Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times*

Vadim V. Goremykin, Sabine Hansmann, and William F. Martin

Key words: Seed plants, molecular phylogeny, chloroplast DNA.

Abstract: Fifty-eight homologous protein sequences from the completely sequenced chloroplast genomes of *Zea mays*, *Oryza sativa*, *Nicotiana tabacum*, *Pinus thunbergii*, *Marchantia polymorpha* and *Poryphyra purpurea* were investigated. Analyzed individually, only 40 of the 58 proteins gave the true, known topology for these species. Trees constructed from the concatenated 14295 amino acid alignment and from automatically generated subsets of the data containing successively fewer polymorphisms were used to estimate the *Nicotiana-Zea* and *Pinus*-angiosperm divergence times as 160 and 348 million years, respectively, with an uncertainty of about 10%. These estimates based upon phylogenetic analysis of protein data from complete chloroplast genomes are in much better accordance with current interpretations of fossil evidence than previous molecular estimates.

> "Ihre molekularen Schätzungen zum Alter der Angiospermen wurden in der Vergangenheit kontrovers diskutiert. Meinen Sie nicht, daß diese allmählich revidiert werden müssen?"
>
> Comment by F. Ehrendorfer on a presentation by WM, Bonn 1993.

Fossil plants generally accepted as angiosperms appear in Lower Cretaceous deposits (roughly 140 MY in age). Their rise to dominance in younger strata to bring forth roughly 250000 species found among contemporary flora is a "success story" with few if any parallels on the geological record. But a clear answer to the question of what a "flower" is and from which organ(s) of which gymnosperms it was ultimately derived is as unavailable today as it was when Richard von Wettstein commented:

"Eine in neuerer Zeit viel diskutierte Frage ist die nach der Herkunft der Angiospermenblüte. Unmittelbar läßt sich diese von den Blüten der höheren Gymnospermen nicht ableiten; solange diese Ableitung nicht sichersteht, ist sogar die Kluft, welche die Angiospermen von den Gymnospermen trennt, größer als die Kluft zwischen Pteridophyten und Gymnospermen [...]."   VON WETTSTEIN (1914: 447).[1]

An understanding of the timing of angiosperm evolution can help to discriminate between alternative theories concerning their ancestors among fossil forms. With this in mind, several estimates of the minimum age of angiosperms based upon molecular clock calculations for divergence between select extant species have been published over the years (RAMSHAW & al. 1972; MARTIN & al. 1989, 1993; WOLFE & al. 1989; BRANDL & al. 1992; SAVARD & al. 1994). Estimates for angiosperm age that vastly exceeded the Lower Cretaceous appearance of flowering plants were highly controversial (CLEAL 1989, CRANE & al. 1989, CLEGG 1990) but due to the problems of identifying putative angiosperms or their ancestors preserved in Pre-Cretaceous sediments (CORNET 1986, CRANE 1988), it has been difficult on the basis of available data to dismiss such claims as absurd or to empirically refute them (CRANE 1993, CRANE & al. 1995).

The idea that contemporary angiosperm lineages may have undergone diversification long before their appearance in the fossil record has gained some acceptance in recent years to the extent that molecular estimates for the date of the "monocot-dicot" divergence at roughly 200 MY B.P. have even been used as calibration times for molecular clock estimates of the timing of other events in the history of land plants (SAVARD & al. 1994). Those familiar with the problem know that all previous molecular estimates of angiosperm age are tenuous for one reason or the other, mostly because relatively few genes were used and because a relatively large sampling error is involved. But molecular data is assuming an increasingly prominent role in integrated approaches to plant evolution, and the discrepancy between molecular and fossil evidence for angiosperm age is an important issue.

Motivated by several years of controversy concerning previous molecular estimates of angiosperm age, we have readdressed the problem by identifing and analyzing the set of 58 cpDNA-encoded proteins common to five land plants and one outgroup for which complete chloroplast genome sequences are available. Here we show that complete chloroplast genome sequences are valuable and easily manageable tools for plant phylogenetics. Based upon this relatively large data set (14295 amino acids per species), revised molecular estimates for the age of the last common ancestor of tobacco and the graminaceaous monocots rice and maize are presented, as are estimates of the age of the last common ancestor of black pine and angiosperms.

---

[1] "A question that has recently received considerable discussion concerns the origin of the angiosperm flower. It cannot be directly derived from the flowers of higher gymnosperms; as long as this derivation is not soundly understood, the chasm dividing angiosperms from gymnosperms is even greater than that between pteridophytes and gymnosperms [...]." Translation by the authors.

