. fluorescens    UGUUCUGUGACGAGUAGUGGCAUUG-GAA-CAC-CUGAUCCCAUCCCGAAC-UCAGAGGUGAAACGAUGCAUCGCC-GAU
B. subtilis       --UUUGGUGGCG-AUAGC-GAAGAG-GUCACAC-CCGUUCCCAUACCGAAC-ACGGAAGUUAAGCUCUUCAGCGCC-GAU
C. pasteurianum   --UCCAGUGUCU-AUGAC-UUAGAG-GUAACAC-UCCUUCCCAUUCCGAAC-AGGCAGGUUAAGCUCUAAUGUGCU-GAU

CONSERVED:                G        C      G     CAC C    C C U C GA C  C G  GU AA C        G
                          '                                ' ' ' '' '          '' '        '

                                     1111111111111111111111111111111111111111111111
                  88888888899999999990000000000111111111122222222223333333333444444
                  12345678901234567890123456789012345678901234567890123456789012345

X. laevis         --UAGUA-CUUGGAUGGGAGA--CCGCCUGGGAAUACC---AGG-UGUCGU-AGGCUUU
D. melanogaster   --UAGUA-CUUAGAUGGGGGA--CCGCUUGGGAACACC---GCG-UGUUGU-UGGCCU-
Tomato            --UAGUA-CUAGGAUGGGUGA--CCCCCUGGGAAGUCC---UCG-UGUUGC-AUCCU--
Sunflower         --UAGUA-CUAGGAUGGGUGA--CCCCCUGGGAAGUCC---UCG-UGUUGC-ACCU---
S. cerevisiae     C-GAGUA-GUGUAGUGGGUGA--CCAUACGCGAAACUC---AGG-UGCUGC-AAUCU--
P. membranaefac.  C-GAGUA-GUGUAGAGGGCGA--CCAUACGCGAAACUC---AGG-UGCUGC-AAUC---
C. cohnii         --UAGUA-CUGGGGUGGGGGA--CCGCCCGGGAAGUCCUUAGGG-UGCUGUCAG-CU--
E. coli           GGUAGUG-------UGGGGUCU-CCCCAUGCGAGAG---UAGGG-AACUGCCAGGCAU-
P. fluorescens    GGUAGUG-------UGGGGUUU-CCCCAUGUCAAGA---UCUCG--ACCAUAGAGCAU-
B. subtilis       GGUAGUC-------GGGGGUUU-CCCCCUGUGAGAG---UAGGA-CGCCGCCAAGC---
C. pasteurianum   GGUACUG-------CAGGGGAAGCCCUGUGGAAGAG---UAGGU-CGACGCUGGGU---

CONSERVED:          UA                         G  A
                                               '
```

B

```
                  00000000011111111112222222222333333333344444444445555555555666666666677777777778
                  12345678901234567890123456789012345678901234567890123456789012345678901234567890

Rat               CGACUCUUAGCGGUGGAUCACUCGGCUCGUGCGUCGAUGAAGAACGCAGCGCUAGCψGCGAGAAUUAAUGUGAAUUGCAG
D. melanogaster   -AACUCUAAGCGGUGGAUCACUCGGCUCAUGGGUCGAUGAAGAACGCAGCA--AACUGUGCGUCAUCGUGUGAACUGCAG
Broad bean        UGACUCUCGGCAACGGAψAUCUAGGCUCUUGCAUCGAUGAAGAACGUAGCG--AAAUGCGAUACUUGGUGUGAAU GCAG
S. cerevisiae     AAACUUUCAACAACGGAUCUCUUGGUUCUCGCAUCGAUGAAGAACGCAGCG--AAAUGCGAUACGUAAUGUGAAψUGCAG
N. crassa         AAACUUUCAACAACGGAUCUCUUGGUUCUGGCAUCGAUGAAGAACGCAGCG--AAAUGCGAUAGGUAAUGUGAAUUGCAG
C. cohnii         CAACUUUCAGCAGUUGAUUCCUUGGUUCAGACCUCGAUGAAGGGCACUGCG--AAA-GUGAAUGGC-AUGUGAA-UGCAG

CONSERVED:          ACU U   C   GGAU  CU GG UC  GC UCGAUGAAGAACG AGCG  A  UG GA    U  UGUGAA UGCAG
                    ''' '   '    '''  '' '' ''     '''''''''  '   ''   '   ' '        '''''' '''''

                                     1111111111111111111111111111111111111111111111111111111111111111
                  88888888899999999990000000000111111111122222222223333333333444444444455555555556666
                  12345678901234567890123456789012345678901234567890123456789012345678901234567890123

Rat               GACACAUUG-AUCAUCGACACUUCGAACGCAC-UUGCGGCCCC--GGGUUCCUC-CCGGGG-CUACGCCUGUCUGAGCGUCGCU-
D. melanogaster   GACACAU-GAA-CAUCGACAUUUUGAACGCAUAUCGCAGUCCA--UGCUG--UG-CUUGGA-CUACAUAUGUUGGAGGGUUGUA-
Broad bean        AAUCCCGUGAACCAUCCAGUCUUUGAACGCAAGUUGC-CCCGAUGCCAUUA-GG-UUGAGGGC-ACGUCUGCCUGGGUGUCACAU
S. cerevisiae     AAUUCCGUGAAUCAUCGAAUCUUUGAACGCACAUUGC-GCCCC-UUGG-UAUUC-CAGGGGGC-AUGCCUGUUUGAGCGUCAUUU
N. crassa         AAUUCAGUGAAUCAUCGAAUCUUUGAACGCACAUUGC-GCUCG-CCAG-UAUUC-UGGCGAGC-AUGCCUGUUCGAGCGUCAUUU
C. cohnii         GCAUCCGGGAAUUGAGAGCUUCUUGAAUGCAUACUGCUCCAGC--UGACUUGUC-AGUUGGA-UA-AUCUUCCUCAGGGACUA--

CONSERVED:         A  C   G A CAU GA   UU GAACGCA  UUGC                      G  C A   CUG   G G GUC
                      '   ' '           ' ''' '''    ''                      '    '           ' '
```

Fig. 3 A and B. Alignments of *C. cohnii* 5S and 5.8S RNAs adapted from published alignments. The conserved residues were tabulated using all known sequences, not just those shown here. The marked conserved residues occur without exception; those unmarked differ in only one or a closely related group of sequences. The complete names for those abbreviated in the figure are as follows. **A** *Xenopus laevis, Drosophila melanogaster, Saccharomyces cerevisiae, Pichia membranaefaciens, Crypthecodinium cohnii, Escherichia coli, Pseudomonas fluorescens, Bacillus subtilis, Clostridium pasteurianum;* **B** same as above, plus *Neurospora crassa*

have been examined contain at least one pseudou (Erdmann 1980; Tanaka et al. 1980), although the position of this modification varies among plants, animals, and fungi. At one of the three positions at which pseudouridine has been observed in other 5.8S sequences (position 18, Fig. 3B), *C. cohnii* 5.8S RNA has a uridine, while at the remaining two (positions 57 and 75, Fig. 3B), the alignment indicates a gap in the *C. cohnii* sequence. The latter would perhaps be likely positions for pseudouridine residues in the dinoflagellate molecule, if they exist. However, we do not detect gaps in the *C. cohnii* sequence ladder at these points, while in a similar study on wheat 5.8S RNA (data not shown), we consistently identified two gaps in the sequence ladder which correspond to known pseudouridine positions in the broad bean 5.8S sequence (Tanaka et al. 1980).

