J Mol Evol (2001) 53:214–224
DOI: 10.1007/s002390010211

JOURNAL OF MOLECULAR EVOLUTION

Molecular Evolution of Calmodulin-Like Domain Protein Kinases (CDPKs) in Plants and Protists

Xiaorong S. Zhang,[1] Jung H. Choi[2]

[1] Department of Biology and Life Sciences, Savannah State University, Savannah, GA 31404, USA
[2] School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received: 11 July 2000 / Accepted: 18 April 2001

Abstract. Many genes for calmodulin-like domain protein kinases (CDPKs) have been identified in plants and Alveolate protists. To study the molecular evolution of the CDPK gene family, we performed a phylogenetic analysis of CDPK genomic sequences. Analysis of introns supports the phylogenetic analysis; CDPK genes with similar intron/exon structure are grouped together on the phylogenetic tree. Conserved introns support a monophyletic origin for plant CDPKs, CDPK-related kinases, and phosphoenolpyruvate carboxylase kinases. Plant CDPKs divide into two major branches. Plant CDPK genes on one branch share common intron positions with protist CDPK genes. The introns shared between protist and plant CDPKs presumably originated before the divergence of plants from Alveolates. Additionally, the calmodulin-like domains of protist CDPKs have intron positions in common with animal and fungal calmodulin genes. These results, together with the presence of a highly conserved phase zero intron located precisely at the beginning of the calmodulin-like domain, suggest that the ancestral CDPK gene could have originated from the fusion of protein kinase and calmodulin genes facilitated by recombination of ancient introns.

Key words: Calmodulin-like domain protein kinase (CDPK) — Phylogenetic analysis — Introns — Molecular evolution — Calmodulin — Calcium/calmodulin dependent protein kinases (CaMK)

Correspondence to: J.H. Choi; *email:* jung.choi@biology.gatech.edu

Introduction

Calcium-dependent protein kinases play important roles in cellular signaling processes involving calcium as a second messenger. In animal cells, transient increases in intracellular calcium activate calcium/calmodulin-dependent protein kinases (CaMK) (Schulman 1993) or calcium/phospholipid-dependent protein kinases (PKC) (Azzi et al. 1992). In plant cells, however, calcium activates calmodulin-like domain kinases (CDPKs) that do not require calmodulin or phospholipids (Harmon et al. 1987; Roberts 1993; Roberts and Harmon 1992), and thus differ from both the CaMK and PKC families prevalent in mammalian cells.

A CDPK commonly contains three functional domains: a protein kinase catalytic domain, a carboxyl-terminal calmodulin-like domain with four EF-hands as Ca^{++} binding sites, and a junction domain between the kinase and the calmodulin-like domain. Many CDPKs also possess an amino-terminal domain of variable length and sequence, which may play a role in establishing functional diversity within the CDPK family (Stone and Walker 1995). The kinase and junction domains of CDPKs share significant homology with the mammalian CaMK catalytic and regulatory domains, respectively, whereas the C-terminal calmodulin-like domains of CDPKs resemble calmodulin, a ubiquitous and highly conserved calcium-binding protein (Harper et al. 1991; Suen and Choi 1991). The sequence similarity of CDPK to CaMK and calmodulin proteins has led to the speculation that CDPK evolved by fusion of an ancestral CaMK gene with a calmodulin gene (Harper et al. 1991; Suen and Choi 1991).

Many CDPKs have been cloned and characterized from a variety of plant species, including soybean (Harper et al. 1991), carrot (Suen and Choi 1991), *Arabidopsis* (Harper et al. 1993; Urao et al. 1994), rice (Breviario et al. 1995; Kawasaki et al. 1993), corn (Berberich and Kusano 1996; Estruch et al. 1994; Saijo et al. 1997), liverwort (Nishiyama et al. 1999), mungbean (Botella et al. 1996), and black spruce (Perry and Bousquet 1998). Members of this family with no functional EF-hands, called CDPK-related protein kinases (CRKs), have also been found in carrot (Lindzen and Choi 1995), *Arabidopsis*, *Tradescantia*, and maize (Furumoto et al. 1996; Lu et al. 1996). CDPKs have also been isolated and characterized from Alveolate protists, including *Plasmodium falciparum* (Farber et al. 1997; Gardner et al. 1998; Zhao et al. 1993), *Eimeria* (Bumstead et al. 1995; Dunn et al. 1996), and *Paramecium tetraurelia* (Kim et al. 1998). No CDPK sequences have been found in animals or fungi.

As of December 1999, more than seventy different CDPK sequences, including 28 genomic sequences, have been deposited in GenBank. Clearly, CDPKs comprise a large multigene family containing many members in plants and protists, which may have evolved to play diverse cellular and physiological roles in plant/protist growth and development.

We have analyzed the molecular evolution of plant and protist CDPK genomic sequences. Intron positions were tabulated to corroborate the phylogenetic tree. The results indicate a common origin for plant and protist CDPKs, and identify introns that may predate the divergence of plants from protists. Furthermore, introns shared between protist CDPKs and calmodulin genes support an evolutionary relationship between calmodulin genes and the calmodulin-like domains of CDPKs.

Materials and Methods

Members of CDPK, CaMK, and calmodulin gene families were identified from public databases using the BLAST search algorithm (Altschul et al. 1990). Previously identified and characterized protein sequences were used as query sequences to search protein databases (e.g. SWALL, SWISS-PROT, and PDB). Genomic sequences of each gene family were identified and all sequences were visually inspected before undergoing further analysis. The selected genomic sequences of each gene family and their corresponding protein sequences were retrieved and compiled.

Multiple sequence alignments were generated with Clustal W (Thompson et al. 1994) software. The alignment was then manually edited to correct obvious misalignments and remove sections of ambiguous alignment. Phylogenetic trees were constructed by neighbor-joining (Saito and Nei 1987) or maximum parsimony methods, using the PHYLIP version 3.5c software package distributed by J. Felsenstein, Department of Genetics, University of Washington, Seattle. Phylogenetic bootstrapping (Felsenstein 1985) was performed to estimate the confidence level of the phylogenetic hypothesis. One hundred bootstrap replicates were conducted. TreeView (Page 1996) was used to draw gene trees.

Compiled genomic sequences were used to compare intron positions among members of each family. Intron data from annotated genomic sequences in databases were used for intron comparison. However, some intron splice sites, which had been predicted incorrectly, were corrected by reference to alignment of exon regions.