Material and methods

Sequence retrieval and alignment. Complete sequences of maize (MAIER & al. 1995; X86563), rice (HIRATSUKA & al. 1989; X15901), tobacco (SHINOZAKI & al. 1986; S54304), black pine *(Pinus thunbergii* WAKASUGI & al. 1994; D17510), liverwort *(Marchantia polymorpha,* OHYAMA & al. 1986; X04465) and *Porphyra purpurea* (REITH & MUNHOLLAND 1995; U38804) chloroplast genomes were retrieved from GenBank and EMBL data bases. Starting with the maize sequence, TFASTA searches were performed with all encoded proteins and annotated open reading frames. The corresponding homologues from the other genomes were retrieved. For those cases in which homologues were not found by this search, TFASTA searches were performed against the six-frame translations of the complete cpDNA nucleotide sequences as a control. Homologues were prealigned in individual files with the PHYLIP program of the GCG package (GENETICS COMPUTER GROUP 1994) then realigned and written into PHYLIP format with CLUSTALW (THOMPSON & al. 1994). For individual analyses, these were analyzed directly with the corresponding Dayhoff matrix distance and neighbor-joining options of the appropriate PHYLIP programs. For concatenate analyses, the individual interleaved PHYLIP input files were concatenated and edited by hand to produce the final single-file alignment for six OTUs of 14295 amino acids each.

Sorting of amino acid positions according to variability. To sort positions of the alignment into descending order of positional variability, the program SORTAL was written in C. In brief, the number of amino acid states is counted at each position. Each position is assigned a priority value according to its variability, invariant positions receiving the highest priority, the most polymorphic positions receiving the lowest. Simultaneously with the reading and counting process, positions are sorted into a new file according to their priority using a standard queue data structure. After sorting, the positions are rewritten into interleaved PHYLIP infile format. In this six species case, invariant (6-0 state) positions appear first in the sorted alignment, followed in order by autapomorphies (5-1), informative sites (4-2, 3-3), three-state (4-1-1, 3-2-1, 2-2-2), four-state (3-1-1-1, 2-2-1-1), five-state (2-1-1-1-1) and six-state sites (1-1-1-1-1-1). Gaps are counted as being a different amino acid from every other amino acid at the position, thus a position containing four identical amino acids and gaps in the other two OTUs is sorted into the three-state (4-1-1) class.

Since in all existing methods for distance estimation between amino acid sequences each site is assumed to evolve independently, unsorted and sorted versions of the same alignment yield identical distance estimates and topologies. An additional program was written in C that iteratively deletes the most variable positions of the sorted alignment by a value specified by the user and produces the desired PHYLIP distance estimate and a neighbor-joining (SAITOU & NEI 1987) tree for each step of the variability removal process. Distance estimates, topologies, and NJ branch lengths are recorded for further analysis. Executables of these programs for UNIX and OSF1 operating systems and the concatenated alignment are available on request. All sequence comparisons were performed on a DEC 3000. Other statistical analyses and plot graphics were performed with the KALEIDAGRAPH program (Abelbeck Software) for Apple computers.

Results

Individual analyses. Fifty-eight homologuous protein sequences were found to be present in each of the six cpDNA genomes investigated (Table 1), these were extracted and aligned. Numbers of amino acid substitutions per site between proteins were estimated using the Dayhoff matrix. Since the protein sequences

were obtained by direct translation from the DNA sequence, they will not always be identical to the in vivo translation products because of RNA-editing in chloroplast transcripts. However, MAIER & al. (1995) recently estimated that there are only about 25 edited sites in the complete maize chloroplast genome. So although the degree of chloroplast transcript editing is not known for all OTUs considered here, editing will only have an effect on our analyses in those cases where edited and non-edited sites are directly compared. And since the total fraction of edited sites is probably very low in our data set, on the order of 25/14295 or 0.0017 of the amino acid positions, we have neglected the effects of editing here. Values of numbers of substitution per site thus indicate the sum of both amino acid substitutions per site plus the very small fraction of apparent substitutions per site resulting from differentially edited transcripts in pairwise comparisons.

Neighbor joining trees were constructed for each coding region individually. For each of the 58 proteins, the Kimura and Dayhoff distances gave identical topologies (data not shown). Only 40 proteins gave the true topology (Fig. 1a), 18 did not (Table 1). For six proteins giving incorrect NJ-topologies, the correct topology was found among the shortest parsimony trees; these proteins are psbF (1 out of 27 trees), atpH (1/6), psbC (1/1), psbI (1/4), atpB (1/1), and psbK (1/1). For the remaining 12 proteins giving incorrect NJ topologies, parsimony (PROTPARS) analysis for the alignment also did not find the correct tree among shortest trees. These results suggest that the incorrect topologies in such cases is a property of the individual data sets (insufficient phylogenetic information), rather than the method of phylogenetic inference.
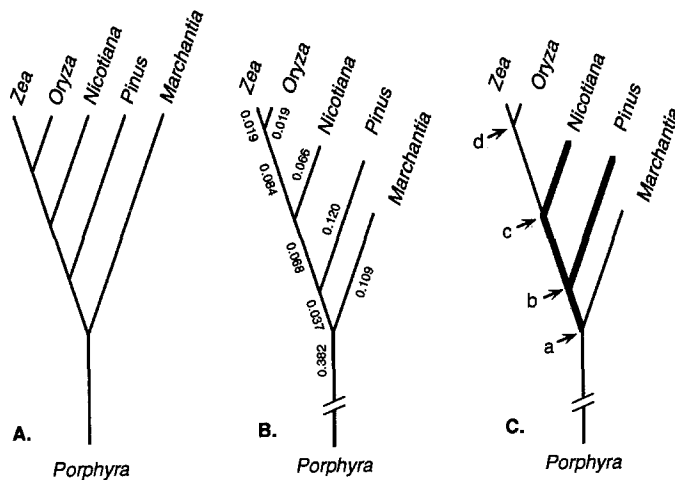

