Plant Syst. Evol. 242: 119–135 (2003)
DOI 10.1007/s00606-003-0056-4

Plant Systematics and Evolution
Printed in Austria

The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis** – structural and phylogenetic analyses

V. Goremykin[1], **K. I. Hirsch-Ernst**[2], **S. Wölfl**[3], and **F. H. Hellwig**[1]

[1] Institut für Spezielle Botanik, Universität Jena, Jena, Germany
[2] Zentrum Pharmakologie und Toxikologie, Universität Göttingen, Göttingen, Germany
[3] Klinik für Innere Medizin, Universität Jena, Jena, Germany

Received February 19, 2003; accepted June 3, 2003
Published online: November 17, 2003

Abstract. The nucleotide sequence of the complete chloroplast genome of a basal angiosperm, *Calycanthus fertilis*, has been determined. The circular 153337 bp long cpDNA is colinear with those of tobacco, *Arabidopsis* and spinach. A total of 133 predicted genes (115 individual gene species, 18 genes duplicated in the inverted repeats) including 88 potential protein-coding genes (81 gene species), 8 ribosomal RNA genes (4 gene species) and 37 tRNA genes (30 gene species) representing 20 amino acids were identified based on similarity to their homologs from other chloroplast genomes. This is the highest gene number ever registered in an angiosperm plastome. *Calycanthus fertilis* cpDNA also contains a homolog of the recently discovered mitochondrial ACRS gene. Since no gene transfer from mitochondria to the chloroplast has ever been documented, we investigated the evolutionary affinity of this gene in detail. Phylogenetic analysis of the protein-coding subset of the plastome suggests that the ancient line of Laurales emerged after the split of the angiosperms into monocots and dicots.

Key words: *Calycanthus fertilis*, angiosperms, chloroplast genome, evolution, phylogeny, substitution rates.

Introduction

The chloroplast DNA has been widely and successfully employed by many plant biologists for the reconstruction of plant evolution (Kellogg 1998, Sytsma and Hahn 1997). One particularly intractable and long standing problem in this field is the origin and early diversification of the angiosperms (Beck 1976, Friis et al. 1987, Crane et al. 1995, Kenrick 1999). In the past many authors have tried to resolve angiosperm phylogeny at higher taxonomic levels with only one or a few molecular markers (Parkinson et al. 1999, Qiu et al. 1999, Soltis et al. 1999). As an outcome of studies of this type the root of the angiosperms was declared to be finally identified. However, these studies have not led to a consistent picture of the early diversification of angiosperms. As Graham and Olmstead (2000) noted in their analysis of 17 chloroplast genes "it is premature to place confidence in the *Amborella* rooting of the angiosperms, in this

**Calycanthus fertilis* Walter var. ferax (Michy.) Rehder is a synonym of *C. floridus* L. var. *glaucus* (Willd.) Torr. & A. Gray.
Data deposition: The sequence reported in this paper has been deposited in the EMBL database (accession no. AJ428413).

or other published studies with fewer characters available".

Beside the rooting problem the sequence of early evolutionary splittings within the angiosperms and their dating are uncertain. Taking the monocots (Liliopsida) as an example, the position of this group remains unresolved and receives little support in recent publications. They appear either as a sister group to Laurales (Parkinson et al. 1999) or to *Ceratophyllum*, while together with *Ceratophyllum* forming a sister group to all other angiosperms except the ANITA grade (Qiu et al. 1999). In Soltis et al. (Soltis et al. 1999) they are in an unresolved clade with other Magnoliidae. Graham and Olmstead (2000) found them again as sister to *Ceratophyllum* and both as sister to eudicots (Rosopsida). For some years, *Ceratophyllum* appeared as a sister to all other living angiosperms (Les et al. 1991, Chase et al. 1993, Qiu et al. 1993), but later its systematic position became uncertain.

In theory, reliability of phylogeny estimations improves with the amount of molecular data analysed (Lecointre et al. 1994). The need for larger datasets has been recognised in recent publications and the trend to use multiple markers in a dataset has emerged (Soltis et al. 1999, Graham and Olmstead 2000). The power of datasets comprising many chloroplast regions to resolve phylogenetic relationships was demonstrated by Martin et al. (1998) in their study of plastid evolution. The accumulation of chloroplast genomic data should lead to a better estimate of the age of angiosperms and their phylogenetic relationships.

In this paper we present the chloroplast genome of *Calycanthus fertilis* Walt., a shrub with deciduous leaves and rather large dark purple flowers growing in North America from Pennsylvania and Ohio in the North to Georgia and Alabama in the South. Certain parts of the plant yield medicinal extracts (Chant 1978). This plant belongs to the family Calycanthaceae, which emerged from the most ancient split in the order Laurales, one of the oldest known groups of flowering plants (Renner 1999). The fossilized calycanthaceous flower *Virginianthus calycanthoides* was described in the Potomac Group from the Upper Albian (Friis et al. 1994), whereas the pollen of Laurales can be traced back to the Aptian (Muller 1984). It is feasible to conclude that the origin of the Laurales dates as far back as the beginning of the Cretaceous. In the APG (1998) classification of the angiosperms, Laurales belong to Magnoliopsida, the basal paraphyletic assemblage, which comprises woody "palaeotrees" and herbaceous "palaeoherbs", many of which were proposed to be "the most ancient angiosperms" in the course of the last century. Although basal angiosperms have been subject to intensive phylogenetic studies, no complete chloroplast genome sequence of a member of this group has been published as yet. This is the gap we intend to close with this publication.

Materials and methods

DNA extraction. Fresh leaves of *Calycanthus fertilis* were harvested from a plant growing at the Botanical Garden of the University of Jena, Germany. A voucher specimen has been disposited at the Herbarium Haussknecht (JE). Total DNA was extracted using the CTAB-based method (Murray and Thompson 1980) and purified with Qiagen columns (Qiagen) according to the manufacturer's protocol.

Amplification strategy. The ARRANGE script was written in Perl. It analyzes local pairwise alignments in SIM format (Huang and Miller 1991) and rearranges all found aligned regions from one sequence according to the sequence pattern of the other sequence. The chloroplast genomic sequences of *Pinus* (Wakasugi et al. 1994), *Oenothera* (Hupfer et al. 2000), *Zea* (Maier 1995) and *Rice* (Hiratsuka et al. 1989) were rearranged with the help of the ARRANGE script so that the resulting sequences were colinear with the plastome sequences of *Nicotiana* (Shinozaki et al. 1986), *Arabidopsis* (Sato et al. 1999) and *Spinacia* (Schmitz-Linneweber et al. 2001). The gene order of the last three genomes was found to be predominant in land plants (Downie and Palmer 1992). The resulting sequences and three colinear chloroplast genomes were then aligned with the CLUSTALW program

(Thompson et al. 1994). The regions of poor alignment were then realigned with REALIGN (CLUSTALW-embedded script) to produce a final ~200.000 bases long genomic alignment of good quality.

Since conserved regions between all seven species in the resulting alignment are rare, we employed a long-range PCR strategy to completely cover a chloroplast genome with PCR products. Such PCR imposes rather strict requirements on PCR-primers so that a manual primer choice is not a good option. As we found no suitable computer program for primer selection in an alignment, a new program was written in Perl/TK. It performs the necessary computations – Tm, GC-content, free energy of possible internal and external primer-dimerisation and of secondary structures according to the nearest-neighbor model (Borer et al. 1974, Rychlik and Rhoads 1990), scans potential primer-binding sites for repetitive stretches and homooligomers and checks if the terminal primer sequences are present in the alignment between the binding sites to avoid the generation of shortcut products. With the help of the above program we completely covered the chloroplast genome of *Calycanthus fertilis* with PCR products ranging in size from ~4 to ~20 kb. The inverted repeat regions in cpDNA were amplified separately, each with two PCR products overlapping in the middle of the repeat, and stretching out to flanking sequences of the respective single-copy regions.

