The Plant Journal (1999) **20**(3), 317–332

Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily

Blake C. Meyers[1,2,*,†], Allan W. Dickerman[3,†], Richard W. Michelmore[1], Subramoniam Sivaramakrishnan[1,4], Bruno W. Sobral[3] and Nevin D. Young[5,*]

[1]*Department of Vegetable Crops, University of California, Davis, CA 95616, USA,*
[2]*DuPont Agricultural Products – Genomics, PO Box 6104, Newark, DE 19714–6104, USA,*
[3]*National Center for Genome Resources, 1800 A S. Old Pecos Trail, Santa Fe, NM 87505, USA,*
[4]*International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Asia Center, Patacheru 502 324, Andhra, Pradesh, India, and*
[5]*Department of Plant Pathology and Department of Plant Biology, 495 Borlaug Hall, University of Minnesota, St. Paul, MN 55108, USA*

Summary

The nucleotide binding site (NBS) is a characteristic domain of many plant resistance gene products. An increasing number of NBS-encoding sequences are being identified through gene cloning, PCR amplification with degenerate primers, and genome sequencing projects. The NBS domain was analyzed from 14 known plant resistance genes and more than 400 homologs, representing 26 genera of monocotyledonous, dicotyledonous and one coniferous species. Two distinct groups of diverse sequences were identified, indicating divergence during evolution and an ancient origin for these sequences. One group was comprised of sequences encoding an N-terminal domain with Toll/Interleukin-1 receptor homology (TIR), including the known resistance genes, *N*, *M*, *L6*, *RPP1* and *RPP5*. Surprisingly, this group was entirely absent from monocot species in searches of both random genomic sequences and large collections of ESTs. A second group contained monocot and dicot sequences, including the known resistance genes, *RPS2*, *RPM1*, *I2*, *Mi*, *Dm3*, *Pi-B*, *Xa1*, *RPP8*, *RPS5* and *Prf*. Amino acid signatures in the conserved motifs comprising the NBS domain clearly distinguished these two groups. The *Arabidopsis* genome is estimated to contain approximately 200 genes that encode related NBS motifs; TIR sequences were more abundant and outnumber non-TIR sequences threefold. The *Arabidopsis* NBS sequences currently in the databases are located in approximately 21 genomic clusters and 14 isolated loci. NBS-encoding sequences may be more prevalent in rice. The wide distribution of these sequences in the plant kingdom and their prevalence in the *Arabidopsis* and rice genomes indicate that they are ancient, diverse and common in plants. Sequence inferences suggest that these genes encode a novel class of nucleotide-binding proteins.

Received 14 July 1999; revised 16 September 1999; accepted 20 September 1999.
*For correspondence (fax +530 752 9659;
e-mail bcmeyers@vegmail.ucdavis.edu; neviny@tc.umn.edu
†Equal contributions were made by the first two authors.

Introduction

The nucleotide binding site (NBS) is a common protein element essential for the catalytic activity of various prokaryotic and eukaryotic proteins. This functional domain, which occurs in several related structural forms, is required for ATP- and GTP-binding (Saraste *et al.*, 1990; Walker *et al.*, 1982). The primary sequence of the NBS is so distinct that protein sequences can be assigned to separate subgroups based on the conserved motifs found within the domain (Saraste *et al.*, 1990; Traut, 1994). These motifs are conserved in many, if not all, nucleotide binding proteins (Bourne *et al.*, 1991). The most common conserved motif is the phosphate-binding loop or 'P-loop' (also referred to as 'motif A' or 'Walker A', after Walker *et al.*, 1982), found in both ATP- and GTP-binding proteins (Saraste *et al.*, 1990). Additional conserved sites can also occur in nucleotide binding proteins (Traut, 1994) and these motifs are often used to classify sequences.

Genes that encode an NBS-containing sequence are common in plant genomes. For example, the superfamily of GTP-binding proteins is prevalent in plants, including signal transducing proteins such as the small GTP-binding proteins (reviewed in Bourne *et al.*, 1991; Downward, 1990) and the G-protein family (Kaziro *et al.*, 1991), as well as translational factors such as the *Arabidopsis tufM* (Kuhlman and Palmer, 1995). The majority of plant disease resistance genes (R-genes) cloned to date, including at least 14 genes from six plant species, also encode a predicted NBS region attached to a C-terminal leucine-rich repeat (LRR) of variable length (reviewed in Baker *et al.*, 1997; Bent, 1996; Hammond-Kosack and Jones, 1997). In the case of these R-genes, nucleotide binding has been

predicted based on sequence similarity only; their biochemical function(s) are yet to be demonstrated. NBS sequences of R-genes have been recognized by the presence of at least five conserved domains including a P-loop, indicating that they are related to the ATP- and GTP-binding superfamily of proteins (Grant *et al.*, 1995; Lawrence *et al.*, 1995; Ori *et al.*, 1997). Initial comparisons suggested that R-genes comprise of at least two groups. One group, including *N*, *L6*, *RPP5*, *M* and *RPP1*, encodes proteins containing a Toll/Interleukin-1 receptor homology region (TIR) N-terminal to the NBS (Anderson *et al.*, 1997; Botella *et al.*, 1998; Lawrence *et al.*, 1995; Parker *et al.*, 1997; Whitham *et al.*, 1994). Other R-genes do not encode a TIR domain, although sequence analysis predicted a leucine zipper motif in the N-terminal region of *RPS2*, *RPM1*, *Prf*, *Mi*, *RPS5* and *RPP8*, but not *Xa1*, *Dm3* or *PiB* (Bent *et al.*, 1994; Grant *et al.*, 1995; McDowell *et al.*, 1998; Meyers *et al.*, 1998; Milligan *et al.*, 1998; Mindrinos *et al.*, 1994; Salmeron *et al.*, 1996; Song *et al.*, 1995; Wang *et al.*, 1999; Warren *et al.*, 1998). In *Arabidopsis*, subsets of R-genes signal through either *EDS1* or *NDR1*; initial studies suggest that there may be an association between signaling pathways and structural features of the R-genes (Aarts *et al.*, 1998b).