Table 1. CDPK genomic sequences used in this study

Gene Name	Organism	AC[a]	Chromosome
T805.150	*A. thaliana*	AL021890	4
or FIN20.40		AL022140	
F8K4.14	*A. thaliana*	AC004392	1
T4B21.11	*A. thaliana*	AF118223	4
T4B21.12	*A. thaliana*	AF118223	4
T4B21.13	*A. thaliana*	AF118223	4
T4B21.15	*A. thaliana*	AF118223	4
T19J18.7	*A. thaliana*	AF149414	4
CDPK1	*M. polymorpha*	AB017515	
CDPK6 or	*A. thaliana*	U20625	4
F9D16.120		AL035394	
CDPK	*Z. mays*	L27484	
F5J6.5	*A. thaliana*	AC002329	2
F23E12.130	*A. thaliana*	AL022604	4
F20D10.350	*A. thaliana*	AL035538	4
CDPK9	*A. thaliana*	U20626	5
F11F19.20	*A. thaliana*	AC007017	2
F25A4.29	*A. thaliana*	AC008263	1
T06D20.24 or	*A. thaliana*	U90439	2
T11A07.4		AC002339	
CDPK19	*A. thaliana*	U20627	
U54615	*A. thaliana*	U54615	
		AF049236[b]	
T28P16.1 or	*A. thaliana*	AC007169	2
T9H9.2		AC007071	
T13L16.9	*A. thaliana*	AC003952	2
T19K4.200	*A. thaliana*	AL022373	4
T3K9.9	*A. thaliana*	AC004261	2
T20E23.130	*A. thaliana*	AL133363	3
MAL3P3.17	*P. falciparum*	Z98547	3
PCAPK-A	*P. tetraurelia*	AF009560	
PFCPK or	*P. falciparum*	X67288	2
PFB0815W		AE001419	
PCAPK-B	*P. tetraurelia*	AF009561	

[a] GenBank accession number.
[b] Gene name is not given.

Results

CDPK Genomic Sequences and Their Chromosomal Distribution

Database searches identified 26 CDPK genomic sequences and two CRK genomic sequences (T3K9.9 and T20E23.130) from plants and protists. Twenty-two sequences are from *Arabidopsis*, and the remaining six sequences are from liverwort (*Marchantia polymorpha*), corn (*Zea mays*), malarial protist (*Plasmodium falciparum*), and *Paramecium tetraurelia* species. The gene name, accession number, plant/protist species, and chromosomal location (if available) for each of the sequences are shown in Table 1.

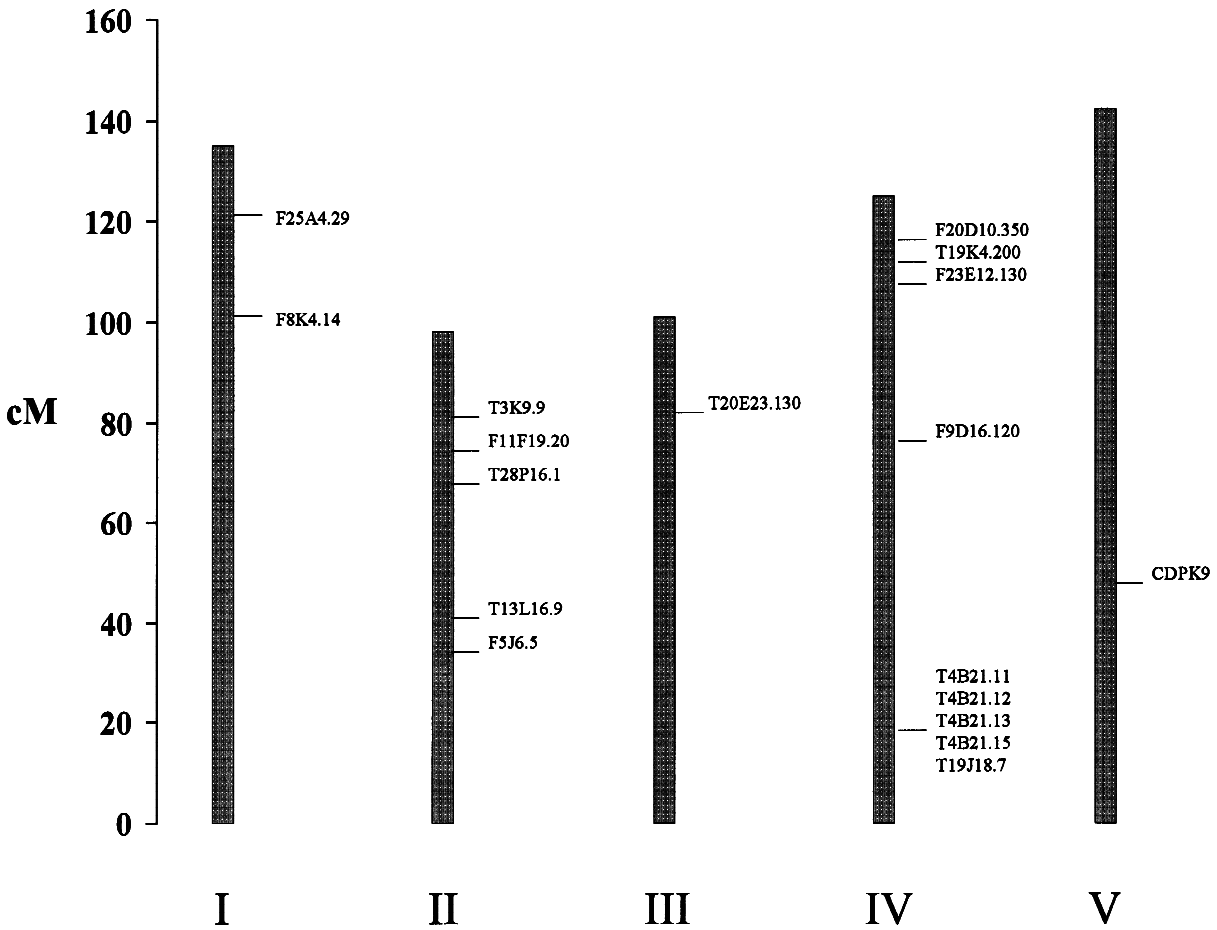

Fig. 1. Distribution of CDPK genes in the *Arabidopsis* genome. The genetic distances (cM) are indicated on a scale bar at left. The chromosome numbers are indicated on the bottom.

Among the 18 *Arabidopsis* CDPK genes with known chromosomal locations, most were found on chromosomes II and IV. Five are located on chromosome II, nine on chromosome IV, two on chromosome I, and one each on chromosomes III and V (Fig. 1). It is not surprising that most CDPK genes were found on chromosome II and IV, because these are the chromosomes most extensively sequenced thus far in *Arabidopsis.*

One CDPK gene cluster on chromosome IV contains five genes: T4B21.11, T4B21.12, T4B21.13, T4B21.15, and T19J18.7 (Fig. 1). They are organized in tandem and in the same transcriptional orientation.

