FEBS Letters 387 (1996) 94–98 FEBS 17088

Evolution of legumin genes: loss of an ancestral intron at the beginning of angiosperm diversification

K.-P. Häger*, B. Müller, C. Wind, S. Erbach, H. Fischer

University of Bayreuth, Department of Plant Ecology and Systematics, Universitätsstr. 30, D-95440 Bayreuth, Germany

Received 20 March 1996; revised version received 22 April 1996

Abstract The polymerase chain reaction was used to survey gymnosperm legumin genes. Characterization of 46 cloned amplificates, differing in sequence and size (1.2–1.6 kb), revealed the ubiquitous occurrence of legumin genes and their organization in small subfamilies in the 22 species investigated. The 3′ portions of the genes, coding for the legumin β-polypeptides, show a highly conserved intron/exon structure divergent from those of angiosperms: an additional intron (intron IV) uniformly interrupts the region coding for the C-terminal part of the β-polypeptides. Phylogenetic analysis of the respective coding sequences as well as the organization of the *Magnolia* B14 legumin gene also investigated here both indicate that intron IV is ancestral and was lost during early angiosperm evolution. Taking into account the intron/exon structures from all legumin genes known, our results suggest that legumin genes evolved by subsequent loss of introns, providing also further evidence for a common origin of legumins and vicilins.

Key words: Seed storage protein; Legumin; Evolution; Intron/exon structure; Gymnosperm; *Magnolia*

1. Introduction

Legumin genes, coding for the most widely distributed group of seed storage proteins [1], show a fairly uniform sequence organization: three introns, strictly conserved in position, interrupt the α- (introns I and II) and β-polypeptide (intron III) encoding regions. A few legumin genes are divergent in number but not in position of introns: Introns I and II have been found in the HaG3 and in the cruA genes from sunflower and oil seed rape, respectively; introns II and III are present in the B subfamily genes from field bean and in J/K genes from pea [2–4].

The coexistence of two- and three-intron genes has led to the assumption that legumin genes evolved by addition [2], or conversely, by loss of introns [3]. The latter scenario gained support from the recently described exceptional five exon/four intron legumin genes from *Ginkgo biloba* which may be viewed as more ancient states of legumin gene structure [5]: the genes contain introns I–III located at exactly the same conserved positions as known from legumin genes of angiosperms, and an additional intron downstream from intron III interrupts the sequence coding for the β-polypeptide C-terminal region. Although divergent from this pattern, a legumin gene from *Welwitschia* (EMBL Z50780 [6]) does contain an intron matching the *Ginkgo* legumin gene intron IV. This may indicate intron IV as being ancestral, or both introns have been gained independently in these species which are separated by a wide phylogenetic gap.

Substantiated assessment of intron IV as either ancestral or acquired is essential for the conclusive reconstruction of legumin gene evolution during seed plant phylogeny. Moreover, recognition of homology in intron/exon patterns will contribute to the decision whether the structural similarities observed between legumin- and vicilin-type storage proteins result from common origin or from convergent evolution [5,7–13].

We have amplified and characterized legumin encoding sequences from 22 species representing four conifer families as well as a *Magnolia* gene coding for the most ancient angiosperm legumins known [14]. The intron/exon organization of the amplificates is reported, a phylogenetic analysis of the sequences is presented and implications for legumin gene evolution are discussed.

2. Materials and methods

2.1. Plant materials

Leaves from individuals of the species given in Table 1 were collected in the Botanic Gardens of Munich, Bayreuth and Edinburgh. Plant materials were snap frozen in liquid nitrogen and stored at −70°C until use.