Sequences with homology to the NBS of R-genes are currently being isolated from plants by three different methods. The original members of this group were identified as parts of disease resistance genes. The *Arabidopsis* genome sequencing initiative (http://genome-www.stanford.edu/Arabidopsis/) has uncovered additional genomic and cDNA sequences related to the NBS-LRR class of R-genes. Finally, R-gene candidate sequences have been amplified using degenerate oligonucleotide primers designed from conserved amino acids in the NBS motifs. Primers designed from the conserved P-loop and 'kinase' or 'GLPL' motifs have resulted in the isolation of numerous NBS homologs from a variety of plant species (Kanazin *et al.*, 1996; Leister *et al.*, 1996; Shen *et al.*, 1998; Yu *et al.*, 1996). Significantly, the genetic positions of these sequences are frequently at or near R-gene loci, indicating that these NBS sequences may be, or at least are related to, the linked R-genes (Kanazin *et al.*, 1996; Leister *et al.*, 1996; Shen *et al.*, 1998; Yu *et al.*, 1996). Although NBS sequences related to plant R-genes are being frequently discovered as part of genome initiatives and from amplification using PCR with degenerate primers, it is not currently known how exclusive this type of NBS sequence is to resistance gene products, and therefore it is difficult to infer function on the basis of sequence alone.

In this paper, we collected and analyzed hundreds of plant NBS sequences in order to examine NBS sequences isolated by the three methods described above. This enabled us to address a variety of questions regarding the evolution and function of the NBS domain in plants related to R-genes: (i) What proportions of plant genomes are comprised of NBS-encoding sequences? (ii) What are the evolutionary relationships among such NBS sequences? (iii) How many distinct groups of sequences have been isolated from plants and are they monophyletic? (iv) What are the characteristics of these groups? (v) How diverse are the members of these groups? (vi) Do groups identified by sequence analysis correlate with other characteristics of cloned R-genes? (vii) Are cladistic or sequence analyses useful to infer whether new sequences are likely to be resistance genes? (viii) Similarly, are there groups of related sequences that do not contain known R-genes and therefore may encode genes of different functions? (ix) In addition, does variation between the aligned sequences provide information on functional domains or conserved residues in the NBS?

Results

NBS sequences identified by database searches and PCR

Four hundred and eighty-one plant NBS sequences were identified from three sources: known R-genes, related NBS-encoding genes in public databases, and sequences isolated by PCR using degenerate oligonucleotide primers (Table 1). At the time of our analysis 14 R-genes had been cloned from plants and shown to be members of the NBS-LRR class of resistance genes (Table 1). A total of 146 related full-length sequences from *Arabidopsis* BAC and PAC clones were identified by BLAST analysis of the NR database. Little or no sequence-related bias is expected among these genes as they represent randomly generated genomic sequences from over 50% of the genome. Among dbGSS (BAC-end sequences), dbEST (expressed sequence tags) and dbHTGS (unfinished genomic sequencing), a total of 121, 32 and 3 sequences, respectively, were identified, largely from rice and *Arabidopsis*. Another 50 genomic clones came from a wide variety of plant species, isolated by means other than random genomic sequencing. These sequences included cloned genes of known function. Finally, 129 sequences derived from PCR amplification with degenerate primers fell into several categories; those recovered by BLAST searches of the NR database, both published and unpublished, as well as 25 unpublished NBS sequences identified by us or co-operating laboratories (GenBank accession numbers AF186623–AF186644). Due to the few primer sequences used, these collections of PCR-derived sequences potentially represent biased samples.

Information about these NBS-encoding sequences has been deposited and organized at the NCGR web site (http://www.ncgr.org/rgenes). This specialized, thematic database of plant NBS sequences includes links to the underlying database records and source, BLAST scores relating the NBS sequences to known R-genes, organisms, map positions in *Arabidopsis* when known, and graphic descriptions

Table 1. Summary of sequences in public databases showing homology to known plant R-genes

GENUS	BAC[a]	EST[b]	GSS[b]	HTGS[b]	PCR[c]	Genomic [d]	Total	Known R-genes
Arabidopsis	146	15	45	3	20	19	248	*RPM1, RPS2, RPP1, RPP5, RPP8*
Avena					2		2	
Brassica		2			2		4	
Cajanus					7		7	
Cicer					7		7	
Glycine		3			11		14	
Helianthus					2		2	
Hordeum					12		12	
Irvingia			1				1	
Lactuca					2	10	12	*RGC2B (Dm3)*
Linum						2	2	*L6, M*
Lycopersicon		6	1		1	6	14	*I2C, Mi (Meu1), PRF*
Medicago		1					1	
Nicotiana						1	1	*N*
Oryza		2	74		23	5	104	*Xa1, PiB*
Pennisetum					2		2	
Phaseolus					4	1	5	
Pinus					1		1	
Populus		1					1	
Prunus					2		2	
Saccharum		1					1	
Solanum					11	4	15	
Sorghum					5		5	
Triticum					6	1	7	
Vigna					1		1	
Zea		1			10	1	12	
Total	146	32	121	3	129	50	481	