Construction of a 5S RNA Phylogeny

We have constructed an evolutionary tree of 5S RNA sequences from a difference matrix obtained from a comparison of *C. cohnii* 5S RNA with other known 5S sequences, aligned as in Fig. 3A. The difference matrix is presented in Table 1, along with the same matrix corrected for multiple substitution at each site using the equation of Holmquist (1972) for a sequence of length 120. Three different methods of selecting the best tree topology have been employed. As explained in Methods, two of these are modifications of published procedures in which various topologies are examined and the one which minimizes a particular property of the topology is chosen:

(1) percent-standard-deviation from the observed difference matrix of the difference matrix reconstructed from the tested topology (Fitch and Margoliash 1967);

(2) sum of the branch lengths of the tested topology (Dayhoff 1978).

These methods are referred to below as the PSD and SBL methods, respectively. We have restricted the number of sequences on our tree to a number small enough (six) to permit an examination of all possible topologies in applying these two methods. The third method is based on a conversion of the difference matrix for unequal evolutionary rates (Klotz et al. 1979), which performs as well as the above methods in picking the correct topology for computer-simulated trees (Blanken et al. submitted), but does not require examination of alternate topologies and is thus not limited to small trees. This conversion involves a subtraction from each element of the difference matrix of the distances of the two sequences being compared from a fixed reference point on the tree (see Methods for details). This method will be referred as the PDA method.

When applied to the corrected difference matrix in Table 1, these methods yield the evolutionary trees shown in Fig. 4A and 4B. These two trees differ only in the order of branching of the higher plants and the dinoflagellates, and in the branch lengths. The PDA method selects the tree in Fig. 4A, using either vertebrates, gram-negative, or gram-positive bacteria as the fixed reference sequences. The PSD and SBL methods choose the tree in Fig. 4B, while selecting the tree in Fig. 4A as the next most likely topology. In fact, the percent-standard-deviation and the sum of the branch lengths for the tree in Fig. 4A are insignificantly smaller than the values for the tree in Fig. 4B. Using a somewhat different alignment based on a different secondary structure model (Schwartz and Dayhoff 1978) of a similiar set of sequences (rye, *S. carlbergensis*, and *B. megaterium* replacing tomato, *S. cerevisiae*, and *B. subtilis*, respectively), the tree in Fig. 4A is chosen unambiguously by all three methods.

The equation of Holmquist (1972) that was used to correct the matrix in Table 1 for multiple substitutions assumes that completely random substitution has occurred at every site in the molecule during its evolution. This assumption is unlikely to be strictly true. At the very least, it does not take into account insertions and deletions which the alignments (Fig. 3A) indicate have occurred frequently in 5S RNA evolution, and it is probable that functional constraints cause cer-

Table 1. Observed and corrected difference matrices for 5S RNA sequences[a]

	1	2	3	4	5	6	7	8	9	10	11
1) *X. laevis*	0	36.2	72.5	79.5	58.1	64.0	122.1	89.7	101.3	139.2	47.3
2) *D. melanogaster*	30	0	82.0	77.1	64.0	66.1	114.6	84.5	101.3	122.1	45.7
3) *S. cerevisiae*	50	54	0	5.1	74.8	82.0	139.2	98.3	79.5	118.2	66.1
4) *P. membranaefaciens*	53	52	5	0	87.1	89.7	139.2	92.5	77.1	114.6	66.1
5) Tomato	43	46	51	56	0	7.2	118.2	95.3	101.3	144.0	56.2
6) Sunflower	46	47	54	57	7	0	122.1	101.3	104.4	134.6	62.0
7) *P. fluorescens*	67	65	71	71	66	67	0	49.0	60.0	107.7	107.7
8) *E. coli*	57	55	60	58	59	61	38	0	40.8	92.5	74.8
9) *B. subtilis*	61	61	53	52	61	62	44	33	0	54.4	82.0
10) *C. pasteurianum*	71	67	66	65	72	70	63	58	41	0	130.2
11) *C. cohnii*	37	36	47	47	42	45	63	51	54	69	0

[a] The observed number of differences is given in the bottom left. The calculated mutational differences are in the upper right.

