Volume 301, number 2, 127–131 FEBS 10942 April 1992
© 1992 Federation of European Biochemical Societies 00145793/92/$5.00

Hypothesis

Controversy on chloroplast origins

Peter J. Lockhart[a], David Penny[a], Michael D. Hendy[b], Christopher J. Howe[c], Timothy J. Beanland[c] and Anthony W.D. Larkum[d]

[a]*Molecular Genetics Unit,* [b]*Department of Mathematics, Massey University, Palmerston North, New Zealand,* [c]*Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QW, UK* and [d]*School of Biological Sciences, University of Sydney, Sydney, NSW 2006, Australia*

Received 20 January 1992; revised version received 3 March 1992

Controversy exists over the origins of photosynthetic organelles in that contradictory trees arise from different sequence, biochemical and ultrastructural data sets. We propose a testable hypothesis which explains this inconsistency as a result of the differing GC contents of sequences. We report that current methods of tree reconstruction tend to group sequences with similar GC contents irrespective of whether the similar GC content is due to common ancestry or is independently acquired. Nuclear encoded sequences (high GC) give different trees from chloroplast encoded sequences (low GC). We find that current data is consistent with the hypothesis of multiple origins for photosynthesic organelles and single origins for each type of light harvesting complex.

Evolution; Chloroplast origins; Substitutional bias; Phylogenetic inference

1. INTRODUCTION

Two main hypotheses have been proposed for the evolutionary relationship of photosynthetic organelles to photosynthetic prokaryotes. On the basis of the pigment-protein composition of light harvesting complexes (LHCs) and thylakoid stacking, it has been suggested that chloroplasts with different LHCs derive from separate events of endosymbiosis [1–3]. This hypothesis also receives support from analyses of nuclear-coded sequences which indicate that several eukaryotic protohost cells [4–10] and eubacterial lineages (giving rise to proto-endosymbionts) [11–14] have been involved in endosymbioses.

However, comparative analyses of most chloroplast-coded sequences suggest that only one eubacterium has been involved in endosymbiosis. That is, chloroplast coded sequences (with the exception of the rubisco-large subunit (*rbcL*) data set [12,14]; see Discussion) indicate that the origin of photosynthetic organelles in monophyletic. The inference from these data are that the original endosymbiont was characterized by or later developed several types of LHC. This second hypothesis follows from nucleotide and amino acid gene trees which show all classes of photosynthetic organelles joining as more closely related to each other than to any photosynthetic prokaryote (e.g. Fig. 1 and [16–18]).

In this communication we present an hypothesis to explain this contradiction of chloroplast origins inferred from sequences of different genomic compartments (nuclear and chloroplast). This hypothesis is based on the observation that methods of reconstructing evolutionary trees select the wrong tree when sequences have independently acquired the same high GC or AT content (compositional bias) [13,19]. The hypothesis is consistent with the evidence from sequences (both nuclear and chloroplast encoded), the performance of tree building methods under conditions of different GC contents, and pigment protein composition and ultrastructure of LHCs.

We propose that chloroplast located sequences have independently gained a high AT content and that consequently trees built from these sequences indicate fewer endosymbiosis events than actually took place.

2. TREE INFERENCE METHODS CAN SELECT THE WRONG TREE

It is known that unequal rates of evolution can lead parsimony and distance tree building methods to select the wrong tree even for very long sequences (referred to as 'Felsensteins paradox' or 'attraction of long edges' [20,21]). We have shown theoretically that a similar problem can arise where 2 lineages have independently increased in GC or AT content [19]. In Fig. 2 the effect

Correspondence address: P. Lockhart, Molecular Genetics Unit, Massey University, Palmerston North, New Zealand. Fax: (64) (6) 350 5637.

