A Superfamily of Proteins with Novel Cysteine-Rich Repeats[1]

Zhixiang Chen*

Department of Microbiology, Molecular Biology, and Biochemistry, University of Idaho, Moscow, Idaho 83844–3052

One of the most intriguing discoveries from the studies of the last several years is the presence of a large number of genes encoding receptor-like protein kinases (RLKs) in plants. In Arabidopsis, there are at least 340 genes that code for proteins, each consisting of an N-terminal signal sequence, an extracellular receptor domain, a single transmembrane domain, and a C-terminal cytoplasmic Ser/Thr protein kinase domain (The Arabidopsis Genome Initiative, 2000). Plant RLKs are classified according to sequence motifs in the putative extracellular receptor domains (Hardie, 1999). One of the largest and best studied classes of RLKs is characterized by the Leu-rich repeat (LRR) known to be involved in protein-protein interaction (Walker, 1994). The second class includes the family of S-domain RLKs (SRKs) of Brassicaceae with high similarity to S-locus glycoproteins (SLGs) involved in the self-incompatibility response (Nasrallah et al., 1994). An S-domain contains a characteristic array of Cys residues and other conserved motifs (Walker, 1994). The third class, represented in Arabidopsis by LECRK1 and LRK1, contain a lectin-like extracellular domain that may bind oligosaccharides, such as the elcitors derived from breakdown of the cell wall (Herve et al., 1996). The wall-associated protein kinases represent a fourth class of RLKs that have extracellular domains containing sequence repeats related to mammalian epidermal growth factors (He et al., 1996). Other Arabidopsis RLKs contain additional types of extracellular domains. For example, PR5K has an extracellular domain related to the PR5 proteins that accumulate in the extracellular space in response to infection by microbial pathogens (Wang et al., 1996). Light repressible receptor protein kinase is an Arabidopsis RLK that has an extracellular domain with a novel Leu zipper motif (Deeken and Kaldenhoff, 1997).

Several groups have reported very recently the isolation of a group of RLK genes in Arabidopsis that are induced by pathogen infection and treatment with reactive oxygen species or salicylic acid (Czernic et al., 1999; Du and Chen, 2000; Ohtake et al., 2000). It is intriguing that these RLK genes are within a tandem array of 20 RLK genes on chromosome IV. The extracellular domains of these RLKs showed little similarity with those of other classes of isolated RLKs and share limited sequence homology among each other. However, all these RLK proteins contain two copies of the C-X8-C-X2-C motif in their extracellular domains (Fig. 1). A fourth Cys residue is usually also found at the C-terminal side of the C-X8-C-X2-C motif but its position varies slightly among repeats. The C-X8-C-X2-C repeat is a novel motif structurally distinct from the Cys-rich region of S-locus glycoproteins and SRKs. The conserved Cys residues in these extracellular domains of RLKs may participate in the formation of the three-dimensional structure of the protein through disulfide bonds. In an alternate manner, they may form zinc finger motifs as found in many DNA-binding transcription factors. Both disulfide bonds and zinc fingers are known to mediate protein-protein interactions, a critical step in the activation of many animal receptor protein kinases upon ligand binding (Hardie, 1999).

It is interesting that Arabidopsis genome sequence search revealed that the C-X8-C-X2-C motif is present in a large number of proteins that can be classified into two groups. In addition to the 20 RLKs encoded by the tandem array of genes on chromosome IV, there are at least 22 other genes coding for RLK proteins that contain copies of the novel Cys-rich repeat (CRR) in their extracellular domains (Fig. 1). Thus, there are at least 42 CRR RLKs (CRKs), constituting one of the largest classes of RLKs in Arabidopsis (Fig. 1). The PK20-1 from common bean is also an RLK containing two copies of the C-X8-C-X2-C repeats (Lange et al., 1999; Fig. 1), indicating that CRKs are present in other plants. In addition, there are at least 60 genes in Arabidopsis encoding non-kinase proteins that contain copies of the CRR motif (Fig. 2). Almost all of these proteins have an N-terminal signal peptide sequence but no hydrophobic transmembrane domains at C terminus (Fig. 2A), suggesting that they are CRR secretory proteins (CRRSPs). A 33-kD rice protein secreted by suspension-cultured rice cells also contains two copies of the CRR (Fig. 2). Thus, in Arabidopsis, there are more than 100 genes that code for proteins containing the novel CRR motifs, making them one of the largest protein superfamilies.