[a]BAC = sequences isolated from BAC or PAC clones sequenced in the *Arabidopsis* genome initiative.
[b]EST, GSS and HTGS databases are described in Experimental procedures.
[c]PCR = sequences isolated using degenerate primers to amplify R-gene homologs.
[d]Genomic = sequences isolated from genomic clones, including cloned R-genes.

of motif organization (see below). A complete copy of the file can be downloaded in a tab-delimited format and analyzed on the user's machine with a typical spreadsheet or database program. We anticipate that the database will be updated periodically with regular curation.

These 481 NBS sequences originated from 26 different plant genera (Table 1). Eight of these genera came from monocots and 17 from dicots. Nine angiosperm families (one monocot and eight dicot) are represented. One sequence originated from conifers. Substantial numbers of NBS sequences from *Arabidopsis* and *Oryza* already reside in public databases due to genome sequencing initiatives; 50.4% of the 481 sequences came from *Arabidopsis* and 23.4% came from *Oryza*. A few of the *Arabidopsis* sequences were redundant (see below). The majority was from the genomic sequence of ecotype Col-0; some cloned R-genes and degenerate PCR products were isolated from other *Arabidopsis* ecotypes. In addition, the majority of the *Oryza* sequences are short fragments from BAC-end sequencing and therefore could only be used for analysis of genomic copy number (see below). Among the 481 sequences, there were also significant contributions from *Lycopersicon*, *Solanum*, *Glycine* and *Hordeum* (Table 1). Three non-plant sequences were identified, including two mammalian homologs of APAF-1 and a sequence from *Streptomyces* (accession number P25941). *Apaf-1* encodes a regulator of cell death with sequence similarity to plant R-genes (Van der Biezen and Jones, 1998).

Phylogenetic analyses revealed two distinct subfamilies

The complete set of sequences was filtered to remove partial and redundant sequences. Removal of these sequences simplified this analysis while retaining the overall level of sequence diversity. To make comparisons between sequences consistent, only the region between the P-loop and GLPL motifs was used for the phylogenetic analysis. Because of this constraint, sequences without full-length NBS domains were not included in the analysis. These sequences were primarily PCR products, but also included ESTs and BAC-end sequences that were trun-

cated. For example, all of the *Glycine* sequences uncovered in Yu *et al*. (1996), and some of the *Zea* sequences from Collins *et al*. (1998) were not included in the final dataset. This parsed set of NBS-encoding sequences was comprised of 248 members, of which 90 came from the *Arabidopsis* genome initiative, 30 from known R-genes or other genomic sources (including additional *Arabidopsis* sequences), and 113 were products of PCR amplification. Two alignment approaches were used that produced essentially similar alignments: an iterative CLUSTALW alignment (Thompson *et al*., 1994) and a hidden-Markov model alignment (Krogh *et al*., 1994). Because of the overall similarity of the alignment from the two algorithms (data not shown), we concluded that the alignments were robust in spite of numerous sequence differences in regions of high variation.

A neighbor-joining tree (Saitou and Nei, 1987) constructed for the parsed and aligned NBS sequences confirmed that these sequences fall into two well-supported, distinct groups (Figure 1a,b). This supports conclusions drawn previously from sequence comparisons and genetic data that indicate TIR-encoding R-genes comprise a distinct and more closely related class (Aarts *et al*., 1998b; Baker *et al*., 1997). These two major groups differed for sequences both N- and C-terminal to the NBS region (discussed in detail below in the context of the motifs identified by MEME analysis). The phylogenetic trees and further analyses of these two groups (TIR and non-TIR) were performed separately because the preponderance of data supports such a distinction and analysis of our entire set of sequences produced a tree with a well-supported split between the two groups (data not shown).

Both the TIR and non-TIR trees have long branch lengths and closely clustered nodes, reflecting a high level of sequence divergence (Figure 1a,b). Bootstrapping provided an estimate of the confidence for each branch point. The nodes closest to the branch tips were most highly supported, although increased support would probably be found for more of the internal nodes if the number of sequences was reduced. The trees are robust, however, as phylogenetic analysis using both distance and parsimony algorithms produced similar trees (data not shown). Furthermore, the major motifs identified monophyletic groups, providing further evidence that the trees are correct (see below). The distribution of motifs on the tree indicates that the TIR and non-TIR groups are mutually exclusive. Sequence P25941 from *Streptomyces* is used as an outgroup to root the trees. The non-plant proteins APAF-1 and CED-4 were not used in the phylogenetic analysis because they are more distantly related to plant NBS-encoding R-genes than the *Streptomyces* sequence (data not shown). It could not be determined which of the TIR and non-TIR predates the other, relative to the outgroup. Other than several *Arabidopsis*-specific branches, PCR-derived sequences were found in most branches on the trees, indicating that the degenerate primer approach successfully isolates a wide variety of sequences.

