Gene duplication and the evolution of photosynthetic reaction center proteins

Peter J. Lockhart[a,*], Michael A. Steel[b], Anthony W.D. Larkum[c]

[a]*Lehrstuhl für Pflanzenökologie und Systematik, Universität Bayreuth, 95440 Bayreuth, Germany*
[b]*Department of Mathematics, University of Canterbury, Christchurch, New Zealand*
[c]*School of Biological Sciences, University of Sydney, Sydney, NSW, Australia*

Received 5 February 1996; revised version received 29 March 1996

Abstract We investigate the evolutionary relationships between photosynthetic reaction center proteins (D1, D2, L and M) and demonstrate that the pattern of nucleotide substitution in these is more complicated than has been assumed in previous phylogenetic analyses. We show that there are serious violations of methodological assumptions in previous published studies. We conclude that there is equal support for hypotheses indicating (i) a single gene duplication of an ancestral reaction center protein followed by diversification and (ii) two independent gene duplications giving rise to proteins in oxygenic and anoxygenic systems.

Key words: Reaction center protein; Photosynthesis origin

1. Introduction

In oxygenic photosynthesis, where the reaction center pigment is chlorophyll, two proteins D1 and D2 form the binding site for the special pair of chlorophylls (P680). D2 binds the primary quinone acceptor Q_A and D1 binds a second quinone acceptor Q_B. Electron transfer between P680 and QA takes place largely via D1. In the homologous pheophytin-quinone ('Q' type) reaction center of anoxygenic photosynthesis, where bacteriochlorophyll is the photosynthetic pigment (P870), protein M carries out the role of D2. Here protein L has the same the role as D1 (Fig. 1).

Based on the functionally similar roles common to L and D1, M and D2 it was thought that an ancient gene duplication originally gave rise to two proteins (X and Y in Fig. 2A) in an ancestral reaction center. These then diverged to give rise to the extant types. Based on this interpretation a strong expectation has been that comparative studies of protein sequences would find D1 and L most closely related and D2 and M most closely related (as in Fig. 2A) [1]. However, sequence studies have not supported this expectation. Analyses have favored evolutionary trees which place together D1 and D2; L and M as most closely related (e.g. as in Fig. 2B) [2-4]. If correct, this suggests that two independent gene duplications may have been responsible for the evolution of the extant proteins.

In this communication we report an investigation of the pattern of sequence evolution in the photosynthetic genes studied. Our results demonstrate the complexity of sequence evolution for reaction center proteins and also the difficulty in making reliable inferences from these data.

2. Materials and methods

We have reconstructed, for the taxa shown in Table 1, the alignment studied by Beanland [3] for a region of similarity extending over 123 amino acids and which is similar to other published alignments (e.g. [1]). On these aligned data we have estimated the proportion of codons which are free to vary using different subsets of sequences. We did this using the method of Sidow et al. [5]. To enumerate the number of constant and variable sites in different codon positions we used the programs PREPARE [6] and MEGA [7].

We have also investigated the effect of alignment order in the progressive multiple alignment of reaction center protein sequences. Specifically, we were interested in testing whether phylogenetic reconstruction for reaction center proteins could be biased as a result of the alignment order [8,9]. For this purpose we studied four sequences: *Spinacea oleracea* D1, D2 and *Rhodopseudomonas viridis* L, M. We first made pairwise alignments for all six possible pairs using Multalin4 [10]. For the three possible orders of alignment we then aligned pairs using the profile alignment procedure from CLUSTALV [11]. We used the PAM matrices implemented in both programs, with gap penalties 2–8 in the initial pairwise alignments and then accepted the default options for the profile alignment in CLUSTALV. This approach resulted in 21 (3×7) final multiple alignments for reaction center proteins. On each of these alignments support for the three possible bifurcating trees was evaluated using protein parsimony and the sites test implemented in PHYLIP3.5 [12].

