Tree Genetics & Genomes (2006) 3:49–59
DOI 10.1007/s11295-006-0051-3

ORIGINAL PAPER

The chloroplast genome of mulberry: complete nucleotide sequence, gene organization and comparative analysis

V. Ravi · Jitendra P. Khurana · Akhilesh K. Tyagi · Paramjit Khurana

Received: 18 April 2006 / Revised: 8 June 2006 / Accepted: 16 June 2006 / Published online: 23 September 2006
© Springer-Verlag 2006

Abstract The complete nucleotide sequence of mulberry (*Morus indica* cv. K2) chloroplast genome (158,484 bp) has been determined using a combination of long PCR and shotgun-based approaches. This is the third angiosperm tree species whose plastome sequence has been completely deciphered. The circular double-stranded molecule comprises of two identical inverted repeats (25,678 bp each) separating a large and a small single-copy region of 87,386 bp and 19,742 bp, respectively. A total of 83 protein-coding genes including five genes duplicated in the inverted repeat regions, eight ribosomal RNA genes and 37 tRNA genes (30 gene species) representing 20 amino acids, were assigned on the basis of homology to predicted genes from other chloroplast genomes. The mulberry plastome lacks the genes *infA*, *sprA*, and *rpl21* and contains two pseudogenes *ycf15* and *ycf68*. Comparative analysis, based on sequence similarity, both at the gene and genome level, indicates *Morus* to be closer to *Cucumis* and *Lotus*, phylogenetically. However, at genome level, inclusion of non-coding regions brings it closer to *Eucalyptus*, followed by *Cucumis*. This may reflect differential selection pressure operating on the genic and intergenic regions of the chloroplast genome.

Keywords *Morus* · Chloroplast genome · PCR · Comparative analysis · Phylogeny

Introduction

Complete chloroplast genome sequences are available for various vascular and non-vascular land plants, green and red algae, and secondary algal lineages (Odintsova and Yurina 2003; Ohta et al. 2003; Sugiura et al. 2003; Wolf et al. 2003, 2005; Asano et al. 2004; Goremykin et al. 2004, 2005; Hagopian et al. 2004; Kim and Lee 2004; Shahid Masood et al. 2004; Kim et al. 2006; Pombert et al. 2005; Saski et al. 2005; Steane 2005; Turmel et al. 2005) (http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/plastids_tax.html). Among the dicots, complete chloroplast genome sequences are available for at least one representative from each subclass except Hamamelidae. Mulberry is a dicot (Magnoliopsida) belonging to the subclass Hamamelidae (Family: Moraceae; Order: Urticales) (http://plants.usda.gov/), which is the smallest subclass (in terms of taxa) among the dicots. The plastome sequence of mulberry, therefore, represents the first case study from this subclass. Mulberry (*Morus indica*) is a perennial tree or shrub cultivated extensively in East, Central, and South Asia for silk production; its foliage is the sole food for the domesticated silkworm (*Bombyx mori*). Sixty-eight species of mulberry are found mostly in Asia, mainly China and Japan, and continental America. They are poorly represented in Africa and Europe and virtually absent in Australia. Besides silk production, mulberry is used for its medicinal properties (foliage), as fodder (foliage), as a source of firewood (pruned shoots), for preparation of jams (fruit) and also as a shade plant for intercropping between rows of other plants like tea and coffee. Although there are

Electronic supplementary material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s11295-006-0051-3 and is accessible for authorized users.

Communicated by Y. Tsumura

V. Ravi · J. P. Khurana · A. K. Tyagi · P. Khurana (✉)
Interdisciplinary Centre for Plant Genomics (ICPG)
and Department of Plant Molecular Biology,
University of Delhi South Campus,
New Delhi 110 021, India
e-mail: param@genomeindia.org

a few reports on phylogenetic studies involving mulberry, these are restricted to only a few genes. A complete repertoire of genes would thus help in establishing the position of mulberry in the tree of life. We present in this paper the complete organization of the mulberry chloroplast genome along with some interesting comparative analyses with other known plastid genomes.

Materials and methods

Chloroplast DNA was isolated from mature green leaves kept in darkness for 48–72 h by the sucrose gradient method (Palmer 1986) followed by purification by cesium chloride–ethidium bromide density gradient centrifugation. As chloroplast DNA could not be separated from the nuclear DNA on the gradient, a long PCR-based strategy was employed to amplify the DNA fragments of interest. Primers were designed from conserved regions of genes from complete chloroplast genomes using Gene Runner (version 3.05) (http://www.generunner.com). Multiple alignments were performed using the ClustalW (Thompson et al. 1994) component of BioEdit sequence alignment editor (Hall 1999). PCR products were generated using the Eppendorf Triple Master PCR System. Products were gel-purified using the QIAquick gel extraction kit (QIAGEN). Products above 3 kb were sheared using HydroShear™ (GeneMachines) and the end-repaired fragments cloned in linearized and dephosphorylated (*Sma*I and bacterial alkaline phosphatase-treated) pUC19 vector (MBI Fermentas).

Smaller products were sequenced by primer walking. Sequencing reactions were performed with Big Dye Terminators™ (ver. 3.1, Applied Biosystems) and products analyzed on ABI Prism 3700 automated sequencers. Vector trimming, assembly, and editing of sequences was done using Phred-Phrap (Ewing and Green 1998) and Consed (Gordon et al. 1998). Annotation of protein-coding, rRNA and tRNA genes was done by Dual Organellar GenoMe Annotator (Wyman et al. 2004) (http://bugmaster.jgi-psf.org/dogma/).

Further confirmation of gene predictions, gene locations, intron–exon boundaries, locations of single-copy regions, inverted repeats, and IR expansion/contraction was done using different programs of NCBI BLAST (Altschul et al. 1990) (http://www.ncbi.nlm.nih.gov/BLAST/). All plastid sequences were obtained from NCBI except *Populus* (available at http://genome.ornl.gov/poplar_chloroplast/). Global alignments were performed using VISTA (Mayor et al. 2000), which uses AVID (Bray et al. 2003) as the alignment program. Global alignment-level phylogenetic tree was generated using MAVID/AMAP (Bray and Pachter 2004). The tRNAs were confirmed using tRNAscan-SE ver. 1.21 (Lowe and Eddy 1997). Phylogenetic analysis was done using TREECON version 1.3b (Van de Peer and De Wachter 1994), Tree-Puzzle 5.1 (Strimmer and von Haeseler 1996) (available online at http://bioweb.pasteur.fr/intro-uk.html) and PHYLIP (the PHYLogeny Inference Package, v3.6; Felsenstein 1989). Neighbor joining (NJ) tree was created using distance calculation based on Kimura (1983) and Tajima and Nei (1984) as implemented in the TREECON package. Dayhoff et al. (1978) and Jones–Taylor–Thornton (JTT; Jones et al. 1992) substitution models were employed for phylogenetic reconstruction in case of Tree-puzzle. Protein parsimony method was also tested using PROTPARS from the Phylip (Felsenstein 1989) package.