Phylogenetic Analyses of CDPK Genes

To determine the evolutionary and functional relationships among CDPK genes, a phylogenetic analysis was performed using the kinase catalytic domain. Fig. 2 is a neighbor-joining tree, rooted by designating the protist sequences as the outgroup. The phylogenetic tree of CDPK sequences from plants and protists forms six well-defined subgroups (A–F). The tree generated by neighbor-joining is also supported by the maximum parsimony method (data not shown).

The phylogenetic analysis revealed several aspects of the molecular evolution of CDPK family. First, all protist CDPK sequences are grouped together (subgroup A) on the phylogenetic tree, as expected. Second, plant CDPK genes form two major branches on the phylogenetic tree, a result strongly indicated by high bootstrap support. One branch (subgroup B) contains four sequences, including two CDPK sequences (T13L16.9, T19K4.200) and two CRK sequences (T3K9.9, T20E23.130). The other branch (including subgroups C, D, E, and F) contains all other plant CDPK sequences. Third, CDPK1 from liverwort is grouped together in subgroup E with CDPK from corn and CDPK6 from *Arabidopsis*. This observation indicates that the common ancestor of this subgroup predates the origin of vascular plants. Finally, CDPKs encoded by the five genes (T4B21.11, T4B21.12, T4B21.13, T4B21.15, and T19J18.7) clustered on chromosome IV (Fig. 1) are all in subgroup D. Therefore, they may share a recent common ancestor.

Comparison of CDPK Amino Acid Sequences

All the proteins encoded by these CDPK genes contain three characteristic domains: kinase catalytic domain,

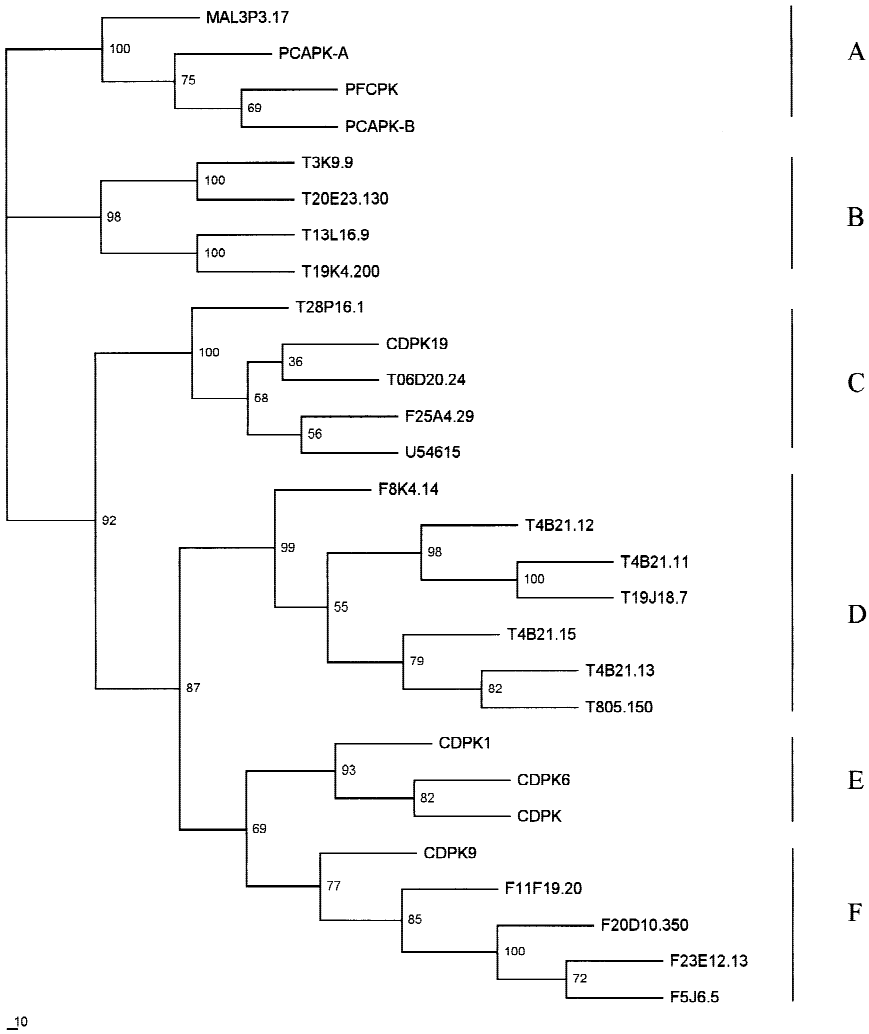

Fig. 2. Phylogenetic tree generated from the alignment of kinase domains of CDPK proteins. Six subgroups (A, B, C, D, E, and F) are defined. The numbers indicate the percent bootstrap replicates supporting the node. Root of the tree is arbitrarily placed.

calmodulin-like domain, and junction domain. Alignment of all sequences shows an overall amino acid sequence similarity (data not shown). A matrix showing the percent divergence between CDPKs from plants and protists is shown in Fig. 3. Plant CDPK sequences, including the CRK sequences, generally share higher amino acid sequence identity with each other (52% average pairwise divergence, or 48% average pairwise identity) than with protist sequences (70% average divergence, or 30% average identity). The protist sequences are more diverged (64% average pairwise divergence, or 36% average pairwise identity); they are nearly as different from each other as they are from plant CDPKs. Multiple sequence alignments of the three funtional domains of all the sequences listed in Fig. 3 indicate that the catalytic kinase domain sequences are more conserved (52% average pairwise identity) than the junction domains (42% average pairwise identity) or the calmodulin-like domains (36% average pairwise identity).

Some CDPKs also contain amino-terminal domains of various lengths, ranging from 5 to 129 bp. Most of the CDPK N-terminal domains share no significant sequence similarity with each other. However, among *Arabidopsis* CDPKs, significant sequence identity is found between T06D20.24 and CDPK19 (47%) of subgroup C, between T13L16.9 and T19K4.200 (50%) of subgroup B, and between the two CRK sequences T3K9.9 and T20E23.130 (57%) of subgroup B. Sequence homology also exists among the N-terminal domains of F5J6.5, F23E12.130, and F20D10.350 of subgroup F (60% average pairwise identity). Most sequences in subgroup D, including T19J18.7. T4B21.11, T4B21.12, T4B21.13, and T4B21.15, again share high sequence identity in their N-terminal domains (49% average pairwise identity). Obviously, the sequences sharing high homology in N-terminal domain have resulted from the most recent duplications as indicated by the phylogenetic analysis (Fig. 2).