Products predicted from known disease resistance genes were found in distantly related clades distributed throughout both trees. There are only three major clades that did not include a product from a known R-gene. Therefore, most NBS-encoding genes were similar to at least one known resistance gene and consequently may encode resistance gene products of as yet unknown specificity. It is possible that some of these genes may have diverged to encode functions other than disease resistance, particularly in the clades that currently lack a known resistance gene product. However, because no function other than disease resistance has been attributed to any of these genes, this seems unlikely.

The TIR group (Figure 1a) contained several distinct subgroups of sequences, demonstrating recent diversification within a single species and within closely related plant families. This tree included the predicted products of known resistance genes *L6*, *M*, *N* and three linked genes from the *RPP1* cluster. Physical mapping information (Figure 1a) reveals that several types of duplication and divergence events have occurred. For example, the *RPP5* family at location 4.0548 on the *Arabidopsis* map has expanded locally on chromosome IV to 10 members, but branch lengths within this family suggest that either some members have changed more rapidly, or some duplication events are more ancient. These sequences are members of a much larger *Arabidopsis* group that includes at least 50 members. Because this group was monophyletic yet physically spread over four of the five *Arabidopsis* chromosomes, NBS-encoding sequences seem to have diversified, duplicated and moved extensively within this species; this is discussed in greater detail below. Several subgroups in the TIR group contained members from closely related species yet were also present multiple times within a single species; this indicates that NBS-encoding sequences have diverged both prior to and since speciation. Conservation of intron position had also been previously noted for the TIR group of cloned resistance genes (Hammond-Kossack and Jones, 1997; Parker *et al*., 1997). Although the TIR group contained the single *Pinus* sequence, no monocot sequences were found in this tree.

The second group of sequences was those lacking a TIR region; this included the products of the NBS-encoding genes *Prf*, *RPM1*, *RPS2*, *I2*, *Mi*, *Pi-B*, *Xa1*, *RPP8*, *RPS5* and *Dm3*, and was subdivided roughly into two subgroups (Figure 1b). Overall, branch lengths within these subgroups were comparable to those of the TIR tree, suggesting the level of sequence conservation within non-TIR sequences is similar to that of the TIR group. One well-supported subgroup contained *RPS2*, *RPS5* and *Dm3*; the other loosely defined subgroup included *Prf*, *Mi*,

I2, *Xa1*, *RPP8* and *RPM1*. Both subgroups contained several species- or monocot-specific clades; however, both also contained numerous paraphyletic branches. The monocot sequences in the non-TIR group were not monophyletic, indicating that the ancient ancestor of monocots and dicots contained numerous non-TIR resistance genes. An ancient origin for the non-TIR family is supported by a lack of intron conservation: *Prf*, *Mi* and *I2* all contain introns in and near the 5′ untranslated region, while *Prf* contains an additional intron within the encoded LRR (Milligan *et al.*, 1998; Salmeron *et al.*, 1996; Simons *et al.*, 1998); *Xa1* contains two introns located 5′ of the encoded NBS (Yoshimura *et al.*, 1998); *RPP8* contains three introns, two of which within the NBS-encoding region (McDowell *et al.*, 1998); *Dm3* contains seven introns (Meyers *et al.*, 1998); there are no introns in *RPM1* and *RPS2* (Hammond-Kosack and Jones, 1997).

Shorter NBS sequences were also recovered from the databases that originally had been obtained using degenerate primers to amplify sequences internal to those used in our phylogenetic analyses. We examined these sequences to determine if different types of sequences had been obtained due to the use of different primers. Nineteen sequences were obtained that had been amplified using primers designed between the P-loop and the GLPL (see below) motif (Collins *et al.*, 1998; Yu *et al.*, 1996). Phylogenetic analysis using these sequences showed no significant differences from the rest of the sequences we analyzed (data not shown); 17 were of the TIR class and two of the non-TIR class. There was no evidence that the use of degenerate primers for different motifs isolated a novel set of sequences. However, these sequences represent only a small subset of the total predicted for any single genome; therefore, combinations of different primers will probably be necessary to isolate a comprehensive range of sequences.

NBS sequences are abundant in plant genomes

To predict the total number of NBS-encoding sequences in *Arabidopsis* we analyzed the prevalence of such sequences in the BAC and PAC clones so far sequenced in the *Arabidopsis* genome project. Approximately 67.6×10^7 base pairs (~52%) of the *Arabidopsis* genome sequence had been deposited into the NR database at the time of our analysis. After the elimination of redundant sequences, 120 predicted gene products from *Arabidopsis* with partial or full homology to the NBS or TIR domains encoded by plant R-genes. Ninety of these NBS-encoding sequences included NBS motifs characteristic of TIR-associated sequences (see below) and 27 were non-TIR NBS-encoding sequences, while three could not be classified. Of the 90 with TIR features, 23 *Arabidopsis* sequences encoded only a TIR domain but no NBS region; no gene of known function has been characterized in plants with such a sequence. Based on a current estimate of the *Arabidopsis* genome of 1.30×10^8 bp (http://genome-www3.stanford.edu), and assuming a similar distribution of genes in the other ~50% of the genome, approximately 200 NBS-encoding genes are present in *Arabidopsis* (approximately 150 of the TIR-type and 50 of the non-TIR type). This would represent close to 1% of all *Arabidopsis* genes, assuming a total of 21 000 genes in the genome (Bevan *et al.*, 1998).