Results

Plastome organization

The chloroplast genome of *Morus indica* (GenBank Accession Number DQ226511) is a circular double-stranded DNA of 158,484 bp with an overall A+T content of 63.63%. The plastome harbors a pair of identical inverted repeat regions (IR_A and IR_B), which are 25,678 bp each. The inverted repeats are separated by a large single-copy (LSC) region of 87,386 bp and a small single-copy (SSC) region of 19,742 bp. The positions of all the genes identified in the mulberry plastome and category-wise distribution of these genes are presented in Fig. 1 and Table 1, respectively. More than half of the plastome is composed of coding regions (90,532 bp; 57.12%) with the peptide-coding regions accounting for the major portion (78,681 bp; 49.65%) followed by ribosomal RNA genes (9,050 bp; 5.71%) and transfer RNA genes (2,801 bp; 1.77%). The remaining 42.88% (67,952 bp) is covered by intergenic regions (46,923 bp; 29.61%) and a total of 20 introns (21,029 bp; 13.27%) present within 18 genes (counting the IR genes only once) (see Supplementary Table S2).

Codon usage

The codon usage (Supplementary Table S3) of the mulberry plastome strongly reflects the AT bias. Majority of the codons (70.75%) end in A or T. Even the stop codons are biased with 73.49% ending in A or T. The A+T content of the IR region is the least and amounts to 57.08% reflecting the low A+T content of the ribosomal RNA genes (44.62%), whereas the A+T content in the LSC and SSC regions is 65.88 and 70.65%, respectively. There are in total 26,179 codons, which represent the total coding capacity of the mulberry plastome; 2,803 (10.71%) of these are for amino acid leucine, 2,281 (8.71%) for isoleucine, 1,990

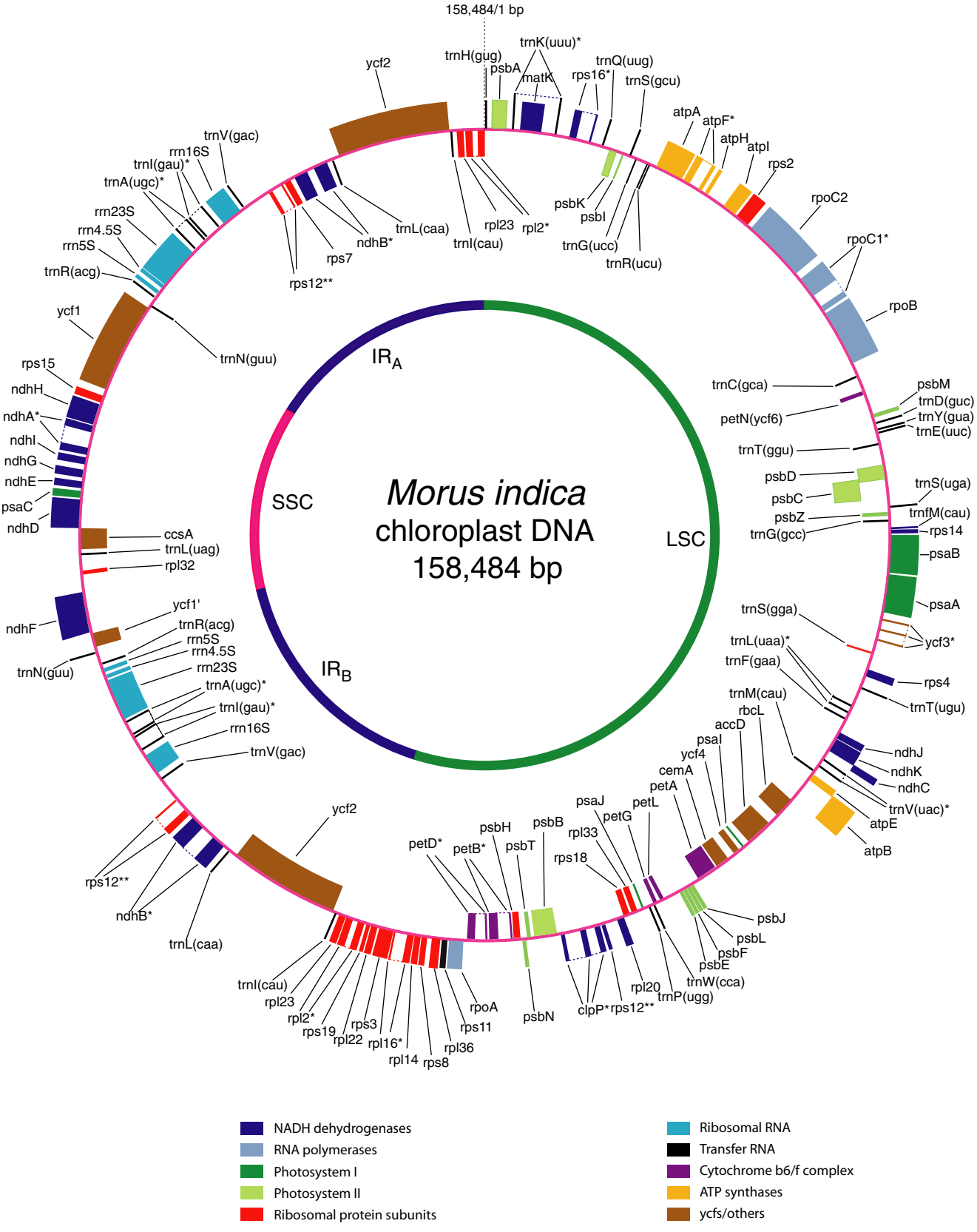


Fig. 1 Organization of the mulberry chloroplast genome. The IR_A/LSC junction (JLA) represents the start of the genome (base position 1). Genes drawn outside and inside the *circle* represent those on the positive and negative strands, respectively. Genes joined by *dotted lines* represent intron-containing genes. Different categories of genes are color-coded. *intron containing gene, **genes showing trans splicing