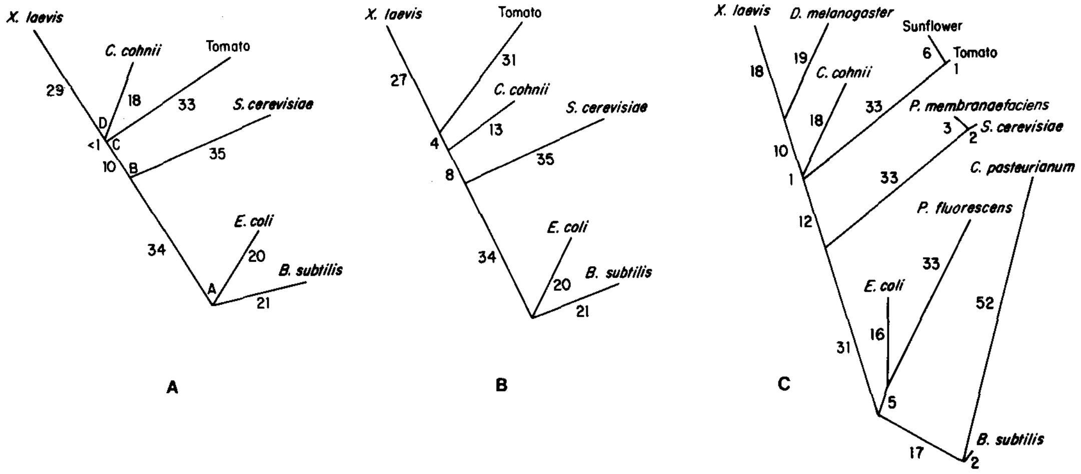

Fig. 4 A-C. 5S RNA phylogenies. The branch lengths are the least-squares number of mutational differences calculated from the corrected difference matrix in Table 1

tain sites to mutate faster than others. Thus, it was important to demonstrate that the calculated branching order of the 5S tree is not dependent upon the use of this correction. In fact, the uncorrected difference matrix in Table 1 yields the same trees as shown in Figs. 4A and 4B, although as expected, with shorter branch lengths. In addition, if we assume a sequence length of 110 in applying the Holmquist correction instead of 120, to take into account the possibility that about 10 positions in the molecule are held invariant by functional constraints, as suggested by current 5S RNA alignments (Schwartz and Dayhoff 1978; Hori and Osawa 1979), the same results are again obtained. Finally, we have employed a different correction which assumes that the frequency of a transition is twice as likely as either transversion that can occur at a given site (see Appendix) as was suggested for 5S RNA evolution (Sankoff et al. 1976). Again, the topologies are the same as shown in Fig. 4, indicating that this result is not likely to be an artifact of the method used to correct the observed difference matrix for multiple substitutions.

In Fig. 4C, we have employed the PDA method and the corrected matrix in Table 1 to construct a more extensive tree representing all of the sequences aligned in Fig. 3A. This tree has the same topology for the groups represented as that recently presented by Schwartz and Dayhoff (1978) using their criterion of minimum absolute sum of branch lengths and examining a large number (although presumably not all) off the possible topologies. It demonstrates that 5S RNA trees do in fact group together organisms that are known to be closely related. In addition, it places *C. cohnii* at the same position as shown in Fig. 4A. This tree was constructed using three *Bacillus* species (*B. subtilis, B. megatorium*, and *B. stearothermophilus*) as the reference point in applying the PDA method. If human and *Xenopus* are used instead, the same tree is obtained except that *C. cohnii* is placed just below the plant-animal divergence as in Fig. 4B. As with the smaller trees in Figs. 4A and 4B, these two topologies are essentially indistinguishable by the criteria of the SBL and PSD methods.

These trees confirm the qualitative impression from the last section of a close relationship between the 5S RNA of *C. cohnii* and other eukaryotic groups. Although there is disagreement on the order of divergence of the dinoflagellate and higher plant branches, they all agree in placing dinoflagellates higher on the eukaryotic line leading to plants and animals than the fungi. As will now be explained, this implies that the absence of histones in dinoflagellates is due to secondary loss rather than primitive origin.

We assume that the common ancestor of all the organisms on the trees in Fig. 4 can only be placed on one of the branches below the yeast node (the A-B internodal branch, the A-*E. coli* branch, or the A-*B. subtilis* branch in Fig. 4A. This assumption provides the direction in time for the eukaryotes on the trees shown in Fig. 4 with plants, animals, and dinoflagellates branching off higher on the tree, and therefore later in time, than the fungi. Such an assumption is necessary because there is no information in the difference matrix on the position of the real ancestor for the sequences involved, and it is reasonable in view of the more ancient fossil record and metabolic diversity of the Prokaryota as well as the probable anaerobicity of the pre-biotic environment (Dobzhansky et al. 1977). Since typical eukaryotic chromatin structure occurs in plants, animals, and fungi, than either it occurred as well in their common ancestor at node B in Fig. 4A, or it evolved independently in the fungi and in higher eukaryotes. This latter alternative is highly unlikely in view of the similarities between the histones and the fine-structure of the nucleosomes in these three groups (Horgen and Silver 1978). It follows that if the dinoflagellates diverged later than the fungi, as indicated by the 5S RNA sequences, then they evolved from a histone-containing eukaryote. (The unlikely placement of the ancestor on any of the eukaryotic branches (above node B in Fig. 4A) leads to the equally interesting conclusion that the prokaryotes originally contained histones but lost them at some point.)

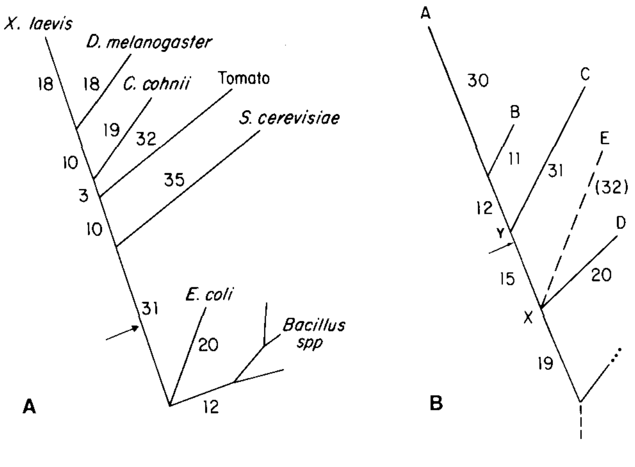

Fig. 5 A and B. The optimum 5S RNA tree for the sequences shown calculated using the PDA method with three *Bacillus* sequences as ancestor: *B. subtilis, B. stearothermophilus* and *B. megaterium* (the latter two are aligned in Hori and Osawa (1979)). **B** The eukaryotic portion of a simulation of the tree in **A**. The ancestral sequence is represented by the dashed line at the base of the tree. The dotted line is the attachment point for the prokaryotic-like sequences not shown

Computer Simulations of 5S RNA Evolution

Since the predicted topologies of the 5S RNA trees in Fig. 4 are not particularly sensitive to differences in sequence alignment, perhaps the major obstacle to their acceptance is the possibility that the close homology observed between the dinoflagellate sequence and the higher eukaryote sequences is due to convergence. Although the only definitive way one can counter this objection is to examine molecular sequences in addition to 5S RNA, we have attempted to provide an estimate of the likelihood that dinoflagellates actually diverged before the fungi and that the homology between *C. cohnii* 5S RNA and the sequences of higher plants and animals is due to random convergence instead of phylogenetic relatedness. This estimate is based on computer simulations of 5S RNA-like evolutionary histories. In these simulations, we began with an arbitrarily chosen 120 nucleotide ancestral sequence (see Methods) and introduced random mutations in this molecule along a set of paths which represents the actual 5S RNA tree for a group of sequences. Based on the secondary structure model for 5S RNA of the general type suggested by Fox and Woese (1975), we began with and maintained throughout the simulations self-complementarity in one-half of each of the simulated sequences. Based on the rate of evolution in the helical and non-helical regions of the molecule predicted by this model and calculated from a comparison of a large set of aligned 5S sequences, we introduced on average 1.5 times as many substitutions into the helical regions as in the non-helical regions. (The secondary structure model on which these calculations were based has not been proven conclusively, but appeared to be the best estimate of universally occurring base-pairing schemes available.)