Fig. 1. Rooted 5-species trees for the taxa considered in this paper. A The true biological tree. B The NJ-tree of Dayhoff distances on the basis of the complete concatenated 14295 amino acid data set. Branch lengths are indicated. The Porphyra branch is not drawn to full length for convenience. C The strategy for estimating Nicotiana–Gramineae and angiosperm–Pinus divergence. Nodes a, b, c, and d denote branches referred to in the text. The portion of the tree shown in bold lines indicates branches used to estimate divergence times

To provide a measure of relative substitution rates across the 58 genes, we compared the total length of all branches in land plant trees (regardless of topology) scaled in terms of substitutions per site (Table 1). The land plant tree length was used rather than the total tree length because some cpDNA-encoded proteins entail paralogy across the rhodophyte-chlorophyte boundry, the most notable example being *rbc*L (DELWICHE & PALMER 1996); the *a*-por branch in trees for such paralogs contributes disproportionately to total tree length and should be disregarded. By this criterium, the most slowly evolving of the proteins analyzed was *psb*F, which has identical sequences in maize, rice and tobacco genomes. The fastest was *rpl*32, which shows roughly 10% divergence in the maize-rice comparison. Notably, one of the more slowly evolving proteins and a widely used gene for plant phylogeny, *rbc*L, did not yield the correct six species topology (Table 1) (for a discussion of the limitations of *rbc*L, see MANHART 1994 and GOREMYKIN & al. 1996).

Table 1. Protein sequences used in this study. Gene names correspond with those given in MAIER & al. (1995). [1] Land plant tree length refers to the length in substitutions per site of all branches except that bearing *Porphyra* in the six-species NJ-trees of Dayhoff matrix distances. [2] Yes and no indicate whether the true branching order for the species investigated (*Porphyra*(*Marchantia*(*Pinus*(*Nicotiana*(*Oryza-Zea*))))) was obtained in the NJ-tree for the respective coding region. L is the number of positions in the six species alignment. [3] Total tree length for these genes is more than four times greater than land plant tree length, indicating a disproportionately long *Porphyra* branch, suggesting possible paralogy between *Porphyra* and land plants surveyed (such paralogy is known to exist for *rbc*L, see text)

Gene name	Land plant tree length[1]	True tree[2]	L
psbF[3]	0.0359	no	44
atpH[3]	0.0432	no	82
petB	0.0919	no	215
psbD	0.0935	yes	353
petD	0.1121	yes	180
psbC	0.1209	no	487
psbL	0.1485	yes	38
psbA	0.1490	yes	360
psbE	0.1506	yes	84
psaB	0.1511	yes	735
psbI	0.1538	no	54
psaA	0.1548	yes	754
psaC	0.1747	no	81
psbB	0.1847	yes	509
rbcL[3]	0.1889	no	490
orf29	0.2092	no	29
atpB	0.2360	no	498
rps12	0.2442	yes	124
psbN	0.2447	yes	43
psbT	0.2447	yes	35
psbJ	0.2584	no	40

Table 1 (continued)

Gene name	Land plant tree length[1]	True tree[2]	L
petG	0.2631	yes	37
atpA	0.2918	yes	507
atpI	0.3192	yes	250
petA	0.3718	yes	321
rps7	0.4283	yes	157
orf62	0.4449	no	62
rpl36	0.4487	yes	37
rpl16	0.4519	yes	143
rpl14	0.4564	yes	124
rps14	0.5571	yes	103
rpl2	0.6064	yes	295
psbH	0.6299	yes	79
rps19	0.6359	yes	94
rps11	0.6531	yes	146
rpoB	0.6833	yes	1170
rps2	0.6863	yes	245
psaJ	0.6962	no	45
rps4	0.7857	no	202
rpoC1	0.8243	yes	716
rps8	0.8347	yes	137
orf185	0.8461	yes	185
rps18	0.8466	no	171
orf31	0.8694	no	31
atpE	0.8719	yes	139
rpl33	0.8920	yes	66
rps3	0.9317	yes	264
psbK	0.9356	no	62
psaI	0.9576	yes	52
rpl23	0.9656	yes	112
atpF	0.9750	yes	188
rpl20	1.0010	no	129
rpoA	1.1338	yes	343
rpoC2	1.2204	yes	1626
orf320	1.2275	yes	325
rpl22	1.3279	no	184
cemA	1.3504	yes	232
rpl32	1.5189	yes	81
		Total	14295

