S9-012

Phylogenetic reconstruction of ancient photosynthetic lineages using chlorophyll and bacteriochlorophyll biosynthetic genes

LS Jermiin[1], RE Blankenship[2], PJ Lockhart[3], AWD Larkum[4]

[1]*School of Biological Sciences, University of Sydney, NSW 2006, Australia. Email: jermiin@bio.usyd.edu.au*

[2]*Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287-1604, USA. Email: blankenship@asu.edu*

[3]*Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand. Email: p.j.lockhart@massey.ac.nz*

[4]*School of Biological Sciences, University of Sydney, NSW 2006, Australia. Email: alark@mail.bio.usyd.edu.au*

Keywords: photosynthesis, evolution, maximum likelihood, model averaging, compositional heterogeneity

Introduction

Photosynthesis is an ancient process. Geochemical evidence of organisms that resemble cyanobacteria is ~2.5 billion years old (Summons et al., 1999). It is generally assumed that more primitive photosynthetic bacteria have existed as far back as 3.5 billion years ago (Schopf, 1993), so gaining knowledge about these organisms and their evolutionary history is an extremely difficult challenge.

Sequence data provide a powerful means to discern phylogenetic affinities but there may be many problems associated with analysis of photosynthetic organisms. The small sub-unit ribosomal RNA (SSU-rRNA) has been widely used in this regard and, in the case of the photosynthetic organisms, produces a tree, where photosynthesis emerges separately on several branches (Woese, 1987). This is considered unlikely given the complexity of photosynthesis.

Recently, Xiong et al., (2000) used genes involved in the synthesis of chlorophyll (Chl) and bacteriochlorophyll (BChl) to trace the origin and evolution of photosynthesis. The results suggested that purple bacteria were the first to diverge from the photosynthetic lineage and that the oxygenic cyanobacteria and plastids are the most recently emerged lineage. Interestingly, it was also suggested that *Heliobacteria* was a sister group to the oxygenic forms.

We have re-examined the sequence data studied by Xiong et al., (2000) using maximum-likelihood, surveying and model averaging methods with the aim to trace the origin and evolution of photosynthesis.

Material and methods

We have obtained sequence data from: *Chlorobium tepidum*, *Chloroflexus aurantiacus*, *Heliobacillus mobilus* and *Rhodobacter capsulatus* (anoxygenic photosynthetic bacteria); *Synechocystis* PCC6803 (a cyanobacteria); *Chlorella vulgaris* and *Porphyra purpurea* (photosynthetic eukaryotes); and *Klebsiella pneumoniae* and *Azotobacter vinelandii*

(non-photosynthetic nitrogen fixing eubacteria). Genes involved in reduction of the B ring in chlorin (X, Y, and Z genes) were obtained from *Chlorobium*, *Chloroflexus* and *Rhodobacter* whereas genes involved in reduction of the D ring in porphyrin (L, N, and B genes) were obtained from the seven photosynthetic organisms. Genes involved in the reduction of di-nitrogen (the H, D, and K genes), used as a reference in the phylogenetic analysis, were obtained from *Klebsiella* and *Azotobacter*.

Amino acid sequences were used in the phylogenetic analysis because they provide more resolution than nucleotide sequences, when the third codon sites of the latter are ignored. This is a reasonable assumption given the time over which the sequences have evolved. All the amino acid sequences were obtained from Genbank (www.ncbi.nlm.nih.gov).

The three sets of homologous amino acid sequences (Homologue 1: XLH; Homologue 2: YND; Homologue 3: ZBK) were aligned with Clustal W (Thompson et al., 1994) and the result was then refined by visual inspection using GDE (Smith et al., 1994). This yielded three alignments with different characteristics.

Initially, assessment of compositional heterogeneity was done on the three alignments of varied sites using a method developed by Jermiin et al., (2002). The method compares all pairs of sequences and produces a ζ score for each pair; the distribution of the ζ scores is then charted and can be compared with the distribution of ζ scores from other data sets; a bell-shaped ζ-score distribution with a non-positive mean and a variance that is less than or equal to 1.0 indicates compositionally homogenous data (the ζ score increase with the increasing level of compositional heterogeneity; randomly generated ζ scores are rarely larger than 2.35).

Following the assessment of compositional heterogeneity, we conducted a phylogenetic analysis using ProtML (Adachi and Hasegawa 1996). Using the JTT-F model of amino acid substitution and the Kishino-Hasegawa test, we searched tree-space exhaustively and found many phylogenetic trees that did not differ significantly from the most likely tree. Since this set of 'good trees' implies there is uncertainty about the phylogenetic model, we used a model averaging approach (Jermiin et al., 1997) to generate a consensus tree.