In Fig. 5B is shown the eukaryotic portion of one of several simulations we have generated of the actual 5S RNA tree shown in Fig. 5A. (The real tree in Fig. 5A was selected using the PDA method, with the *Bacillus* sequences as the common ancestor, and agrees in topology with the trees shown in Fig. 4 for the sequences shared, although the lengths of the branches held in common are slightly different due to the different sequence sets used in the two cases.) In the simulations, the ancestral sequence was arbitrarily placed midway on the branch separating eukaryotes and prokaryotes (the arrow in Fig. 5A) and sets of random mutations were introduced according to the branch lengths on the actual 5S RNA tree. Using the generated sequences A-D, which represent the two animals, tomato, and yeast sequences respectively, and the prokaryote-like sequences not shown, the tree topology was back-calculated using the methods described above. If the selected topology corresponded to the real 5S RNA topology for these sequences, and the branch lengths were also in reasonable agreement, then it was used to test the significance of the dinoflagellate placement on the real tree. To do this, we determined the frequency that a sequence which diverges from point X in Fig. 5B (the simulated yeast node) and evolves to an extent comparable to the number of mutations separating the actual dinoflagellate sequence and the yeast node (32), is placed by the tree construction methods described earlier at or above the expected position of the dinoflagellate sequence on the simulated tree (i.e. 13 mutations above the yeast node, at the arrow in Fig. 5B). The dotted line in Fig. 5B represents the hypothetical dinoflagellate branch. It is placed at the yeast node instead of below this point (as the Mesokaryote hypothesis suggests) to make this test of the likelihood that dinoflagellate divergence occurred prior to fungal divergence as stringent as possible. We found that in only two cases out of 50, a sequence simulated from the yeast node was placed 13 mutations higher on the tree, using the SBL and PSD criteria to select the branching point. However, we must also consider the equally likely 5S tree, not shown for the sequences in Fig. 5, in which *C. cohnii* diverges just below the plant-animal split at 11 instead of 13 mutations above the yeast node. In only 3 out of 50 trials is sequence E placed this high above node X.

We have also simulated sequences in a similar fashion from the plant-animal node (Y in Fig. 5B) to test the significance of the placement of yeast below this point.

Thus, we introduced 46 mutations in the nodal sequence Y for each of many trials and scored the number of the resulting sequences that were placed 10 mutations below node Y. Out of 75 such trials, only 2 were placed this low. In this set of simulations, as well as those just mentioned above, we employed several different simulated 8-membered trees of the type shown in Fig. 5B.

While these simulations involve completely random substitution at any given site, we have also carried out simulations which are probably more realistic in which we impose the non-random equal transition-transversion probability mentioned in the last section. In this case, 0 of 100 sequences simulated from node X are placed 11 mutations higher than X, and 0 out of 50 simulated from node Y are placed 10 mutations below Y. Admittedly, even these simulations do not consider functional convergence or artifacts of alignment, and they may depend to some extent on the general validity of the secondary structure model we asssumed in designing them; however, we believe that they indicate with high confidence (in the range of 95%) that the placement of *C. cohnii* above yeast on the 5S trees is not the result of *random* sequence convergence, and is therefore deserving of further attention.

The Phylogeny of 5.8S RNA

We have attempted to obtain an independent confirmation of the branching order on the eukaryotic side of the 5S RNA trees in Fig. 4 by also constructing a phylogeny from the 5.8S RNA sequences. In Table 2 is shown the 5.8S RNA difference matrix calculated from the alignments in Fig. 3B and a corrected matrix obtained using the Holmquist equation assuming an effective sequence length of 140 nucleotides. This length was chosen instead of the true sequence length of about 160 nucleotides because of the occurrence of 20 residues in the three long regions at 34–42, 69–74, and 76–80 (Fig. 3B) that are invariant in all of the known sequences. Although it is not certain that these positions are absolutely conserved, since the data set is limited, this seems likely in view of the high degree of substitution found in the rest of the molecule. Using all three methods for calculating optimum trees on the corrected matrix in Table 2 gives the tree shown in Fig. 6A. Note that it has the same order of branching as the eukaryotic portion of the 5S tree in Fig. 4A. The same result is obtained if:

(1) the uncorrected matrix in Table 2 is used;

(2) we reduce the sequence length in applying the Holmquist correction to 125 to include the shorter three residue-long invariant sequence stretches (Fig. 3B);

(3) we employ our alternate correction for multiple substitutions that assumes an equal transition-transversion probability.

However, if we attempt to include all of the invariant sequences in correcting for multiple substitutions, which results in an effective sequence length of 105, we find that it is not possible to obtain a tree with positive branch lengths that accurately clusters the different animal and fungal sequences on the tree, suggesting that this is an over-correction.