The concatenated data set. The 58 alignments were concatenated, providing an alignment of 14295 amino acid positions per species. A neighbor-joining tree for the Dayhoff distances between the concatenated sequences is shown in Fig. 1. From the resulting branch lengths it was evident that *Marchantia* has a lower substitution rate than the other four land plants surveyed, and that the

graminaceous monocots have a higher substitution rate. These lineage-specific rate deviations, i.e. the slow rate in *Marchantia* (WOLFE & al. 1987) and the increased rate in graminaceaous monocots (BOUSQUET & al. 1992) were previously noted in analyses of other data. But the length of the branches connecting node *b* with *Pinus* (*b*-pin, 0.1195) and *Nicotiana* (*b*-tob, 0.1342), respectively, differ only slightly, the former being only 11% shorter than the latter. This suggests that the average substitution rates along the branches *b*-tob and *b*-pin are quite similar. This conclusion is supported by the average branch length ratios found for the 40 proteins that yield correct topologies. For those 40 proteins, the average ratio of branch lengths *a*-pin/*a*-tob was 1.059 and the corresponding value for *b*-tob/*b*-pin was 1.081.
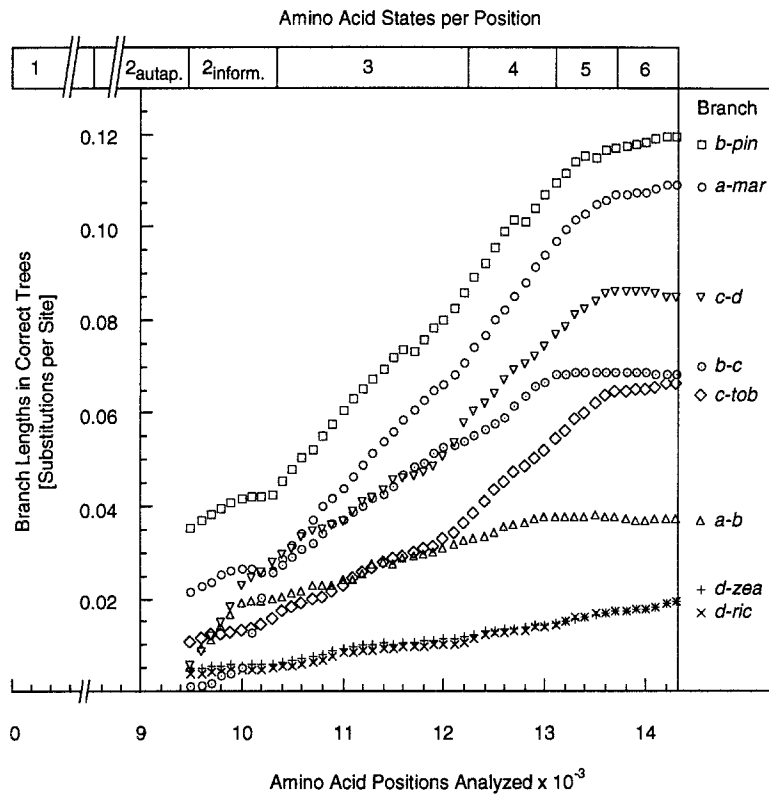

Fig. 2. Relationship between branch length and amount/type of amino acid positions analyzed from cpDNA-encoded proteins. The concatenated 58-protein alignment was sorted according to positional variability. Dayhoff distances and an NJ-tree were constructed from the complete data set (rightmost points), the 100 distal positions were deleted and a new NJ-tree costructed in successive iterations until the true topology was no longer obtained (9500 positions considered, leftmost points). Upper and lower abcissas give the ranges of different classes of amino acid positions analyzed and the total number of positions considered, respectively. Points in the graph represent the NJ-branch lengths in substitutions per site for the respective data sets. Branches are as designated in Fig. 1. The length of the *a-Porphyra* branch is not shown. ric *Oryza*; zea *Zea*; tob *Nicotiana*; pin *Pinus*; mar *Marchantia*

Sorting of positions by variability. Highly variable positions or regions of uncertain alignment can affect phylogenetic results. Although this is probably not a serious problem for the majority of the cpDNA-encoded proteins in comparisons of land plants, it was of interest to investigate the behaviour of trees as a function of the amount of positional variability in the alignment used for inference. A program was designed that automatically sorts the alignment into classes of increasing positional variability. Positions in the concatenated 14295 amino acid data set were sorted according to variability. The sorted alignment consists of 6848 invariant, 2552 autapomorphic 2-state, 969 informative 2-state, 1835 3-state, 941 4-state, 557 5-state and 593 6-state positions. The most variable positions of the complete alignment were removed in steps of 100 positions each and the NJ-tree of Dayhoff distances was examined for each increasingly conservative data set.

The correct NJ-topology was obtained for every data set that contained a fraction of the informative sites, all trees constructed from subsets consisting of only autapomorphic and invariant sites gave incorrect topologies. Branch lengths of resulting NJ-trees were plotted for each data subset which gave the correct topology. As expected, branch lengths tend to increase in a regular, more or less linear manner with increasing overall positional variability contained in the respective data set analyzed (Fig. 2). For the most variable positions (5 and 6 states), only the very short branches connecting *Zea* and *Oryza* tend to increase. This is because the most variable positions preferentially possess differences between these closely related species. Most of the total length of the *c*-tob branch in the tree for the complete data set derives from the subset of 4- and 5-state positions. Of the 2552 autapomorphies found in the 14295 positions analyzed, *Oryza* possess 29, *Zea* 40, *Nicotiana* 87, *Pinus* 294, *Marchantia* 173 and *Porphyra* 1929.