Results

The sites that could be aligned were decomposed into subsets according to the variation at the sites. Table 1 shows that the three homologues differ substantially with respect to the number of unvaried sites. The relative proportion of unvaried sites (in relation to the number of sites without gaps) ranges from 19.2% in Homologue 1 to 2.7% and 0.9% in Homologues 2 and 3. This variation explains (to some degree) the difficulty in aligning Homologues 2 and 3, and suggests that the three homologues could have evolved under different selective pressures, which, in turn, may have implications on the phylogenetic analysis.

Table 1: Number of Amino Acids in Different Subsets of The Three Homologues

Homologue	Aligned Sites	Sites Excl. Gaps	Varied Sites	Unvaried Sites
1 (XLH)	293	239	193	46
2 (YND)	500	364	354	10
3 (ZBK)	620	321	318	3

Varied sites in the three alignments were used to examine compositional heterogeneity. Figure 1 shows the distributions of ζ scores for the three alignments. Assuming that the three homologues have had the same evolutionary history and that there are no lineage-specific differences in the distribution of sites that are free to vary, the result in figure 1 shows, that there is little evidence of compositional heterogeneity in Homologue 1. On the other hand, Homologues 2 and 3 are clearly compositionally heterogeneous, with the larger ξ scores due to the comparisons between *Chlorobium* Z and most of the other Z, B, and K sequences and between many of the Y, N, and D sequences.

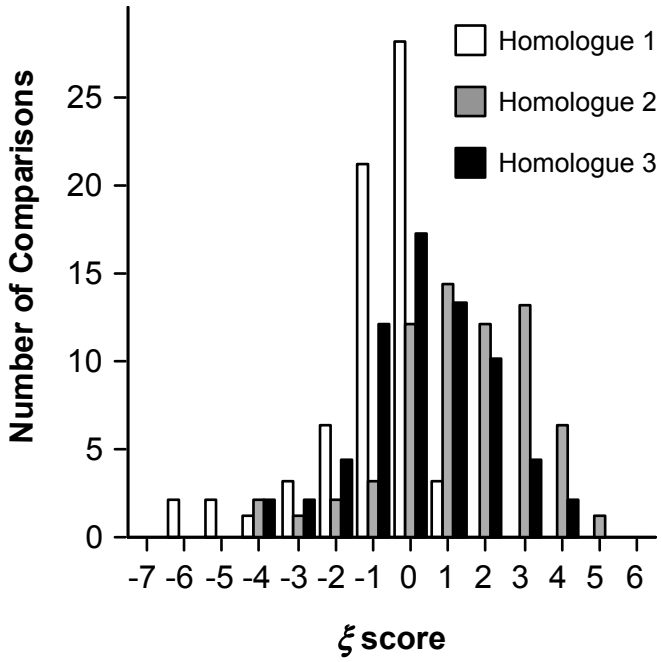

Fig. 1 ξ-score distributions for the three homologues, based on the varied sites (ξ is an approximately standard normal)

This result suggests that the X, L, and H genes may have evolved under stationary conditions; that many of the Y, N, and G genes have evolved under non-stationary conditions; and that many of the Z, B, and K genes may have evolved under stationary conditions (the Z gene from *Chlorobium* is the only notable exception). Therefore, we consider Homologue 1 to be more suitable for phylogenetic analysis than the other two homologues.

Figure 2A shows the most likely (ML) tree inferred from the L, X, and H gene products. Clearly, the three sets of genes cluster together. Within the L cluster *Heliobacillus* is at the root, but with low significance indicating the possibility of a trichotomy at the base. However, figure 2B shows that this pattern is not upheld when the model uncertainty is considered. Here *Heliobacillus* groups weakly with the oxygenic phototrophs, and these two groups separate strongly from the other photosynthetic bacteria. However, clearly it is possible that there is a trichotomy or even a quadrotomy at the base of the L genes.