Table 1 Genes present in the mulberry plastome

Category	Gene names
Ribosomal RNAs	*rrn16*, *rrn23*, *rrn4.5*, *rrn5*
Transfer RNAs	*trnA-UGC*, *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnG-GCC*, *trnG-UCC*, *trnH-GUG*, *trnI-CAU* *trnI-GAU*, *trnK-UUU*, *trnL-CAA*, *trnL-UAA*, *trnL-UAG*, *trnM-CAU*, *trnN-GUU*, *trnP-GGG*, *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*, *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC*, *trnV-UAC*, *trnW-CCA*, *trnY-GUA*
Proteins of small ribosomal subunit	*rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps12*, *rps14*, *rps15*, *rps16*, *rps18*, *rps19*
Proteins of large ribosomal subunit	*rpl2*, *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*
Subunits of RNA polymerase	*rpoA*, *rpoB*, *rpoC1*, *rpoC2*
Subunits of NADH-dehydrogenase	*ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Subunits of Photosystem I	*psaA*, *psaB*, *psaC*, *psaI, psaJ*
Subunits of Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
Large subunit of Rubisco	*rbcL*
Subunits of cytochrome *b/f* complex	*petA*, *petB*, *petD*, *petG*, *petL*, *petN*
Subunits of ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*
Acetyl-CoA carboxylase	*accD*
Cytochrome *c* biogenesis	*ccsA*
Maturase	*matK*
Protease	*clpP*
Envelope membrane protein	*cemA*
Conserved hypothetical chloroplast reading frames	*ycf1*, *ycf2*, *ycf3*, *ycf4*

(7.6%) for serine, and 1,750 (6.68%) are for glycine. One-third of the total codons are represented by these four amino acids. Again, most of the codons end in A or T for these four amino acids (66.82% for leucine, 80.62% for isoleucine, 68.54% for serine, and 74.51% for glycine). The least number of codons (excluding stop codons) are for cysteine (307; 1.17%) and these too are with an AT bias (76.87% ending in T).

Comparison with other plastomes

The gene content, order, and organization of the *Morus indica* chloroplast genome is similar to that of higher plants except for the inverted repeat/single copy (IR/SC) junction regions (see IR expansion/contraction section). Genes absent in the mulberry plastome are *infA* (translation initiation factor 1), *sprA* (small plastid RNA), and *rpl21* (ribosomal protein large subunit 21). Two non-functional genes *ycf15* and *ycf68* (Supplementary Table S4), are also present in the mulberry plastome. The *ycf15* gene is 5′-truncated with respect to *Nicotiana*, *Atropa*, *Panax*, and *Nymphaea* genes (determined by TBLASTN). When compared to the *Panax* and *Nymphaea*, the reading frame of mulberry reveals stop codons and an insertion corresponding to 20 amino acids (Supplementary Fig. S1c). The *ycf68* gene on the other hand is a non-truncated pseudogene in mulberry having accumulated stop codons in its reading frame. Upon close scrutiny, it was observed that in maize and rice, there are two 'AAAC' units one after another in a region before the first stop codon of mulberry. One of the 'AAAC' units was missing from the mulberry *ycf68* gene, which was causing the frameshift and the resulting stop codons. When the four bases were added, the frame was restored (Supplementary Fig. S1a,b). In *Pinus*, although four bases 'CAAA' are missing in the same region, an additional stretch of four bases 'TGTG' restores the frame and, hence, does not have stop codons.

Eighteen genes in the mulberry plastome contain one or two introns. This is identical to the *Panax*, *Cucumis*, and *Calycanthus* plastome. In comparison, there are 17 intron-containing genes in the *Spinacia* plastome and 15 in the tobacco plastome. Supplementary Table S2 summarizes the sizes of exons and introns for each gene. Five of these introns, *rpl2*, *ndhB*, *rps12*, *trnI-GAU*, and *trnA-UGC*, are located within the IR regions. Genes *clpP* and *ycf3* have two introns each. A total of 30 tRNA gene species coding for all 20 amino acids were identified from the *M. indica* chloroplast genome. The numbers and kinds of tRNA genes from *M. indica* are identical to well characterized vascular plant chloroplasts. The codon usage of the *M. indica* chloroplast genome and the anticodons present in the 30 tRNA species are summarized in Supplementary Table S3.

IR expansion/contraction

The borders between the two inverted repeats (IR_A and IR_B) and the two single-copy regions (LSC and SSC) usually differ in plastomes of various species. The length variations in chloroplast genomes of different plant groups are often due to expansions and contractions in the IR regions. These expansions and contractions result in the truncation of genes at or near the boundaries (e.g., *rps19*