Table 2. Observed and corrected differences matrices for 5.8S sequences[a]

	1	2	3	4	5	6
1) Rat	0	56.7	77.5	56.7	67.5	130.9
2) *D. melanogaster*	44	0	98.2	88.5	98.2	130.9
3) Broad bean	55	64	0	51.7	56.7	138.1
4) *S. cerevisiae*	44	60	41	0	13.8	106.2
5) *N. crassa*	50	64	44	13	0	114.8
6) *C. cohnii*	75	75	77	67	70	0

[a] The matrices are arranged as in Table 1

Unfortunately, the selection of Fig. 6A as the optimum 5.8S tree is somewhat alignment-dependent. We have constructed two different alignments that vary from the one shown in Fig. 3B between positions 118–148. This is the region where the most substitution has occurred and there are no highly conserved residues to facilitate alignment. Although these alternate alignments also result in the selection of the Fig. 6A tree as the optimum tree, one of them indicates that the tree shown in Fig. 6B is not significantly inferior by the SBL and PSD criteria. Because these alignments make greater use of gaps without revealing a significantly larger number of homologies, they are probably inferior to that shown in Fig. 3B. However, since there is no way of being sure, we can only conclude that the dinoflagellate lineage joins the rest of the tree either just above or just below the plant-animal divergence.

Since prokaryotic ribosomes do not contain 5.8S RNA, it is not possible to establish from 5.8S sequences alone the point of earliest time on the tree. However, if we use the result of the 5S tree in placing the root between the fungal divergence and the plant-animal split, it can be seen that the 5.8S trees in Fig. 6 agree with the 5S trees in Fig. 4 in placing the dinoflagellate divergence just before or just after the plant-animal divergence[1]. This rooting position is also supported by two recent tRNA phylogenies (Larue et al. 1979) and is at least consistent with cytochrome c data (Fitch 1976).

However, there is an alternative explanation that must be considered. If we root the 5.8S trees as just suggested, it implies that fungal 5.8S RNAs have evolved 3-fold less than the animal and plant molecules and > 6–

[1] To see this clearly, after attaching the ancestral branch to the fungal-plant internodal branch in the tree in Fig. 6A, or the fungal-*C. cohnii* internodal branch in Fig. 6B, rotate the fungal branch counter-clockwise by 90°. This will establish the same time orientation as shown in the 5S trees in Fig. 4

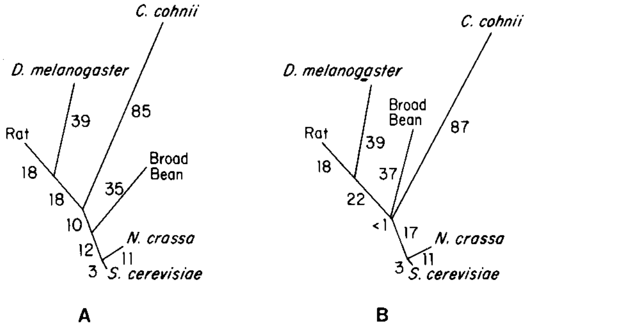

Fig. 6 A and B. The optimum 5.8S RNA phylogenies calculated from the corrected difference matrix in Table 2

fold less than the dinoflagellate molecule – a great disparity in evolutionary rates. Studies on closely related animal groups argue in favor of more equivalent evolutionary rates in different lineages (Wilson 1977) and if we extend this idea of equal rates to the much longer evolutionary periods represented in the trees shown here, we would be obliged to connect the common ancestor of all the 5.8S sequences to the dinoflagellate branch. This would make the divergence of the dinoflagellates the earliest event on the tree; however, it would be inconsistent with several other molecular phylogenies which have been constructed, in the following way. If the point of earliest time on either of the 5.8S trees in Fig. 6 is placed on the *C. cohnii* branch, then the tree in Fig. 6A will imply that the fungi diverged from the plant lineage after the plant-animal split, while the tree in Fig. 6B will indicate a coincident divergence of plants, animals, and fungi[2]. Both of these alternatives are at odds with our 5S RNA trees and the two recently reported tRNA trees (Larue et al. 1979) which place the fungal divergence well below the plant-animal split, and our simulations indicate that in the case of the 5S trees, this is a highly significant result. In addition to this objection, the disparity in the 5.8S tree branch lengths of the two animal sequences in Fig. 6, which differ > 2-fold, and the two fungal sequences which differ > 3-fold indicates that significantly unequal rates of evolution do exist for 5.8S RNAs even among closely related species. Thus, there is no compelling reason to invoke equivalent evolutionary rates over the much greater interval represented by the evolution of all the eukaryotes represented on the 5.8S trees shown in Fig. 6.

Discussion

5S RNA primary structure may be of great usefulness in determining the phylogenetic relationships between the eukaryotic protists, bacteria, and higher plants and animals. The molecule is found in the large subunits of prokaryotic and eukaryotic ribosomes, and the length and most probably the general secondary structure have been conserved throughout evolution. Examination of sequences aligned according to the secondary structure model proposed by Fox and Woese (1975) reveals the occurrence of short regions in the molecule conserved in both prokaryotes and eukaryotes (Hori and Osawa 1979), as well as a degree of homology between the most divergent sequences that is significantly higher than that expected for unrelated sequences. In addition, there are indications of similar functions for prokaryotic and eukaryotic 5S RNA in tRNA binding (Erdmann 1976) and ribosome subunit association (Azad 1979). All of these considerations justify the construction of a phylogenetic tree based on 5S RNA sequence comparisons.

The 5S RNA of the dinoflagellate *C. cohnii* lacks three different regions that are highly conserved in all prokaryotic but not eukaryotic 5S RNA molecules. In addition, the secondary structure proposed in Fig. 2A features a loop-hairpin structure closely resembling that suggested for other eukaryotic 5S RNAs (Vigne and Jordan 1977; Hori and Osawa 1979), and dissimilar from that predicted for most prokaryotic moelcules (Fox and Woese 1975). These features, along with the extensive sequence homology with 5S RNAs from higher eukaryotes indicate a very distant relationship between dinoflagellates and prokaryotes. The evolutionary trees in Fig. 4 provide a graphic representation of this fact, and further suggest that the fungi diverged from the eukaryotic lineage earlier than the dinoflagellates. This result is interesting because it implies a loss of histones and nucleosomal chromatin structure at some point in dinoflagellate evolution.