Cloning and sequencing. Long-range PCR products were purified by electrophoresis through gels prepared with low-melting agarose. Following agarose digestion with agarase, DNA in the resulting solution was directly subjected to fragmentation and subcloning employing the TOPO Shotgun Subcloning kit (Invitrogen), according to the manufacturer's protocol. Recombinant plasmids were isolated from clones using the Montage Plasmid Miniprep kit (Millipore). The resulting plasmid DNA was prepared for sequence analysis with the Big Dye Terminator sequencing kit (Applied Biosystems) according to the manufacturer's protocol. Automated sequencing was performed on ABI 3100, ABI 377 (Applied Biosystems) and MegaBACE 1000 (Amersham/Pharmacia Biotech) sequencers.

Sequence assembly. ABI-reads and MegaBACE-reads were base-called with the PHRED program (Ewing et al. 1998, Ewing and Green 1998). Sequence masking and assembly was performed with the STADEN Package (Staden et al. 2000, 2001) on a Linux Pentium III PC. The sequencing data were accumulated until 8x coverage for all PCR fragments was achieved. The remaining gaps were closed by PCR.

Results and discussion

General genome properties. The chloroplast genome of the basal angiosperm *Calycanthus fertilis* is a circular 153337 bases long DNA molecule (see Fig. 1) with a 86948 bp large single-copy region (LSCR) and a 19799 bp small single-copy region (SSCR) separated by 23295 bp long inverted repeats. It is colinear to the chloroplast genomes of *Nicotiana*, *Arabidopsis* and *Spinacia* in respect to both gene order and overall homology (Fig. 2).

One difference in the *Calycanthus fertilis* plastome structure in comparison to the other known angiosperm cpDNA molecules is a shorter inverted repeat (IR) region. We found this to be due to two indels at the border of the IR – one 1.5 kb long indel in the inverted repeat A (IRA) at the border to LSCR which includes a complete copy of the *rpl*2 gene, and another 0.7 kb long indel in the IRB at the border to SSCR (relative to the cpDNA of tobacco). The last indel is situated completely within an ORF which varies in length in different species (ORF 350 in tobacco cpDNA, ORF 482 in *Spinacia*), which is identical to the start region of the large *ycf*1 ORF. Since this 3'-truncated "mirror" *ycf*1 ORF in the *Calycanthus fertilis* cpDNA is 70% shortened, one can conclude that it codes for no product. The shifting of the borders of the inverted repeats is a well-known phenomenon (Sugita et al. 1984, Moon and Wu 1988, Maier et al. 1990, 1995, Hupfer et al. 2000). Since the plastome we present in the paper is the first completely sequenced among Magnoliopsida, it is uncertain if the shorter repeat region represents an ancestral state of cpDNA in angiosperms. It seems, however, noteworthy, that the only completely sequenced cpDNA of a gymno-

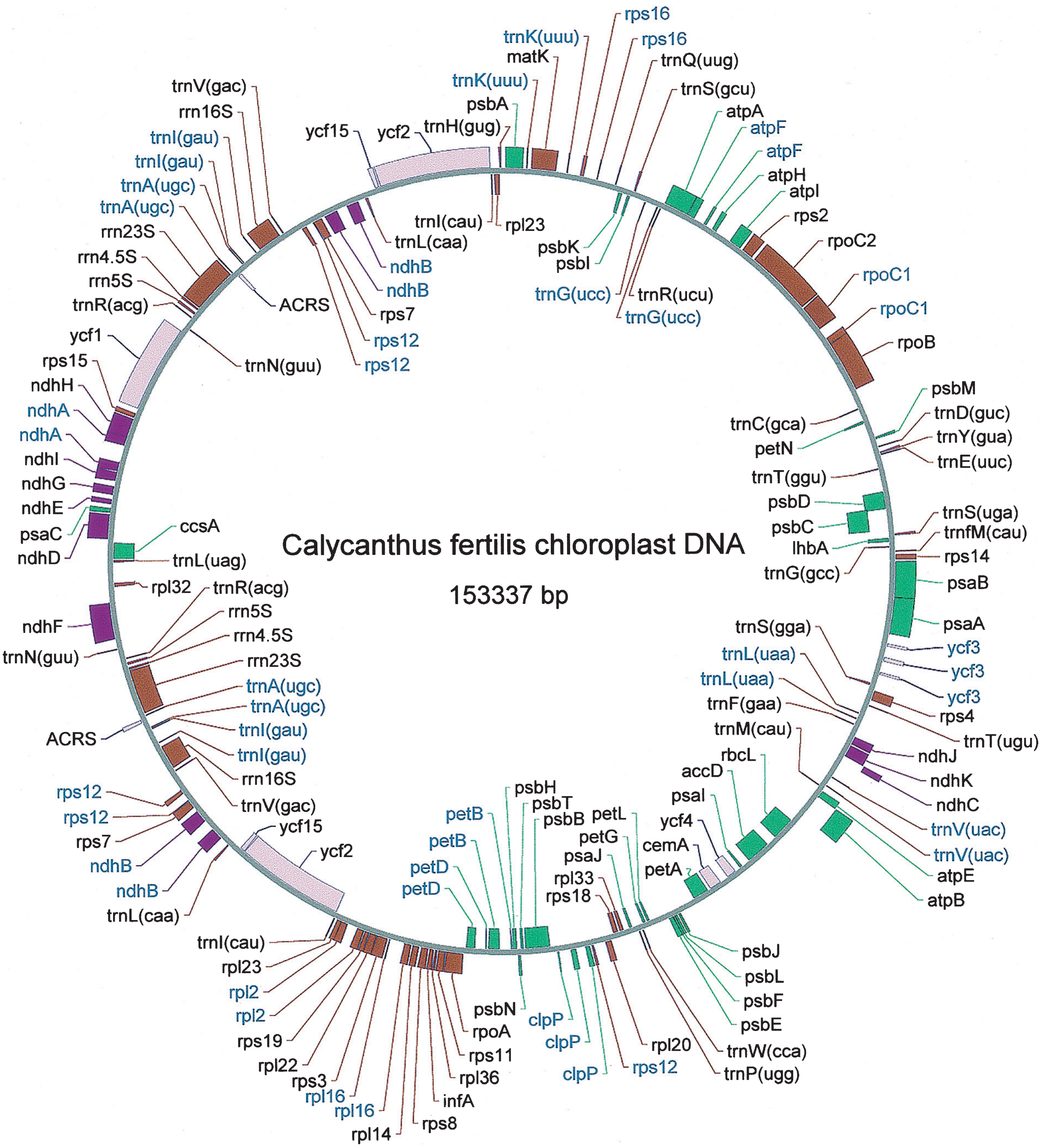

Fig. 1. Gene map of the chloroplast genome of *Calycanthus fertilis*. The topmost part of the map corresponds to the start and the end of the EMBL sequence entry. The genes shown inside the circle are transcribed clockwise, those outside the circle are transcribed counter-clockwise. The genes of the genetic apparatus are shown in red, photosynthesis genes are shown in green, and genes of NADH dehydrogenase are shown in violet. Pink are the ORFs, *ycf*s and genes of unknown function. Intron-containing genes are represented by their exons. Their names are given in blue. In the cases in which two genes overlap, one of them is shifted off the map to show its position