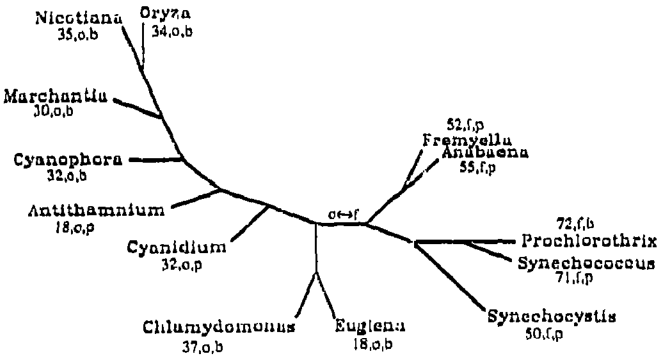

Fig. 1. The most parsimonious tree using 1st and 2nd codon position data for *psbA* (photosystem II D 1 protein gene) from 13 oxygenic-photosynthetic taxa [15]. The tree supports a single endosymbiotic event for all chloroplasts but that a phycobilin-Chl-*b* change occurred 3 times. Maximum likelihood and Neighbor Joining (Jukes/Cantor distance) trees [15] similarly indicate a monophyletic origin for the organelle sequences (results not shown). Alignment of sequences and tree reconstruction used the approach given elsewhere [13]. Indicated in the figure are (%GC) contents at the 3rd codon position in the compared sequences, whether the sequences are from organelles (o) or free living prokaryotes (f) and whether LHCs are of the Chl-*a*/*b* (b) or Chl-*a*/phycobilin (p) type. The sequences were from organelles (with Chl-*a*/*b* LHCs; indicated as o,b): *Nicotiana tabacum*, *Marchantia polymorpha*, *Chlamydomonas reinhardtii*, *Euglena gracilis*, (with Chl-*a*/phycobilin LHCs; indicated as o,p) *Cyanophora paradoxa*, *Antithamnium sp.*, *Cyanidium caldarium*, and from free-living photosynthetic bacteria: (with Chl-*a*/phycobilin LHCs) *Synechocystis sp. 6803*, *Synechococcus sp.* PCC7942, *Anabaena sp.* PCC7120, *Fremyella diplosiphon* and (with Chl-*a*/*b* LHCs) *Prochlorotrix hollandica*. Sequences were obtained from GenBank(ver. 69).

of such a bias in nucleotide substitution processes to mislead inference has been calculated for 4 taxa. In this example two unrelated taxa have independently acquired the same GC content. We show the range of values under which parsimony is expected to select the correct tree and conversely the range of values it selects the wrong tree. For this method (and presumably others) as both the central branch length decreases and as the GC content of the 2 unrelated sequences increases, the unrelated sequences of similar composition are joined rather than the sequences that last shared a common ancestor.

A possible example of this problem in tree building is demonstrated in Fig. 3 with the same 7 taxa for (3A) *rbcL* (rubisco-large subunit gene, which is chloroplast encoded in all photosynthetic organelles) and (3B) *rbcS* (rubisco-small subunit gene, which is nuclear encoded in higher plants). The two parsimony consensus trees [15] shown are contradictory in that they indicate different origins for the photosynthetic organelle ('cyanelle') of *Cyanophora paradoxa* and the evolution of light harvesting complexes. Tree reconstruction with maximum likelihood and neighbor joining methods [15] also gives similar results to parsimony (trees not shown).

With all three methods trees were reconstructed from 100 bootstrap samples [22]. That is, for each set of aligned sequences, sampling of columns with replacement was carried out to generate a further 100 'bootstrap' data set. This approach identifies the most stable relationships in the tree (those which occur most often in the 100 samples). The results obtained using the bootstrap, give strong support for a chloroplast-cyanelle affinity with the *rbcL* data and strong support for a cyanelle-cyanobacterial affinity with the *rbcS* data (see Fig. 3 and legend).

The bootstrap indicates that there is sufficient information in both *rbcL* and *rbcS* data sets to make it unlikely that the trees would change were more se-