Discussion

We analyzed the deduced amino acid sequences of 58 proteins encoded in the completely sequenced chloroplast genomes of *Zea mays, Oryza sativa, Nicotiana tabaccum, Pinus thunbergii, Marchantia polymorpha* and *Porphyra purpurea*. The true phylogeny for these six species is known (Fig. 1a). We wished to use the data and known topology to estimate the divergence times for *Nicotiana* and the two graminaceous and for the divergence of the conifer and angiosperm lineages. Previous studies had revealed lineage-specific rate deviations for cpDNA-encoded genes. Among the taxa surveyed here, *Marchantia* cpDNA was previously shown to have a lower substitution rate than the other four land plants surveyed here (Wolfe & al. 1987), and the graminaceous monocots were shown to have a higher substitution rate (Bousquet & al. 1992). The overall rate of substitution as estimated by branch lengths in NJ-trees of Dayhoff distances was found to differ by only about 10% for the *Nicotiana* and *Pinus* lineages. Thus, if the geological age of the event corresponding to node *a* in Fig. 1c is known (land plant diversification), one can estimate the approximate geological age of the event corresponding to node *b* (angiosperm-conifer divergence) by simply calculating the ratio of branch length *ab* (d_{ab}) to the average length of the *a*-pin and *a*-tob branches ($d_{ab}/0.5(d_{a\text{-pin}} + d_{a\text{-tob}})$). The same applies to branch *ac* for estimating divergence

time for *Nicotiana* and *Zea-Oryza*. This approach is independent of the rate fluctuations on the grass and *Marchantia* branches.

The age of the calibration point (node *a*) corresponding to the divergence time for the lineage leading to the liverwort *Marchantia* from that leading to higher plants is only approximately known. Fossil plants with liverwort-like features were recently reported lower Devonian (~ 400 MY in age) deposits, the associated spores of these plants share similarities with specimens of Ordivician age (EDWARDS & al. 1995). Dessication resistant spores were recently reported from Lower Silurian (~440 MY) deposits (TAYLOR 1995). The origins of vascular plants can be traced into the Middle Silurian (~ 420 MY) (STEWART & ROTHWELL 1993, SHEAR 1991, CAI & al. 1996), providing the latest point in time at which the bryophyte and spermatophyte lineages must have been distinct. From these considerations, we chose a calibration point of 450 MY for the *Marchantia*-vascular plant divergence. This value may be somewhat too ancient or too recent, but it is unlikely to be wrong in either direction by more than 10%.

Using that calibration point, branch lengths in the NJ tree for the complete 14295 amino acid data set suggest that the divergence time for conifers and angiosperms is 348 MY, that for *Nicotiana* and the *Zea-Oryza* lineage is 160 MY. These estimates reflect relative branch lengths in the NJ-tree and can vary depending on the amount of data upon which the tree is based. Therefore, the same calculations were performed for each of the data subsets obtained through sequential deletion of the most variable positions. The results of those calculations are shown in Fig. 3. As seen in the figure, both divergence time estimates tend to increase with increasing variability considered in the data (i.e. the ratios of the corresponding branch lengths change), whereby the rate of increase is greater for the estimate of *Nicotiana*-grass divergence than for conifer-angiosperm divergence. The increase observed when 5- and 6-state positions are considered is not due to
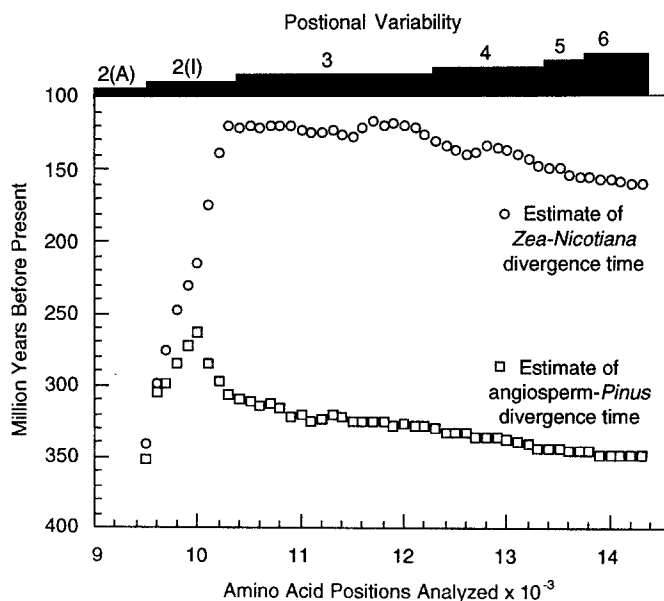

Fig. 3. Relationship between estimates of divergence time and amount/type of amino acid positions analyzed from cpDNA-encoded proteins. Divergence time was estimated as described in the text from each tree represented as a set of vertical points in Fig. 2.

reduction in length of the *ab* and *cd* branches, but rather to increase in length of the *a*-tob and to a lesser extent *a*-pin branches for this fraction of the data (Fig. 2).

Since the NJ-method derives least-square estimates of branch lengths (Saitou & Nei 1987), the standard errors of branch lengths in Fig. 1b can be estimated by Li's method (Li 1989) using the branch lengths shown as d_{ij}. These standard errors will be somewhat too conservative (too great), because the Dayhoff estimates of numbers of substitutions per site are slightly greater than estimates obtained with Li's method. By this approach, the lengths of branches *ab* and *ac* are 0.03693 ± 0.0016 and 0.1050 ± 0.0028, respectively. These (still too conservative) standard errors are very small due to the large number of sites compared, and suggest an approximate branch length uncertainty from the data of about 4% and 3% respectively. Since this is much less than the uncertainty of the calibration time used (450 MY±10%), we attached an uncertainty of 10% to estimates of divergence time.