The trees in Fig. 4 were selected by three independent methods for choosing the optimum tree relating a set of molecular sequences. The results are not restricted to a particular 5S RNA alignment. Nor are they dependent on the use of a correction for reverse and duplicate mutations: the same trees are obtained whether the uncorrected difference matrix is used or if instead, it is first corrected assuming either completely random substitution or a preference for transitions over transversions. The computer simulations of 5S RNA evolution indicate that the placement of the dinoflagellate sequence higher on the eukaryotic branch than the fungi is not likely to be the result of random convergence, although the probability that it is, is far from vanishingly small. Although the simulations are not completely realistic, particularly because they don't include insertion and deletion mutations which cause alignment problems, we feel that they provide a valuable indication of the probability of choosing an incorrect topology on the basis of chance alone. However, they do not incorporate functional constraints on 5S RNA

[2] To see this clearly, turn the trees in Fig. 6 upside-down, place the ancestral branch on the *C. cohnii* branch, and rotate the new *C. cohnii* branch and the plant and animal branches upward so that they point in the same direction as the fungal branch

evolution (which are largely unknown), and the possibility exists that the similarities observed between dinoflagellate and higher eukaryotic 5S RNAs may merely reflect common functional requirements. One argument against this is that phylogenies we have obtained from sequence comparisons limited to either the helical or nonhelical regions of the 5S RNA molecules (as defined by the structural model of Fox and Woese) are the same as those presented in Results for the whole sequences (data not shown). These regions might be expected to experience different functional constraints on nucleotide replacement, and the fact that both regions show the greatest homology with the animal sequences argues for evolutionary relatedness versus functional convergence. However, this argument depends on the validity of the secondary structural model employed in making these calculations, which has not been demonstrated conclusively.

The only definitive way to rule out the possibility of convergence and alignment artifacts is to demonstrate the same phylogeny using other molecular sequences. While the 5.8S trees in Fig. 6 are consistent with the 5S results, in lacking prokaryotic representation, the 5.8S phylogenies cannot be considered as strong independent confirmation of the 5S results, and other sequences will have to be determined. Some appropriate sequences will be mentioned below.

It is important to consider also the non-sequence evidence that has accumulated concerning the phylogeny of dinoflagellates and fungi. There are a number of properties of the latter which suggest their primitive state. These include:

(1) the lowest DNA contents among eukaryotes, which overlap with the prokaryotic range (Cavalier–Smith 1978);

(2) the clustering of the 5S with the large rRNA cistrons, as in bacteria and in contrast to higher animals and at least two eukaryotic protists (Philippsen et al. 1978; Tartof 1975; Tonnesen et al. 1976; Marco and Rochaix 1980);

(3) the occurence of homologous recombination upon DNA transformation (Hinnen et al. 1978), as in bacteria but in contrast to higher eucaryotes (Wigler et al. 1979);

(4) the virtual absence of polyunsaturated fatty acids (Ragan and Chapman 1978);

(5) the absence of flagella and centrioles;

(6) a saprophytic mode of nutrition.

On the other hand, the occurrence of chitin and glycogen in the fungi has often been cited as evidence for a divergence of this group following the plant-animal split. However, the similarity between glycogen and the floridean starch found in red algae and the occurence of chitin in some diatoms weakens these arguments (Demoulin1975).

Dinoflagellates show a number of affinities with other flagellate groups. They are very similar to the euglenoids in the ultrastructure of their chloroplasts and flagella (Taylor 1976) and in possessing permanently condensed chromosomes (Leedale 1968), and both resemble the kinetoplastids in flagellar structure (Taylor 1976). The mechanism of mitosis, at least as is occurs in the parasitic dinoflagellate *Syndinium*, is very similar to that seen in the trichomonads and hypermastigids (Ris and Kubai 1974; Hollande 1974). They resemble the chromophyte algae (golden brown, brown algae and diatoms) in the occurrence of c–chlorophyllide, acetylenic xanthophylls and an extraplastidic reserve carbohydrate (Taylor 1976; Ragan and Chapman 1978). These latter features have generally resulted in the placement of dinoflagellates on the chromophyte branch in a number of protistan phylogenies which have beeen proposed (Cavalier–Smith 1975; Taylor 1976; Ragan and Chapman 1978); however, their many distinctive features such as the lack of histones, the unique structure of their cell covering and the mechanism of mitosis in the free-living species earn them a very early divergence from this lineage following its split from the chlorophyte (green algae).

The order of divergence of the fungi and the various flagellate groups with which the dinoflagellates show affinities is disputed in these phylogenies. However, flagellates are invariably placed higher on the eukaryotic line leading from the prokaryotes than are the red algae. which are almost universally believed to be the most primitive eukaryotes. This is based on the absence in red algae of flagella and centrioles and the occurrence of phycobiliproteins in phycobilisomes as found elsewhere only in the prokaryotic blue-green algae (Ragan and Chapman 1978). Interestingly, red algal interphase chromatin is typically decondensed (McDonald 1972) as in nearly all eukaryotes, and there are indications of histone-like basic chromosomal proteins (Duffus et al. 1973). If the latter can be substantiated, then the classical protistan phylogenies, like our own, will also imply a secondary loss of histones in the dinoflagellates (e.g. see Cavalier-Smith 1975). It has long been suggested that the red algae and the higher fungi may be closely related (see Demoulin (1975) for a recent discussion). This would explain the primitive features of the latter mentioned above, in particular the absence of flagella and centrioles, and would be consistent with both our 5S RNA phylogeny and the tRNA phylogeny reported by Larue et al. (1979) which places *Euglena gracilis* at the same position with respect to the higher eukaryotes, fungi, and prokaryotes as that occupied by the dinoflagellates on our 5S tree. It must be noted however, that the cytochrome c trees which have been reported disagree with these phylogenies and place the fungal divergence above that of the flagellates *Euglena gracilis* and *Crithidia oncopelti* (Fitch 1976; Schwartz and Dayhoff 1978). This could be due to any of the following:

(1) the cytochrome c tree is in error due to convergence;

(2) the prokaryotic cytochromes used in the tree are not orthologous with the eukaryotic proteins;

(3) the 5S and tRNA trees are both affected by convergence and agree either by accident or due to interrelated functional constraints.

Only additional sequence determinations are likely to resolve this discrepancy. Of particular interest will be the sequences of dinoflagellate tRNA and cytochrome c and *Euglena* 5S RNA.