2.2. PCR amplification and cloning of legumin gene fragments

DNA from leaves was isolated essentially as described in [15]. Three pairs of primers (given as forward/reverse primers) were alternatively used for PCR-amplification of legumin gene fragments: pair A: 5′RTKGTNTTYCCNGGRTGYCCNGAGACTT3′ and 5′ACHTCCTSHGGCATYGCYTTSARHACHGA3′; pair B: 5′TKGTHTTYCCYGGRTGYCCHGAGACTT3′ and 5′CATYACNACYTCCTSHGGCATYSCCTTCA3′; pair C: 5′ACCGYCGGTGTGGYCGYGTGASKTT3′ and 5′TCCCTTGTGGTCGACTACAAAGAAGG3′ (N = ACGT, H = ACT, R = AG, K = GT, Y = CT, S = GC). Amplifications of gymnosperm legumin genes were performed in 25 μl reaction volumes containing 20 mM Tris-HCl pH 8.4, 50 mM KCl, 1.5–3.0 mM $MgCl_2$, 2 μM of each of the respective primers, 0.2 mM of each dNTP, 0.5–2.0 units Taq polymerase and 50–200 ng of genomic DNA; when primer pair A was used for amplification of Taxodiaceae legumin genes, the concentration of the forward primer was decreased to 1 μM. After template denaturation at 94°C, 30 cycles of amplification were carried out with a cycling regime of 30 s denaturation at 96°C, 40–60 s annealing at 55°C and 2.0–2.5 min extension at 72°C. Finally, an additional polymerization step was performed at 72°C for 3–10 min. *Magnolia* legumin genes were amplified as above, except for the following modifications: 10 mM Tris-HCl pH 8.3, 0.12 μM reverse primer, 0.1 mM of each dNTP, 2.5 units Taq polymerase and 1 μg genomic DNA. An initial 37°C annealing step was included (30 s), and an annealing temperature of 68°C was used during the 30 cycles performed.

PCR products were electrophoretically separated on 0.8% agarose gels and isolated using the Jetsorb kit (Genomed). Isolates were ligated into the pGEM-T vector (Promega) or phosphorylated, blunted and ligated into the *Eco*RV site of pKS^+ (Stratagene). Plasmids were introduced into *E. coli* XL1-blue and recombinants were identified by blue/white screening [16].

*Corresponding author. Fax: (49) (921) 552786.

PII S0014-5793(96)00477-2

2.3. Sequencing of cloned PCR products

Standard techniques were used for plasmid isolation, digestion with restriction enzymes and religation [16]. The nucleotide sequences of recombinants were determined by the dideoxynucleotide method [17] using the T7 Sequencing Kit (Pharmacia) according to the manufacturer's instructions. Double stranded sequencing was performed using T3, universal (Pharmacia) and internal sequencing primers.

2.4. Alignment and tree construction

The derived amino acid sequences of legumin β-polypeptides were aligned manually using the Sequence Editor and Analysis Program, version 3.0.7 (G.J. Olsen, 1992). Based on the corresponding nucleotide sequence alignment, phylogenetic analysis was performed with the program package PHYLIP (version 3.5c, J. Felsenstein, Dept. of Genetics, University of Washington, Seattle). Computing was carried out on an Apple Macintosh IIfx and on a VAX 6000-310.

3. Results and discussion

3.1. Intron IV is ubiquitous in legumin genes of gymnosperms

In order to investigate a broad range of gymnosperms we designed two pairs of PCR primers, based on a sequence alignment comprising legumin genes of *Ginkgo* [5] and legumin cDNAs of Pinaceae (3 species; EMBL accession nos. Z11486, X63192, L07484), Taxaceae, Taxodiaceae and Cupressaceae (one species each; unpublished data, [18,19]). Primer pair A considers all sequences, whereas the less degenerate pair B was derived from the sequences of the latter three taxa only. The 3′ termini of the forward primers match 45 bp downstream of intron I, those of the reverse primers 32 (A) and 39 (B) bp downstream of intron IV as found in the *Ginkgo* legumin genes [5]. These are the 3′ and 5′ most well-conserved sequence positions, in contrast to the very terminal regions which are too variable to allow deduction of primers with a reasonable degree of degeneracy. Of the *Magnolia* legumins the B14 gene was chosen for amplification; the primers cover positions 259–281 and 1413–1435 as given in [14].

Legumin encoding sequences were amplified from genomic DNA of the gymnosperms given in Table 1. As judged from analysis by agarose gel electrophoresis, the amplifications resulted in one to three major fragments each, their size varying between 1.2 and 1.6 kb. Southern hybridization using probes described in [5] or derived from legumin cDNAs [18,19] and terminal sequencing of the cloned PCR fragments established the identity of specific PCR products. From each of the species investigated, one to three (see Table 1) legumin encoding amplificates differing in size and sequence were characterized. The regions coding for VRI, VRII and HVR [20] as well as the intervening sequences account for the heterogeneity in size. Although there are diagnostic differences in sequence indicating the presence of two or three legumin subfamilies in each of the conifer families examined, the sequences share high overall similarities, suggesting a quite conservative mode of legumin gene evolution in gymnosperms.