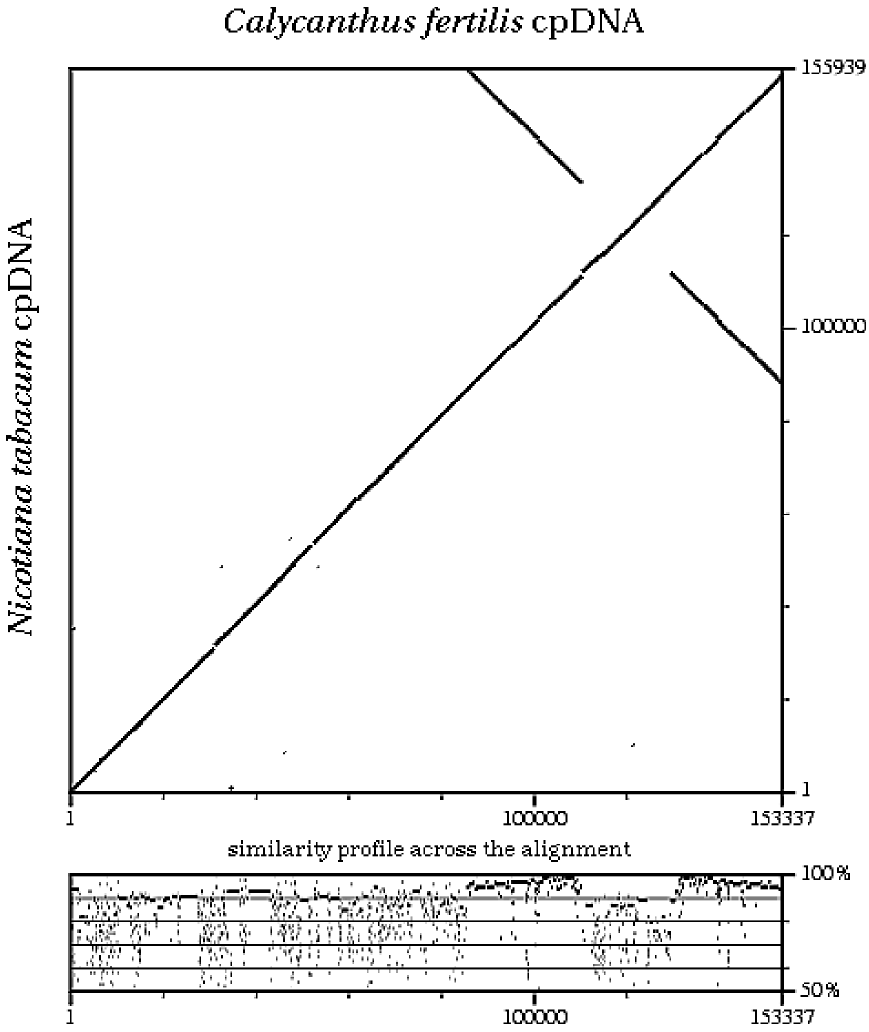

Fig. 2. Dot-plot of SIM alignments of *Nicotiana tabacum* and *Calycanthus fertilis* chloroplast genomes. Dots in the second box below represent the identity scores (uncorrected p-distances) of the corresponding alignment regions above

sperm, *Pinus thunbergii,* has yet more extensively reduced inverted-repeat regions.

GC content of the genome. The overall GC content of the chloroplast genome of *Calycanthus fertilis* is 39.3%, similar to that of the other spermatophytes studied. Different functional regions in the genome however have their distinct base composition biases, with the GC content ranging from 55,3% to 29.3%. The first number is characteristic of the rRNA genes, which together comprise 5.9% of the whole plastome (9038 bases). The tRNA genes which span 1.8% of the plastome (2796 bases in total) show a similar GC content (53.3%). A subset of protein-coding genes in the chloroplast genome of *Calycanthus fertilis* is 78665 bases long (*ycf*s included), which corresponds to the overall plastome coding capacity of 51.3%, and has an intermediate GC content of 39.1%. There is a significant difference in GC content between introns and intergenic spacer regions. On the whole, introns (19432 bases in total, 12.7% of the genome length) show a GC content (39.4%) close to the overall one. Yet spacers (45266 bases in total, 29.5% of the genome length) are AT-richer with an average GC content of 35.4%. This discrepancy is obviously due to the lower necessity for spacers to form stable secondary structures. The different overall compositional biases of inverted repeats (43.6% G + C), LSCR (38.2% G + C) and SSCR (34.0% G + C) are largely due to the different GC contents of uncoding sequences there, but in each ot these regions spacers still have a lower GC content as compared to the introns (respectively, 41.3%/45.5% (IR), 34.1%/36.9% (LSCR), 29.3%/34.3% (SSCR)). An obvious conclusion from the above observations is that combining the RNA genes, introns and spacers with low GC content together into one "non-coding" dataset for phylogenetic analysis could be misleading. Though there are methods compensating for compositional biases *between* different sequences, we are not aware of any substitutional model, which would compensate for a different compositional bias within each sequence.

Gene content. A total number of genes encoded by the *Calycanthus fertilis* chloroplast genome is at least 133. It was estimated with our BLAST-embedded script by blasting the complete genome sequence against our local nucleotide BLAST (Altschul et al. 1997) database of all annotated genes from all known plastid genomes followed by a number of postprocessing steps. The script produced an annotated sequence in GENBANK format. Additionally, we have verified the tRNA hits with the tRNAscan-SE program (Lowe et al. 1997). Then, to exhaust the possibilities of similarity search, we have blasted all translated hypothetical ORFs in *Calycanthus fertilis* cpDNA larger than 60 bases (3758 ORFs together, bacterial genetic code) against the non-redundant Swissprot and TREMBL databases. With one exception, all found matches differing from those already annotated were hypothetical ORFs (non-ycf ORFs) themselves.

The exception is a 171 bases long ORF in an intron within the tRNA-Ala(ugc) gene, which should encode a 56 amino acids long protein. On the amino acid level this hypothetical protein shares 90% similarity to the 56 amino acids long mitochondrial ACRS protein from *Citrus jambhiri* recently reported by Ohtani et al. (2002), which is also located within the tRNA-Ala intron. It is interesting to notice in this respect that the group 2 intron-containing tRNA-Ala(ugc) gene is a distinctive feature of the cpDNAs of the land plants. Manhart and Palmer (1990) used the gain of this intron and that of the adjacent tRNA-Ile(gau) gene in cpDNA to pinpoint the green algae *Coleochaete, Nitella* and *Spirogyra* as belonging to the line ancestral to the land plants. Apart from these algal copies, the nonredundant GENBANK database contains more than 40 full and partial copies of the tRNA-Ala(ugc) introns from chloroplast DNA of angiosperms, gymnosperms, ferns and bryophytes. By contrast, the mitochondrial copies of the tRNA-Ala(ugc) intron (or its parts) in the non-redundant GENBANK database are from three angiosperm genera – *Citrus*, *Helianthus* and *Phaseolus*. Our TBLASTX search against that database revealed that *Spinacia oleracea* and *Oenothera lamarckiana* plastomes have intact ACRS ORFs. The only mitochondrial sequence with an intact ACRS ORF present in GENBANK is that of *Citrus jambhiri,* reported by Ohtani et al. (2002).

The phylogenetic distances between the mitochondrial sequence from rough lemon and all its chloroplast homologs are small – under 0.1 subst./site (both with Jukes-Kantor (1969) and Kimura (1980) distances). The ACRS ORF from the *Calycanthus fertilis* chloroplast genome actually shows more affinity to the *Citrus jambhiri* mitochondrial ACRS coding sequence (dJK 0.075) than to that from chloroplasts of more closely related *Spinacia oleracea* (dJK 0.082). Taking into account the above considerations, one could suggest that the ACRS gene was relatively recently transferred from the chloroplasts to the mitochondria.