Intron Evolution in CDPK Genes

The coding regions of all CDPK genes contain introns, ranging from five to eleven (Table 2). Introns were numbered consecutively from N-terminal to C-terminal of aligned CDPK protein sequences. Introns 1–16 are

Gene

1. T805.150/FIN20.40	-																											
2. F8K4.14	34	-																										
3. T4B21.11	41	45	-																									
4. T4B21.12	36	41	43	-																								
5. T4B21.13	20	31	38	32	-																							
6. T4B21.15	28	38	43	38	16	-																						
7. T19J18.7	39	43	27	41	35	43	-																					
8. CDPK1	44	45	49	45	39	44	48	-																				
9. CDPK6/F9D16.120	43	43	50	46	42	46	49	35	-																			
10. CDPK	41	42	47	44	37	44	50	32	35	-																		
11. F5J6.5	48	49	56	47	44	48	52	39	44	40	-																	
12. F23E12.130	50	49	56	47	46	49	52	39	47	40	12	-																
13. F20D10.350	43	45	55	46	40	45	53	36	42	40	13	15	-															
14. CDPK9	47	47	55	49	44	47	54	39	45	40	36	36	35	-														
15. F11F19.20	48	58	62	59	57	60	64	53	57	54	53	50	49	51	-													
16. F25A4.29	56	54	61	54	53	53	60	47	52	46	51	52	48	49	59	-												
17. T06D20.24/T11A07.4	55	55	59	54	54	52	61	51	54	49	53	51	48	50	60	39	-											
18. CDPK19	58	55	59	54	51	53	59	50	53	48	52	52	48	50	60	35	27	-										
19. U54615	53	54	60	53	50	50	60	47	53	46	50	50	47	49	60	34	39	33	-									
20. T28P16.1/T9H9.2	54	58	60	59	56	57	61	55	56	50	56	58	53	56	63	47	43	44	43	-								
21. T13L16.9	48	61	63	63	58	59	65	60	57	55	61	62	59	58	65	62	64	63	63	66	-							
22. T19K4.200	49	62	64	64	59	61	67	61	61	56	64	65	61	62	68	64	66	66	63	64	20	-						
23. T3K9.9	68	68	71	69	65	67	70	65	67	64	68	67	65	64	71	65	67	67	67	69	56	58	-					
24. T20E23.130	66	67	70	69	65	65	69	65	63	61	67	67	64	63	70	65	65	65	65	67	54	55	28	-				
25. MAL3P3.17	72	71	68	68	69	69	69	71	70	67	69	71	68	67	72	69	71	72	72	75	74	72	76	78	-			
26. PCAPK-A	67	68	70	69	67	68	72	66	66	68	67	67	68	67	74	68	68	68	67	69	73	73	76	76	65	-		
27. PFCPK/PFB0815W	70	68	72	68	69	69	72	69	69	67	70	70	67	67	75	71	71	71	71	72	72	73	76	75	67	63	-	
28. PCAPK-B	70	68	70	70	68	69	71	68	69	68	68	68	68	67	75	68	69	68	67	67	69	71	76	76	68	62	59	-
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28

Fig. 3. Degree of divergence between CDPK homologs. Values indicate percent difference (100 minus percent identity) for the complete sequences. The small triangular region on the bottom right includes pairwise comparisons between protist CDPK homologs. The rectangular region on the bottom contains the pairwise comparisons between plant and protist CDPK homologs. The remaining values are pairwise comparisons between plant CDPK proteins.

within the kinase domain. Intron 17 is in the junction domain. Intron 18 is positioned exactly on the border between the junction and calmodulin-like domains. The remaining introns 19–29 are located in the calmodulin-like domain.

For plant CDPK genes, the patterns of intron distribution can be readily mapped on the phylogenetic tree (Table 2). All subgroups of plant CDPK and CRK sequences share a set of four introns (introns 11, 16, 18, and 21; Table 2) at exactly the same positions (same phase in the same codon). This indicates that all plant CDPK and CRK sequences share a common ancestor. However, the sequences of the introns are highly diverged and most homologous introns are not alignable.

Three additional introns (introns 13, 26, and 28) are shared among CDPK genes from liverwort, maize, and *Arabidopsis* (Table 2), which suggests that these seven introns were present in the ancestral gene before the separation of bryophytes and higher plants. Intron 26 is retained by all members in subgroups D and E, but lost in subgroups C and F. CDPK genes sharing intron 9 are clustered in subgroup C. Clearly, CDPK genes clustered together on the phylogenetic tree have similar intron/exon structure.

The deepest branching subgroup B is well defined by the presence of introns 3, 6, 8, 10, 20, and 27, and the absence of introns 13, 26, and 28. Among members in subgroup B, the two CDPK sequences share an additional intron (intron 14), which is absent in the two CRK sequences.

Three subgroup B introns, introns 6, 14, and 27, coincide in both codon and phase with introns in protist CDPK genes: PCAPK-B, MAL3P3.7, and PFCPK, respectively. The occurrence of conserved introns suggests a monophyletic origin for plant and protist CDPKs.

In contrast to the overall conservation of intron locations among plant CDPK genes, intron locations among the protist CDPK genes are highly diverged. Of fifteen introns present in protist CDPK genes, only intron 15 in the *P. tetraurelia* PCAPK-A gene coincides with an intron in the *P. falciparum* PCAPK-B gene in codon, but not in phase (Table 2). If intron "frame-shifting" or "sliding" is discounted (Stoltzfus et al. 1997), then these must be considered nonhomologous introns.