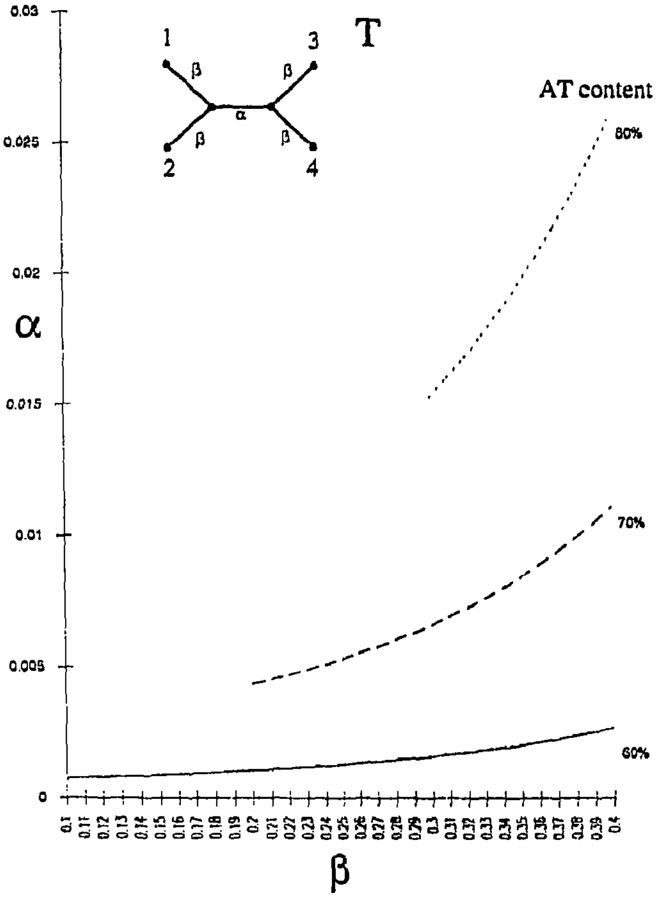

Fig. 2. Calculation for edge lengths and base composition differences needed to mislead a parsimony analysis in a 4 taxon case. The Jukes Cantor one parameter model [15] for nucleotide changes on the tree T was modified so that the AT frequency is Γ at vertices 1 and 3, but 50% at the remaining vertices. The probability that there is a nucleotide mutation per site is α on the internal edge and β on each of the remaining edges. With Γ = 50%, parsimony is consistent [20,21] but for larger values of Γ, parsimony can be inconsistent, particularly for small values of α. The figure shows those values of α and β for which inconsistency occurs for Γ = 60%, 70%, and 80%. Points (α,β) below the lines for each Γ are the values for which parsimony will be inconsistent. Conversely points above the line show when parsimony is consistent. For these values, the longer the sequences, the greater the likelihood that parsimony will select the incorrect tree. This occurs typically for small values of α, larger values of β, and particularly with higher proportions of Γ.

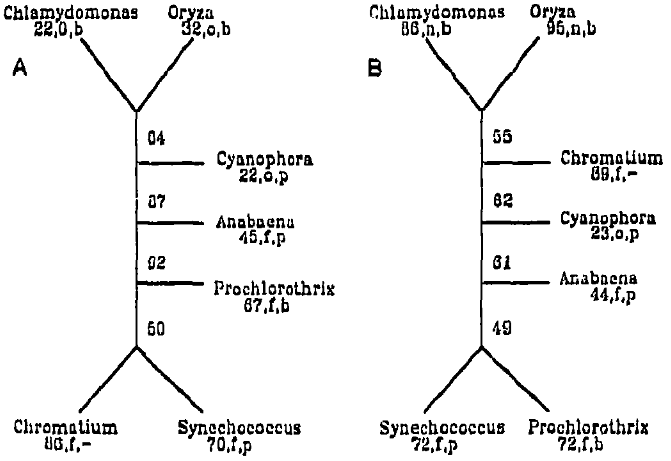

Fig. 3. Consensus Parsimony trees [15] using 1st and 2nd codon position data for 100 bootstrap samples for (A)*rbcL* (B)*rbcS*. Base composition (%GC) at the 3rd codon position of compared sequences is indicated as is whether the sequences are nuclear encoded (n), organelle encoded (o) or from free living prokaryotes (f). The number of times species grouped as monophyletic is also indicated. Similar trees and results were obtained with neighbor joining (Jukes/Cantor distance) and a maximum likelihood method. In the *rbcL* data set the cyanelle grouped with the chloroplast encoded homologues in 89 (neighbor joining) and 89 (maximum likelihood) of the 100 bootstrap trees. With the *rbcS* data set these methods grouped the cyanelle with the photosynthetic bacteria to the exclusion of the plant nuclear encoded homologues in 71 (neighbor joining) and 61 (maximum likelihood) of the 100 bootstrap trees. The sequences were from the cyanelle of *Cyanophora paradoxa*, plants: *Nicotiana tabacum* and *Chlamydomonas reinhardtii*,from oxygenic photosynthetic bacteria: *Synechococcus sp.* PCC6301, *Anabaena sp.* PCC7120 and *Prochlorothrix hollandica* and from a non-oxygenic photosynthetic bacteria: *Chromatium vinosum* which is used as the out-group. The sequences were obtained from GenBank(ver. 69).