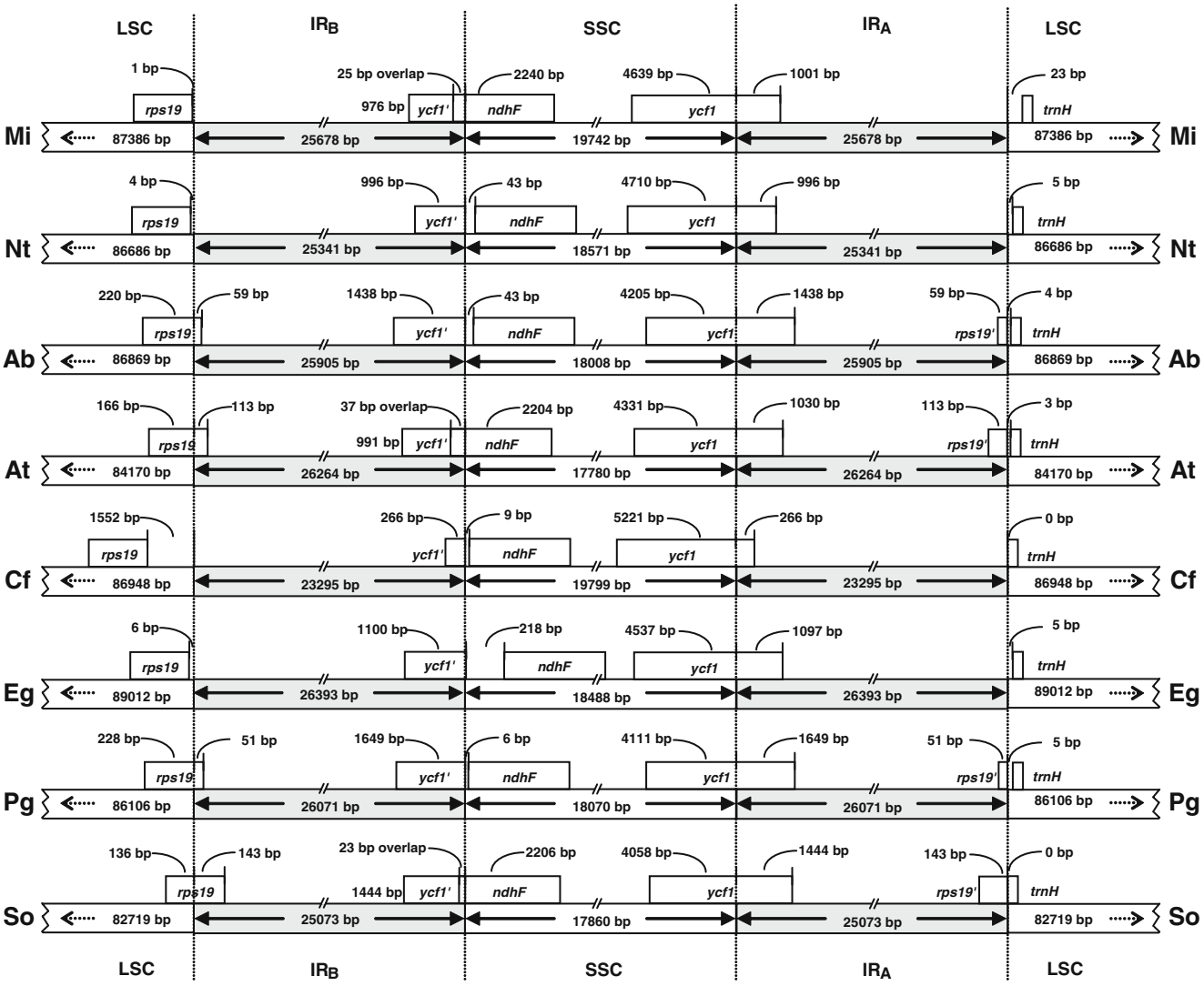


Fig. 2 Detailed view of the inverted repeat-single copy (IR/SC) border regions from various dicot species with respect to the genes located at or near the boundaries. Gene names suffixed by a single quotation mark represent pseudogenes. The figure is not to scale and only shows relative changes at or near the IR-SC boundaries. Mi: *Morus indica*, Nt: *Nicotiana tabacum*, Ab: *Atropa belladonna*, At: *Arabidopsis thaliana*, Cf: *Calycanthus fertilis*, Eg: *Eucalyptus globulus*, Pg: *Panax ginseng*, So: *Spinacia oleracea*

and *ycf1*). Comparisons of IR boundaries reveal that these changes are not uncommon in higher plants. Figure 2 shows the detailed view of the IR/SC border positions with respect to adjacent genes of *Morus*, *Eucalyptus*, *Calycanthus*, *Panax*, *Nicotiana*, *Atropa*, *Spinacia*, and *Arabidopsis*. A close-up of the border regions reveals differences even between *Nicotiana* and *Atropa*—members of the same family.

The IR_A/SSC borders are located in the 3′ region of the *ycf1* gene and create the *ycf1* pseudogenes at the IR_B/SSC border with lengths of 996 bp and 1,438 bp in *Nicotiana* and *Atropa*, respectively. The same is the case with *Morus* where a *ycf1* pseudogene of 1,001 bp is produced at the IR_B/SSC border. The situation is, however, different in the sense that this *ycf1* pseudogene has an overlap of 25 bp with the *ndhF* gene. The IR_B/SSC border passes through the *ndhF* gene in *Morus* just as it does in *Arabidopsis*. The IR/LSC borders, unlike their SSC-counterparts, show much less variation between different genera. The IR_A/LSC borders are usually located downstream of the non-coding region of *trnH-GUG*. The IR_B/LSC borders are either located upstream of the non-coding region or within the coding region of *rps19*.

In *Morus*, a single base pair separates the *rps19* gene from the IR_B/LSC border and thus the gene marks the end of the LSC. Thus, no *rps19* pseudogene is created at the other border. A similar situation is seen in *Nicotiana*, where 4 bp separate the *rps19* from the IR_B/LSC border, and in *Eucalyptus,* where there are 6 bp in between. For other genera the case is different and the IR_B/LSC borders pass

through the *rps19* gene to create *rps19* pseudogenes of lengths 59 bp, 113 bp, 51 bp, and 143 bp in *Atropa*, *Arabidopsis*, *Panax*, and *Spinacia*, respectively. On the other extreme is *Calycanthus*, where the *rps19* gene is separated from the IR_B/LSC border by 1,552 bp. The expansions/contractions of IR, as observed in the IR/SSC borders, are probably mediated by gene conversion and recombinational repair of double-strand breaks (Goulding et al. 1996).

Comparison by means of global and local alignments

Comparison of the mulberry plastome sequence with other plastomes at the global level using VISTA brought out *Eucalyptus* (a tree species) as the closest relative, followed by *Panax*, *Populus* (another tree species), *Cucumis*, *Nicotiana*, and *Atropa*. Interestingly, *Acorus*, which is thought to be one of the ancient lineages of monocots, appears to be closer to the dicots. Morphological similarities with dicots have already been reported for *Acorus* (Grayum 1987), and the VISTA alignments further reinforce the fact. A similar kind of result was obtained when concatenated local alignments obtained by BLAST were compared (Fig. 3).

In this case also, *Eucalyptus* came out as the closest relative of *Morus*, followed by *Cucumis*, *Panax*, *Populus*, *Nicotiana*, and *Atropa*. *Acorus*, again, was found to be close to the dicot cluster. Both VISTA and local alignment comparisons clearly bring out three distinct groups—dicots, monocots, and lower plants (including algae). The vista-plot patterns produced are remarkably group-specific and each group shows nearly identical patterns among themselves. The gap regions in *Oenothera* and *Lotus* represent the ~54 and ~51 kb inversions, respectively (Hupfer et al. 2000; Kato et al. 2000) in the large single-copy regions of both the plastomes. A similar gap (~28 kb) representing a large single-copy inversion is seen in the monocots (Doyle et al. 1992).

Phylogenetic positioning of mulberry

Amino acid sequences from 43 protein-coding genes common to 26 genera were used to create a concatenated data set. The total alignment length was 12,299 positions long. Another alignment of 9,727 positions was created after removal of gaps. Similar topologies were obtained with distance-based, maximum parsimony and maximum likelihood approaches using the models specified in the methods section (Fig. 4a–c). In all the cases, mulberry paired up with *Cucumis*, which was highly similar at the global and local alignment levels, too. Bootstrap values of 100 and 53 were obtained in case of NJ and ML trees, respectively. With PROTPARS, the 9,727-position-long alignment gave 44 and 46 trees out of a total of 100 having *Morus/Cucumis* and *Morus/Lotus* pairs, respectively (almost 1:1). The remaining ten trees had *Lotus/Cucumis* or *Lotus/Arabidopsis* pairs. The 12,299-position-long alignment, however, gave a very strong support for *Morus/Cucumis* (89 out of 100 trees).