Previous molecular estimates of angiosperm age have suggested that the diversification of the group long preceded their widespread appearance in Lower Cretaceous flora (Fig. 4a), but estimates based upon relatively small numbers of genes are potentially subject to a large sampling error. For example, the previous estimates based upon only one nuclear gene (Martin & al. 1989), one nuclear plus one chloroplast gene (Martin & al. 1993, Savard & al. 1994) a few chloroplast genes plus one nuclear gene (Wolfe & al. 1989) or five tRNA sequences (Brandl & al. 1992) should be considered as provocative, but they are in need of revision. The present estimate of 160±16 MY for *Nicotiana – Gramineae* divergence is based upon the study of amino acid sequences from 58 genes of chloroplast DNA and provides the most recent molecular estimate for this divergence time obtained to date. It suggests that the separation of these lineages, neither of which likely represent early-branching lineages within the flowering plants (Endress 1994, Crane & al. 1995), preceded the Lower Cretaceous appearance of a continuous record of fossil angiosperm diversification by less than 30 MY, if at all.

The estimate we obtained for the divergence time of the angiosperm and conifer lineages is 350±35 MY (Fig. 4a). This molecular estimate is in very good agreement with the relatively well-known age of the progymnosperms (Beck 1960a, b, 1970), a group of plants critical to understanding of seed plant evolution that existed for only a brief window of time during the Upper Devonian and Lower Carboniferous. Adopting the widely accepted view that the last common ancestor of *Pinus* and angiosperms was in fact a member of the progymnosperms (Ehrendorfer 1976, 1991: 712–731; Stewart & Rothwell 1993: 366–437; Taylor & Taylor 1993: 721–735), the molecular estimate for this divergence time would be in very good agreement with many current interpretations of fossil evidence.

Although based on a large amount of data and to a large extent congruent with fossil evidence, it is still difficult to tell how reliable these estimates might be. Both are topology-dependent, since the positions of terminal branches affect the length of the internal branches from which the estimates derive, and these positions may change with the addition of further taxa for the 58 protein data set. Also, the estimate for the age of contemporary angiosperm lineages is not to be confused with the complex question of the nature (and hence age) of possible angiosperm ancestors preserved in the fossil record that might have attained a similar level of

A

Figure A: Comparison chart with "Epoch", "Molecular Estimates of Divergence Time", and "Fossil Evidence" columns.

Epoch / time scale (left axis, Myr):
- 0
- Tertiary
- 66
- Cretaceous
- 144
- Jurassic
- 208
- Triassic
- 245
- Permian
- 282
- Carboniferous
- 360
- Devonian
- 408
- Silurian
- 438
- Ordovician
- 505
- Cambrian
- 570

Legend:
- Gramineae-Nicotiana
- Angiosperm-Conifer

Column labels: Angiosperms, Progymnosperms

Genes Compared	1	9	5	2	58
Calibration Time for Liverwort-Seed Plant Divergence Used	b)	400	400	450	450
Reference	1	2	3	4	this paper

B

Figure B: Rooted topology diagram with taxa: Coniferales, Gnetales, Cycadales, Ginkgoales, Angiosperms. A "?" marks an uncertain node.

Fig. 4. Summary of results. *A* Comparison of molecular estimates of divergence time and fossil evidence. [a]Strikingly angiosperm-like plants have been found in Triassic deposits, but there is currently no consensus as to whether or not they represent "true" angiosperms (see text). [b]Calibration times used in that study were from the vertebrate fossil record. References: 1 MARTIN & al. (1989), 2 WOLFE & al. (1989), 3 BRANDL & al. (1992), 4 SAVARD & al. (1994). Note that several calibration points were used by SAVARD & al. (1994) one of them being the 200 MY estimate for *Nicotiana-Zea* divergence calculated by WOLFE & al. (1989). *B* Rooted topology for extant seed plant groups recently obtained by GOREMYKIN & al. (1996) and CHAW & al. (1997) (see text). Dotted lines indicate that the topology suggested by those studies is not yet confirmed through analysis of large data sets

organization. Notable exceptions of Pre-Cretaceous fossils such as the monocot-like plant *Sanmiguelia* (CORNET 1986) and the dicot-like plant *Pannaulika* (CORNET 1993, FRASER & al. 1996) provide convincing evidence for the existence of strikingly angiosperm-like characters in Triassic flora, but the phyletic affinities

and angiospermous nature of such plants are still debated (CRANE 1988, 1993; CRANE & al. 1995; STEWART & ROTHWELL 1993: 381, 441–443).