The list of all genes found in the genome under study is provided in Table 1. We found the cpDNA of *Calycanthus fertilis* to contain more known genes than any other completely sequenced angiosperm genome. Every gene, RNA- or protein-coding, previously annotated in the chloroplast genomes of angiosperms is on the *Calycanthus fertilis* cpDNA, with the exception of *spr*A, a hypothetical

Table 1. List of genes found in the *Calycanthus fertilis* plastome

Protein-coding genes							
Photosynthesis genes							
Rubisco	*rbc*L						
Photosystem I	*psa*A	*psa*B	*psa*C	*psa*I	*psa*J		
Photosystem II	*psb*A	*psb*B	*psb*C	*psb*D	*psb*E	*psb*F	p*sb*H
	*psb*I	*psb*J	*psb*K	*psb*L	*psb*M	*psb*N	*psb*T
Light-harvesting protein	*lhb*A						
Cytochrome b/f	*pet*A	*pet*B	*pet*D	*pet*G	*pet*L	*pet*N	
ATP synthase	*atp*A	*atp*B	*atp*E	*atp*F	*atp*H	*atp*I	
Membrane protein	*cem*A						
Cytochrome c biogenesis	*ccs*A						
NADH dehydrogenase	*ndh*A	*ndh*B	*ndh*C	*ndh*D	*ndh*E	*ndh*F	
	*ndh*G	*ndh*H	*ndh*I	*ndh*J	*ndh*K		
Acetyl-CoA carboxylase	*acc*D						
ATP-dependent protease	*clp*P						
Genetic system genes							
Small subunit	*rps*2	*rps*3	*rps*4	*rps*7	*rps*8	*rps*11	*rps*12
	*rps*14	*rps*15	*rps*16	*rps*18	*rps*19		
Large subunit	*rpl*2	*rpl*14	*rpl*16	*rpl*20	*rpl*22	*rpl*23	*rpl*32
	*rpl*33	*rpl*36					
RNA polymerase	*rpo*A	*rpo*B	*rpo*C1	*rpo*C2			
Translation initiation factor IF-1	*inf*A						
Intron maturase	*mat*K						
Conserved ORFs	*ycf*1	*ycf*2	*ycf*3	*ycf*4	*ycf*15	ACRS	
RNA genes							
Ribosomal RNAs	*rrn*23S		*rrn*16S		*rrn*5S		
	*rrn*4.5S						
Transfer RNAs	*trn*A(ugc)		*trn*C(gca)		*trn*D(guc)		
	*trn*E(uuc)		*trn*F(gaa)		*trn*G(gcc)		
	*trn*G(ucc)		*trn*H(gug)		*trn*I(cau)		
	*trn*I(gau)		*trn*K(uuu)		*trn*L(caa)		
	*trn*L(uaa)		*trn*L(uag)		*trnf*M(cau)		
	*trn*M(cau)		*trn*N(guu)		*trn*P(ugg)		
	*trn*Q(uug)		*trn*R(acg)		*trn*R(ucu)		
	*trn*S(gcu)		*trn*S(gga)		*trn*S(uga)		
	*trn*T(ggu)		*trn*T(ugu)		*trn*V(gac)		
	*trn*V(uac)		*trn*W(cca)		*trn*Y(gua)		
Intron-containing genes							
Genes with 2 introns	*rps*12, *clp*P, *ycf*3						
Genes with 1 intron	*atp*F, *pet*B, *pet*D, *ndh*A, *ndh*B, *rpl*16, *rpl*2, *rpo*C1, *rps*16, *trn*A(ugc), *trn*G(ucc), *trn*I(gau), *trn*K(uuu), *trn*L(uaa), *trn*V(uac)						

small RNA gene suggested to facilitate maturation of pre-16S rRNA (Vera and Sugiura 1994). Later it was found that this hypothetical gene plays no role in the process (Sugita et al. 1997), so it could be that this region is of no functional importance. 30 tRNA gene species present in the *Calycanthus fertilis* plastome are enough to code for all amino acids. These are the same tRNA species found in the plastomes of other angiosperms. The

Table 2. Genes found in the *Calycanthus fertilis* cpDNA, but absent in other completely sequenced chloroplast genomes of angiosperms

	*acc***D**	*inf***A**	*ycf***1**	*ycf***2**	*rpl***22**	*rpl***23**
Calycanthus	+	+	+	+	+	+
Triticum	–	+	–	–	+	+
Zea	–	+	–	–	+	+
Oryza	–	+	–	–	+	+
Nicotiana	+	–	+	+	+	+
Arabidopsis	+	–	+	+	+	+
Oenothera	+	–	+	+	+	+
Lotus	+	–	+	+	–	+
Spinacia	+	+	+	+	+	–

plastome of *Calycanthus fertilis* has 15 one-intron genes and 3 two-intron genes (Table 1). We did not find the genome under study to possess *chl*B, *chl*L, *chl*N, *psa*M and *ycf*12 genes, which are coded by cpDNA of the gymnosperm *Pinus thunbergii*. Taking into account the age of Laurales, one can conclude that these genes were probably lost during transition from gymnosperms to the angiosperms. The genes found in the *Calycanthus fertilis* cpDNA, but absent in other completely sequenced plastomes of angiosperms are summarized in Table 2.

The plastome of *Calycanthus fertilis* has 5 conserved open reading frames with no known function (*ycf*s). One way to determine if an open reading frame encodes a protein is to examine the ratio of the rate of synonymous substitutions (dS) to the rate of the nonsynonymous substitutions (dN) in the frame. All known protein-coding chloroplast genes of angiosperms belong to the core cell machinery and have synonymous substitution rates significantly higher than the nonsynonymous ones – from 83 times (*atp*H) to 3.8 times (*cem*A) higher. The maximum likelihood estimation (YN00 program, Yang 1997) of the dS/dN ratios in the codon-based alignments of the *ycf*s from the plastome under study with their angiosperm homologs revealed that the only frame under strong pressure to maintain the amino acid composition of the encoded protein is *ycf*3 (dS/dN = 25,4). We found the dS/dN values for other frames to be much lower: 3.1 for *ycf*4, 2.5 for *ycf*1, 1.7 for *ycf*15 and 1.3 for the *ycf*2.

Recently, Schmitz-Linneweber et al. (2001) found that the *ycf*15 primary transcript is not spliced in spinach. This finding led them to the conclusion that *ycf*15 does not encode any protein. The 7000 bp long *ycf*2, the largest conserved ORF in the cpDNA, shows an even lower dS/dN ratio than *ycf*15. The sliding frame analysis performed with the YN00-embedded script revealed that large regions within *ycf*2 have a dN rate even higher than the dS rate and are therefore under no evolutionary pressure to maintain the amino acid composition of the corresponding hypothetical protein. These data seem to suggest that *ycf*2 is a non-coding DNA region. The fact that it has no eubacterial ortholog (Martin et al. 1998) seems to corroborate this. Its importance for the chloroplast, which was demonstrated by Drescher et al. (2000) in their knock-out experiments, could have other reasons.

Two genes in cpDNA of *Calycanthus fertilis* – *ndh*D and *psb*L were found to have ACG codons at translation initiation sites, which is unknown to function among eucaryotes and bacteria. Based on the observation that the ACG codon is a valid initiation codon in the Sendai virus, Hiratsuka et al. (1989) suggested that ACG could still be a valid initiation codon in chloroplasts. Later, however, it was discovered that the ACG initiation codon is transformed to the AUG codon by RNA editing in these genes (Maier et al. 1995,