A representative selection of the nucleotide and deduced amino acid sequences, confined to the corresponding β-polypeptides, is given in the alignments of Figs. 1 and 2, respectively. The polypeptide sequences show between 74.5 and 98.6% identity to each other and there is 55.2–60.1% identity when they are compared with a prototype legumin β-polypeptide of *Pisum* [21]. Comparisons with sequences of several legumin cDNAs, and in particular with those originating from the same species or members of the same conifer families investigated here (see EMBL accession numbers given above and [18,19]), reveal three intervening sequences in the legumin gene fragments characterized. The position of the 5′ most introns relative to the amino acid sequences deduced (not shown) clearly verify them as homologous to intron II usually found in the α-polypeptide encoding regions of legumin genes of angiosperms and of *Ginkgo* as well. The legumin β-polypeptide encoding regions (Fig. 1) show an unusual however uniform intron/exon pattern which precisely matches that hitherto known only from *Ginkgo*: introns homologous to intron III (*Ginkgo* and angiosperms) and intron IV (*Ginkgo*) interrupt the 5′ and 3′ portion of the β-polypeptide coding region, respectively. The alignment in Fig. 2 shows the exact conservation of intron positions, also in relation to the $PX_{14}G$ and FYLAG motifs [10,11] highly conserved in angiosperms and also recognizable in the gymnosperm sequences deduced here. Likewise highly conserved, intron III falls between two codons (phase O intron), while intron IV separates the first two bases of a codon from the remaining base (phase 2 intron) (Fig. 1).

With respect to intron IV there is only little variability in size, generally ranging between 75 and 90 bp; notable extremes are the 148 bp introns in amplificates obtained from *Cunninghamia lanceolata* (not shown). These are relatively small sizes compared to other legumin gene introns ranging

Table 1
Species investigate for legumin gen intron IV

Pinaceae		Taxodiaceae	
Picea abies	+,A,2	*Sequoia sempervirens*	+,A,3
Picea glauca	+,A,2	*Sequoiadendron giganteum*	+,A,2
Pinus ponderosa	+,A,3	*Metasequoia glyptostroboides*	+,A,3
Pinus parviflora	+,A,2	*Taxodium distichum*	+,B,2
Larix occidentalis	+,A,2	*Athrotaxis cupressoides*	+,B,3
Larix kaempferi	+,A,2	*Cryptomeria japonica*	+,B,1
Cedrus deodara	+,A,2	*Cunninghamia lanceolata*	+,B,3
Pseudolarix amabilis	+,A,2	*Glyptostrobus pensilis*	+,B,1
Pseudotsuga menziesii	+,A,2	*Taiwania cryptomerioides*	+,B,2
Cupressaceae		Taxaceae	
Calocedrus decurrens	+,A,1	*Taxus baccata*	+,A,2
Thuja plicata	+,A,2		
Juniperus virginiana	+,A,2	Magnoliaceae	
		Magnolia salicifolia	−,C,1

Presence/absence of legumin gene intron IV is indicated by ±; A–C refer to the PCR primer pairs used (see Section 2), and 1–3 to the number of different amplificates analyzed.