Two aspects of dinoflagellate nuclear organization deserve additional comment. First, it was originally proposed that dinoflagellate mitosis is the most primitive among all eukaryotes when it was not possible to detect an association between the segregating chromosomes and the extranuclear microtubules normally associated with the dividing nucleus (Kubai and Ris 1969). It was believed that the observed membrane attachments of the chromosomes were solely responsible for their movements, and that this was an indication of a close relationship with the Prokaryota. Although this situation is still reported for certain free-living species (Cachon et al. 1979), chromosome association with spindle microtubules has been reported for a free-living *Amphidinium* sp. (Oakley and Dodge 1974) and the parasitic dinoflagellates *Syndinium* (Ris and Kubai 1974) and *Oodinium* (Cachon and Cachon 1977). In the former, this interaction appears to span the nuclear membrane while in the latter two organisms it occurs at nuclear membrane pores in a fashion very similar to that found in the polymastigids. Thus, although the nuclear membrane certainly plays a role in dinoflagellate mitosis, its importance is not unusual among the protists. For example, in the higher fungi, the membrane also persists and is the locus of the spindle pole bodies which organize an entirely intranuclear spindle apparatus (Heath 1978). We feel it is not unreasonable to suggest that dinoflagellate mitosis, at least as it occurs in the parasitic dinoflagellates, is intermediate between the completely closed mitosis of the fungi and the open mitosis of higher eukaryotes, and that the peculiarities observed in certain free-living dinoflagellate mitoses are reductive. In fact, in *Oodinium*, the kinetochores undergo a reductive transition during the sporogenetic mitoses that generate free-living swarmers from the ectoparasitic plasmodium (Cachon and Cachon 1977).

The parasitic dinoflagellates are of additional interest because two species of the family *Syndiniaceae* have been reported to have typical amounts of basic chromosomal proteins, as indicated by histochemical analysis. Moreover, they lack the typical fibrillar ultrastructure of free-living dinoflagellate chromosomes, although the chromatin is permanently consensed (Ris and Kubai 1974; Hollande 1974). In fact, even the latter feature is not universal – in the dinoflagellates *Oodinium*, *Blastodinium* and *Noctiluca*, chromatin condensation exists only during sporogenesis (Cachon and Cachon 1977; Soyer 1971; Soyer 1972). These exceptions to the peculiar nuclear features of most dinoflagellates are consistent with our proposal that the typical free-living state is not a primary one.

In conclusion, although it is clear from their fossil record (Loeblich 1974) and the great diversity of their form and function that the dinoflagellates are an ancient group, the characteristics of their DNA sequence organization (Hinnebusch et al. 1980), the 5S RNA sequence, and the affinities they exhibit with other flagellates argue that these are not the most primitive of the eukaryotes, and that the absence of histones in the free-living species is secondary. If this proposal can be strengthened by additional molecular analysis it will imply that at least among the protists, the histone-associated subunit structure of eukaryotic chromatin may only serve the structural role of DNA packaging and can be dispensed with entirely under special circumstances.

Acknowledgements. The authors are grateful to Professor Paul Doty for the use of facilities, Debra Peattie and Forrest Fuller for advice and gifts of materials in RNA sequencing, and Nancy Hinnebusch for help in preparing the manuscript. This research was supported by NSF grants PCM-8017146 and PCM-76-11158 to L.C.K.

Appendix

Correction of the Difference Matrix for Multiple Substitutions Assuming an Equal Probability of Transitions and Transversions. For a particular site in the molecule, the probability that after m mutations the site will be unchanged is the following:

$$S(m) = I_A P_{AA}(m) + I_U P_{UU}(m) + I_G P_{GG}(m) + I_C P_{CC}(m),$$

where I_A, I_U, I_G, and I_C are the probabilities that the site is initially A, U, G, or C, respectively, and the $P_{ii}(m)$s are the probabilities that after m mutations, the particular site of type i is unchanged. The quantity $1 - S(m)$ will be the probability that a base change has occurred at the position in question, and the product $L(1 - S(m))$ (where L is the molecule length) will be the expected number of observed differences in the molecule as a whole after m mutations. The problem is to tabulate $1 - S(m)$ for each value of m, such that given the number of observed differences between two sequences, one can provide the corresponding value of m as a probabilistic estimate of the number of underlying mutations that produced this difference.

This problem can be solved by noting that the base changes which occur at each site form a Markov chain. We can write the probabilities P_{ij} for a single-base change in the whole sequence in matrix form as follows:

$$\underset{\sim}{P} = \begin{array}{c} \\ A \\ U \\ G \\ C \end{array} \begin{array}{cccc} A & U & G & C \\ \left[\begin{matrix} (L-1)/L \\ f_{AU}/L \\ f_{GA}/L \\ f_{CA}/L \end{matrix} \right. & \begin{matrix} f_{AU}/L \\ (L-1)/L \\ f_{GU}/L \\ f_{CU}/L \end{matrix} & \begin{matrix} f_{AG}/L \\ f_{UG}/L \\ (L-1)/L \\ f_{CG}/L \end{matrix} & \left. \begin{matrix} f_{AC}/L \\ f_{UC}/L \\ f_{GC}/L \\ (L-1)/L \end{matrix} \right] \end{array}$$

where f_{ij} is the probability that a single base change occurring anywhere in the molecule that begins with a base of type i will arrive at a base of type j. f_{ij}/L is the probability that, after a single mutation, a particular site in the molecule which is of type i, will be converted to a type j site, and $(L - 1)/L$ is the probability that the mutation will not occur at the particular site in question. Markov theory states that after m mutational steps the probability of any given site existing in state j after beginning in state i, the $P_{ij}(m)$ from above, will be the ijth element of the matrix $\underset{\sim}{P}^{m}$. Thus, to find the quantities $P_{AA}(m)$, $P_{UU}(m)$, etc., we raise the matrix $\underset{\sim}{P}$ to the mth power and take the left to right diagonal elements.

To incorporate an equal transition-transversion frequency, we set $f_{AG} = f_{AC} + f_{AU}$, $f_{UC} = f_{UA} + f_{UG}$, etc. The values of the I_{ii}s were set equal to one another at 0.25.

Interestingly, after generating a table of observed differences (the $1 - S(m)$ values) versus expected mutational differences (m) using this model, and comparing them with the comparable values calculated from the Holmquist (1972) model (in which all of the f_{ij}s are equal), it is found that the differences between the two are very slight, at least for sequences of length 100–150. For example, for a sequence length of 120, 71 observed differences (the maximum number for the 5S RNA sequences considered in Chapter IV) implies 139 mutational differences using the Holmquist correction and 143 using that derived here (< 3% difference). The disparity between the two lessens as the number of observed differences gets smaller, so that the above example is the largest difference for the 5S RNA calculations. For the 5.8S RNA sequences, where 77 was the greatest number of observed differences between any of the sequences compared, the difference in the two corrections is only about 4%. Thus, whether we assume a mutational process in which transitions are twice as favored as transversions, or that base-changes are completely random in nature, will make no significant difference in the final results.