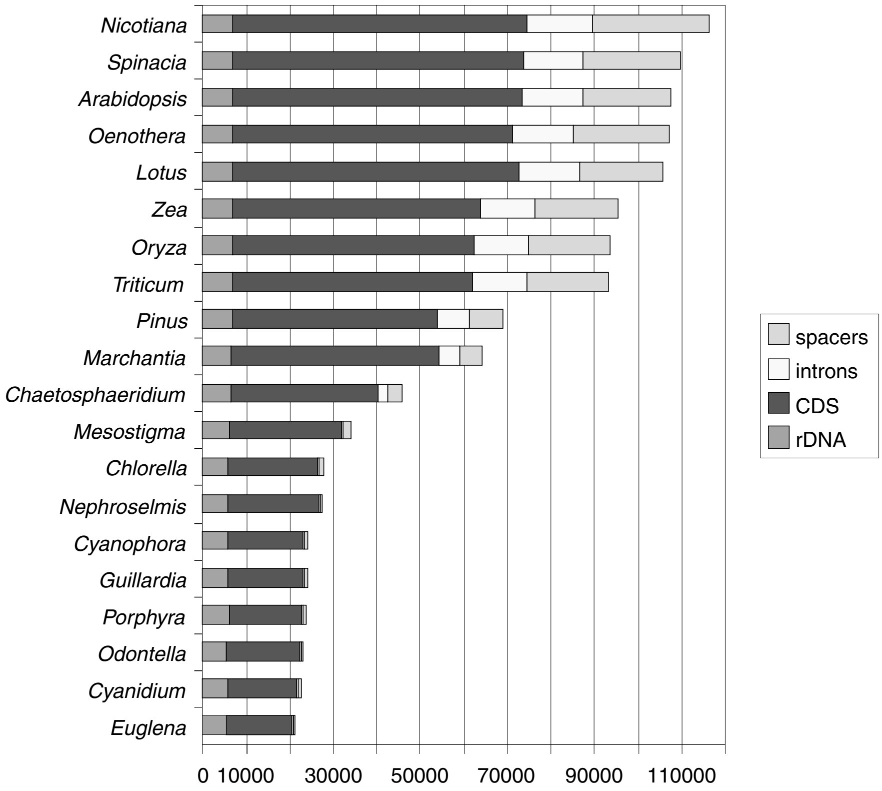

Fig. 3. Concatenated lengths of non-redundant local pairwise alignments of cpDNA of *Calycanthus fertilis* to known cpDNAs of other species. Only alignments with dJK no more than 0.4 subst./site are shown. Different functional regions in the alignments are colour-coded as shown in the legend. The subset of introns partially overlaps with the subset of the protein-coding genes, since *Mat*K and ORF1 are within the intron sequences. The sequences used in this analysis are published by: Ohyama et al. (1986) – *Marchantia*; Schinozaki et al. (1986) – *Nicotiana*; Hiratsuka et al. (1989) – *Oryza*; Hallick et al. (1993) – *Euglena*; Wakasugi et al. (1994) – *Pinus*; Maier et al. (1995) – *Zea*; Stirewalt et al. (1995), direct submission to GENBANK – *Cyanophora*; Reith and Munholland (1995) – *Porphyra*; Kowallik et al. (1995) – *Odontella*; Wakasugi et al. (1997) – *Chlorella*; Sato et al. (1999) – *Arabidopsis*; Turmel et al. (1999) – *Nephroselmis*; Douglas and Penny (1999) – *Guillardia*; Hupfer et al. (2000) – *Oenothera*; Ogihara (2002) – *Triticum*; Kato et al. (2000) – *Lotus*; Lemieux et al. (2000) – *Mesostigma*; Glockner et al. (2000) – *Cyanidium*; Schmitz-Linneweber et al. (2001) – *Spinacia*; Turmel et al. (2002) – *Chaetosphaeridium*

Hirose et al. 1999). Taking this into account, the presence of ACG codons at the translation initiation sites can be taken as evidence that RNA editing takes place in the expression of the *Calycanthus fertilis* plastome.

Comparative and phylogenetic analyses. Since a number of research groups is currently involved in sequencing entire chloroplast genomes for phylogenetic purposes, it is of practical interest to estimate the amount of genetic material in the cpDNAs which can be used for phylogeny reconstruction at different taxonomical levels. This will help to decide whether it is feasible to sequence entire plastid genomes for the given task.

To address this, we performed pairwise local alignments of the *Calycanthus* cpDNA without IRA with published cpDNA sequences from major groups of photosynthetic plants using the SIM program (Huang and Miller 1991). A graphical representation of a SIM alignment is given in Fig. 2. All redundant local alignments were removed from the output files with the help of our script. In the case of complete overlaps, the longer continuous alignment was retained and shorter alignments were discarded. In the case of partial overlaps, the overlapping region was assigned to the alignment with the higher identity score. The resulting files were analyzed with our SIMEX program. It counts bases in the local alignments of different identity classes (e.g. 100%, 99%) and sums up the resulting numbers in the user-defined identity intervals. The identity percentages were then transformed into Jukes-Cantor distances to provide a rough measure of phylogenetic divergence between different similarity ranges. Comparison between Magnoliopsida and Rosopsida revealed that more than 100 kb (114.5–103.8 kb) of the non-redundant plastome sequence is within dJK 0–0.4 subst./site range (Fig. 3). (The range was chosen because SIM alignments within it mostly consist of homologous sequences.) 93.6–91.5 kb of the cpDNA sequence of the grasses (*Zea*, *Oryza*, *Triticum*) is also within the above range in comparison with *Calycanthus* cpDNA, though the lengths of the most conservative parts in these alignments decrease. This trend continues in the alignments with *Pinus* and *Marchantia* plastomes. In cpDNA of *Marchantia*, only 64 kb (~50%) is within the range. Comparisons with chlorophyte algae demonstrate that only 46–27 kb of their cpDNA can be safely used for phylogeny reconstruction involving flowering plants. In the alignments with Rhodophytes, Euglenophytes, Cryptophytes and Glaucocystophytes this number drops to 23.5–21 kb.

Another function of our SIMEX program allowed us to estimate the proportions of different functional elements in the pairwise genomic alignments. Most conspicuously, we found just a few good local alignments between introns and spacers from cpDNA of *Calycanthus* (which together are 64146 bases long) and the algal plastomes. The total length of the intron alignments between the genome under study and algal cpDNAs ranges from 1980 bases (*Chaetosphaeridium*) to 34 bases (*Nephroselmis*). Spacer regions are somewhat better represented in the alignment with the charophyte alga *Chaetosphaeridium* (3482 bases) and streptophyte alga *Mesostigma* (1882 bases). In alignments with other algae they are 921–374 bases long. The subset of RNA genes in the alignments shows minimal length variations – from 6789 (*Nicotiana*) to 5299 bases (*Euglena*); these are the most conserved elements of the plastid genomes.

The largest subset of the local genomic alignments consists of the sequences of protein-coding genes. The subset of the protein-coding sequences in the alignments of land plants with dJK > 0.4 is from 68187 (*Calycanthus-Nicotiana*) to 51121 (*Calycanthus-Marchantia*) bases long. In the comparisons with algae the total length of the protein-coding subset within 0–0.4 dJK range shrinks to 33947–16447 bases.

Jukes-Cantor distance estimate (or any other total distance measure) can serve only as a rough approximation of the length of the *coding* sequences which can be safely used for phylogenetic analysis. Maximum likelihood estimation of the synonymous (dS) and non-synonymous (dN) distances with the yn00 program shows that the substitutional saturation at the synonymous positions is a factor to be considered when analysing chloroplast protein-coding genes. The lowest mean dS value (of pairwise dS estimations between all OTUs) we registered in alignments of 46 proteins common to 22 completely sequenced plastomes of photosynthetic plants was 1.5 [subst./site] (*psb*L gene). Mostly the dS values are much higher. Practically, it means that any phylogenetic analysis of the unmasked coding sequences from cpDNA from across the whole breadth of the plant kingdom can be misleading. The results from the series of the dN/dS estimations for 61 protein-coding genes common to 12 plastomes of the land plants (presented on Fig. 4) also warrant caution for the use of the unmasked coding sequences of chloroplast genes in comparisons involving