Relationships with CaMK and Calmodulin Families

Seven CaMK genomic sequences were identified from rat, fission yeast, *Aspergillus nidulans*, and *Caenorhabditis elegans*. CaMK sequences are highly diverged, except for three *C. elegans* sequences that share 92% amino acid identity with each other. The proteins encoded by these CaMK genes are readily alignable with the kinase and regulatory domains of CDPK members, with only a few ambiguities (data not shown). However, they generally share less than 40% amino acid sequence identity with CDPK kinase and junction domains (33% average pairwise identity). One intron is conserved among fission yeast, rat, and nematode CaMKs, indicating that this is an evolutionarily older intron. However, most introns show restricted phylogenetic distribution, indicating

Table 2. Distribution of introns in CDPK genes

		Intron No.[a]													
Gene	Subgroup	1 (II)	2 (0)	3 (0)	4 (0)	5 (II)	6 (II)	7 (II)	8 (II)	9 (0)	10 (0)	11 (I)	12 (0)	13 (1)	14 (0)
T805.150/FIN20.40	D											+		+	
F8K4.14					+							+		+	
T4B21.11												+		+	
T4B21.12												+		+	
T4B21.13												+		+	
T4B21.15												+		+	
T19J28.7												+		+	
CDPK1	E											+		+	
CDPK6/F9D16.120						+						+		+	
CDPK														+	
F5J6.5	F											+		+	
F23E12.130												+		+	
F20D10.350												+		+	
CDPK9												+		+	
F11F19.20												+		+	
F25A4.29	C											+		+	
T06D20.24/T11A07.4										+		+		+	
CDPK19										+		+		+	
U54615												+		+	
T28P16.1/T9H9.2		+								+		+			
T13L16.9	B			+			+		+		+	+			+
T19K4.200				+			+		+		+	+			+
T3K9.9				+			+		+		+	+			
T20E23.130				+			+		+		+	+			
MAL3P3.7	A														+
PCAPK-A			+					+							
PFCPK/PFB0815W															
PCAPK-B							+						+		

[a] The number in parenthesis indicates the intron phase.
[b] Two protist introns numbered 16 coincide with each other in position, but not in phase. Please see text for details.
[c] The protist intron in this column coincides in position (but not in phase) with the intron shared by three plant genes.
[d] This sequence is truncated in the C-terminal. Therefore, the information for this intron is not available for this gene.

more recent origins (data not shown). No introns found in these CaMK genes coincide in position with any introns in CDPK genes.

Forty calmodulin genomic sequences were identified from a variety of species, including protist, plant, fungal, and animal species. They are highly conserved (82% average pairwise identity, with a maximum 100% identity) and share at least 41% amino acid identity with each other. Their protein sequences are readily alignable with the calmodulin-like domains of CDPK, although they share significantly lower amino acid sequence identity with CDPK calmodulin-like domains (31% average pairwise identity). The length and position of four EF-hands are highly conserved between two families.

The phylogenetic distribution of introns in the calmodulin gene family is highly restricted (Table 3), again indicating recent origins of most calmodulin introns. For example, five introns (B, H, I, L, and N) show animal-specific distribution, four introns (C, D, J, and M) are fungal-specific, and three introns (F, G, and K) are present only in protist species. However, potentially old introns were also found, which are conserved among species across the kingdoms. For example, intron A is conserved across animal, fungal, and protist species. Intron E is shared by both plant and fungal species.

The intron positions were also compared between CDPK and calmodulin families. Intron J (Table 3), shared by seven fungal calmodulin sequences, coincides with intron 22 (Table 2) from a *P. falciparum* CDPK gene. These seven fungal calmodulin genes represent six fungal species from the phylum Ascomycota, including *Magnaporthe grisea* (rice blast fungus), *Neurospora crassa*, *Colletotrichum trifolli*, *Aspergillus nidulans*, *Aspergillus oryzae*, and *Ajellomyces capsulatus*. In addition, an intron shared by eleven animal calmodulin sequences coincides with a *P. tetraurelia* CDPK intron (intron 25, Table 2). These animal calmodulin sequences are from *Homo sapiens* (human), *C. elegans*, *Halocynthia roretzi* (sea squirt), *Branchiostoma lanceolatum* (lancelet), *Drosophila melanogaster* (fruit fly), *Rattus*

Table 2. Extended

Gene	Subgroup	Intron No.[a] 15[b]	16 (I)	17 (0)	18 (0)	19 (0)	20 (0)	21 (0)	22 (I)	23 (0)	24 (0)	25 (0)	26 (II)	27 (0)	28 (0)	29 (0)
T805.150/FIN20.40	D		+		+			+					+		+	
F8K4.14			+		+			+					+		+	
T4B21.11			+		+			+					+		+	
T4B21.12			+		+			+					+		+	
T4B21.13			+		+			+					+		+	
T4B21.15			+		+			+					+		+	
T19J28.7			+		+			+					+		+	
CDPK1	E		+		+			+					+		+	
CDPK6/F9D16.120			+		+			+					+		+	
CDPK			+		+			+					+		+	
F5J6.5	F		+		+			+							+	
F23E12.130			+		+			+							+	
F20D10.350			+		+			+							+	
CDPK9			+					+							+	
F11F19.20			+		+			+							[d]	
F25A4.29	C		+		+			+							+	
T06D20.24/T11A07.4			+					+							+	
CDPK19			+		+			+							+	
U54615			+		+			+							+	
T28P16.1/T9H9.2			+		+										+	
T13L16.9	B		+		+		+	+						+		
T19K4.200			+		+		+	+						+		
T3K9.9			+		+		+	+						+		
T20E23.130			+		+		+	+						+		
MAL3P3.7	A									+						
PCAPK-A		+				+					+					
PFCPK/PFB0815W									+					+		+
PCAPK-B		+		+			+[c]					+				

norvegicus (rat), *Gallus gallus* (chicken), and *Aplysia california* (California sea hare).

Plant Phosphoenolpyruvate Carboxylase Kinases (PEPCKs)

Plant PEPCKs have protein kinase catalytic domains that resemble CDPKs, but lack the junction domains or calmodulin-like domains (Hartwell et al. 1999). The *Arabidopsis* genome contains four genes encoding PEPCK or related kinases (accession nos. AAF99758, AAF63784, AAF88079, AAF88093). Genes for all four of these kinases contain an intron at the same codon and phase as CDPK intron 16 (Table 2), and two of these contain another intron in the same codon and phase as CDPK intron 11 (Table 2).

Discussion

Conserved Introns Support a Monophyletic Origin of Plant and Protist CDPK Genes

Database searches have identified over seventy different CDPKs from vascular plants, liverwort, moss, green algae, and Alveolate protist species. A recent phylogenetic analysis of the catalytic domains of these sequences suggested that plant and algal CDPKs have a common origin, but could not resolve whether CRKs and protist CDPKs evolved separately (Harmon et al. 2000). The results from this analysis of conserved intron positions add another dimension that strongly supports a monophyletic origin for plant and protist CDPKs and CRKs.