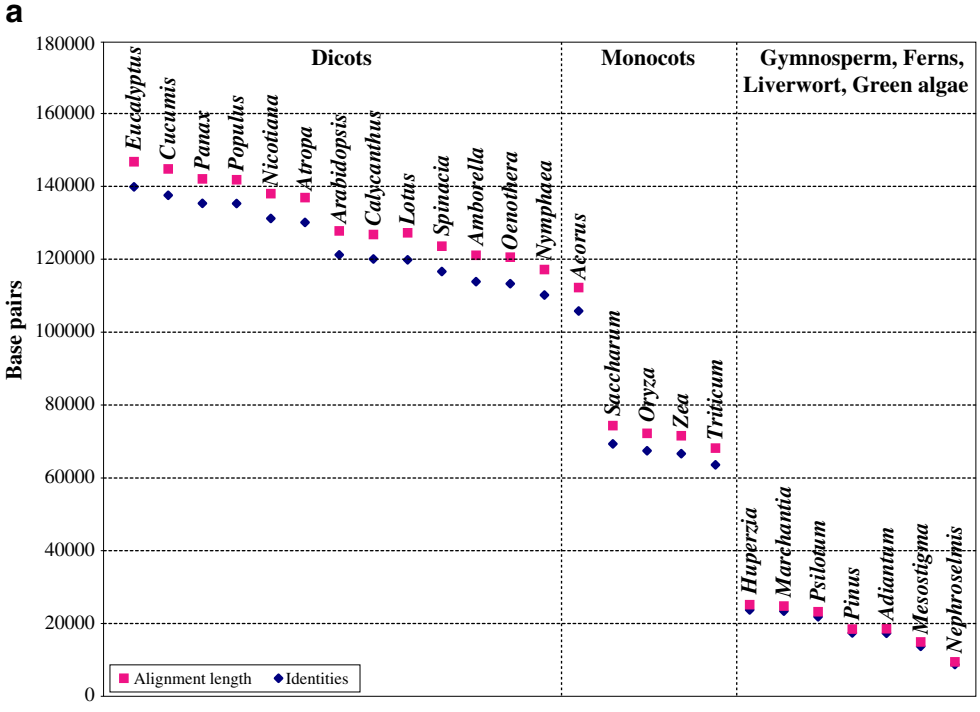
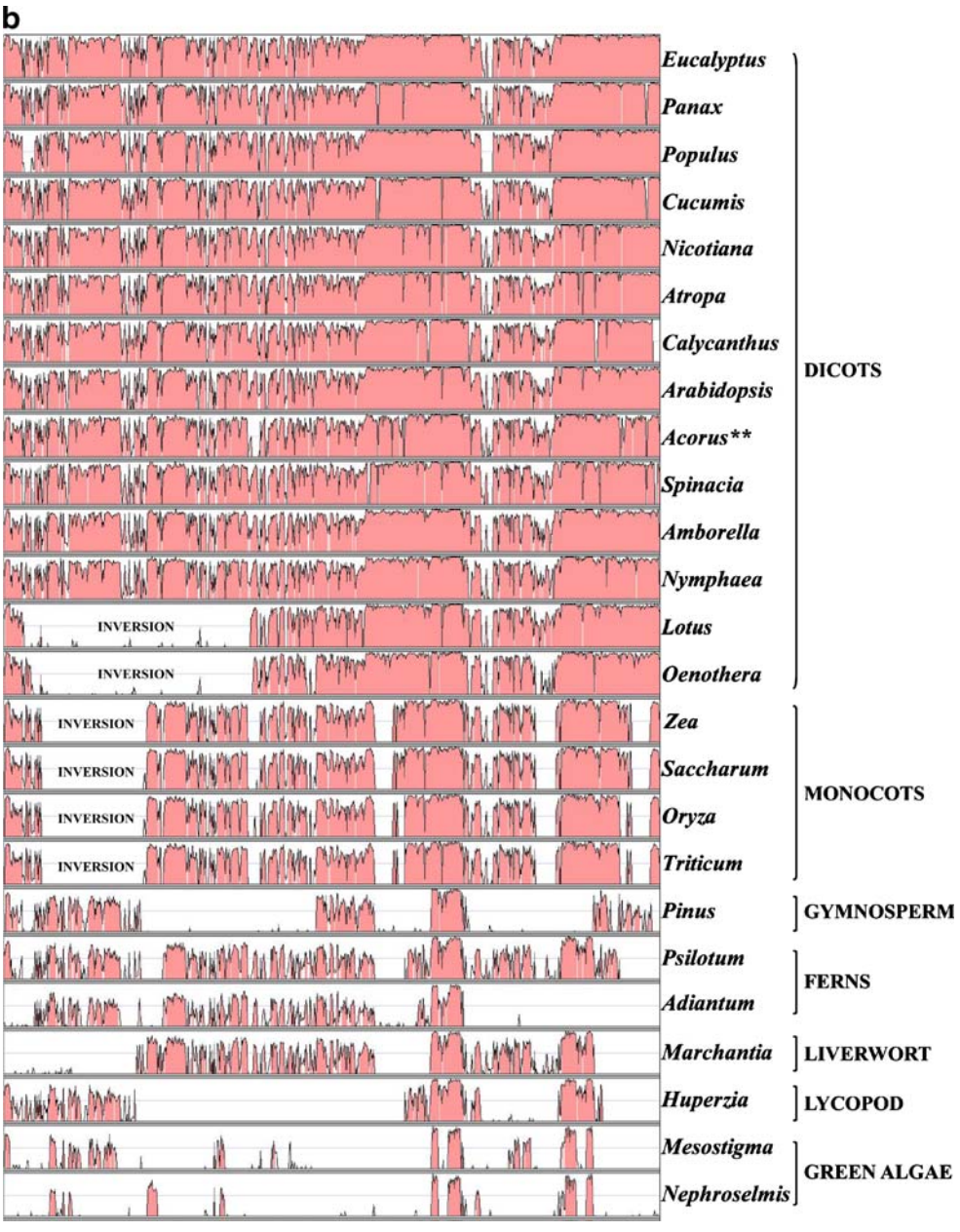
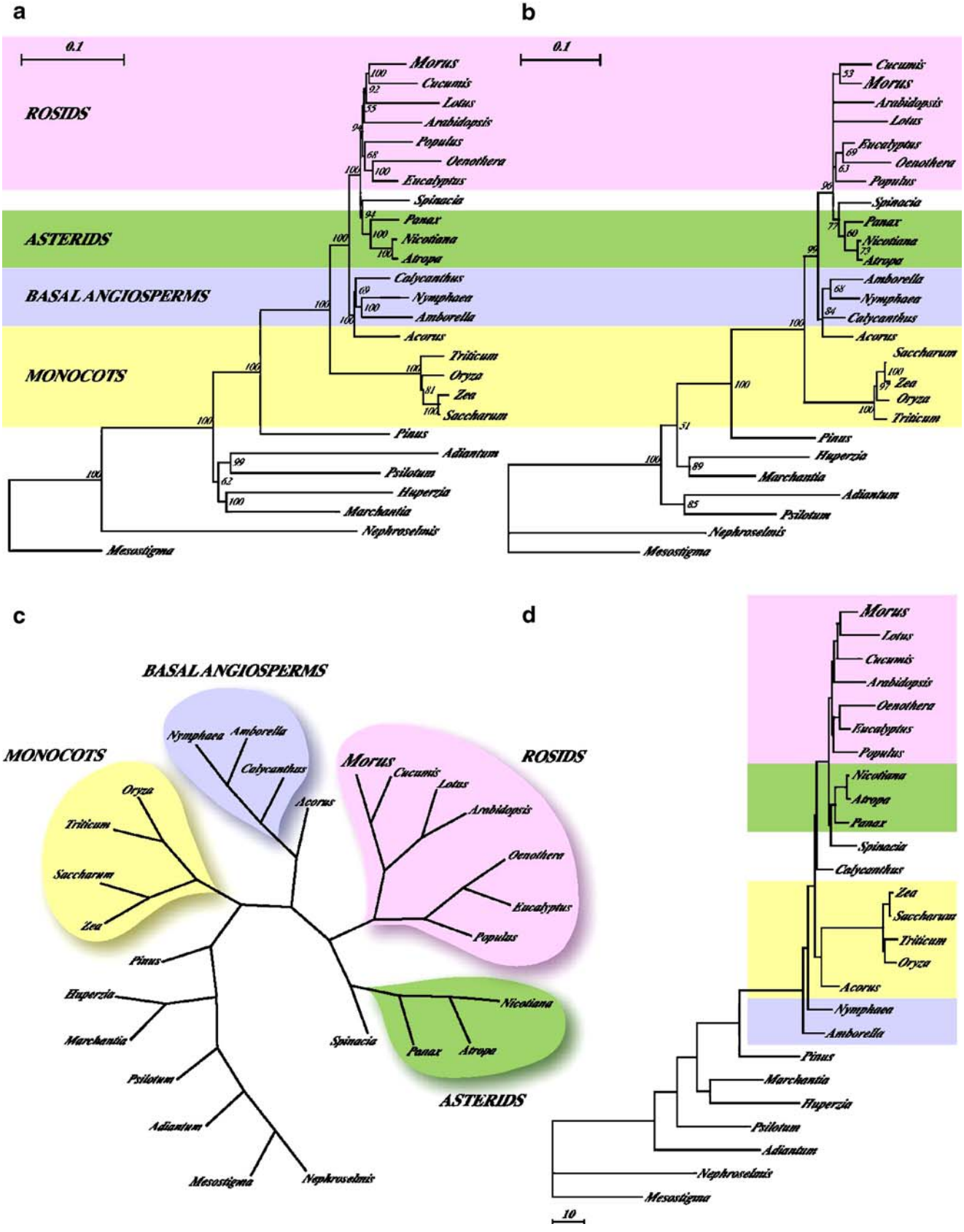