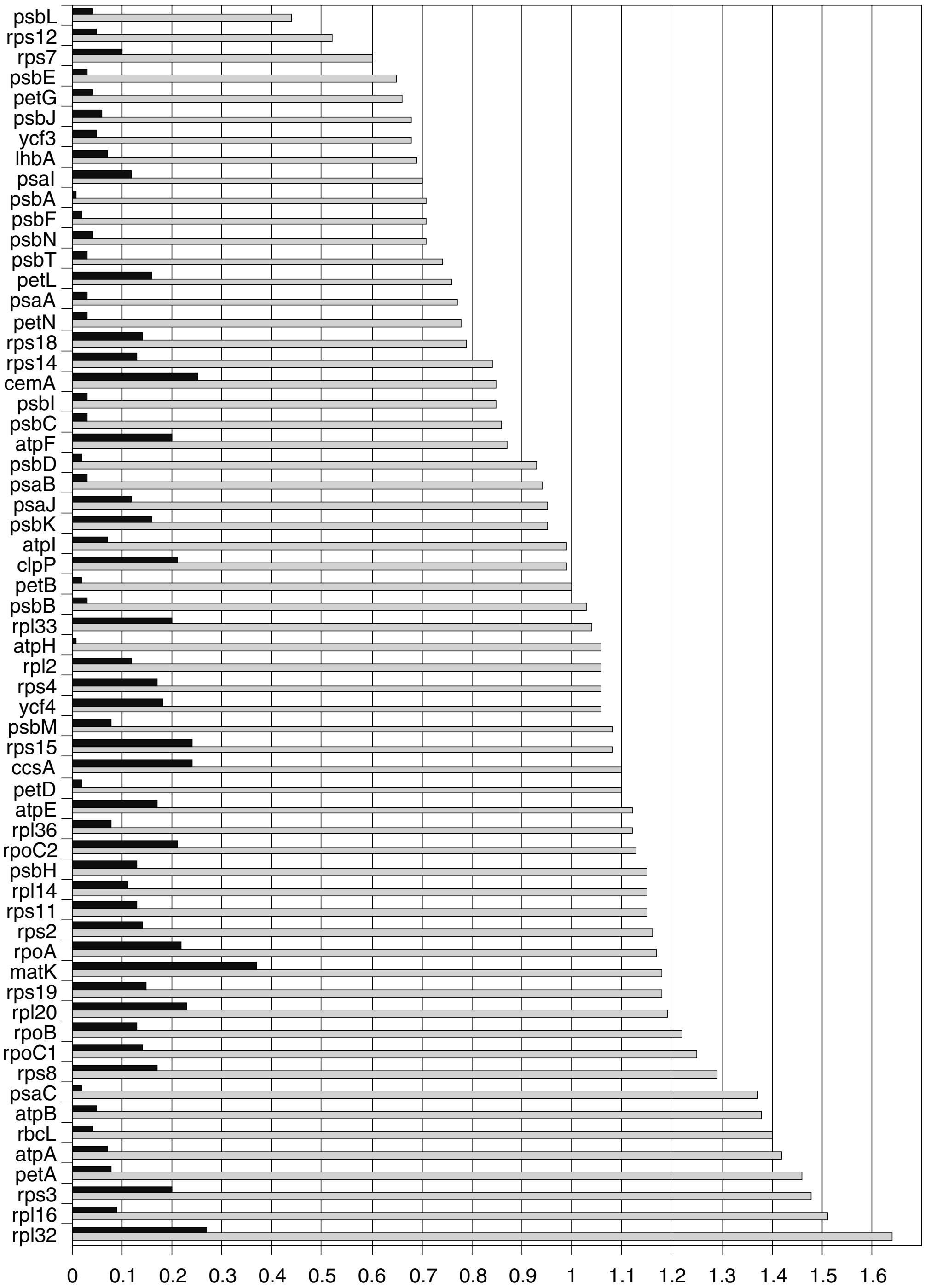

Fig. 4. Mean rates of synonymous and nonsynonymous substitutions for 61 proteins common to 12 land plant genomes determined with the yn00 program. The black bars represent the rate of the nonsynonymous substitutions. The gray bars represent the rate of synonymous substitutions

major land plant groups. 44 of the 61 protein-coding genes analysed have mean dS values among the land plant homologs above 0.8 subst./site, 8 have dS values within the range of 0.8–0.7, 7 genes are within the dS 0.7–0.6 range, and for only two genes these values are lower than 0.6 subst./site (*psb*L and *rps*12). The dS values we registered between *Pinus thunbergii* OTU and 9 angiosperm OTUs are also too high. Of 61 proteins 45 have mean dS values for these comparisons higher than 0.84, 6 are within the dS 0.6–0.8 range, 6 are within the dS 0.5–0.6 range and only four (*rps*7, *rps*12, *psa*I and *psb*L) have dS values lower than 0.5 subst./site. The proteins with lower dS rates are short molecules. These data suggest that the analysis of the unmasked coding sequences from cpDNA is often a poor choice for the investigation of the affinity between different angiosperm taxa and the outgroups necessary to establish the phylogenetic position of angiosperms in respect to other land plant lineages and to address the problem of the "basalmost angiosperm lineage".

At the same time the non-synonymous distances we registered for the 61 land plant genes should pose no problem for phylogeny reconstruction. The highest mean dN value among the 12 land plant OTUs is 0.45 subst./site, detected in the *rps*16 alignment. *Mat*K has 0.35 and all other dN estimates are well below 0.3 subst./site. Even in the comparisons in the 46 gene series from the whole spectrum of the plant kingdom (performed without *mat*K and *rps*16 sequences, since they are missing in some genomes) the dN values for only 4 genes exceed the 0.4 mark. These genes are *atp*F (0.63), *rpo*A (0.47), *rpo*C2 (0.45) and *ycf*4

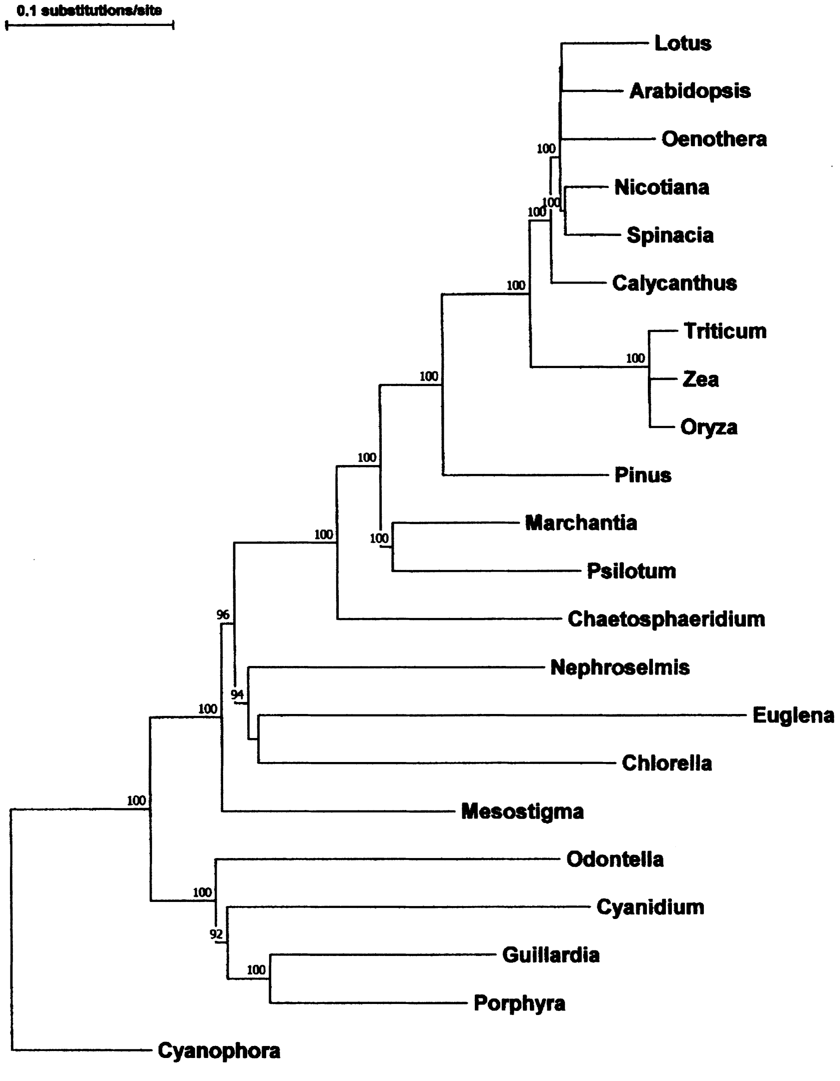

Fig. 5. Neighbor-Joining tree built from Kimura (1983) distances derived from analysis of the data set built from 46 genes. The branches with BP support below 80/100 are shown as unresolved

(0.41). For the above reasons we performed the phylogenetic analyses with the translated sequences. After manual edition we concatenated the alignments of 46 proteins common to all completely sequenced plastomes of photosynthetic plants to create a 11485 positions long total alignment. We produced another 14422 amino acids long alignment of good quality with 61 protein-coding genes from plastomes of land plants.