A similar estimate of NBS-encoding genes was performed with rice genomic sequences. This estimate was made in September 1999, from a database consisting of BAC-end sequences at Clemson University (www.genome.clemson.edu). A total of 41 862 rice BAC-end sequences were analyzed that comprised 3.1×10^7 bp. Seventy-four NBS-encoding sequences were identified. All were of the non-TIR type. These 74 sequences clustered into 54 unique groups. Sequences with greater than 96% identity in the first 400 bp were assumed to be the same gene. Because only 353 bp of the rice BAC-ends is estimated to be high quality sequence (www.genome.clemson.edu), we used two values for the number of base pairs analyzed: the 3.1×10^7 bp in the database, and the 1.48×10^7 bp of high quality sequence. By extrapolation,using an estimated genome size of 4.3×10^8 bp for rice (Arumuganathan and Earle, 1991), this suggests that there are 750–1550 non-TIR NBS-encoding sequences in the rice genome. This is four to eight times the total number predicted for *Arabidopsis*, and 15 to 31 times the estimated number of non-TIR sequences in the *Arabidopsis* genome. No TIR-encoding sequences were detected by BLASTX searches with this domain. In addition, we searched the DuPont database of wheat, maize, rice and soybean EST sequences (unpublished results), which contained more than 750 000 sequences at the time of the search (Table 2). Seventy-one unique EST sequences were identified from wheat, maize and rice with similarity to NBS-encoding sequences; however, none of these had homology to the TIR class of NBS-encoding sequences. In contrast, a search of approximately 183 000 DuPont soybean ESTs identified 24 NBS sequences, of which eight were similar to the NBS-encoding sequences of the TIR type (Table 2). Using the TIR domain only (excluding the NBS), BLASTX analysis of the DuPont database identified 41 soybean ESTs but none from wheat, maize and rice. Therefore, non-TIR NBS sequences are highly abundant in the rice genome, but TIR-encoding sequences are rare, absent, or have diverged beyond recognition in monocots.

Analysis of the physical map information demonstrated that NBS-encoding sequences tended to cluster in the *Arabidopsis* genome. We examined BAC and PAC clones to determine if sequences were near one another or randomly distributed through the genome; two or more NBS-encoding genes that mapped to the same position on