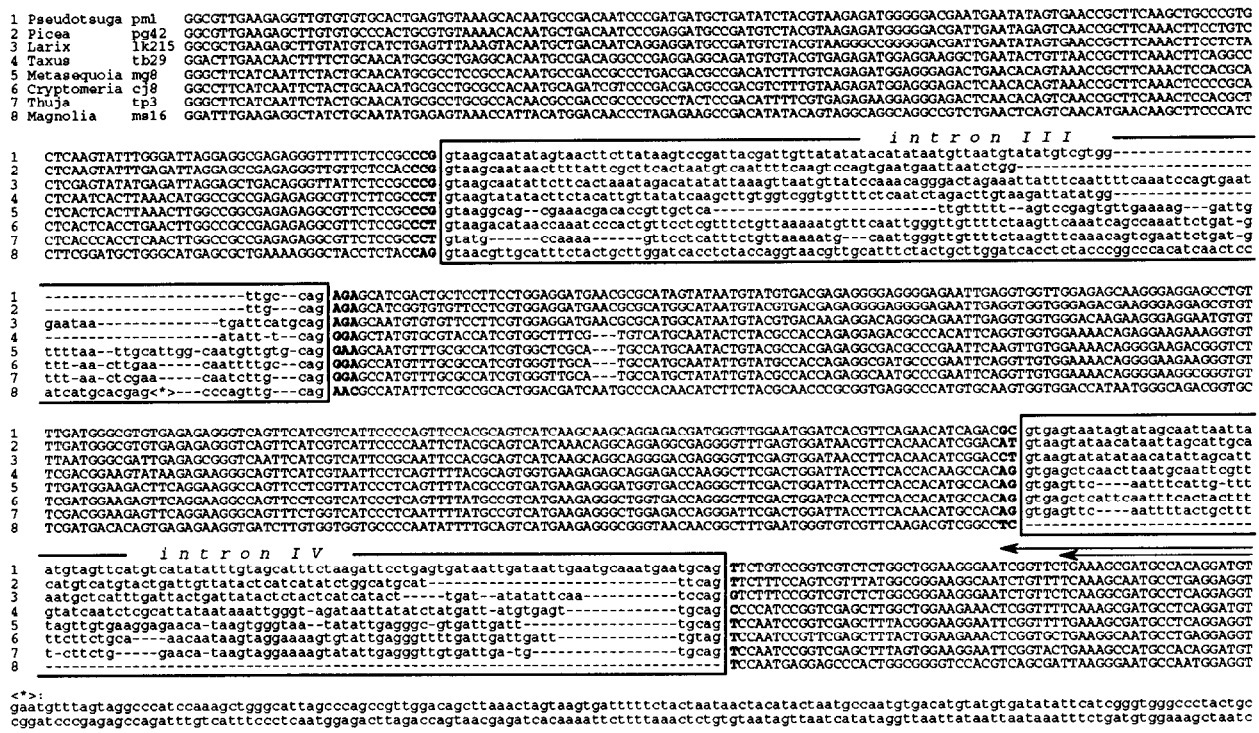

Fig. 1. Multiple alignment of β-polypeptide coding sequences from legumin genes of gymnosperms and *Magnolia*; sequences are derived from cloned PCR amplificates. *Picea* refers to *P. glauca*, *Larix* to *L. kaempferi* (see also Table 1). Introns are boxed, flanking codon(s) printed in bold. Intron sequences were only aligned for *Metasequoia*, *Cryptomeria*, and *Thuja*; gaps (dashes) have been introduced to maximize alignment. Upper and lower horizontal arrows mark positions of A and B reverse primers (see Section 2), respectively.

up to 600 bp in soybean. Like all the introns found, introns IV contain splice sites that obey the GT/AG boundary rule [22], and neglecting a few exceptions (7 out of 46 different amplificates), the very 5′ and 3′ terminal sequences of introns IV are fairly well conserved. They can be described by the consensus sequences **gt**(ag)ag(ct).. and ..t(tcg)**cag**, respectively, and are in good agreement with the plant consensus sequences [23]; the exon sites of the boundary sequences are less conserved. Altogether, introns IV show very limited sequence conservation, but the situation is different when amplificates representing members of one legumin subfamily are compared. Exemplarily, the sequences of amplificates cj8 (from *Cryptomeria japonica*, Taxodiaceae) and tp3 (from *Thuja plicata*, Cupressaceae) are given in Fig. 1, with the respective introns also brought into alignment: The high degree of sequence identity found for the coding sequences (91.0%) here