The evolutionary processes through which extant seed plants are related are still unknown. Getting the simple, rooted, five taxon phylogeny for angiosperms, cycads, gnetophytes, *Ginkgo* and conifers right with the help of large and robust molecular data sets would be of enormous value. It would also help in understanding of the evolution of fossil gymnosperms and seed fern allies, because a fixed topology for extant representatives could aid in the interpretation of cladistic analyses of characters in extinct groups. In previous studies using that approach, the gnetophytes have received increasing interest as possible extant sisters to the angiosperms (DOYLE & DONOGHUE 1986, NIXON & al. 1994). The discovery of double fertilization(-like) processes in gnetophytes (FRIEDMANN 1990, 1992, 1994) – that however give rise to additional embryos instead of endosperm – have contributed significantly to interest in that group (for a discussion, see CRANE & al. 1995).

But two recent molecular phylogenetic studies that included representatives from all five extant seed plant lineages lineages do not suggest that gnetophytes and angiosperms are sisters. Data from a ∼ 500 bp region of chloroplast DNA (*cpITS*) (GOREMYKIN & al. 1996) and from roughly 1.4 kb of nuclear 18S rDNA (CHAW & al. 1994, 1997) provided the same rooted topology for the five major groups of extant seed plants (shown in Fig. 4b), whereby the rDNA data gave more meaningful bootstrap support for the three internal branches. Surprisingly, both studies provided evidence to suggest that extant gymnosperms represent a common line of descent and that no contemporary gymnosperm group is the sister group to the angiosperms. These new molecular results need to be critically inspected on the basis of large amounts of additional data since the relationships they suggest between extant gymnosperms find, to our knowledge, no precedent in any previous taxonomic treatment. If a possible monophyly of extant gymnosperms as tentatively suggested by recent molecular data (Fig. 4b) is ultimately borne out by additional data, the gnetophytes could be excluded from the list of candidates for angiosperm ancestors. They would then by necessity be sought in the more traditional domain encompassing pteridosperms and relatives (EHRENDORFER 1976). In time, complete sequences for chloroplast genomes will permit us to more boldly draw the dotted lines of evolutionary history that interconnect extant seed plants.

References

BECK, C. B., 1960a: Connection between *Archaeopteris* and *Callixylon*. – Science **131**: 1524–1525.
– 1960b: The identity of *Archaeopteris* and *Callixylon*. – Brittonia **12**: 351–368.
– 1970: The appearance of gymnospermous structure. – Biol. Rev. **45**: 379–400.

BOUSQUET, J., STRAUSS, S. H., DOERSKEN, A. H., PRICE, R. A., 1992: Extensive variation in the evolutionary rate of *rbc*L gene sequences among seed plants. – Proc. Natl. Acad. Sci. USA **89**: 7844–7848.

BRANDL, R., MANN, W., SPRINZL, M., 1992: Estimation of the monocot-dicot age through tRNA sequences from the chloroplast. – Proc. Roy. Soc. London B **249**: 13–17.

CAI, C., OUYANG, S., WANG, Yi., FANG, Z., RONG, J., GENG, L., LI, X., 1996: An Early Silurian vascular plant. – Nature **379**: 592.

CHAW, S.-M., SUNG, H.-M., LONG, H., ZHARKIKH, A., LI, W.-H., 1994: Phylogeny of the major subclasses of angiosperms and date of the monocot-dicot divergence. – Amer. J. Bot. **81**: S146.

– ZHARKIKH, A., SUNG, H.-M., LAU, T.-C., LI, W.-H., 1997: Molecular phylogeny of gymnosperms and seed plant evolution: analysis of 18S rRNA sequences. – Molec. Biol. Evol. **14**: 56–68.

CLEAL, C., 1989: Evolution in hidden forests? – Nature **339**: 16.

CLEGG, M., 1990: Dating the monocot-dicot divergence. – Trends Ecol. Evol. **5**: 1–2.

CORNET, B., 1986: The leaf venation and reproductive structures of a Late Triassic angiosperm, *Sanmiguelia lewisii*. – Evol. Theory **7**: 231–309.

– 1993: Dicot-like leaf and flowers from the Late Triassic tropical Newark supergroup rift zone, USA. – Mod. Geol. **19**: 81–99.

CRANE, P. R., 1988: Review of CORNET, B., The leaf venation and reproductive structures of a Late Triassic angiosperm, *Sanmiguelia lewisii*. – Taxon **36**: 788–789.

– 1993: Time for angiosperms. – Nature **366**: 631–632.

– DONOGHUE, M. J., DOYLE, J. A., FRIIS, E. M., 1989: Critique of MARTIN & al. 'Angiosperm Origins'. – Nature **342**: 131.

– FRIIS, E. M., PEDERSEN, K. R., 1995: The origin and early diversification of angiosperms. – Nature **374**: 27–33.

DELWICHE, C. F., PALMER, J. D., 1996: Rampant horizontal transfer and duplication of Rubisco genes in eubacteria and plastids. – Molec. Biol. Evol. **13**: 873–882.

DOYLE, J. A., DONOGHUE, M. J., 1986: Seed plant phylogeny and the origin of the angiosperms: an experimental cladistic approach. – Bot. Rev. **52**: 321–431.

EEDWARDS, D., DUCKETT, J. G., RICHARDSON, J. B., 1995: Hepatic characters in the earliest land plants. – Nature **374**: 635–636.