Other than the conserved CRR motifs, the overall sequence homology among these CRR proteins is not

[1] This work was supported in part by the U.S. National Science Foundation (grant no. MCB–9905976).

* E-mail zchen@uidaho.edu; fax 208–885–6518.

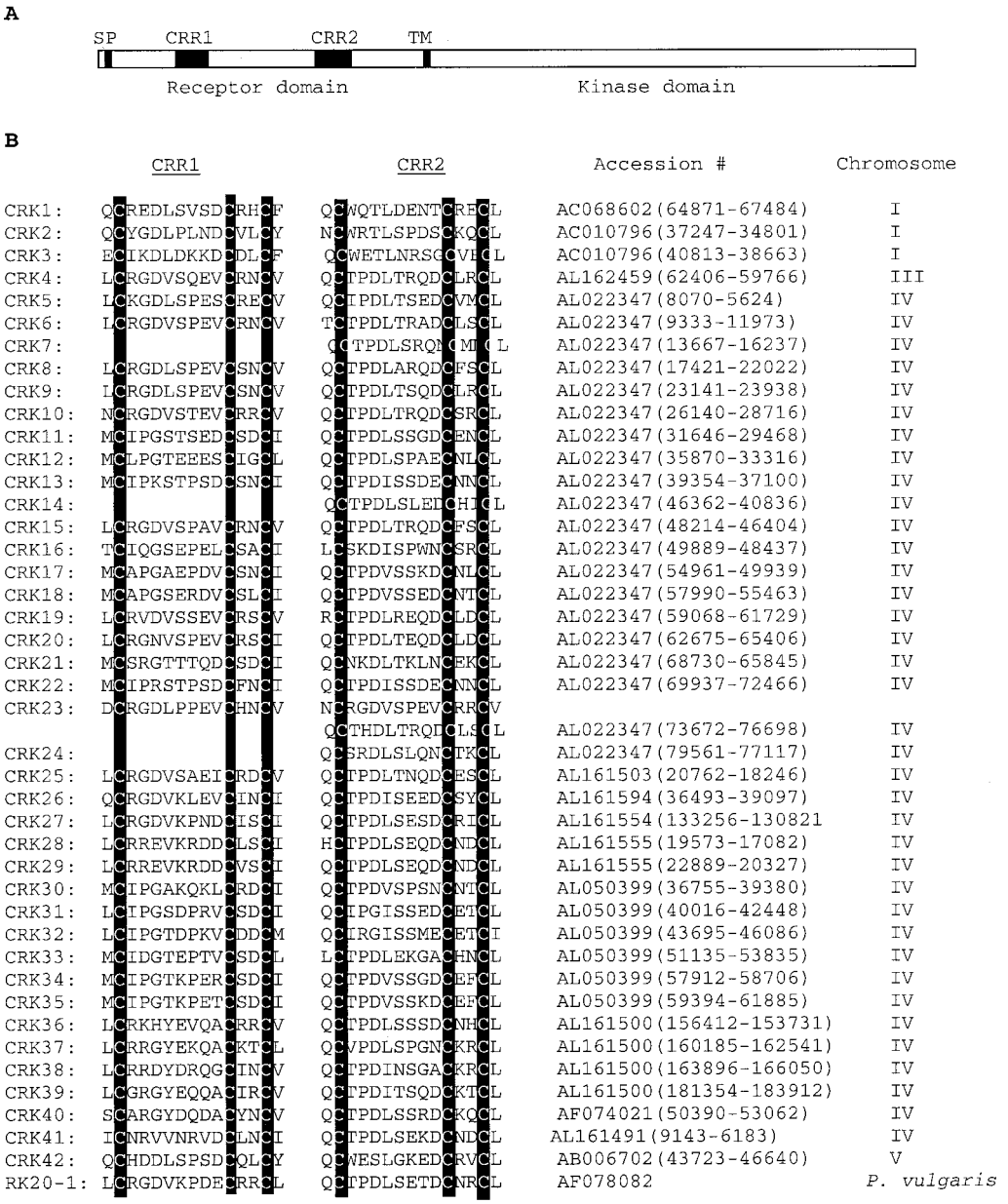

Figure 1. The CRK protein superfamily. A, Schematic diagram of the structures of common CRK proteins. SP, Signal peptide; TM, transmembrane domain. B, Multiple alignment of the CRR motifs from 42 Arabidopsis CRKs and the RK20-1 from common bean (*Phaseolus vulgaris*). The locations of the Arabidopsis genes are indicated by the starting and stopping positions of the coding regions on the bacterial artificial chromosome clones according to the annotated genome sequence. The conserved Cys residues in CRR motifs are highlighted.

particularly high, even among those encoded by the genes from the same tandem arrays. The exception is the 13 CRRSPs (CRRSP42–54) encoded by an array of genes on chromosome IV that have almost identical protein sequences (Fig. 2B). Thus, there appears to be a similar number of genes encoding distinct CRKs and CRRSPs (42 and 47, respectively) in Arabidopsis. It is also interesting to notice that although there is a tandem array of genes encoding 20 distinct CRKs (CRK5 to CRK24) on chromosome IV, a similar tandem array of genes encode 23 distinct CRRSPs (CRRSP16 to CRRSP38) on chromosome III (Fig. 2). Recent genome sequence analysis also revealed that the majority of the Arabidospsis genome is represented in duplicated segments, supporting that the plant, like maize (*Zea mays*), had a tetraploid ancestor (The Arabidopsis Genome Initiative, 2000). The presence of the approximately equal number of genes encoding distinct CRKs and CRRSPs may also result from duplication of the Arabidopsis genome.

Most of the genes coding for CRKs and CRRSPs are organized in tandem arrays in Arabidopsis. On chro-