The rooted phylogenetic tree derived from genomic sequences (Fig. 2) agrees well with the unrooted tree of Harmon et al. (2000) constructed from both cDNA and genomic sequences. Both trees show a branch (subgroup B of Fig. 2) containing CRK and CDPK sequences separated from all other plant CDPKs. Comparison of introns strongly supports this deep phylogenetic branching with several subgroup B-specific introns. The remaining subgroups of plant CDPKs share a similar pattern of introns, with subgroups differing in only one or two introns. We have analyzed additional *Arabidopsis* CDPK and CRK genes recently entered into the databases, but did not find any new subgroups or unexpected patterns of introns (data not shown). We were unable to identify any sequences for chimeric calcium/calmodulin-dependent kinases (Patil et al. 1995) in the *Arabidopsis* genome.

Table 3. Distribution of introns in calmodulin genes

		Introns[a]													
GenBank AC	Organism	A (0)	B (I)	C (0)	D (0)	E (I)	F (I)	G (0)	H (0)	I (I)	J[b] (I)	K (0)	L[c] (0)	M (I)	N (I)
AC006536	*H. sapiens*	+	+							+			+		+
X52606, X52607, X52608	*H. sapiens*	+	+							+			+		+
U94725, U94726, U94728	*H. sapiens*	+	+							+			+		+
U12022	*H. sapiens*	+	+							+			+		+
X14265	*R. norvegicus*	+	+							+			+		+
L00096, L00097, L00098, L00099, L00100, L00101	*G. gallus*	+	+							+			+		+
X13931, X13932, X05117	*R. norvegicus*	+	+							+			+		+
X13833, X13834, X13835	*R. norvegicus*	+	+										+		+
AB018797	*H. roretzi*	+	+						+				+		+
AB018796	*H. roretzi*	+	+						+				+		+
AJ001092	*B. lanceolatum*								+				+		
AJ001093	*B. lanceolatum*								+						
AF016429	*C. elegans*	+	+												+
X05948, X05949, X05950, X05951	*D. melanogaster*	+								+					+
X64653, X64655	*A. california*	+	+							+					
AC004261	*A. thaliana*					+									
L34546, D45848	*A. thaliana*					+									
Z97336	*A. thaliana*					+									
AC004261	*A. thaliana*					+									
AC005623	*A. thaliana*					+									
D45848	*A. thaliana*					+									
X67273	*A. thaliana*					+									
M73711	*A. thaliana*					+									
X60737	*M. domestica*					+									
Z12827, Z12828, L18914	*O. sativa*					+									
AF064456	*O. sativa*					+									
AF103729	*M. grisea*	+		+		+					+			+	
AF104986	*M. grisea*	+		+		+					+			+	
L02964	*N. crassa*	+		+		+					+			+	
U15993	*C. trifolli*	+		+		+					+			+	
J05543	*A. nidulans*	+		+		+					+			+	
D44468	*A. oryzae*	+		+		+					+			+	
AF072882	*A. capsulatus*	+		+		+					+			+	
L05572, M96933	*P. carinii*				+										
U91643	*P. ostreatus*				+										
Z95395, M16475	*X. pombe*	+													
M59349, X56950, M59770, M99442	*P. falciparum*						+								
AF020781	*S. microadriaticum*							+				+			
M20729	*C. reinhardtii*						+					+			
M64089	*D. discoideum*	+						+							

[a] Introns are lettered consecutively from N-terminal to C-terminal of aligned calmodulin protein sequences. Only conserved introns are included in this table.
[b] Intron J coincides in codon and phase with intron 22 of CDPKs (Table 2).
[c] Intron L coincides in codon and phase with intron 25 of CDPKs (Table 2).

The presence of four highly conserved introns among all subgroups of plant CDPKs indicates that all plant CDPK and CRK genes have a common ancestry (Table 2). Moreover, this lineage most likely includes plant PEPCKS, because two of these conserved introns are also found in *Arabidopsis* PEPCK genes. As the conserved introns are also present in a liverwort CDPK gene, they must predate at least the divergence of vascular plants.

Intron analysis suggests further that plant and protist CDPKs also share a common ancestor. Particularly, subgroup B of plant CDPK and CRK sequences shares three common introns with protist CDPKs (Table 2). The simplest explanation is that these three introns originated in the common ancestor of plant and protist CDPKs. According to this hypothesis, these introns were lost in the common ancestor of plant CDPK subgroups C, D, E, and F. The fact that these putative ancestral introns are not

conserved among the four protist CDPK genes could also be explained by random loss of these introns. These protist CDPK sequences are as divergent from each other as they are from plant CDPKs (Table 3). Therefore, it is not surprising that they have no introns in common.

The alternative hypothesis would require that protist CDPK genes recently gained these three introns. It seems unlikely that three of sixteen protist introns were acquired in exactly the same positions (same codon and phase) as the plant subgroup B introns. Analysis of additional protist CDPK genes is required to determine if the putative ancestral introns are indeed shared among other Alveolate protists, as would be expected if they are ancestral, or whether they are unique occurrences, as would be expected if they are recent independent insertions.

Intron Distributions Reveal Recent Loss and Gain of Introns

Intron comparison revealed recent intron loss and gain within CDPK genes. Some individual introns (e.g. 11, 13, 18, and 21), which are shared by most plant CDPK genes, have been lost from CDPK, T28P16.1, CDPK9/T06D20.24, and T28P16.1, respectively. These introns have been precisely excised, leaving the coding region intact. Such precise excision of introns has been attributed by Fink (1987) to the reverse transcription of partially processed transcript and reinsertion of the DNA product back into the genome. Novel introns (e.g. 1, 4, and 5) were also found within a number of plant CDPK genes (T19J18.7, T28P16.1, F8K4.14, and CDPK6). The presence of an intron in a particular plant CDPK gene, and its absence from all other plant CDPK genes, can be explained most simply by its recent insertion into the gene.

The identification of new introns in CDPK genes agrees with the introns-late theory (or insertional theory of introns), which postulates that ancient genes existed as uninterrupted exons and that intron insertions occurred relatively recently during the evolution of eukaryotic lineages (Cavallier-Smith 1978; Rogers 1985; Rogers 1989; Palmer and Logsdon 1991; Cavallier-Smith 1985). Likewise, introns show highly restricted phylogenetic distribution in CaMK and calmodulin genes as well.

However, evolutionarily older introns are also identified within CDPK, CaMK, and calmodulin gene families, which are shared by a variety of species across the kingdoms. The conservation of intron positions among genes from diverse phyla seems to support the introns-early theory (or exon theory of genes), which proposes that introns are ancient and recombination within introns generated new exon combinations, and thus new genes (Blake 1978; Doolittle 1978; Gilbert 1978; Gilbert 1987). Therefore, it seems that, in the case of CDPK genes, some of introns are ancient, whereas others have probably arisen by recent insertions.