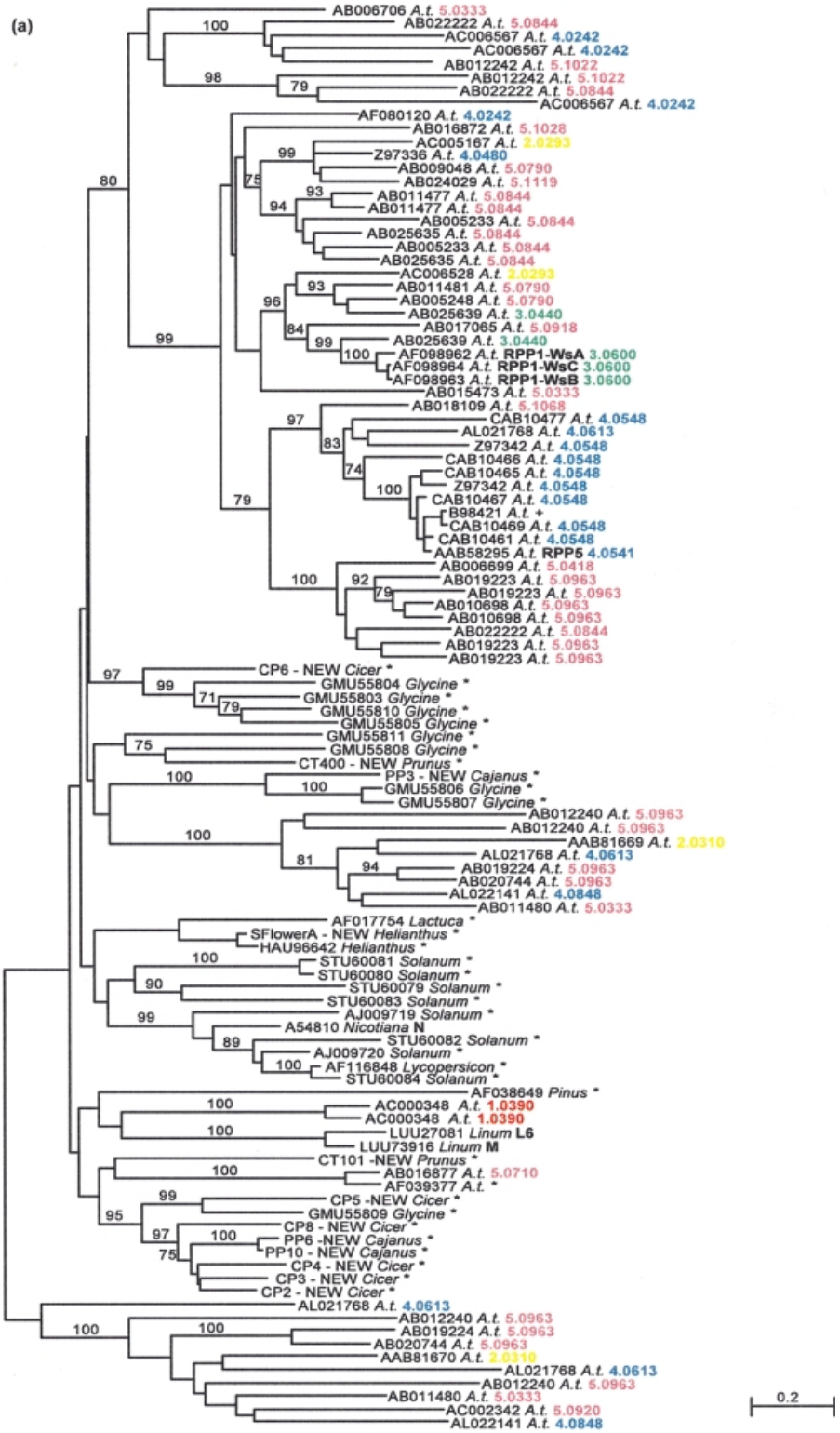
(a)
AB006706 A.t. 5.0333
AB022222 A.t. 5.0844
AC006567 A.t. 4.0242
AC006567 A.t. 4.0242
AB012242 A.t. 5.1022
AB012242 A.t. 5.1022
AB022222 A.t. 5.0844
AC006567 A.t. 4.0242
AF080120 A.t. 4.0242
AB016872 A.t. 5.1028
AC005167 A.t. 2.0293
Z97336 A.t. 4.0480
AB009048 A.t. 5.0790
AB024029 A.t. 5.1119
AB011477 A.t. 5.0844
AB011477 A.t. 5.0844
AB005233 A.t. 5.0844
AB025635 A.t. 5.0844
AB005233 A.t. 5.0844
AB025635 A.t. 5.0844
AC006528 A.t. 2.0293
AB011481 A.t. 5.0790
AB005248 A.t. 5.0790
AB025639 A.t. 3.0440
AB017065 A.t. 5.0918
AB025639 A.t. 3.0440
AF098962 A.t. RPP1-WsA 3.0600
AF098964 A.t. RPP1-WsC 3.0600
AF098963 A.t. RPP1-WsB 3.0600
AB015473 A.t. 5.0333
AB018109 A.t. 5.1068
CAB10477 A.t. 4.0548
AL021768 A.t. 4.0613
Z97342 A.t. 4.0548
CAB10466 A.t. 4.0548
CAB10465 A.t. 4.0548
Z97342 A.t. 4.0548
CAB10467 A.t. 4.0548
B98421 A.t. +
CAB10469 A.t. 4.0548
CAB10461 A.t. 4.0548
AAB58295 A.t. RPP5 4.0541
AB006699 A.t. 5.0418
AB019223 A.t. 5.0963
AB019223 A.t. 5.0963
AB010698 A.t. 5.0963
AB010698 A.t. 5.0963
AB022222 A.t. 5.0844
AB019223 A.t. 5.0963
AB019223 A.t. 5.0963
CP6 - NEW Cicer *
GMU55804 Glycine *
GMU55803 Glycine *
GMU55810 Glycine *
GMU55805 Glycine *
GMU55811 Glycine *
GMU55808 Glycine *
CT400 - NEW Prunus *
PP3 - NEW Cajanus *
GMU55806 Glycine *
GMU55807 Glycine *
AB012240 A.t. 5.0963
AB012240 A.t. 5.0963
AAB81669 A.t. 2.0310
AL021768 A.t. 4.0613
AB019224 A.t. 5.0963
AB020744 A.t. 5.0963
AL022141 A.t. 4.0848
AB011480 A.t. 5.0333
AF017754 Lactuca *
SFlowerA - NEW Helianthus *
HAU96642 Helianthus *
STU60081 Solanum *
STU60080 Solanum *
STU60079 Solanum *
STU60083 Solanum *
AJ009719 Solanum *
A54810 Nicotiana N
STU60082 Solanum *
AJ009720 Solanum *
AF116848 Lycopersicon *
STU60084 Solanum *
AF038649 Pinus *
AC000348 A.t. 1.0390
AC000348 A.t. 1.0390
LUU27081 Linum L6
LUU73916 Linum M
CT101 -NEW Prunus *
AB016877 A.t. 5.0710
AF039377 A.t. *
CP5 -NEW Cicer *
GMU55809 Glycine *
CP8 - NEW Cicer *
PP6 -NEW Cajanus *
PP10 - NEW Cajanus *
CP4 - NEW Cicer *
CP3 - NEW Cicer *
CP2 - NEW Cicer *
AL021768 A.t. 4.0613
AB012240 A.t. 5.0963
AB019224 A.t. 5.0963
AB020744 A.t. 5.0963
AAB81670 A.t. 2.0310
AL021768 A.t. 4.0613
AB012240 A.t. 5.0963
AB011480 A.t. 5.0333
AC002342 A.t. 5.0920
AL022141 A.t. 4.0848
0.2

Figure 1a.