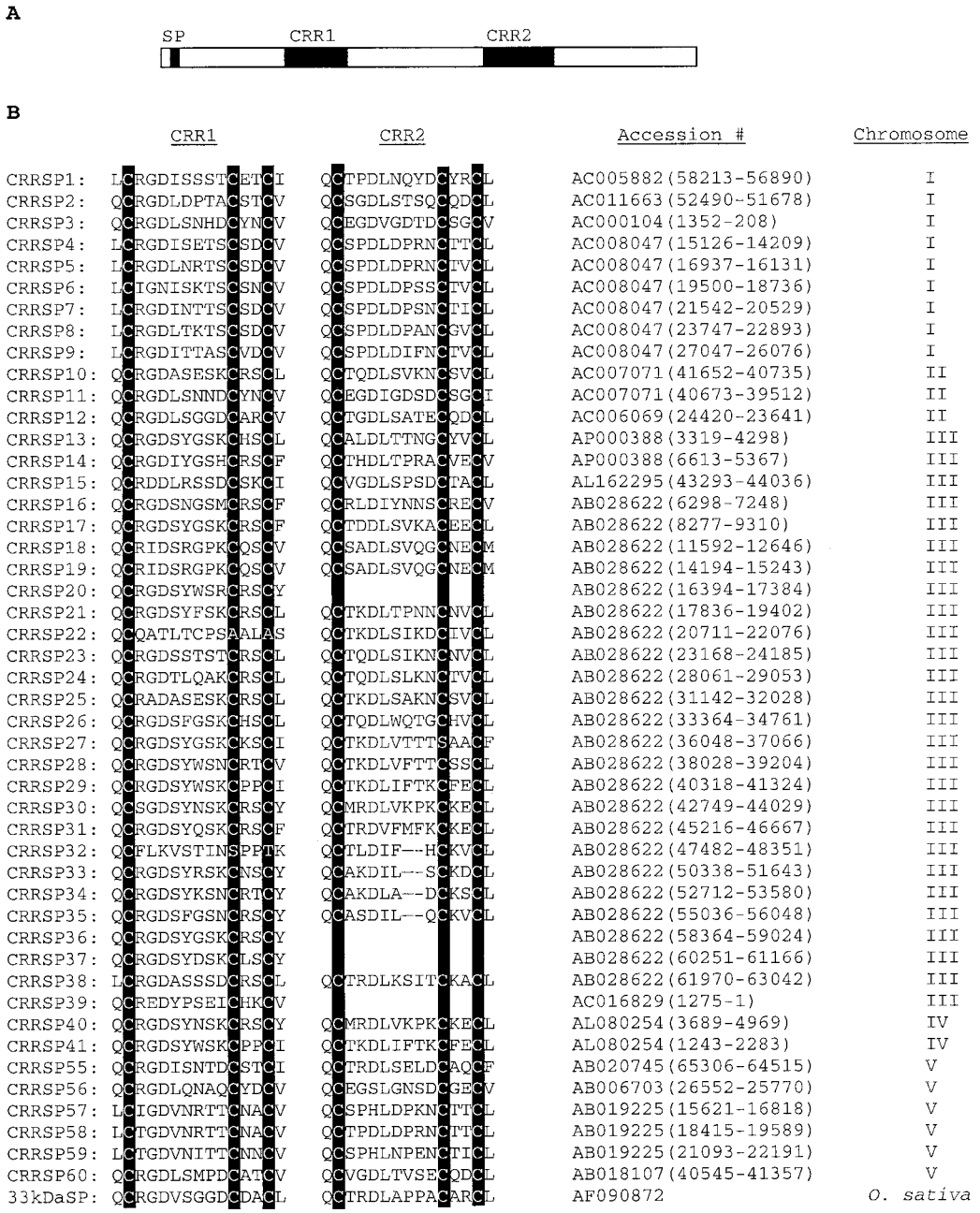

Figure 2. The CRRSP superfamily. A, Schematic diagram of the structures of common CRRSPs. SP, Signal peptide. B, Multiple alignment of the CRR motifs from 47 Arabidopsis CRRSPs and the rice (*Oryza sativa*) 33-kD secretory protein. The amino acid sequences of CRRSP42 to CRRSP54, encoded by a tandem array of genes on chromosome IV (accession no. AL080253) are almost identical to that of CRRSP41 and are not shown here. The locations of the Arabidopsis genes are indicated by the starting and stopping positions of the coding regions on the BAC clones according to the annotated genome sequence. The conserved Cys residues in CRR motifs are highlighted.

mosome IV, there are two large tandem arrays, one coding for 20 CRKs (CRK5 to CRK24) and the other encoding 13 CRRSPs (CRRSP42 to CRRSP54; Figs. 1 and 2). Chromosome III contains another gene array coding for 23 CRRSPs (CRRSP16 to CRRSP38; Fig. 2). Genome sequence analysis revealed DNA elements associated with retrotransposons flanking or within these gene arrays. For example, a gene coding for a protein similar to the copia-like retrotransposon Hopscotch polyprotein from maize is present within the gene array for CRK5 to CRK24 on chromosome IV. Sequences similar to non-long terminal repeat (LTR) retrotransposons are also identified in the regions within the gene array coding for CRK30 to CRK35 on chromosome IV. Sequences coding for proteins similar to non-LTR retrotransposon reverse

transcriptases are present in the regions flanking the two large tandem arrays coding for CRRSPs on chromosmes III and IV. In addition, a large number of DNA repeats, including those similar to LTRs, are present within the tandem arrays. These observations suggest that the large tandem arrays of genes coding for CRR proteins have probably evolved from retrotransposon-mediated gene duplication.