quence data available. The bootstrap test also indicates the extent of the difference between the two trees of Fig. 3. Close examination of this figure shows that it is the placement of the single *Chromatium* edge which distinguishes alternative hypotheses of origin for the cyanelle. The stability of this edge in the trees of Fig. 3A and B is indicated by the bootstrap. In the *rbcS* data set there is almost no evidence for the placement of this edge at a position which would support common ancestry for the cyanelle and chloroplasts. For example, with parsimony, the partition (grouping) which includes *Chromatium* + *Synechococcus* + *Prochlorothrix* occurs in 92% of the bootstrap *rbcL* trees (Fig. 3A). This partition does not occur at all in any of the 100 *rbcS* bootstrap trees. Similarly, the partition *Chromatium* + *Synechococcus* occurs in 50% of the *rbcL* bootstrap trees but only in 2% of the *rbcS* bootstrap trees.

These observations with the bootstrap also illustrate the difference between the concepts of convergence and consistency in tree building. The bootstrap provides information on whether the sequences are long enough for a method to converge to a single tree [19,20]. It does not provide evidence for consistency of a tree building method (give indication of whether the selected tree is the true historical tree). When the fit between model and data is poor, as in the *rbcL* and *rbcS* data sets (Table I), methods of inference may be inconsistent and give the wrong tree [13,19]. Our example highlights this point because our results suggest that the presence of the different GC contents between nuclear and chloroplast sequences are sufficient to lead inference methods to converge to different trees.

In Figs. 1 and 3 the GC biases at 3rd codon positions for compared sequences are given. The values indicate the direction of substitutional biases which are most evident at codon position 3 but also present at 1st and 2nd codon positions in these taxa [13]. When trees are reconstructed from 1st and 2nd codon position data

Table I

The fit between model and predicted data [10] for the trees in Figs. 1 and 3. χ^2 values indicate a poor fit between model and data. The Hadamard (discrete Fourier) transform estimates the optimal parameters for an evolutionary model and uses these estimates to predict the frequencies of partitions in the data [19,20,23]. This is done for each of the (2^{n-1}) bipartitions. The number of predicted and observed bipartitions (averaged over the 7 ways of forming them [19]) for edges supporting the trees (in the tree; 'i-tree') in Figs. 1 and 3 are given together with those supporting other trees ('o-trees'). 'Constant' are the number of patterns where all taxa have the same character state. 'Pendant' are the number of changes on terminal edges and present in all trees. These results show that these data do not fit a tree under the standard assumptions, presumably because the condition of constant nucleotide frequencies is not met. Little is known about the consistency of tree building methods when the standard assumptions are not met

Nucleotide position	Sum of entries		Number of bipartitions
	Predicted	Observed	
psbA, 13 taxa, 662 columns			
Constant	1029.13	938.25	1
Pendant	96.09	88.25	13
in-Tree	27.40	25.25	10
o-Trees	6.37	106.75	4072
Sum	1159.00	1158.50	4096
Value of χ^2, 3 d.f. 1590			
rbcL, 7 taxa, 1353 columns			
Constant	1272.66	1264.00	1
Pendant	206.96	156.00	7
in-Tree	62.83	33.50	4
o-Trees	19.55	107.00	52
Sum	1561.00	1561.00	64
Value of χ^2, 3 d.f. 417			
rbcS, 7 taxa, 170 columns			
Constant	172.18	167.25	1
Pendant	75.53	51 75	7
in-Tree	27.66	13.00	4
o-Trees	22.63	65.50	52
Sum	298.00	297.50	64
Value of χ^2, 3 d.f. 97			

these biases cause the close placement of species with similar GC contents but with different LHCs (as shown in Figs. 1 and 3).

3. EARLIER STUDIES ON THE CYANELLE

Previously, we [13] examined in detail the phylogeny of the Chl-*a* and phycobilin-containing photosynthetic organelle (cyanelle) of *Cyanophora paradoxa* since, despite its pigment composition and membrane ultrastructure being similar to cyanobacteria and red algal chloroplasts, some authors had suggested it shared the same endosymbiont as chl-*a*/*b* containing (green) chloroplast [17,25]. We proposed that green chloroplasts and the cyanelle had independently acquired a similar AT composition thus causing these groups to be spuriously brought together in tree analyses [13]. Such an hypothesis would account for the discrepancy between the general biochemical evidence and the evidence from DNA sequences which indicate different origins for the cyanelle.