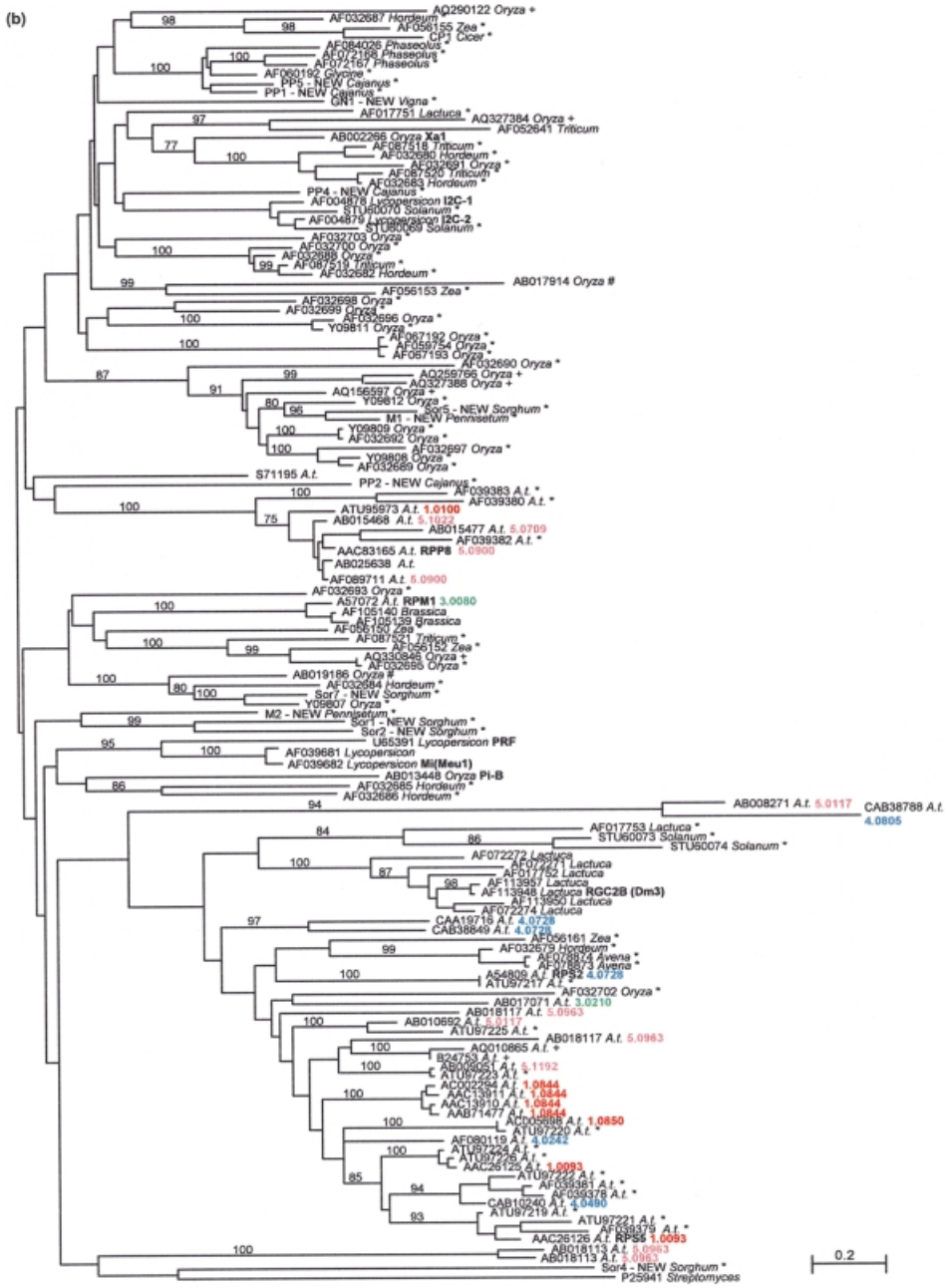

Figure 1. Phylogenetic relationship of TIR-class and non-TIR-class sequences.

Neighbor-joining trees from distance matrices were constructed according to Kimura's two-parameter method by using the aligned amino acid sequences between the P-loop and GLPL domains. The branch lengths are proportional to genetic distance in PAM units. IC, GenBank numbers, or unique names are given for each sequence followed by genus names in italics; *Arabidopsis* sequences are identified by *'A.t.'*. *Arabidopsis* sequences followed by a number are mapped; numbers refer to the genetic map location of the closest marker (the first digit is the chromosome number; digits after the decimal refer to the location on that chromosome to the tenth of a centiMorgan, such that 5.0963 indicates chromosome V, map location 96.3 cM). Colored map locations distinguish chromosomes. Map data were obtained from GenBank and AtDB (http://genome-www3.stanford.edu/). Names of known resistance gene products are indicated in bold. Sequences derived from PCR using degenerate primers are denoted by *; + indicates a GSS sequence; # indicates an EST sequence. Bootstrap values are indicated on branches for nodes supported with >70% of 100 replicates. Sequences without an accession number and followed by 'NEW' are novel PCR-derived NBS sequences. The prefix indicates the species from which they were derived: PP = *Cajanus cajan,* SFlower = *Helianthus annuus,* M = *Pennisetum glaucum,* GN = *Vigna subterranea,* CT = *Prunus persica,* Sor = *Sorghum bicolor,* CP = *Cicer arietinum.* All sequences are approximately 170 amino acids in length. These sequences were submitted to GenBank with accession numbers AF186623–AF186644. SflowerA was generated by M. Ayele Gedil, M.B. Slabaugh and S.J. Knapp (Oregon State University, Corvallis). CT101 and CT400 was generated by C. Thurmann and F. Bliss (University of California – Davis).

(a) TIR-class sequences.

(b) Non-TIR-class sequences.

Table 2. ESTs related to the NBS of plant R-gene products

	ESTs in DuPont database[a]				ESTs in public databases	
	Wheat	Maize	Rice	Soybean	*Arabidopsis*[b]	Rice[c]
Non-TIR	13	35	23	16	1	2
Non-TIR, clustered[d]	11	18	18	12	1	2
TIR	0	0	0	8	13	0
TIR, clustered[d]	0	0	0	8	11	0
Total ESTs searched	127 247	389 733	114 162	183 663	37 745	40 499

[a]Unpublished results.
[b]http://www.cbc.umn.edu/
[c]http://www.ncbi.nlm.nih.gov/dbEST/
[d]ESTs were clustered on the basis of BLASTN comparisons. To avoid over-estimating prevalence due to sequencing errors and multiple sequences from the same gene, sequences with more than 80% identity over a minimum overlap of 50 bp were considered potentially identical.