A neighbor-joining tree from Kimura distances (Kimura 1983) (as implemented in the TREECON package (Van de Peer and De Wachter 1994)) based on the alignment of 46 chloroplast proteins is presented in Fig. 5. On that tree *Calycanthus fertilis* appears as sister to eudicots. *Marchantia* and *Psilotum* form a clade at the base of the land plant clade, which in turn forms the sister group with the charophyte alga *Chaetosphaeridium* (Turmel et al. 2002). All these branches receive 100/100 bootstrap proportion support. Neither the tree topology, nor the bootstrap support for these branches changed when we applied Tajima-Nei (1984) (TREECON implementation) and Dayhoff distances (Dayhoff et al. 1978) (PHYLIP implementation (Felsenstein 1989)). The NJ trees built with the above three distance methods from the concatenated alignment of 61 proteins from land plants were all topologically identical with the land plant cluster of the tree presented in Fig. 5. The position of *Calycanthus fertilis* on all three land plant trees still received 100/100 BP support. The angiosperm topology became more stable: the clade (*Oenothera* (*Lotus*, *Arabidopsis*)) appeared as a sister group to the *Nicotiana*/*Spinacia* branch on all trees built from this dataset. The branch *Lotus*/*Arabidopsis* received bootstrap support above 90%.

Among recent works, only Qiu et al. (1999) and Barkman et al. (2000) found divison of monocots and dicots prior to the branching off of Laurales. In these papers both the position of monocots and the lauralean-eudicot clade received no significant support and hence were not discussed. Often though Laurales form a group with other woody magnoliids and Piperales s. l. (Mathews and Donoghue 1999, Qiu et al. 1999, Zanis et al. 2003).

A probable interpretation of our data and the data recently published by other authors is that the paraphyletic Magnoliidae are divided into the basalmost flowering plants and the “archaeodicot” group, comprising Laurales, Magnoliales, Piperales and Winterales.

The sister group relationship of *Psilotum* and *Marchantia* was recovered with highest bootstrap support in our analyses. Yet it contradicts some of the recent molecular analyses based on a relatively small amount of characters. A similar phenomenon has been noticed before: of 46 chloroplast proteins used in the analyses by Stöbe et al. (1999) only 2 recovered the topology unambigously supported by the dataset of their concatenated alignments. The authors suggested that it was due to the fact that the phylogenetically relevant characters are dispersed across the whole dataset and are not concentrated in one gene. Our current analysis shows that homoplasy in chloroplast coding sequences is yet another problem to consider. Many previous studies included unmasked coding sequences of the chloroplast genes which are saturated with substitutions at the synonymous codon positions. For example, Pryer et al. (2001) used unmasked *rbc*L, *atp*B and *rps*4 sequences which all exibited a dS rate above 1 subst./site in comparisons among land plants. Our analyses also have a weak point: poor taxon sampling. We are, however, confident that, as more chloroplast genome sequences will become available, the phylogenetic history of land plant diversification will be revealed.

On the whole, one can conclude that in phylogenetic comparisons involving angiosperms and bryophytes more than a half of the cpDNA sequence is of little use in deducing phylogenetic relationships. In the absence of the rearrangement-based method of phylogeny reconstruction, this suggests to us that for studies of the land plant phylogeny on the whole, accumulating a set of markers may have a priority over the complete chloroplast genome sequencing. In comparisons across the whole plant kingdom one would have no other option but to employ a pre-selected set of molecular markers. The value of the complete plastome sequence data increases with further analyses of more related taxa. So, approximately 70% of the cpDNA sequence can be used in phylogenetic analysis for comparisons involving different classes of angiosperms. This emphasizes the importance of the chloroplast genomic data for phylogeny reconstruction within flowering plants.

We would like to express our gratitude to the German Research Society (DFG) for providing financial support. We thank William Martin for helpful and interesting discussion.

References

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Angiosperm Phylogeny Group (1998) An ordinal classification for the families of flowering plants. Ann. Missouri Bot. Gard. 85: 531–553.

Barkman T. J., Chenery G., McNeal J. R., Lyons-Weiler J., Ellisens W. J., Moore G., Wolfe A. D., dePamphilis C. W. (2000) Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc. Natl. Acad. Sci. U S A. 2000 97: 13166–13171.

Beck C. B. (1976) Origin and early evolution of angiosperms: a perspective. In: Beck C. B. (ed.) Origin and early evolution of angiosperms. Columbia University Press, New York, London, pp. 1–10.

Borer P. N., Dengler B., Tinoco I. (1974) Stability of ribonucleic acid and double-stranded helices. J. Mol. Biol. 86: 843–853.

Chant S. R. (1978) Calycanthaceae. In: Heywood V. H (ed.) Flowering Plants of the World. Elsevier International Projects Ltd., Oxford, pp. 35–36.

Chase M. W., 41 others (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plasid gene *rbc*L. Ann. Missouri Bot. Gard. 80: 528–580.

Crane P. R., Friis E. M., Pedersen K. R. (1995) The origin and early diversification of angiosperms. Nature 374: 27–33.

Dayhoff M. O., Schwartz R. M., Orcutt B. C. (1978) A model of evolutionary change in proteins. In: Dayhoff M. O. (ed.) Atlas of protein sequence and structure, vol. 5. National Biochemical Research Foundation, Washington DC, pp. 345–352.

Douglas S. E., Penny S. L. (1999) The plastid genome of the cryptophyte alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. J. Mol. Evol. 48: 236–244.

Downie S. R., Palmer J. D. (1992) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis P. S., Soltis D. E., Doyle J. J. (eds.) Molecular systematics of plants. Chapman and Hall, New York, London, pp. 14–35.

Drescher A., Ruf S., Calsa T. Jr, Carrer H., Bock R. (2000) The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 22: 97–104.

Ewing B., Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–184.

Ewing B., Hillier L., Wendl M. C., Green P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175–185.

Felsenstein J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.

Friis E. M., Chaloner W. G., Crane P. R. (1987) Introduction to angiosperms. In: Friis E. M., Chaloner W. G., Crane P. R. (eds.) The origins of angiosperms and their biological consequences. Cambridge University Press Cambridge, New York, Port Chester, Melbourne, Sydney, pp. 1–15.

Friis E. M., Eklund H., Pedersen K. R., Crane P. R. (1994) *Virginianthus calycanthoides* gen. et sp. nov.: a calycanthaceous flower from the Potomac Group (Early Cretaceous) of eastern North America. Int. J. Pl. Sci. 155: 772–785.

Glockner G., Rosenthal A., Valentin K. (2000) The structure and gene repertoire of an ancient red algal plastid genome. J. Mol. Evol. 51: 382–390.

Graham S. W., Olmstead R. G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Amer. J. Bot. 87: 1712–1730.

Hallick R. B., Hong L., Drager R. G., Favreau M. R., Monfort A., Orsat B., Spielmann A., Stutz E. (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. Nucl. Acids Res. 21: 3537–3544.

Hiratsuka J., Shimada H., Whittier R., Ishibashi T., Sakamoto M., Mori M., Kondo C., Honji Y., Sun C. R., Meng B. Y., Li Y. Q., Kanno A., Nishizawa Y., Hirai A., Shinozaki K., Sugiura M. (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. 217: 185–194.

Hirose T., Kusumegi T., Tsudzuki T., Sugiura M. (1999) RNA editing sites in tobacco chloroplast transcripts: editing as a possible regulator of chloroplast RNA polymerase activity. Mol. Gen. Genet. 262: 462–467.

Huang X., Miller W. A (1991) Time-efficient, linear-space local similarity algorithm. Advances in Applied Mathematics 12: 337–357.

Hupfer H., Swiatek M., Hornung S., Hermann R. G., Maier R. M., Chiu W-L., Sears B. (2000) Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable *Euoenothera* plastomes. Mol. Gen. Genet. 263: 581–585.

Jukes T. H., Cantor C. R. (1969) Evolution of protein molecules. In: Munro H. N. (ed.) Mammalian protein metabolism. Academic Press, New York, pp. 21–132.

Kato T., Kaneko T., Sato S., Nakamura Y., Tabata S. (2000) Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. DNA Res. 7: 323–330.

Kellogg E. A. (1998) Who's related to whom? Recent results from molecular systematic studies. Curr. Opin. Plant Biol. 1998 1: 149–158.

Kenrick P. (1999) The family tree flowers. Nature 402: 358–359.