A challenge in the future studies will be to determine the biological functions of the hundreds of RLKs in plants. In Arabidopsis, the functions of only a few LRR RLKs are known; two are involved in regulation of cell division and differentiation in meristems (Torii et al., 1996; Clark et al., 1997), and one appears to be a brasssinosteriod receptor (Li and Chory, 1997; He et al., 2000). A gene encoding an LRR RLK protein in rice mediates resistance response to a bacterial pathogen (Song et al., 1995). There is no genetic analysis that links a mutant phenotype with mutations of genes encoding CRR proteins. Studies with several CRKs from Arabidopsis indicated that they were induced by pathogen infection, reactive oxygen species, and salicylic acid (Czernic et al., 1999; Du and Chen, 2000; Ohtake et al., 2000). The PvPK20-1in the roots of common bean is also differentially regulated in response to pathogens, symbionts, and nodulation factors (Lange et al., 1999). These observations suggest that at least some of the CRR protein superfamily is involved in plant perception and response to biotic and/or abiotic stress signals. In the self-incompatibility in Brassicaceae, the SLG and SRK are encoded by two of the genes at the same S locus (Stein et al., 1996; Dixit et al., 2000). Both proteins are coordinately expressed in the stigma and they may act in combination to form a receptor for some component of the pollen grain specified by the S locus (Stein et al., 1996; Dixit et al., 2000). The similar numbers of distinct CRKs and CRRSPs in Arabidopsis raise the possibility that these two groups of proteins may also act in combination in a manner similar to SRKs and SLGs.

Received January 12, 2001; returned for revision February 12, 2001; accepted February 21, 2001.

LITERATURE CITED

Clark SE, Williams RW, Meyerowitz EM (1997) Cell **89:** 575–585

Czernic P, Visser B. Sun W, Savoure A, Deslandes L, Marco Y, Van Montagu M, Verbruggen N (1999) Plant J **18:** 321–327

Deeken R, Kaldenhoff (1997) Planta **202:** 479–486

Dixit R, Nasrallah ME, Nasrallah JB (2000) Plant Physiol **124:** 297–311

Du L, Chen Z (2000) Plant J **24:** 837–848

Hardie DG (1999) Annu Rev Plant Physiol Plant Mol Biol **50:** 97–131

He Z, Wang ZY, Li J, Zhu Q, Lamb C, Ronald P, Chory J (2000) Science **288:** 2360–2363

He ZH, Fujik M, Kohorn BD (1996) J Biol Chem **271:** 19789–19793

Herve C, Dabos P, Galaud JP, Rouge P, Lescure B (1996) J Mol Biol **258:** 778–788

Lange J, Xie ZP, Broughton WJ, Vogeli-Lange R, Boller T (1999) Plant Sci **142:** 133–145

Li J, Chory J (1997) Cell **90:** 929–938

Nasrallah JB, Stein JC, Kandasamy MK, Nasrallah ME (1994) Science **266:** 1505–1508

Ohtake Y, Takahashi T, Komeda Y (2000) Plant Cell Physiol **41:** 1038–1044

Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, Gardner J, Wang B, Zhai WX, Zhu LH et al. (1995) Science **270:** 1804–1806

Stein JD, Dixit R, Nasrallah ME, Nasrallah JB (1996) Plant Cell **8:** 429–445

The Arabidopsis Genome Initiative (2000) Nature **408:** 796–815

Torii KU, Mitsukawa N, Oosumi T, Matsuura Y, Yokoyama R, Whittier RF, Komeda Y (1996) Plant Cell **8:** 735–746

Walker JC (1994) Plant Mol Biol **26:** 1599–1569

Wang XQ, Zafian P, Choudhary M, Lawton M (1996) Proc Natl Acad Sci USA **93:** 2598–2602