3. Results and discussion

Reconstructing histories for the reaction center proteins is made difficult because, whilst reaction center protein sequences show enough regions of local similarity to be regarded as homologous, they show low overall identity [13]. Fig. 3 shows the alignment used by Beanland [3] in his inference of two independent gene duplications for the origin of D1, D2, L and M subunits. This alignment is similar to that of Michel and Deisenhofer [1]. It is also very similar to optimal alignments that we have reconstructed using Multalin4 or CLUSTALV [11]/CLUSTALW [14], when default options for

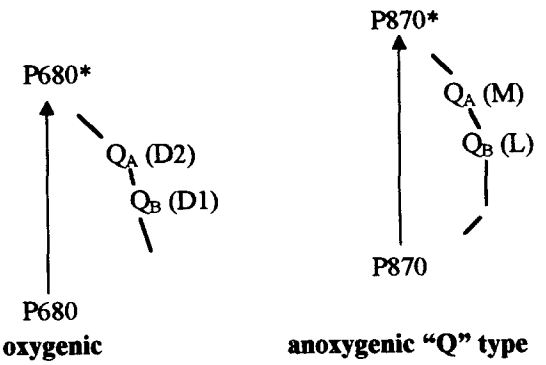

Fig. 1. Schematic diagram showing roles of reaction center proteins.

*Corresponding author. Fax: (49) (921) 552786;
E-mail: Pete.Lockhart@uni-bayreuth.de

PII S0014-5793(96)00383-3

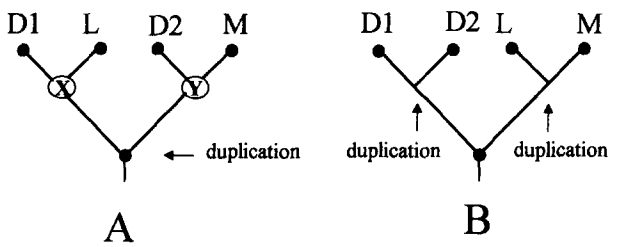

Fig. 2. (A) Expectation for a reconstructed evolutionary tree based on functional similarities between reaction center proteins. (B) Optimal reconstructed tree obtained in recent published studies of reaction center proteins. The number of gene duplications needed on each tree to explain the observed data is shown.

these programs are accepted. The sequence of amino acids is highly conserved between D1 sequences as it is between D2 sequences. Yet amino acid residues are not strongly conserved between D1 and D2 sequences.

3.1. Different evolutionary constraints or alignment error

We can obtain a quantitative measure of the extent of this difference by estimating, for this alignment, the number of sites free to vary between reaction center proteins. The results are shown in Table 2. The results are striking in that they indicate that, on this alignment, there is a large number of invariant sites unique to D1 sequences and a large number of different invariant sites unique to D2. It also appears to suggest that there is a greater proportion of invariant sites common to D1 and D2 than there is between these oxygenic proteins and their anoxygenic homologues. If the sequences are truly evolving by a mechanism of evolution in which the sites are changing identically and independently, then this pattern in the data is not expected. Rather, we would expect similar estimates for sites free to vary between and amongst the different groups. The observed values in Table 2 suggest either that the D1, D2, L, M groups are incorrectly aligned and/or that because of their different biological functions they

```
IGGP-YELIVLHFLLGVACYMGREWELSFRLGMRPWIAVAYSAPVAAATAVFLIYPIGQG Spinacia D1
.........................Y.......................... Marchantia
......Q...C..FI.ICS.........................S...IV..L... Euglena
.............FI.ICA.........Y...........F..........I....... Cyanidium
......Q.V.F...I.IF.....Q....Y........C.......S.............. Synechocystis
L..L.WAFVA..GAFALIGF.L.QF..ARSVQL..YN.I.F.G.I.VFVS......L..S Spinacia D2
L..L.WTFVA..GAF.LIGF.L.QF..ARSVQL..YN.I.F.G.I.VFVS......L..S Marchantia
L..L.WPF.A..GAF.LIGF.L.QF.IAKAVQI..YN.I.F...ISVFVS......L..S Euglena
L..L.WTFTA..GAF.LIGFCL.QF.IARLV.I..YN.I.F.G.I.VFVS......L..A Cyanidium
L..L.WSFVA..GAF.LIGF.L.QF.IARLV.I..YN.I.F.G.I.VFVS...M..L..S Synechococcus
EP.FAWQMT..FATIAFFGW.M.QVDI.MK.D.GYHVPI.FGVAFS.WLVLQV.R..AL. Chloroflexus L
E..F.WQA.TVCA.GAFISW.L..V.I.RK..IGWHVPL.FCV.IFMFCVLQVFR.LLL. Rhodospuedomonas
E..W.WLIATFFLTVSIFAWYMHIYTRAKA..IK.YL.YGFTGAI.LYLVIYI.R.VWM. Chloroflexus M
D..W.WLMAG.FMT.SLGSWWI.VYSRARA..LGTH..WNFA.AIFFVLCIGC.H.TLV. Rhodospeudomonas

SFSDGMPLGISGTFNFMIVFQA-EHNILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLIRETT
.....
.....L....
.....L.....M.....R.
.....V....
GWFFAPSF.VAAI.R.ILF..G.F..WTLN....M.....L.AA.LC.I..AT.ENT.F-.DG
GWFFAPSF.VAAI.R.ILF..G.F..WTLN....M.....L.AA.LC.I..AT.ENT.F-.DG
GWFFAPSF.VAAI.R.ILF..G.F..WTLN....M.....L.AA.LC.I..AT.ENTIF-.DG
.WFFAPSF.VAAI.R.ILFL.G.F..WTLN....M.....IL..A.LC.I..AT.ENT.F-.DG
.WFFAPSF.VA.I.R.ILFL.G.F..WTLN....M.....IL..A.LC.I..AT.ENT.F-.DS
MWHE.FV...MPHLDWVSN.GYRYN.FFYN...AI.IT.L.ASTWLL.C....IL.AAQYRGP
.WGHAF.Y..LSHLDWVNN.GYQYL.WHYN.G..SS.SFL.VNAMALGL..G.IL.VANPGDG
DW.EAPAH..KALLDWTNNVSVRYG.FYYN.....SIFFLL.ST.LL...AGTIWALEKYAAH
.W.E.V.F..WPHIDWLTA.SIRYG.FYYC.W.GFSIGFAY.CG.LF.A..ATILAVARFGGD
```