References

Azad AA (1979) Nucleic Acids Res 7:1913–1929

Cachon J, Cachon M, Salvano P (1979) Arch Protistenk 122: 43–54

Cachon J, Cachon M (1977) Chromosoma 60:237–251

Cavalier-Smith T (1975) Nature 256:463–468

Cavalier-Smith T (1978) J Cell Sci 34:247–278

Crick FHC (1966) J Mol Biol 19:548–555

Dayhoff MO (1978) Atlas of Protein Sequence and Structure, vol 5, suppl 3, Nation Biomed Res Foundation, Washington, DC

Demoulin V (1975) Bot Rev 40:315–345

Dobzhansky T, Ayala FJ, Stebbins GL, Valentine JW (1977) Evolution ch 12, WH Freeman & Co, San Francisco

Dodge JD (1965) Excerpta Med Int Congr Ser 19:339

Dodge JD (1973) The Fine Structure of Algal Cells, ch 3, Academic Press, New York

Duffus JH, Penman CS, Webb NWG (1973) Experientia 29: 632–633

Erdmann VA (1976) Prog Nucleic Acids Res Mol Biol 18: 45–90

Erdmann VA (1980) Nucleic Acids Res 8:r31–r47

Fitch WM (1976) J Mol Evol 8:13–40

Fitch WM, Margoliash E (1967) Science 155:279–284

Fox GE, Woese CR (1975) Nature 256:505–507

Giesbrecht P (1962) Zentralbl Bakteriol Parasitenk Infektionskr Hyg Abl 1: Orig 196:516–519

Hamkalo BA, Rattner JB (1977) Chromosoma 60:39–47

Heath BI (1978) Nuclear Division in the Fungi. Academic Press, New York, p 89

Herr W, Noller HE (1975) FEBS Lett 53:248–252

Hinnebusch AG, Klotz LC, Immergut E, Loeblich AR III (1980) Biochemistry 19:1744–1754

Hinnen A, Hicks JB, Fink GR (1978) Proc Natl Acad Sci USA 75:1929–1933

Hollande A (1974) Protistologica 10:413–451

Holmquist R (1972) J Mol Evol 1:115–133

Horgen PA, Silver JC (1978) Ann Rev Microbiol 32:249–284

Hori H, Osawa S (1979) Proc Natl Acad Sci USA 76:381–385

Klotz LC, Komar N, Blanken RL, Mitchell RM (1979) Proc Natl Acad Sci USA 76:4516–4520

Kubai DF (1973) J Cell Sci 13:511–552

Kubai D, Ris H (1969) J Cell Biol 40:508–528

Larue B, Cedergren RJ, Sankoff D, Grosjean H (1979) J Mol Evol 14:287–300

Leedale GF (1968) In: Buetow DE (ed) The Nucleus in Euglena, ch 5, Academic Press, New York

Loeblich AR Jr (1974) Taxon 23:277–290

Loeblich AR III (1976) J Protozool 23:13–28

McDonald K (1972) J Phycol 8:156–166

Marco Y, Rochaix JD (1980) Mol Gen Genet 177:715–723

Maxam AM, Gilbert W (1980) Methods Enzymol 65:521–522

Maxwell IH, Maxwell F, Hahn WE (1977) Nucleic Acids Res 4:241–246

Nazar RN, Sitz TO, Busch T (1975) J Biol Chem 250:8591–8597

Nishikawa K, Takemura S (1974) J Biochem 76:935–947

Noller HF, Garrett RA (1979) J Mol Biol 132:621–636

Oakley BR, Dodge JD (1974) J Cell Biol 63:322–325

Pavlakis GN, Jordan BR, Wurst RM, Vournakis JN (1979) Nucleic Acids Res 8:2213–2238

Peattie D (1979) Proc Natl Acad Sci USA 76:1760–1764

Philippsen P, Thomas M, Kramer RA, Davis RW (1978) J Mol Biol 123:387–404

Ragan MA, Chapman DJ (1978) A Biochemical Phylogeny of the Protists, Academic Press, New York

Ris H, Kubai DF (1974) J Cell Biol 60:702–720

Rizzo PJ, Burghardt RC (1980) Chromosoma 76:91–99

Rizzo PJ, Nooden LD (1974) Biochim Biophys Acta 349: 402–444

Nishikawa K, Takemura S (1974) J Biochem 76:935–947

Rubin GM (1975) In: Prescott DM (ed) Methods in Cell Biology vol 12. Academic Press, New York, p 45

Sagan L (1967) J Theor Biol 14:225–274

Sankoff D, Cedergren RJ, Lapalme G (1976) J Mol Evol 7: 133–149

Schwartz RM, Dayhoff MO (1978) In: Dayhoff MO (ed) Atlas of Protein Sequence and Structure, vol 5, suppl 3, Nation Biomed Res Foundation, Washington DC

Soyer MO (1971) Chromosoma 33:70–114

Soyer MO (1972) Chromosoma 39:419–441

Tanaka Y, Dyer TA, Brownlee GG (1980) Nucleic Acids Res 8:1259–1272

Tartof KD (1975) Ann Rev Genet 9:355–385

Taylor FJR (1976) J Protozool 23:28–40

Tonnesen T, Engberg J, Leick V (1976) Eur J Biochem 63: 399–407

Tuttle RD, Loeblich AR III (1975) Phycologia 14:1–8

Vigne R, Jordan BR (1977) J Mol Evol 10:77–86

Werner-Schlenzka H, Werner E, Kroger H (1978) Comp Biochem Physiol 61B:587–591

Wigler M, Sweet R, Sim GK, Wold B, Pellicer A, Lacy E, Maniatis T, Silverstein S, Axel R (1979) Cell 16:777–785

Wilson AC, Carlson SS, White TJ (1977) Ann Rev Biochem 46: 573–639

Received October 24, 1980/Revised February 9, 1981