Fig. 3 Plot of concatenated local alignment lengths and identities of several dicot, monocot, and lower plant groups compared to mulberry plastome (**a**). A cut-off of 90% identity was used. VISTA plot (global alignment) comparison of mulberry chloroplast genome with 25 chloroplast genomes (**b**). Y-scale represents the percent identity ranging from 50 to 100%. Genomes are arranged according to the number of conserved bases with respect to mulberry. *Double asterisk*: *Acorus* is a monocot but shows more similarity to dicots in terms of global alignment

Fig. 3 (continued)



There was only one tree with *Morus/Lotus* pairing, and the remaining ten trees had *Lotus/Arabidopsis* pairs. In all the distance-based approaches, *Lotus* appeared sister to *Morus/Cucumis*. *Eucalyptus*, which was closest to mulberry in terms of global and local alignments, however, grouped with *Oenothera* in all phylogenetic analyses. This is no surprise as both belong to the same order Myrtales. *Acorus*, the ancient Liliopsid, grouped with the 'basal angiosperms' and paired up with *Calycanthus* in all cases. The monocots appeared as the sister group to all the angiosperms, which in itself is highly controversial (Goremykin et al. 2003a,b–2005; Soltis et al. 2004; Stefanovic et al. 2004; Martin et al. 2005). Monocots did not appear to be monophyletic in all the above cases. However, when protein maximum likelihood program ProML (Felsenstein 1989) employing the JTT model, as implemented in BioEdit (Hall 1999), was used for phylogeny construction, monophyly for the monocots was observed with *Acorus* as the sister genus to the rest of the monocots (Fig. 4d). In the tree obtained from the 9,727-position-long alignment (ungapped), *Nymphaea* appeared as the basal angiosperm, next to *Amborella*, while in the other tree, obtained with the 12,299-position-long alignment (with gaps), *Amborella*-basal topology was obtained. However, *Morus* paired up with *Lotus* in both


a
0.1
ROSIDS
ASTERIDS
BASAL ANGIOSPERMS
MONOCOTS
Morus
Cucumis
Lotus
Arabidopsis
Populus
Oenothera
Eucalyptus
Spinacia
Panax
Nicotiana
Atropa
Calycanthus
Nymphaea
Amborella
Acorus
Triticum
Oryza
Zea
Saccharum
Pinus
Adiantum
Psilotum
Huperzia
Marchantia
Nephroselmis
Mesostigma
b
0.1
Cucumis
Morus
Arabidopsis
Lotus
Eucalyptus
Oenothera
Populus
Spinacia
Panax
Nicotiana
Atropa
Amborella
Nymphaea
Calycanthus
Acorus
Saccharum
Zea
Oryza
Triticum
Pinus
Huperzia
Marchantia
Adiantum
Psilotum
Nephroselmis
Mesostigma
c
BASAL ANGIOSPERMS
MONOCOTS
ROSIDS
ASTERIDS
Nymphaea
Amborella
Calycanthus
Acorus
Oryza
Triticum
Saccharum
Zea
Morus
Cucumis
Lotus
Arabidopsis
Oenothera
Eucalyptus
Populus
Pinus
Huperzia
Marchantia
Psilotum
Adiantum
Mesostigma
Nephroselmis
Nicotiana
Spinacia
Panax
Atropa
d
Morus
Lotus
Cucumis
Arabidopsis
Oenothera
Eucalyptus
Populus
Nicotiana
Atropa
Panax
Spinacia
Calycanthus
Zea
Saccharum
Triticum
Oryza
Acorus
Nymphaea
Amborella
Pinus
Marchantia
Huperzia
Psilotum
Adiantum
Nephroselmis
Mesostigma
10