Fig. 3. Part of the alignment between reaction center proteins studied by Beanland [3] but including additional taxa.

Table 1
Sequences used in study

Locus	Species	Accession No.
D1 (psbA)		
SPICPPSBA	*Spinacea oleracea*	J01442
CHMPXX	*Marchantia polymorpha*	X04465, Y00686
CLEGCGA	*Euglena gracilis*	X70810
CCPLPSBA	*Cyanidium caldarum*	X52758
SYCNPSBA2	*Synechocystis* sp. PCC6803	X13547
D2 (psbD)		
SICPPHTII	*Spinacea oleracea*	M27308, M12028, M16873
CHMPXX	*Marchantia polymorpha*	X04465, Y00686
CLEGCGA	*Euglena gracilis*	X70810
CYNCPPLAS	*Cyanidium caldarum*	X62578
SSD2P (EMBL)	*Synechococcus* sp. PCC7002	M29659
L+M subunits		
CFXRCG12	*Chloroflexus aurantiacus*	X07847, Y00972, X14979
RVPRCLM	*Rhodopseudomonas viridis*	X03915

are evolving under different processes of constraint. In either case, these data cannot be simply interpreted to indicate phylogeny for reaction center proteins.

3.2. Evidence for frequent insertion/deletion events in reaction center proteins?

In an attempt to better understand these data we have investigated the question of possible alignment error in section 3.3. Here we point out that alignment ambiguity between reaction center proteins could cause a problem for phylogenetic inference if insertion and deletion events have been frequent [9]. One important assumption in the alignment procedure is that the reading frame of the genetic code has always been similarly maintained in the sequences. However, Fig. 4 which shows a portion of the alignment studied by Beanland [3] suggests the presence of a putative frameshift mutation between D1 and D2 proteins. It shows that the residues aspartic acid (D), glycine (G) and isoleucine (I) in D1 are aligned against serine (S), tryptophan (W) and tyrosine (Y) in D2. In terms of their biological properties [14] these aligned residues are functionally very different. Further, the observed triplets occur only once in D1 and D2 sequence. Hence, there is no obvious expectation that such positioning of amino acids between D1 and D2 would arise either as a result of functional convergence or chance occurrence. This apparent frameshift between D1 and D2 does not occur in the most conserved portion of the molecule, nevertheless if our interpretation of such an event is correct then it may suggest that these reaction center proteins have undergone a very large amount of divergence since their early duplication from an ancestral gene. If so, an expectation may be that insertion and deletion events have been very common in the evolution of these proteins.