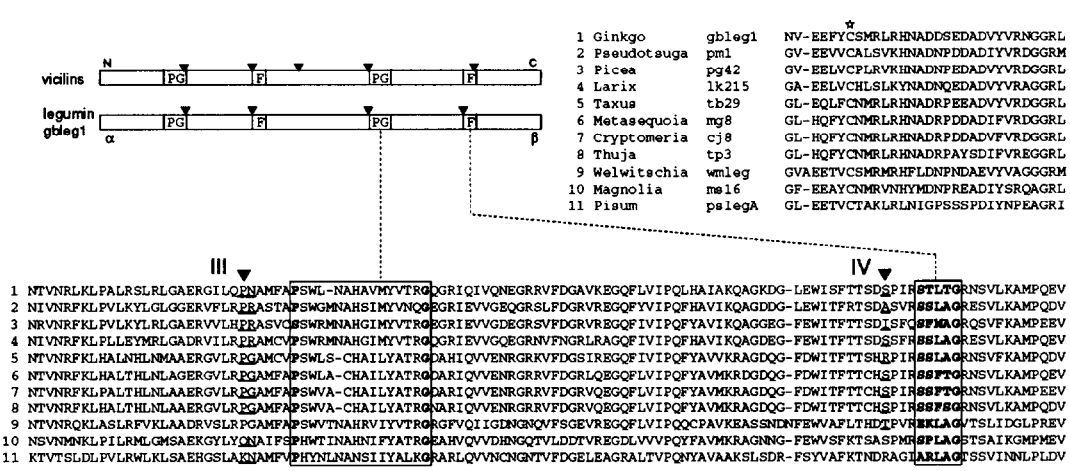

Fig. 2. Legumin β-polypeptides deduced from nucleotide sequences given in Fig. 1, in comparison to those from *Ginkgo* [5], *Welwitschia* [6], and *Pisum* [21]. The asterisk marks cysteines probably involved in α-β-polypeptide linkage and arrowheads indicate positions of introns. Amino acid residues corresponding to codons involved in intron/exon boundaries are underlined. Note the absence of intron III in *Welwitschia* and of intron IV in *Magnolia* and *Pisum*. Boxed regions are homologous to $\mathbf{PX_{14}G}$ (a) and **FYLAG** (b) motifs adjacent to intron positions (if present), likewise conserved as in legumins and vicilins from angiosperms [9,11]. The upper left-hand scheme compares sequence organization of vicilins and *Ginkgo* legumin and assigns homology of introns [5,11]. N and C refer to N- and C-terminal halves of vicilin subunits, and α and β to the respective legumin polypeptides.

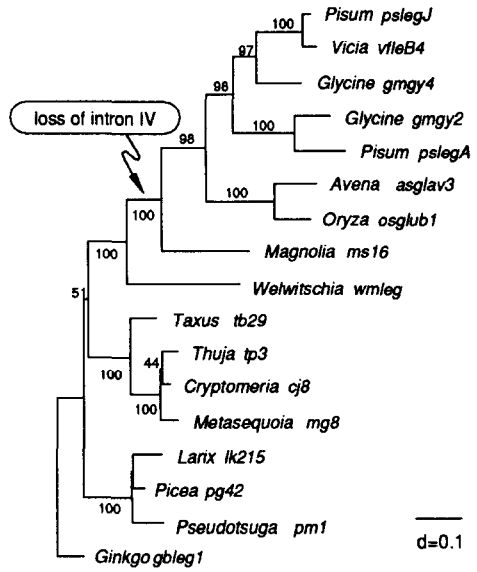

Fig. 3. Reconstruction of legumin gene evolution by the neighbor-joining distance method based on nucleotide sequences coding for β-polypeptides from gymnosperms and angiosperms. For sources of sequences from *Ginkgo*, *Welwitschia*, and *Pisum* pslegA see legend to Fig. 2; conifer and *Magnolia* sequences: this paper; remaining sequences: EMBL accession nos. X54314 (osglub1), X74741 (asglav3), X52863 (gmgy2), X52863 (gmgy4), X03677 (vfleB4), X07014 (pslegJ). The bar denotes 10% dissimilarity; numbers along branches refer to bootstrap values (100 replicates).

extends to the introns as well (introns IV: 90.6%). The apparent conservation of both exon and intron sequences could reflect a relatively recent origin of the species concerned. As this is obviously not the case, we assume that multiple members of the respective legumin gene subfamily are present in each species and are subjected to homogenization [24] leading to a decreased evolutionary rate of individual genes and causing the considerable sequence conservation observed.

3.2. *Legumin gene intron IV is ancestral and was lost during early angiosperm evolution*

Under the conditions described, the amplification of *Magnolia* legumin gene fragments yielded a 2.2 kb product. From four cloned amplificates which were identical in several hundred 3′ and 5′ terminal base pairs, ms16 was analyzed completely. The sequence proved to be highly similar (96.2% identity) to *Magnolia* cDNA B14 [14]. Introns I–III were present, with intron III differing considerably in size from the corresponding gymnosperm legumin gene introns (403 bp vs. <130 bp). All four amplificates lacked intron IV (see also Fig. 1), which is in contrast to the situation found in all gymnosperms but in agreement with the structure of legumin genes from higher angiosperms.