Origin of CDPK Genes

It has been speculated that the ancestral CDPK gene evolved by fusion of ancestral CaMK and calmodulin genes (Harper et al. 1991; Suen and Choi 1991). The basis for this speculation is that the kinase and junction domains of CDPKs share similarity with CaMK kinase catalytic and regulatory domains, respectively, and that CDPK calmodulin-like domains resemble calmodulin. In this study, we found two protist CDPK introns located in the same codon and phase as introns conserved among fungal calmodulin genes and animal calmodulin genes. Again, if the protist CDPK introns are old (found in other protist CDPK genes), they would provide strong support for a common ancestry for CDPK calmodulin-like domains and calmodulin genes. Given the ancient origin of the postulated gene fusion event, loss of these introns in the ancestral plant CDPK and calmodulin genes may account for the absence of these introns in extant plant CDPK and calmodulin genes.

No common introns were found between CDPK and CaMK genes. In fact, the survey of CaMK genomic sequences identified only one potentially old intron conserved between fungal and animal CaMK genes. Given the paucity of conserved introns among these CaMK genes, the lack of conserved introns between CDPK and CaMK genes is not surprising, and should not be construed as evidence against a common origin of CDPK and CaMK kinase domains.

This raises another interesting question: how might CaMK and calmodulin genes have been brought together? According to the exon theory of genes (Gilbert 1987), ancient introns facilitated the recombination of previously evolved functional domains to form new functional proteins. It has been observed that, at least for some complex proteins, exons correlate with protein domains (de Souza et al. 1996; Gilbert and Glynias 1993; Gilbert et al. 1986; Go 1981). Most recently, it was shown that such correlation of intron positions with boundaries of protein domains is due primarily to phase 0 introns (de Souza et al. 1998). Such excess of phase 0 introns in the boundaries of protein domains is expected, because exon shuffling will not disturb the reading frame if the introns are in the same phase. Even though the introns-late theory rejects the notion of shuffling of primordial or pre-eukaryotic exons, it does not deny the possibility of more recent exon shuffling within eukaryotic genes and protein evolution by fusion of domains (Rzhetsky et al. 1997).

In the case of CDPK genes, we found that a phase 0 intron (intron 18, Table 2) is located precisely on the border between the junction domain and calmodulin-like domain in nearly all plant CDPK genes. Therefore, we hypothesize that exon shuffling facilitated by ancestral introns may have played an important role in the origin of CDPK genes. We propose that recombination between a phase 0 intron near the calmodulin N-terminus and a

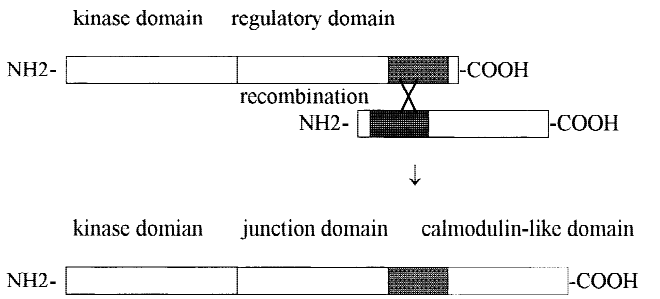

Fig. 4. CDPK origin hypothesis. An intron (gray box) in the C-terminal of an ancestral CaMK and an intron (gray box) in the N-terminal of ancestral calmodulin gene allowed recombination to take place, which resulted in the fusion of ancestral CaMK and calmodulin genes.

phase 0 intron near the C-terminus of the calmodulin-binding region of CaMK formed the ancestral CDPK gene without disrupting the functions of the domains (Fig. 4). Indeed, a phase 0 intron (intron A, Table 3) is located between the first and second codons of calmodulin genes (Table 3). This intron appears to be ancestral because it is conserved among calmodulin genes from animals and fungi. Moreover, a number of phase 0 introns have been found near the calmodulin-binding sequences of CaMK genes. For example, the *C. elegans* CaMK gene unc-43/K11E8.1 has a phase 0 intron between K290 and A291, precisely at the end of the calmodulin-binding region. An intron at the same position can also be found in the *Drosophila* CaMKI gene (CG1495).

This analysis of introns in CDPK genes, calmodulin genes, and CaMK genes suggests that plant CDPKs and CRKs arose from a single common ancestor, that plant and protist CDPKs also share a common ancestry, and that the ancestral CDPK gene originated by exon shuffling facilitated by recombination within phase 0 introns. For each of these interpretations, the abundance of genomic sequences from plants and animals strongly supports ancestral origins of the conserved introns. However, the paucity of protist genomic sequences makes it impossible to determine whether the protest introns at the same positions are ancestral or newly gained. Sequencing of additional protist genomes is urgently needed to answer these and many other questions of molecular evolution.

Acknowledgments. This work was carried out at Georgia Institute of Technology under a Faculty Development Program supported by Georgia Institute of Technology and Savannah State University.

References

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

Azzi A, Boscoboinik D, Hensey C (1992) The protein kinase C family. Eur J Biochem 208:547–557

Berberich T, Kusano T (1996) Cycloheximide induces a subset of low-temperature-inducible genes in maize. Mol Gen Genet 254: 275–283

Blake CCF (1978) Do genes-in-pieces imply protein in pieces? Nature 273:267

Botella JR, Arteca JM, Somodevilla M, Arteca RN (1996) Calcium-dependent protein kinase gene expression in response to physical and chemical stimuli in mungbean (*Vigna radiata*). Plant Mol Biol 30:1129–1137

Breviario D, Morello L, Giani S (1995) Molecular cloning of two novel rice cDNA sequences encoding putative calcium-dependent protein kinases. Plant Mol Biol 27:953–967

Bumstead JM, Dunn PPJ, Tomley FM (1995) Nitrocellulose immunoblotting for identification and molecular gene cloning of *Eimeria maxima* antigens that stimulate lymphocyte proliferation. Clin Diagn Lab Immuol 2:524–530

Cavallier-Smith T (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J Cell Sci 34:247–278

Cavallier-Smith T (1985) Selfish DNA and the origin of introns. Nature 315:283–284

de Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W (1998) Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proc Natl Acad Sci USA 95:5094–5099

de Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W (1996) Intron positions correlate with module boundaries in ancient proteins. Proc Natl Acad Sci USA 93:14632–14636

Doolittle WF (1978) Gene-in-pieces: were they ever together? Nature 272:581–582

Dunn PPJ, Bumstead JM, Tomley FM (1996) Sequence, expression and localization of calmodulin-domain protein kinases in *Eimeia tenella* and *Eimeria maxima.* Parasitology 113:439–448