Kimura M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16: 111–120.

Kimura M. (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, p. 75.

Kowallik K. V., Stoebe B., Schaffran I., Kroth-Pancic P., Freier U. (1995) The chloroplast genome of a chlorophyll a + c- containing alga, *Odontella sinensis*. Plant. Mol. Biol. Rep. 13: 336–342.

Lecointre G., Hervé P., Le H. L., Le Guyader H. (1994) How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. Mol. Phylogenet. Evol. 3: 292–309.

Lemieux C., Otis C., Turmel M. (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature 403: 649–652.

Les D. H., Garvin D. K., Wimpee C. F. (1991) Molecular evolutionary history of ancient aquatic angiosperms. Proc. Natl. Acad. Sci. USA 88: 10119–10123.

Lowe T. M., Eddy S. R. (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 25: 955–964.

Maier R. M., Döry I., Igloi G. L., Kössel H. (1990) The *ndh*H genes of graminean plastomes are linked with the junctions between small single copy and inverted repeat regions. Curr. Genet. 18: 245–250.

Maier R. M., Neckermann K., Igloi G. L., Kossel H. (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J. Mol. Biol. 251: 614–628.

Manhart J. R., Palmer J. D. (1990) The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. Nature 345: 268–270.

Martin W., Stoebe B., Goremykin V., Hansmann S., Hasegawa M., Kowallik K. V. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393: 162–165.

Mathews S., Donoghue M. J. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286: 974–950.

Moon E., Wu R. (1988) Organization and nucleotide sequence of genes at both junctions between the two inverted repeats and the large single-copy region in the rice chloroplast genome. Gene 70: 1–12.

Muller J. (1984) Significance of fossil pollen for angiosperm history. Ann. Missouri Bot. Gard. 71: 419–443.

Murray M. G., Thompson W. F. (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8: 4321–4325.

Ogihara Y., Isono K., Kojima T., Endo A., Hanaoka M., Shiina T., Terachi T., Utsugi S., Murata M., Mori N., Takumi S., Ikeo K., Gojobori T., Murai R., Murai K., Matsuoka Y., Ohnishi Y., Tajiri H., Tsunewaki K. (2002) Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. Mol. Genet. Genomics 266: 740–746.

Ohtani K., Yamamoto H., Akimitsu K. (2002) Sensitivity to *Alternaria alternata* toxin in citrus because of altered mitochondrial RNA processing. Proc. Natl. Acad. Sci. USA 99: 2439–2444.

Ohyama K., Fukuzawa H., Kohchi T., Shirai H., Sano T., Sano S., Umesono K., Shiki Y., Takeuchi M., Chang Z., Aota S., Inokuchi H., Ozeki H. (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322: 572–574.

Parkinson C. L., Adams K. L., Palmer J. D. (1999) Multigene analyses identify three earliest lineages of extant flowering plants. Curr. Biol. 9: 1485–1488.

Pryer K. M., Schneider H., Smith A. R., Cranfill R., Wolf P. G., Hunt J. S., Sipes S. D. (2001) Horsetails and ferns are a group and the closest living relatives to seed plants. Nature 409: 618–622.

Qiu Y-L., Chase M. W., Les D. H., Parks C. R. (1993) Molecular phylogenetics of the Magnoliidae: Cladistic analysis of nucleotide sequences of the plastid gene *rbc*L. Ann. Missouri Bot. Gard. 80: 587–606.

Qiu Y-L., Lee J., Bernasconi-Quadroni F., Soltis D. E., Soltis P. S., Zanis M., Zimmer E. A., Chen Z., Savolainen V., Chase M. W. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402: 404–407.

Reith M. E., Munholland J. (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol. Biol. Rep. 13: 333–335.

Renner S. S. (1999) Circumscription and phylogeny of the Laurales: evidence from molecular and morphological data. Amer. J. Bot. 86: 1301–1315.

Rychlik W., Rhoads R. E. (1990) Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acids Res. 18: 6409–6412.

Sato S., Nakamura Y., Kaneko T., Asamizu E., Tabata S. (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res. 6: 283–290.

Schmitz-Linneweber C., Maier R. M., Alcaraz J. P., Cottet A., Herrmann R. G., Mache R. (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Mol. Biol. 45: 307–315.

Shinozaki K., Ohme M., Tanaka M., Wakasugi T., Hayashida N., Matsubayashi T., Zaita N., Chunwongse J., Obokata J., Yamaguchi-shinozaki K., Ohto C., Torazawa K., Meng B-Y., Sugita M., Deno H., Kamogashira T., Yamada K., Kusuda J., Takaiwa F., Kato A., Tohdoh N., Shimada H., Sugiura M. (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 5: 2043–2049.

Soltis P. S., Soltis D. E., Chase M. W. (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402: 402–403.

Staden R., Beal K. F., Bonfield J. K. (2000) The Staden package, 1998. Methods Mol. Biol. 132: 115–130.

Staden R., Judge D. P., Bonfield J. K. (2001) Sequence assembly and finishing methods. Methods Biochem. Anal. 43: 303–322.

Stöbe B., Hansmann S., Goremykin V., Kowallik K. V., Martin W. (1999) Proteins encoded in sequenced chloroplast genomes: an overview of gene content, phylogenetic information and endosymbiotic gene transfer to the nucleus. In: Hollingsworth P. M., Bateman R. M., Gornall R. J. (eds.) Molecular systematics and plant evolution. Taylor & Francis, London, pp. 327–352.

Sugita M., Kato A., Shimada H., Sugiura M. (1984) Sequence analysis of the junctions between a large inverted repeat and single-copy regions in tobacco chloroplasts. Mol. Gen. Genet. 194: 200–205.

Sugita M., Svab Z., Maliga P., Sugiura M. (1997) Targeted deletion of *spr*A from the tobacco plastid genome indicates that the encoded small RNA is not essential for pre-16S rRNA maturation in plastids. Mol. Gen. Genet. 257: 23–27.

Sytsma K. J., Hahn W. J. (1997) Molecular systematics: 1994–1995. Prog. Bot. 58: 470–499.

Tajima F., Nei M. (1984) Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 1: 269–285.

Thompson J. D., Higgins D. G., Gibson T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

Turmel M., Otis C., Lemieux C. (1999) The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. Proc. Natl. Acad. Sci. USA. 96: 10248–10253.

Turmel M., Otis C., Lemieux C. (2002) The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc. Natl. Acad. Sci. USA 99: 11275–11280.

Van de Peer Y., De Wachter R. (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Applic. Biosci. 10: 569–570.

Vera A., Sugiura M. (1994) A novel RNA gene in the tobacco plastid genome: its possible role in the maturation of 16S rRNA. EMBO J. 13: 2211–2217.

Wakasugi T., Nagai T., Kapoor M., Sugita M., Ito M., Ito S., Tsudzuki J., Nakashima K., Tsudzuki T., Suzuki Y., Hamada A., Ohta T., Inamura A., Yoshinaga K., Sugiura M. (1997) Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. Proc. Natl. Acad. Sci. USA. 94: 5967–5972.

Wakasugi T., Tsudzuki J., Ito S., Nakashima K., Tsudzuki T., Sugiura M. (1994) Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc. Natl. Acad. Sci. USA 91: 9794–9798.

Yang Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13: 555–556.

Zanis M. J., Soltis D. E., Soltis P. S., Mathews S., Donoghue M. J. (2003) The root of the angiosperms revisited. Proc. Natl. Acad. Sci. USA 99: 6848–6853.

Addresses of the authors: Vadim Goremykin (e-mail: Vadim.Goremykin@uni-jena.de), F. H. Hellwig, Institut für Spezielle Botanik, Universität Jena, Philosophenweg 16, D-07743 Jena, Germany. K. I. Hirsch-Ernst, Zentrum Pharmakologie und Toxikologie, Universität Göttingen, Robert-Koch-Strasse 40, D-37075 Göttingen, Germany. S. Wölfl, Klinik für Innere Medizin, Universität Jena, Erlanger Allee 101, D-07740 Jena, Germany.