In order to investigate the evolutionary relationships of legumin genes containing and lacking intron IV, respectively, and thus to assess the direction of its evolution, a phylogenetic analysis was performed: legumin β-polypeptides deduced from amplificates representing the taxa investigated here as well as of *Ginkgo*, *Welwitschia*, monocots and Fabaceae were aligned as in Fig. 2. The corresponding nucleotide sequences were used to infer phylogenetic relationships based on distance matrix analysis, using *Ginkgo* legumin GbLeg1 [5] as an outgroup sequence (Fig. 3). The conifer legumins form two distinct clusters, with the Pinaceae legumins (*Pseudotsuga*, *Picea*, *Larix*) emerging at an earlier stage of gymnosperm evolution and a subsequent lineage bearing the common ancestral gene which gave rise to the legumins from *Taxus*, Taxodiaceae (*Metasequoia*, *Cryptomeria*) and Cupressaceae (*Thuja*). The *Welwitschia* legumin appears to be of relatively recent origin and is most closely related to the common ancestor of *Magnolia* legumin ms16 and all remaining angiosperm legumins. Within angiosperms, the monocot legumins separated early from the remaining angiosperm legumins represented here by those of the Fabaceae; the latter are characterized by a gene duplication giving rise to two legumin subfamilies (pslegA/gmgy2 and gmgy4/vfleB4/pslegJ). We assume that the dendrogram represents a reliable description of legumin gene evolution: the main lineages and the topology found suggest species phylogenies which are in good congruence with those deduced from analyses of rbcL genes [25] and with traditional concepts of plant phylogeny as well [26–28].

Therefore, the gene phylogeny presented provides a solid base to assess the direction of evolution of legumin gene intron IV. Assuming a hypothetical ancestral gene devoid of intron IV would force one to postulate multiple independent gains of intron IV within gymnosperms. However, the parallel acquirement of introns strictly conserved in position and phase is highly improbable in our opinion. Minimizing the number of assumptions required we suggest intron IV to have been present in the legumin gene(s) of the last common ancestor of angiosperms and gymnosperms. Obviously, intron IV has then been retained throughout gymnosperm evolution and has been lost prior to or at the very beginning of angiosperm diversification (Fig. 3); further loss of introns has given rise to the two-intron legumin genes found in several angiosperms.

The fact that loss of legumin gene intron IV meets the beginning of angiosperm evolution may be viewed as coincidental. However, there could be lineage specific selective pressure distinguishing between more and less important introns; intron IV might be significant for gymnosperm legumin pre-mRNA stability or might be involved in other mechanisms of legumin gene expression control at the posttranscriptional level, specifically acting in gymnosperms.

3.3. *Intron IV provides evidence for the common origin of legumins and vicilins and is a potential character for tracing early angiosperm evolution*

Recognition of intron IV as ancestral provides important evidence for the common origin of legumins and vicilins. Based on sequence alignments, secondary structure predictions and X-ray structure analysis such a relationship has been previously assumed [7–10]. However, it is difficult to decide whether the similarities observed reflect a common evolutionary history or convergent development of both protein classes [11–13]; convergence may be caused by shared structural and functional constraints due to synthesis, processing, intracellular transfer, self-assembly and packaging, and mobilization. The above ambiguity may be resolved taking into account the intron evolution of the respective genes. Comparison of intron/exon patterns has revealed that three intron positions are precisely conserved between the genes of both protein families, suggesting homology of legumin gene introns I, II and III with vicilin gene introns I, II and IV [11]. The position of the so far peculiar intron IV of *Ginkgo* legu-

min genes closely matches the position of vicilin gene intron V, which has led to the suggestion of homology of these introns as well [5,11] (see Fig. 2). The recognition of legumin gene intron IV as ancestral rather than gained, based on the results presented here, strongly advocates homology with vicilin gene intron V and thus provides substantial evidence for a common ancestor of legumins and vicilins. It is interesting that in both gene classes it was this 3′ most intron which has been lost coincidently with key events of evolution, namely with the very beginning of angiosperm diversification (four-intron vs. three-intron legumin genes, this study) and with the divergence of dicotyledonous and monocotyledonous plants (five-intron vs. four-intron vicilin genes) [29], respectively.