3.3. Alignment order: its effect on phylogeny reconstruction

In the case of D1, D2, L and M sequences and as shown in Fig. 5, most support for similar alignments to that obtained by Michel and Deisenhofer [1] is found when, in the multiple

Table 2
Estimates of codons free to vary

Sequences	Number of 1st position changes	Number of second position changes	Number of 1st+2nd position changes	Estimates of codons free to vary (%)
D1	30/120	11/120	7	39.3 ± 7.8
D2	30/121	13/121	9	35.8 ± 5.5
D1D2	66/120	52/120	35	81.7 ± 5.4
D1L	92/121	78/121	64	92.7 ± 2.7
D1M	95/121	81/121	68	93.5 ± 2.4
D2L	92/120	75/120	62	92.7 ± 2.8
D2M	89/120	79/120	65	90.1 ± 2.4
ML	92/121	78/121	67	88.4 ± 2.1
D1D2LM	103/120	101/120	91	95.3 ± 1.1

alignment procedure, the oxygenic reaction center proteins (D1, D2) are first joined before adding the nonoxygenic proteins (L, M). Such alignments, when subsequently used for phylogeny reconstruction, always give strong support for trees which suggest D1 and D2 are most closely related. This is true irrespective of the pairwise gap penalty used in the alignment procedure. In contrast, if the alignment order chosen is ((L and D1),(M and D2)); and the weighting for gap penalties is low, the resulting profile alignment can give strong support for trees in which L and D1 (not L and M) sequences are most closely related (Fig. 5; squares). When the alignment order is ((D2 and L),(D1 and M)); and gap penalties are low, support is never found at the 0.05 significance level for a tree which links D1 and M (Fig. 5; triangles).

		D G I
Spinacia	D1	..T GGT AT. ...
Oryza	D1	..T GGT AT. ...
Marchantia	D1	..C GGT AT. ...
Chlamydomonas	D1	..T GGT AT. ...
Euglena	D1	..T GGT AT. ...
Cyanidium	D1	..T GGT AT. ...
Synechocystis	D1	..C GGT AT. ...
		S W Y
Spinacia	D2	... TGG TAT ...
Oryza	D2	... TGG TAT ...
Marchantia	D2	... TGG TAT ...
Chlamydomonas	D2	... TGG TAT ...
Euglena	D2	... TGG TAT ...
Cyanidium	D2	... TGG TAT ...
Synechoccocus	D2	... TGG TAC ...

Fig. 4. In the amino acid alignment of Beanland [3] there is an aspartic acid (D) in position 101 of D1 and this is aligned against a serine (S) in D2. In the universal code the amino acid aspartic acid (D) is encoded by GAT or GAC; amino acid glycine (G) is encoded by GGA or GGT or GGG or GGC; amino acid isoleucine (I) is encoded by ATA or ATT or ATC; amino acid tryptophan (W) is encoded by TGG; amino acid tyrosine (Y) is encoded by TAT or TAC.

3.4. *A functional criterion for evaluating the multiple alignment of reaction center proteins*

Although it is possible to reconstruct an alignment which will favor D1 joining with L the question that arises is whether this is biologically meaningful. Fig. 6 shows the alignment obtained when the Multalin4 gap penalty = 3. It contains numerous gaps and deletions, nevertheless it still maintains many of the essential conserved binding sites described by Michel and Deisenhofer [1] which are also conserved in the alignment of Beanland [3]. It also contains the conserved motifs between (i) M, D2 helix IV and photosystem I B helix X sequences as well as between (ii) L, D1 with photosystem I B helix X [13]. Hence, on the basis of the correct positioning of functionally important residues, the alignment shown in Fig. 6 may not be significantly worse than the optimal one found when the alignment order is ((D1, D2),(L, M));.