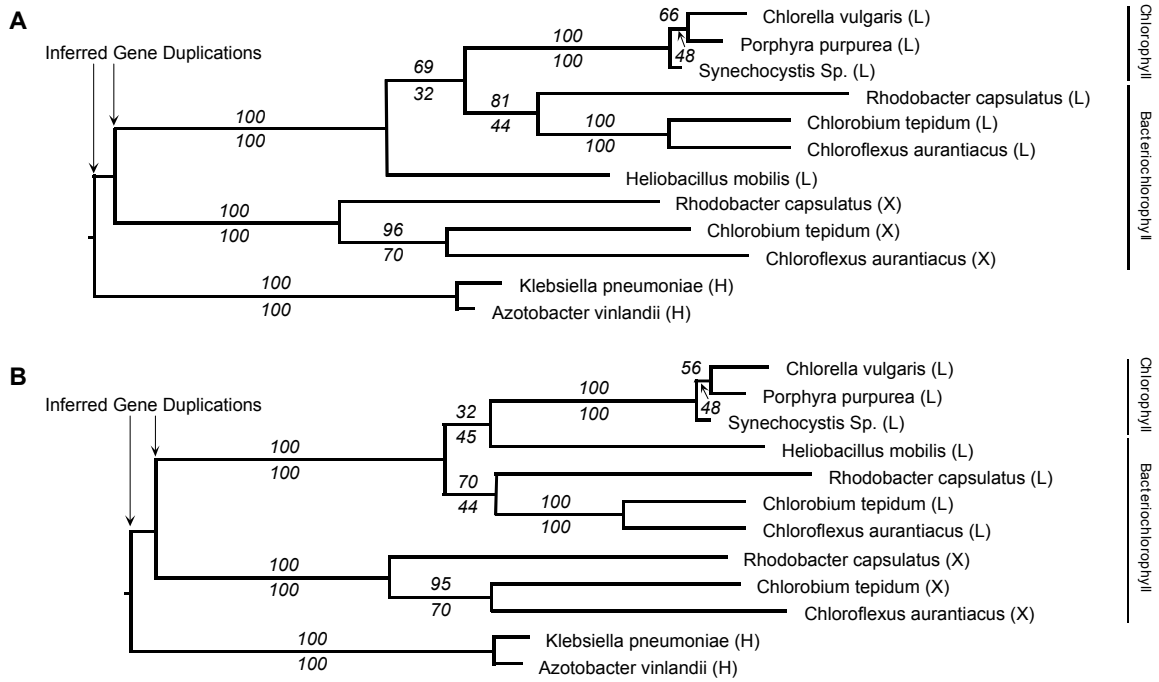

Fig. 2 The most likely tree (A) and the consensus tree (B) based on phylogenetic analysis of the L, X, and H gene products. Values above the edges are local bootstrap probabilities and values under the edges are relative likelihood scores, which were obtained using model averaging across the 70 'good trees' inferred from these data (size of Tree-space = 654,729,075).

Discussion

Xiong et al., (2000) have previously published analyses based on concatenated sequences of Homologues 1, 2, and 3. We analysed the homologues individually and used different alignments. Secondly, we assessed compositional heterogeneity, which can interfere with phylogenetic analysis, and discovered evidence of substantial compositional differences in two of the three homologues; hence we only used Homologue 1. Thirdly, we explored a larger proportion of tree space and discovered evidence of model uncertainty, and have used model averaging to summarise the results. Our approach provides more information than that used by Xiong et al., (2000) because the latter ignored other possible explanations of the data. From our results we draw the following conclusions:

1. *Rhodobacter* is the most diverged of the anoyxygenic photosynthetic bacteria, but there is little evidence to support the conclusion of Xiong et al., (2000) that it is the earliest lineage to diverge from the other photosynthetic forms;

2. There is weak support for the clustering of *Helobacillus* with the Cyanobacteria; it may indicate a distant relationship. The clustering is supported by the lack of X, Y, and Z genes in both of these groups. However, the timing of this loss is uncertain. The presence of a homodimer reaction centre in *Heliobacillus* suggests that the change to a heterodimer for both reaction centres in Cyanobacteria is a relatively recent event;
3. The cyanobacterial lineage is relatively young, and so is that of the plastids. This raises the question of what kinds of organisms preceded these lineages;
4. Our current approach provides a means to identify compositional heterogeneity. There is also evidence of lineage-specific rate-heterogeneity in our data, and this may also be a further factor confounding phylogenetic inference. Saturation of substitutions is another potential problem that needs to be addressed;

Our consensus tree presents a reasonable fit to the current data and appears to be better than the most likely tree. Interestingly, the root-to-tip distances are similar over the whole tree, with the exception of the H genes.

References

Adachi J, Hasegawa M (1996). *Computer Science Monographs Nº 28*. The Institute of Statistical Mathematics, Tokyo

Jermiin LS, Olsen GJ, Mengeren KL, Easteal S (1997). *Molecular Biology and Evolution* **14**, 1296–1302

Jermiin LS, Wilson SR, Easteal S (2002). *Molecular Biology and Evolution* (submitted)

Schopf JW (1993). *Science* **260**, 640–646

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994). *Computer Applications in the Biosciences* **10**, 671-675

Summons RE, Jahnke LL, Hope JM, Logan GA (1999). *Nature* **400**, 554–557

Thompson JD, Higgins DG, GibsonTJ (1994). *Nucleic Acids Research* **22**, 4673–4680

Woese CR (1987). *Microbiological Review* **51**, 221-271

Xiong J, Fischer WM, Inoue K, Nakahara M, Bauer CE (2000). *Science* **289**, 1724–1730