A prediction from our hypothesis was that if cyanelle and green chloroplast homologues of different compositional biases were compared then these species would not group together in gene trees. To test this we compared sequences for green chloroplast coded (AT-rich; for ATP synthese-beta subunit gene: *atpB* and elongation factor Tu gene: *tufA*) and nuclear coded (relatively GC-rich; ATP synthase-delta subunit gene: *atpD*) sequences. We found that, at all 3 codon positions, similar GC contents could be sufficient to mislead inference of cyanelle origins [13].

As shown in [12] and Fig. 3B, analysis of *rbcS* has also been found to support our findings on the cyanelle. Like the *atpD* data set, *rbcS* sequences (also nuclear encoded in Chl-*a*/*b* plants) are relatively GC-rich in Chl-*a*/*b* containing plants and AT rich in the cyanelle. These data sets differ from others used for studying chloroplast origins (e.g. Fig. 1, [12,13,16,17]) since in these latter data sets both cyanelle and organelle homologues have a high AT content. In contrast the GC composition of compared photosynthetic prokaryotes varies greatly from the organelle sequences (e.g. Fig. 1) and the fit between model and data is expected to be very poor as a result of this (e.g. as for the *psbA* sequences: Table I).

4. CHLOROPLAST ORIGINS

The implications from our findings for chloroplast origins are that if the high-AT base composition common to all plant organelle genomes [12] is, like the cyanelle and chl-*a*/*b* organelle genomes, acquired independently then methods of tree analysis will be expected to bring together these AT-rich sequences and suggest few (or one) endosymbiosis event (as in Fig. 1). Such a suggestion at present receives support from analyses of *petF* sequences [11,14]. In these, phycobilin containing organelles other than the cyanelle are also found to join with cyanobacteria and not with Chl-*a*/*b* containing chloroplasts. Since *petF* is also thought to be nuclear located (with the exception of the cyanelle) [14] this is in agreement with our predictions and provides explanation for the discrepancy between these results and those reported for red-algal chloroplast-located sequences (e.g. Fig. 1 and [18]).

At first sight analyses of *rbcS* and *rbcL* data sets also appear to support polyphyletic origins for Chl-*a*/*b* and some Chl-*a*/phycobilin containing organelles other than the cyanelle [12,14]. However, inference from these sequence data (which include chloroplasts in red and green plants) lead to different results from those obtained with 16S rRNA and *psbA* sequences e.g. [18], Fig. 1). Analyses from *rbcLS* data indicate a deep split between phycobilin-containing taxa and suggest that evolution of oxygenic photosynthesis occurred twice [12]. In view of these anomalous results we would caution against uncritical acceptance that *rbcLS* data indicate polyphyletic origins for Chl-*a*/*b* containing and some Chl-*a*/phycobilin organelles. Possible explanations for the lack of congruence between trees has been made [18]. A further suggestion is that the comparison made across the split is between paralagous and not orthologous sequences. This explanation is favoured by the observation that subsets of the data behave similarly in analyses to other sequences from the same taxa [13,24].

5. RESOLVING THE CONTROVERSY

Three general approaches will help resolve the present controversy, and test the hypothesis presented here. These include the development of more robust methods of tree reconstruction that will still be consistent when sequences have different GC/AT composition. We have made some progress in this direction. Determination of additional sequences where homologues have different GC contents in plants with unlike LHCs will also help. Thirdly, an understanding of processes leading to different base compositions would help in modelling the problem for inference methods. (Are there biochemical reasons why photosynthetic organelles have a high AT base composition? [13])

In view of our findings we caution against the uncritical acceptance of analyses which link sequences with similarly biased sequences but unlike LHCs [12,13,16–19,25–28]. Despite the general acceptance of inference from nucleotide sequence data, the robustness of methods of inference with markedly divergent base compositions is undetermined. Present methods are unreliable (see Fig. 2) in these circumstances (i.e. when compared sequences from unrelated taxa have independently acquired similar base compositions) hence inference from sequence data should be corroborated with biochemical and ultrastructural data where available.

Interestingly, the analogous problem of inconsistency in phylogenetic analyses resulting from rate inequalities [20,21] appears similar in effect to that of substitutional biases and this previously may have led to some confusion in interpreting results. It is important to recognise that solutions to these two problems of inference may not be the same. For example, substitutional biases which lead to convergence can mislead evolutionary parsimony [29] which has been designed to cope with convergence resulting from unequal rate effects [30]. Despite improvements in this approach [31] the assumption of unequal (similarly biased) base frequencies in this methodology are unlikely to be met by nucleotide sequence data sets for photosynthetic and other anciently diverged taxa. Therefore, the problem of inference from sequence data under conditions of irregular (differently biased) base compositions for compared taxa appears as a serious and general problem facing those interested in elucidating the relationships between early diverged form of life.