3.5. *Conclusions*

The constraints imposed on photosystem II proteins in developing oxygenic photosynthesis (e.g. binding of manganese) would necessarily have been very different to those on the proteins involved in anoxygenic photosynthesis [1]. If oxygenic and anoxygenic photosynthesis diverged after the initial duplication of an ancestral reaction center homodimeric protein (autogenous hypothesis [16]) then an important consideration for phylogenetic analysis will be the different constraints imposed on the evolution of reaction center

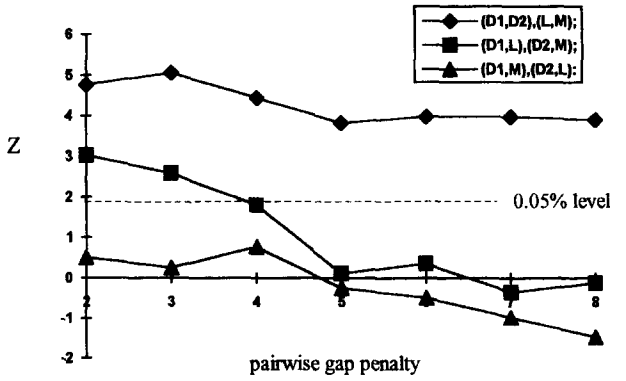

Fig. 5. Z scores calculated using PROTPARS [12]. In each case the plotted values are the (tree length of the hypothesis shown minus the tree length of the shortest other tree)/standard deviation. A value greater than 1.96 would indicate support at the 0.05 significance level.

```
MTAI---LERRESESLWGRFCN-WITSTE--NRLYIGWFGVLMIP-TLLT  D1
MALL--SFERKYR--VRGG--T-LIGGD-----LFDFWVGPYFV--GFFG  L
MTIAVGKFTK-DEKDLFDSM-DDWLRRDRFV--FVG-WSGLLLFPCAYFA  D2
MADYQTIYTQIQARGPHITVSGEWGDNDRVGKPFYSYWLGKI--GDA--Q  M

ATSVFIIAFIAAPPVDIDGIREPVSGSLLYGNNIISGAIIPTSAAIGLHF
VSAIFFI-FLGVSLIGYAASQGPTWDP--FA---IS--INPP----DLK-
LGGWFTGTTFVTSWYTH-GLASSYLEGCNF--LTAAVS-TPANSLAHSLL
IGPIYLGASGIAA-FAF-G-STAILI-ILF-NMAAEVHFDPLQFFRQFFW

YPIWEAAS-----VDEWLYNGGPYELIVLHFLLGVACYMGREWELSFRLG
YGL-GAAP-----L---L-EGGFWQAITVCALGAFISWMLREVEISRKLG
L-LWGPEAQ-G-DFTRWCQLGGLWAFVALHGAFALIGFMLRQFELARSVQ
LGLYPPKAQYGMGIPPLHD-GGWWLMAGLFMTLSLGSWWIRVYSRARALG

MRPW-IAVAYSAPV-AAATAVFLIY-PIGQGSFSDGMPLGISGTFNFMIV
I-GWHVPLAFCVPI-FM-FCVLQVFRPLLLGSWGHAFPYGILSHLDWVNN
LRP-YNAIAFSGPIAVFV-SVFLIY-PLGQSGWFFAPSFGV-AAIFRFIL
LGT-HIAWNFAAAI-FFVLCIGCIH-PTLVGSWSEGVPFGIWPHIDWLTA

FQAEH-NILMHPFHMLGVAGVFGGSLFSAMHGSLVTSSLIRETTENESAN
FGYQYLNWHYNPGHMSSVSFLFVNAMALGLHGGLILS--VANPGDGDKVK
FFQGFHNWTLNPFHMMGVAGVLGAALLCAIHGATV-ENTLFEDGDGANTF
FSIRYGNFYYCPWHGFSIGFAYGCGLLFAAHGATILAVARF-GGD-----

EGYRFGQEEETYNIVAAHGYFGRLI-FQYASFNNSRSLHFFLAAWPVVGI
TA---EHENQ---------YF-RDV-VGYSI--GALSIH-----R--LGL
RAFNPTQAEETYSMVTANRFWS-QI-FGVAFSN-KRWLHFFMLFVPVTGL
REIEQITDRGT-AVERAALFWRWTIGFNATIESVHRWGWFFSLMVMVSA-

WFTALGISTMAFNLNGFNFNQSVVDSQGRVINTWADIINRANLGMEVMHE
-F--LA-SNI-F-LTG-AF--GTIAS-GPF---WT----R---GW---PE
WMSALGVVGLALNLRAYDFVSQEIRAAED--PEFETFYTKNILLNEGIRA
SVGIL-LTGTFVDNW-YLWC---VKHGAA--PDYPAY-------------

RNAH----NFPLDLAAIEAPSTNG-
WWGW-----W-LDI-----PFWS--
WMAAQDQPHENLIFPEEVLPRGNAL
-LPATPDP-ASL--PGA--PK----
```