Finally, the phylogenetic events that took place between the origins of the most advanced gymnosperms and the most primitive angiosperms, represented here by *Welwitschia* and *Magnolia* respectively, are the subject of one of the most controversial issues in plant evolutionary research: neither classical systematic research nor phylogenetic analysis of molecular data was able to resolve unambigously the very beginning and first radiations of angiosperm evolution [28]. This situation requires the use of characters related to discrete events of gain or loss. The legumin gene intron IV, if found in some angiosperm legumin genes, would qualify the respective organisms as positively earlier derived than those in which it is not present any more. Thus, legumin gene intron IV could also be a tool to improve our understanding of seed plant evolution.

Acknowledgements: We thank A. Täuber for skilful technical assistance and we are indebted to Dr. H.J. Sluiman (Royal Botanic Garden Edinburgh) for providing *Athrotaxis* leaves. This study was supported by DFG grant Hä 1925/1-2 and by a GRAKO fellowship ('proteins') to C.W.

References

[1] Shewry, P.R. (1995) Biol. Rev. 70, 375–426.
[2] Shotwell, M.A. and Larkins, B.A. (1989) in: The Biochemistry of Plants, vol.15 (Marcus, A. ed.) pp. 297–345, Academic Press, New York.
[3] Vonder Haar, R.A., Allen, R.D., Cohen, E.A., Nessler, C.L. and Thomas, T.L. (1988) Gene 74, 433–443.
[4] Ryan, A.J., Royal, C.L., Hutchinson, J. and Shaw, C.H. (1989) Nucl. Acids Res. 17, 3584.
[5] Häger, K.-P., Braun, H., Czihal, A., Müller, B. and Bäumlein, H. (1995) J. Mol. Evol. 41, 457–466.
[6] Bäumlein, H., Braun, H., Kakhovskaya, I.A. and Shutov, A.D. (1995) J. Mol. Evol. 41, 1070–1075.
[7] Jackson, P., Boulter, D. and Thurman, D.A. (1969) New Phytol. 68, 25–33.
[8] Argos, P., Narayana, S.V.L. and Nielsen, N.C. (1985) EMBO J. 2, 1111–1117.
[9] Gibbs, P.E.M., Strongin, K.B. and McPherson, A. (1989) Mol. Biol. Evol. 6, 614–623.
[10] Lawrence M.C., Izard, T., Beuchat, M., Blagrove, R.J. and Colman, P.M. (1994) J. Mol. Biol. 238, 748–770.
[11] Shutov, A.D., Kakhovskaya I.A., Braun, H., Bäumlein, H. and Müntz, K. (1995) J. Mol. Evol. 41, 1057–1069.
[12] Doolittle, R.F. (1994) Trends Biosci. 19, 15–18.
[13] Casey, R., Domoney, C. and Ellis, N. (1986) in: Oxford Surveys of Plant Molecular and Cell Biology, vol.3 (Miflin, B.J. ed.) pp. 1–95, Oxford University Press, Oxford.
[14] Fischer, H., Haake, V., Horstmann, C. and Jensen U. (1995) Eur. J. Biochem. 229, 645–650.
[15] Doyle, J.J. and Doyle, J.L. (1987) Phytochem. Bull. 19, 11–15.
[16] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd Edn, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[17] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[18] Wind, C. and Häger, K.-P. (1996) FEBS Lett. 383, 46–50.
[19] Häger, K.-P. and Dank, N. (1996) Plant Sci. 116, 85–96.
[20] Nielsen, N.C., Dickinson, C.D., Cho, T.J., Thanh, V.H., Scallon, B.J., Fischer, R.L., Sims, T.L., Drews, G.N. and Goldberg, R.B. (1989) Plant Cell 1, 313–328.
[21] Lycett, G.W., Croy, R.R.D., Shirsat A.H. and Boulter, D. (1984) Nucl. Acids Res. 12, 4493–4506.
[22] Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853–4857.
[23] Slightom, J.L., Sun, S.M. and Hall, T.C. (1983) Proc. Natl. Acad. Sci. USA 80, 1897–1901.
[24] Dover, G. (1982) Nature 299, 111–117.
[25] Chase, M.W. et al. (1993) Ann. Missouri Bot. Gard. 80, 529–563.
[26] Hart, J.A. (1987) J. Arnold Arbor. 68, 269–307.
[27] Crane, P.R. (1985) Ann. Missouri Bot. Gard. 72, 716–793.
[28] Kubitzki, K. (1972) Ber. Dtsch. Bot. Ges. 85, 259–277.
[29] McHenry, L. and Fritz, P.J. (1992) Plant. Mol. Biol. 18, 1173–1176.