Acknowledgements: This work was supported by the Australian Research Council, the N.S.W. Linnean Society, the Australia New Zealand Foundation and the U.K. Science and Engineering Research Council.

REFERENCES

[1] Raven, P.H. (1970) Science 169, 641–646.
[2] Gibbs, S.P. (1981) Ann. NY Acad. Sci. 193–207.
[3] Whatley, J.M. and Whatley, F.R. (1981) New Phytol. 87, 233–247.
[4] Perasso, R., Baroin, A., Hu Qu, L., Bachellerie, J.P. and Adoutte, A. (1989) Nature 339, 142–144.
[5] Leanaers, G., Maroteaux, L., Michot, B. and Herzog, M. (1989) J. Mol. Evol. 29, 40–51.
[6] Douglas, S.E., Murphy, C.A., Spencer, D.F. and Gray, M.W. (1991) Nature 350, 148–151.
[7] Penny, D. and O'Kelly, C.J. (1991) Nature 350, 106–107.
[8] Ariztia, E., Anderson, R.A. and Sogin, M.L. (1991) J. Phycol. 27, 428–436.
[9] Goldschmidt-Clermont, M. and Rahire, M. (1986) J. Mol. Biol. 191, 421–432.
[10] Viale, A.M., Kobayashi, H. and Akazawa, T. (1989) J. Bacteriol. 171, 2391–2400.
[11] Matsubara, H., Hase, T., Wakabayashi, S. and Wada, K. (1980) in: The Evolution of Protein Structure and Function, (D.S. Sigman and A.B. Brazier, Eds.), pp. 245–266.
[12] Morden, C.W. and Golden, S.S. (1991) J. Mol. Evol. 32, 379–395.
[13] Lockhart, P.J., Howe, C.J., Beanland, T.J. and Larkum, A.W.D. (1992) J. Mol. Evol. 34, 153–162.
[14] Luttke, A. (1991) Endocytobiosis and Cell Res. 8, 75–82.
[15] Felsenstein, J. (1991) PHYLIP 3.4 Manual, Univ. California Herbarium, Berkeley, California.
[16] Turner, S., Burger-Wiersma, T., Giovannoni, S.J., Mur, L.R. and Pace, N.R. (1989) Nature 337, 380–382.
[17] Evrard, J.-L., Kuntz, M. and Weil, J.-H. (1990) J. Mol. Evol. 30, 16–25.
[18] Douglas, S.E. and Turner, S. (1991) J. Mol. Evol. 33, 267–273.
[19] Penny, D., Hendy, M.D., Zimmer, E.A. and Hamby, R.K. (1990) Aust. Syst. Bot. 3, 21–38.
[20] Penny, D., Hendy, M.D. and Steel, M.A. (1991) in: Phylogenetic analysis of DNA sequences (M. Miyamoto and J. Cracraft, Eds.), Oxford University Press, pp. 155–183.
[21] Felsenstein, J. (1978) Syst. Zool. 27, 401–410.
[22] Felsenstein, J. (1985) Evolution 39, 783–791.
[23] Penny, D., Hendy, M.D. and Henderson, I.M. (1987) Cold Spring Harbor Symp. Quant. Bio. 52, 857–862.
[24] Ritland, K. and Clegg, M.T. (1987) Am. Nat. 130, S74–S100.
[25] Giovannoni, S.J., Turner, S., Olsen, G.J., Barnes, S., Lane, D.J. and Pace, N.R. (1988) J. Bacteriol. 170, 3584–3592.
[26] Lockhart, P.J., Beanland, T.J., Howe, C.J. and Larkum, A.W.D. (1992) Proc. Natl. Acad. Sci. USA 89, in press.
[27] Palenik, B. and Haselkorn, R. (1992) Nature 355, 265–267.
[28] Urbach, E., Robertson, D. and Chisholm, S.W. (1992) Nature 355, 267–270.
[29] Lake, J. (1987) Mol. Biol. Evol. 4(2), 167–191.
[30] Lockhart, P.J. (1990) PhD Thesis, University of Sydney.
[31] Sidow, A. and Wilson, A.C. (1990) J. Mol. Evol. 31, 51–68.