Fig. 6. The alignment of reaction center proteins obtained using Multalin4 (gap penalty = 3) [10] and ClustalV [11]. Functionally important positions conserved in all sequences have been underlined. D1 and D2 sequences are from *Spinacea oleracea*. L and M sequences are from *Rhodopseudomonas viridis*.

proteins. That homologous genes of different biological function [17], or even similar function but distantly related [18], can have different sites free to vary has been demonstrated elsewhere. Similarly, differences in the pattern of evolution have also been observed to occur between proteins in anoxygenic and oxygenic systems [17,19]. If such patterns of constraint reflect the historical relationships between reaction center proteins then these will not mislead evolutionary tree reconstruction. However, if the pattern does not reflect the true evolutionary relationships then the evolutionary tree reconstruction of Beanland [3] and Blankenship [4] could have been misled. Even in the absence of alignment error, this might occur simply because of the different requirements for proteins in anoxygenic and oxygenic photosynthesis. This could lead to D1 and D2 sharing fewer sites that vary than L and M sequences. In this case, if there is enough change in L and M sequences, patterns will arise to mislead methods of evolutionary tree selection [17,20].

Our estimates of sites free to vary for reaction center proteins in the published analysis of reaction center proteins make it clear that at present there can be little confidence in the conclusion that two independent gene duplications have occurred in the evolution of reaction center proteins. Unfortunately, the difficulty in correctly aligning these proteins and the additional complexity of their substitution patterns will continue to make phylogenetic inference of their evolutionary origins from primary sequence data very challenging.

Acknowledgements: We thank the New Zealand Public Good Science Fund and the German Alexander von Humboldt Foundation for the financial support of this work.

References

[1] Michel, H. and Deisenhofer, J. (1988) Biochemistry 27, 1-7.
[2] Williams, J.C., Steiner, L.A. and Feher, G. (1986) Proteins Struct. Funct. Genet. 1, 312-325.
[3] Beanland, T.J. (1990) J. Theor. Biol. 145, 535-545.
[4] Blankenship, R.E. (1992) Photosynth. Res. 33, 91-111.
[5] Sidow, A., Ngyen, T. and Speed, T.P. (1992) J..Mol. Evol. 35, 253-260.
[6] Penny, D., Watson, E.E., Hickson, R.E. and Lockhart, P.J. (1993) N.Z. J. Bot. 31, 275-288.
[7] Kumar, S., Tamura, K. and Nei, M. (1993) MEGA: Molecular Evolutionary Genetics Analysis, version 1.01, The Pennsylvania State University, University Park, PA.
[8] Lake, J. (1988) Nature 331, 184-186.
[9] Hein, J. (1994) in: Methods in Molecular Biology, vol. 25: Computer Analysis of Sequence Data, Part II (Griffin A.M. and Griffin H.G. eds.) Humana Press, Totowa, NJ.
[10] Corpet, F. (1988) Nucleic Acids Res. 16, 10881-10890.
[11] Higgins, D.G. Bleasby, A.J. and Fuchs, R. (1991) CABIOS 8, 189-191.
[12] Felsenstein, J. (1993) PHYLIP3.5 Manual, Univ. California Herbarium, Berkeley, CA.
[13] Margulies, M.M. (1991) Photosynth. Res. 29, 133-147.
[14] Thompson, J.D. Higgins, D.G. and Gibson, T.J (1994) Nucleic Acids Res. 22, 4673-4680.
[15] Larkum, A.W.D. (1992) in: Research in Photosynthesis (Murata, N. ed.) vol. III, pp. 475-480.
[16] Lockhart, PJ., Larkum, A.W.D., Steel, M.A., Waddell, P.J. and Penny, D. (1996) Proc. Natl. Acad. Sci. USA 93, in press.
[17] Miyamoto, M.M. and Fitch, W.M. (1995) Mol. Biol. Evol. 12, 503-513.
[18] Buttner, M., Xie, D.-L., Nelson, H., Pinther, W., Hauska, G. and Nelson, N. (1992) Proc. Natl. Acad. Sci. USA. 89, 8135-8139.
[19] Felsenstein, J. (1978) Syst. Zool. 27, 401-410.