ERENDORFER, F., 1976: Evolutionary significance of chromosomal differentiation patterns in gymnosperms and primitive angiosperms. – In BECK, C. B., (Ed.): Origin and early evolution of angiosperms, pp. 220–240. – New York: Columbia University Press.

– 1991: Evolution und Systematik. – In SITTE, P., ZIEGLER, H., EHRENDORFER, F., BRESINSKY, A., (Eds): Lehrbuch der Botanik für Hochschulen ("Strasburger"), 33rd edn, pp. 712–731. – Stuttgart: G. Fischer.

ENDRESS, P., 1994: Floral structure and evolution of primitive angiosperms: recent advances. – Pl. Syst. Evol. **192**: 79–97.

FRASER, N. C., GRIMALDI, D. A., OLSEN, P. E., AXSMITH, B., 1996: A Triassic Lagerstätte from eastern North America. – Nature **380**: 615–619.

FRIEDMANN, W., 1990: Double fertilization in *Ephedra*, a non-flowering seed plant: its bearing on the origin of angiosperms. – Science **247**: 951–954.

– 1992: Evidence for a pre-angiosperm origin of endosperm: implications for the evolution of flowering plants. – Science **255**: 336–339.

– 1994: The evolution of embryogeny in seed plants and the developmental origin and early history of endosperm. – Amer. J. Bot. **81**: 1468–1486.

Genetics Computer Group, 1994: Program manual for Version 8. – 575 Science Drive, Madison, Wisconsin, USA, 53711.

Goremykin, V. V., Bobrova, V. K., Pahnke, J., Troitsky, A. V., Antonov, A. S., Martin, W., 1996: Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbcL* data do not support gnetalean affinities of angiosperms. – Molec. Biol. Evol. **13**: 383–396.

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. – Molec. Gen. Genet. **217**: 185–194.

Li, W.-H., 1989: A statistical test of phylogenies estimated from sequence data. – Molec. Biol. Evol. **6**: 424–435.

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

Manhart, J. R., 1994: Phylogenetic analysis of green plant *rbcL* sequences. – Molec. Phylogeny Evol. **3**: 114–127.

Martin, W., Gierl, A., Saedler, H., 1989: Molecular evidence for pre-Cretaceous angiosperm origins. – Nature **339**: 46–48.

– Lydiate, D., Brinkmann, H., Forkmann, G., Saedler, H., Cerff, R., 1993: Molecular phylogenies in angiosperm evolution. – Molec. Biol. Evol. **10**: 140–162.

Nixon, K. C., Crepet, W. L., Stevenson, D., Friis, E. M., 1994: A reevaluation of seed plant phylogeny. – Ann. Missouri Bot. Gard. **81**: 484–533.

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 organisation deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. – Nature **322**: 572–574.

Ramshaw, J. A. M., Richardson, D. L., Meatyard, B. T., Brown, R. H., Richardson, M., Thompson, E. W., Boulter, D., 1972: The time of origin of the flowering plants determined by using amino acid sequence data of cytochroms *c*. – New Phytol. **71**: 773–779.

Reith, M., Munholland, J., 1995: Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. – Pl. Molec. Biol. Reporter **13**: 333–335.

Saitou, N., Nei, M., 1987: The neighbor-joining method: a new method for reconstructing phylogenetic trees. – Molec. Biol. Evol. **4**: 406–425.

Savard, L., Li, P., Strauss, S. H., Chase, M. W., Michaud, M., Bousquet, J., 1994: Chloroplast and nuclear gene sequences indicate Late Pennsylvanian time for the last common ancestor of extant seed plants. – Proc. Natl. Acad. Sci. USA **91**: 5163–5167.

Shear, W. A., 1991: The early development of terrestrial ecosystems. – Nature **351**: 283–289.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, M., Hayashida, M., 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 tobacco chloroplast genome: its gene organization and expression. – EMBO J. **5**: 2043–2049.

Stewart, W. N., Rothwell, G. W., 1993: Paleobotany and the evolution of plants. – Cambridge: Cambridge University Press.

TAYLOR, T. N., TAYLOR, E. L., 1993: The biology and evolution of fossil plants. – Englewood Cliffs: Prentice & Hall.

TAYLOR, W. A., 1995: Spores in earliest land plants. – Nature **373**: 391– 392.

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.

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 black pine *Pinus thunbergii.* – Proc. Natl. Acad. Sci. USA **91**: 9794–9798.

WETTSTEIN, R. VON, 1914: Phylogenie der Pflanzen. – In HERTWIG, R., VON WETTSTEIN, R. VON, (Eds): Abstammungslehre, Systematik, Paläontologie, Biogeographie, pp. 439–452. – Leibzig: Teubner.

WOLFE, K. H., LI, W.-H., SHARP, P., 1987: Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. – Proc. Natl. Acad. Sci. USA **84**: 9054–9058.

– GOUY, M., YANG, Y.W., SHARP, P., LI, W.-H., 1989: Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. – Proc. Natl. Acad. Sci. USA **86**: 6201–6205.

Addresses of the authors: WILLIAM F. MARTIN (correspondence), email: w.martin@tubs.de, VADIM V. GOREMYKIN, and SABINE HANSMANN, Institut für Genetik, TU Braunschweig, Spielmannstrasse 7, D-38023 Braunschweig, Federal Republic of Germany.