◀ **Fig. 4** Phylogenetic position of *Morus* as inferred from 43 chloroplast proteins. The same topology was obtained with both Neighbor-joining method (**a**) (Kimura, Tajima-Nei, and Poisson correction distance calculation as implemented in TREECON); Maximum Likelihood method (**b**) (Dayhoff and Jones–Taylor–Thornton models of substitution as implemented in Tree Puzzle 5.1) and PROTPARS (**c**) (Protein parsimony method: PHYLIP package, Joseph Felsenstein). A consensus tree is shown in the figure from the 12,299-position-long alignment. Trees with *Morus/Cucumis* pairs were 89 out of a total of 100 trees. Phylograms are displayed for **a** and **b**, while **c** is shown in the form of a dendrogram with color-coded groups. **d** Tree topology obtained with the 12,299-position-long alignment (gapped) using ProML with the JTT model as implemented in BioEdit. Bootstrap values for support are indicated at branch-points. The scale represents the number of substitutions per site. Following are the genes used in the analysis, with the lengths of alignment used in parentheses: *atpA* (512), *atpB* (502), *atpE* (143), *atpF* (187), *atpH* (82), *atpI* (251), *petB* (234), *petG* (38), *psaA* (754), *psaB* (735), *psaC* (81), *psaI* (52), *psaJ* (57), *psbC* (487), *psbD* (354), *psbE* (83), *psbF* (41), *psbH* (88), *psbI* (54), *psbJ* (42), *psbK* (65), *psbN* (46), *psbZ* (*ycf9*) (62), *rpl2* (279), *rpl14* (123), *rpl16* (143), *rpl20* (142), *rpl36* (38), *rpoA* (533), *rpoB* (1153), *rpoC1* (904), *rpoC2* (2041), *rps2* (317), *rps3* (257), *rps4* (211), *rps7* (157), *rps8* (141), *rps11* (147), *rps12* (138), *rps14* (103), *rps18* (171), *rps19* (95), *ycf4* (256)

these cases, *Cucumis* being sister to *Morus/Lotus*. When ProtML with JTT model (Molphy, Adachi, and Hasegawa, http://bioweb.pasteur.fr/intro-uk.html) was used with the 12,299-position-long alignment, 34 out of a total of 47 (72.3%) trees gave *Morus/Cucumis* pairs (tree not shown).

The remaining trees comprised of *Lotus/Morus* pairs (10) and *Cucumis/Lotus* pairs (3). The 9,727 (ungapped)-position-long alignment gave a much stronger support to the *Morus/Cucumis* pair (91.8% trees; 45 out of a total of 49). Only one tree had a *Morus/Lotus* pair and the remaining three trees had *Lotus/Cucumis* pairs. Thus, *Morus* appears to be closer to *Cucumis* phylogenetically, although *Lotus* cannot be ruled out. If we take both the genome level and phylogenetic comparison, then *Cucumis* appears to be the closest. The dendrogram (Fig. 4c) obtained using PROTPARS clearly shows different groups, namely, the dicots, monocots, basal angiosperms, and lower plants. Even among the dicots, there is a clear distinction between the rosid and asterid groups. Finally, a global alignment-level phylogenetic tree was generated using the MAVID/AMAP (Bray and Pachter 2004) multiple-alignment server (http://baboon.math.berkeley.edu/mavid/) and viewed using the ATV applet. *Morus* paired up with *Cucumis* in this case, too (Supplementary Fig. S2).

Discussion

The complete chloroplast genome of mulberry has been determined using a combination of long PCR and shotgun-based approaches using purified chloroplast DNA as a template. The long PCR approach has been used to determine the entire plastome sequence in other organisms like sugarcane (Asano et al. 2004), *Calycanthus* (Goremykin et al. 2003a), *Amborella* (Goremykin et al. 2003b), *Nymphaea* (Goremykin et al. 2004), and *Acorus* (Goremykin et al. 2005) using total DNA as the template. Our approach uses DNA isolated from a chloroplast-enriched preparation, thus making the amplifications even more reliable and decreasing the chances of nuclear-localized plastid DNA (nuptDNA) getting amplified. The chloroplast genome of mulberry is highly similar in organization, gene content, order, and A+T content to other known land plant plastomes (Shinozaki et al. 1986; Sato et al. 1999; Kato et al. 2000; Schmitz-Linneweber et al. 2001, 2002; Odintsova and Yurina 2003; Steane 2005).

The organization of the *rps12* gene exhibits signatures of trans-splicing. To produce mature *rps12* transcript, the 5′-exon present in the large single-copy region and the 3′-exons present in the inverted repeats as duplicates should splice. This discontinuous arrangement of the first exon with respect to the 3′ exons is typical of plant chloroplasts. Unlike tobacco, the mulberry plastome does not contain a small plastid RNA-encoding gene *sprA*. This gene, believed to be involved in 16S rRNA maturation (Vera and Sugiura 1994), has only been found in plastomes of Solanaceae members. Other genes absent from the mulberry plastome are *infA* and *rpl21*.

The *infA* gene is known to have been lost from almost all known rosid plastomes (Millen et al. 2001) having been transferred to the nucleus. The *rpl21* gene is present only in plastomes of ferns and bryophytes. The mulberry plastome also contains two pseudogenes *ycf15* and *ycf68*. The presence of portions of the *ycf15* gene indicates that it is probably a remnant of a functional gene in one of its predecessors. It is believed that *ycf15* is not a protein-coding gene (Schmitz-Linneweber et al. 2001; Steane 2005). The deletion observed in the *ycf68* gene, which causes the frameshift, does not appear to be a sequencing problem, as the coverage and read quality in the concerned region are high.