Estruch JJ, Kadwell S, Merlin E, Crossland L (1994) Cloning and characterization of a maize pollen-specific calcium-dependent calmodulin-independent protein kinase. Proc Natl Acad Sci USA 91: 8837–8841

Farber PM, Graeser R, Franklin RM, Kappes R (1997) Molecular cloning and characterization of a second calcium-dependent protein kinase of *Plasmodium falciparum.* Mol Biochem Parasitol 87:211–216

Felsenstein J (1985) Confidence limits on phologenies: an approach using the bootstrap. Evolution 39:783–791

Fink GR (1987) Pseudogenes in yeast? Cell 49:5–6

Furumoto T, Ogawa N, Hata S, Izui K (1996) Plant calcium-dependent protein kinase-related kinases (CRKs) do not require calcium for their activities. FEBS Lett 396:147–151

Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C, Pederson J, Shen K, Jing J, Aston C, Lai Z, Schwartz DC, Pertea M, Salzberg S, Zhou L, Sutton GG, Clayton R, White O, Smith HO, Fraser CM, Adams MD, Venter JC, Hoffman SL (1998) Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum.* Science 282: 1126–1132

Gilbert W (1978) Why gene in pieces? Nature 271:501

Gilbert W (1987) The exon theory of genes. Cold Spring Harbor Symp Quant Biol 52:901–905

Gilbert W, Glynias M (1993) On the ancient nature of introns. Gene 135:137–144

Gilbert W, Marchionni M, McKnight G (1986) On the antiquity of introns. Cell 46:151–153

Go M (1981) Correlations of DNA exonic regions with protein structural units in haemoglobin. Nature 291:90–93

Harmon AC, Gribskov M, Harper JF (2000) CDPKs—a kinase for every Ca^{2+} signal? Trends Plant Sci 5:154–159

Harmon AC, Putnam-Evans C, Cormier MJ (1987) A calcium-dependent but calmodulin-independent protein kinase from soybean. Plant Physiol 83:830–837

Harper JF, Sussman MR, Schaller GE, Putnam-Evans C, Charbonneau H, Harmon AC (1991) A calcium-dependent protein kinase with a regulatory domain similar to calmodulin. Science 252:951–954

Harper JF, Binder BM, Sussman MR (1993) Calcium and lipid regulation of an *Arabidopsis* protein kinase expressed in *Escherichia coli*. Biochemistry 32:3282–3290

Hartwell J, Gill A, Nimmo GA, Wilkins MB, Jenkins GI, Nimmo HG (1999) Phosphoenolpyruvate carboxylase kinase is a novel protein kinase regulated at the level of gene expression. Plant J 20:333–342

Kawasaki T, Hayashida N, Baba T, Shinozaki K, Shimada H (1993) The gene encoding a calcium-dependent protein kinase located near the *Sbel* gene encoding starch branching enzyme I is specifically expressed in developing rice seeds. Gene 129:183–189

Kim K, Messenger LA, Nelson DL (1998) Ca^{2+}-dependent protein kinases of *Paramecium*-cloning provides evidence of a multigene family. Eur J Biochem 251:605–612

Lindzen E, Choi J (1995) A carrot cDNA encoding an atypical protein kinase homologous to plant calcium-dependent protein kinases. Plant Mol Biol 28:785–797

Lu YT, Hidaka H, Feldman LJ (1996) Characterization of a calcium/calmodulin-dependent protein kinase homolog from maize roots showing light-regulated gravitropism. Planta 199:18–24

Nishiyama R, Mizuno H, Okada S, Yamaguchi T, Takenaka M, Fukuzawa H, Ohyama K (1999) Two mRNA species encoding calcium-dependent protein kinases are differentially expressed in sexual organs of *Marchantia polymorpha* through alternative splicing. Plant Cell Physiol 40:205–212

Page RDM (1996) Tree View. An application to display phylogenetic trees on personal computer. Comp Appl Biol Sci 12:357–358

Palmer JD, Logsdon JM (1991) The recent origins of introns. Curr Opin Genet Dev 1:470–477

Patil S, Takezawa D, Poovaiah BW (1995) Chimeric plant calcium/calmodulin-dependent protein kinase gene with a neural visinin-like calcium-binding domain. Proc Natl Acad Sci USA 92:4897–4901

Perry DJ, Bousquet J (1998) Sequence-tagged-site (STS) markers of arbitrary genes: development, characterization and analysis of linkage in black spruce. Genetics 149:1089–1098

Roberts DM (1993) Protein kinases with calmodulin-like domain: novel targets of calcium signals in plants. Curr Opin Cell Biol 5:242–246

Roberts DM, Harmon AC (1992) Calcium-modulated proteins: targets of intercellular calcium signals in higher plants. Annu Rev Plant Physiol Plant Mol Biol 43:375–414

Rogers J (1985) Exon shuffling and intron insertion in serine proteases genes. Nature 315:458–459

Rogers J (1989) How were introns inserted into nuclear genes? Trends Genet 5:213–216

Rzhetsky A, Ayala FJ, Hsu LC, Chang C, Yoshida A (1997) Exon/intron structure of aldehyde dehydrogease genes supports the "introns-late" theory. Proc Natl Acad Sci USA 94:6820–6825

Saijo Y, Hata S, Sheen J, Izui K (1997) cDNA cloning and prokaryotic expression of maize calcium-dependent protein kinases. Biochem Biophys Acta 1350:109–114

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

Schulman H (1993) The multifunctional Ca^{2+} /calmodulin-dependent protein kinases. Curr Opin Cell Biol 5:247–253

Stoltzfus A, Logsdon JM, Palmer JD, Doolittle WF (1997) Intron "sliding" and the diversity of intron positions. Proc Natl Acad Sci USA 94:10739–10744

Stone J, Walker C (1995) Plant protein kinase families and signal transduction. Plant Physiol 108:451–458

Suen K-L, Choi J (1991) Isolation and sequence analysis of a cDNA clone for a carrot calcium-dependent protein kinase: homology to calcium/calmodulin-dependent protein kinases and to calmodulin. Plant Mol Biol 17:581–590

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

Urao T, Katagiri T, Mizoguchi T, Yamaguchi-Shinozaki K, Hayashida N, Shinozaki K (1994) Two genes that encode Ca^{2+}-dependent protein kinases are induced by drought and high-salt stresses in *Arabidopsis thaliana*. Mol Gen Genet 244:331–340

Zhao Y, Kappes B, Franklin RM (1993) Gene structure and expression of an unusual protein kinase from *Plasmodium falciparum* homologous at its carboxyl terminus with the EF hand calcium-binding proteins. J Biol Chem 268:4347–4354