Moreover, the reads in this region are of two types—one from a whole chloroplast genome shotgun and the other from a PCR-based approach, indicating that the deletion is part of the plastome and not an artifact. If these highly conserved genes are essential for the organism, then it is possible that they have a counterpart in the nuclear genome and are in different stages of degeneration. If not, the conservation of these sequences might be signatures of regulatory regions. C to U transitions in chloroplast genomes are known, but the same is not true for reverse editing (U to C). This phenomenon has been observed in mitochondria and only in the plastome of *Anthoceros* (Kugita et al. 2003a,b). Thus, it does not seem likely that the stop codons get converted to sense codons.

IR expansion/contraction studies in *Atropa* and *Nicotiana* (Kim and Lee 2004) reveal that there is considerable difference even between members of the same family. However, it would be of great interest to compare IR/SC junctions in different varieties of *Morus* along with other members of Moraceae. Similarity at this level could indicate common ancestry and will be a useful evolutionary tool besides the conventional ones.

Plastome comparison at the global and local level brought out a tree species, *Eucalyptus*, as the closest out of the 25 genera taken for comparison. *Cucumis* and *Panax* were the next in line. However, phylogenetic analysis using amino acid sequences from common protein coding genes brought out *Cucumis* as the closest followed by *Lotus*. This was confirmed by several phylogenetic methods, and similar results were obtained with global-alignment level phylogenetic analysis also. Thus, *Cucumis* seems to be the closest to mulberry if results of both methods are combined, i.e., global/local comparison and phylogeny.

In summary, the complete plastome sequence of an economically important plant—mulberry—has been determined, thus filling in the gap representing the sixth subclass of the Magnoliopsida. Global- and local-alignment-based comparison and phylogenetic positioning of mulberry have been presented. It appears to be closest to *Cucumis* and *Lotus* phylogenetically, but at the genome level, *Eucalyptus* appears to be closest. This may be attributed to the non-coding portions, which were not part of phylogenetic analysis. The difference in the results obtained from the gene-level and genome-level studies are probably due to the differences in the selection pressures operating on the coding and non-coding regions. This is the third angiosperm tree species after *Populus* and *Eucalyptus*, whose plastome sequence has been completely deciphered. Studies based on IR/SC junction regions and other variable regions from different *Morus* species would be of great help in systematics. The information thus generated will be also useful for taxonomic analyses of other species of *Morus*, other genera within Moraceae, and other families within the same subclass.

Acknowledgements This work was financially supported by grants received from the Department of Biotechnology (DBT), Government of India. VR acknowledges CSIR for the award of a research fellowship.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K (2004) Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. DNA Res 11:93–99

Bray N, Dubchak I, Pachter L (2003) AVID: a global alignment program. Genome Res 13:97–102

Bray N, Pachter L (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Res 14:693–699

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biochemical Research Foundation, Washington DC, pp 345–352

Doyle JJ, Davis JI, Soreng RJ, Garvin D, Anderson MJ (1992) Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proc Natl Acad Sci U S A 89:7722–7726

Ewing B, Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

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

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003a) The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis*-structural and phylogenetic analysis. Plant Syst Evol 242:119–135

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003b) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol Biol Evol 20:1499–1505

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol 21:1445–1454

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH (2005) Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol 22:1813–1822

Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet 252:195–206

Grayum MH (1987) A summary of evidence and arguments supporting the removal of *Acorus* from the Araceae. Taxon 36:723–729

Hagopian JC, Reis M, Kitajima JP, Bhattacharya D, de Oliveira MC (2004) Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. *liui* provides insights into the evolution of rhodoplasts and their relationship to other plastids. J Mol Evol 59:464–477

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hupfer H, Swiatek M, Hornung S, Herrmann RG, Maier RM, Chiu WL, 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

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

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

Kim J-S, Jung JD, Lee J-A, Park H-W, Oh K-H, Jeong W-J, Choi D-W, Liu JR, Cho KY (2006) Complete sequence and organization of the cucumber (*Cucumis sativus* L. cv. Baekmibaekdadagi) chloroplast genome. Plant Cell Rep 25:334–340

Kim KJ, Lee HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res 11:247–261

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

Kugita M, Kaneko A, Yamamoto Y, Takeya Y, Matsumoto T, Yoshinaga K (2003a) The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. Nucleic Acids Res 31:716–721

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K (2003b) RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res 31:2417–2423

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Martin W, Deusch O, Stawski N, Grunheit N, Goremykin V (2005) Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci 10:203–209

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16:1046

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH (2001) Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645–658

Odintsova MS, Yurina NP (2003) Plastid genomes of higher plants and algae: structure and function (translated from Russian). Mol Biol 37:649–662

Ohta N, Matsuzaki M, Misumi O, Miyagishima SY, Nozaki H, Tanaka K, Shin-I T, Kohara Y, Kuroiwa T (2003) Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. DNA Res 10:67–77

Palmer JD (1986) Isolation and structural analysis of chloroplast DNA. Methods Enzymol 118:167–186

Pombert JF, Otis C, Lemieux C, Turmel M (2005) The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of Chlorophyte lineages. Mol Biol Evol 22:1903–1918

Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, Jansen RK (2005) Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Mol Biol 59:309–322

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 RM, Alcaraz JP, Cottet A, Herrmann RG, Mache R (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Mol Biol 45:307–315

Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG, Maier RM (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of speciation. Mol Biol Evol 19:1602–1612

Shahid Masood M, Nishikawa T, Fukuoka S, Njenga PK, Tsudzuki T, Kadowaki K (2004) The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. Gene 340:133–139

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 DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD, Soltis PS (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci 9:477–483

Steane DA (2005) Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, *Eucalyptus globulus* (Myrtaceae). DNA Res 12:215–220

Stefanovic S, Rice DW, Palmer JD (2004) Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol Biol 4:35

Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Sugiura C, Kobayashi Y, Aoki S, Sugita C, Sugita M (2003) Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus. Nucleic Acids Res 31:5324–5331

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

Thompson JD, Higgins DG, Gibson TJ (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 (2005) The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biology 3:22

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 Appl 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

Wolf PG, Karol KG, Mandoli DF, Kuehl J, Arumuganathan K, Ellis MW, Mishler BD, Kelch DG, Olmstead RG, Boore JL (2005) The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene 350:117–128

Wolf PG, Rowe CA, Sinclair RB, Hasebe M (2003) Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. DNA Res 10:59–65

Wyman S, Jansen R, Boore J (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255