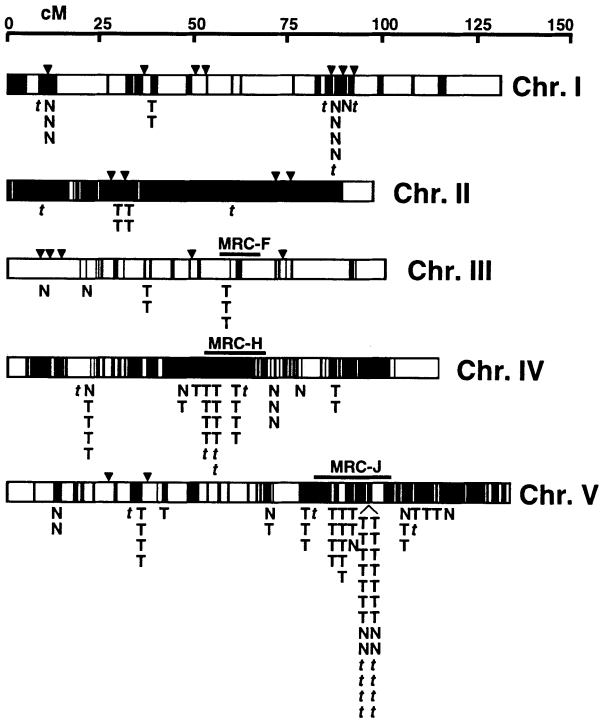

Figure 2. Arabidopsis sequences related to NBS-encoding plant R-genes.
Arrowheads above the chromosome designate the position of known *Arabidopsis* R-genes; regions MRC-F, H and J are clusters of known R-genes (Holub, 1997; Kunkel, 1996). Letters below each chromosome designate the approximate locations of 97 NBS-encoding sequences: 'T' indicates a TIR-type sequence; 'N' indicates a non-TIR sequence. Locations of an additional 22 sequences with similarity to the TIR only of NBS-encoding genes are indicated by a '*t*'. Sequencing progress of *Arabidopsis* at the time of the analysis is indicated by shading on the chromosomes: black indicates regions sequenced and in the databases; white regions had not yet been sequenced. Chromosome lengths are shown in centiMorgans. Modified from http://genome-www3.stanford.edu/cgi-bin/AtDB/Schrom.

the *Arabidopsis* genetic map were considered to be a 'cluster'. The 97 NBS-encoding genes were clustered in 21 clusters of two or more sequences along with 14 'singletons': this is consistent with previous observations (Aarts *et al.*, 1998a; Botella *et al.*, 1997; Speulman *et al.*, 1998). The number of NBS sequences per cluster ranged from two to 18, with an average of 4.9 (Figure 2). *Arabidopsis* chromosomes IV and V contained the largest NBS clusters and the highest number of NBS sequences (21 and 40, respectively). Two clusters between map positions 50 and 60 on chromosome IV contained 12 NBS sequences in total, 11 of which were of the TIR-type (described in part by Bevan *et al.*, 1998). A second highly populated region contained 30 sequences in six clusters on chromosome V (4.6 Mb; map positions 70–102). Thirteen additional open reading frames in this region encoded only a TIR domain but no NBS region. Numerous phenotypically defined resistance loci map to these regions on chromosomes IV and V (Holub, 1997; Kunkel, 1996). Another cluster of resistance loci has been identified on *Arabidopsis* chromosome III using classical genetic techniques (Holub, 1997; Kunkel, 1996); it is likely that sequence analysis of this region will also identify numerous NBS-encoding sequences

The relationship between phylogenetic position and chromosomal location is complex. In several cases, sequences are clustered both physically and phylogenetically, suggesting local duplications within gene families. Among the TIR group, the best examples are physically located at chromosomal positions 1.0390, 5.0963 and 4.0548 (Figure 1a). Several similar examples are found for the non-TIR sequences (e.g. 4.0728 and 1.0844; Figure 1b), although this tree is less populated with informative sequences. Some clusters contain very diverse sequences; for example, the largest cluster, at position 5.0963, includes four non-TIR and 12 TIR sequences. Some clades include both localized and dispersed sequences; for example, the clade containing the large gene family at 4.0548 is most closely related to several sequences on chromosome V.

Table 3. Conserved amino acids of R-genes and sequences used to search databases: motifs present in the N-terminal region of the aligned sequences

Gene or identifier[a]	TIR-1	TIR-2	TIR-3	Pre-P-loop
Consensus[b]:	VFPSFRGEDVRKTFLSH	YASSSWCLDEL	VIPIFYKVDPSDVRKQTGEFG	VRMVGIWG
N	$_{13}$VFLSFRGEDTRKTFTSH	$_{76}$YATSRWCLNEL	$_{100}$VIPIFYDVDPSHVRNQKESFA	$_{208}$VRIMGIWG
RPP5	$_{13}$VFPSFSGVDVRKTFLSH	$_{75}$YASSTWCLNEL	$_{99}$VIPVFYDVDPSEVRKQTGEFG	$_{208}$ARMVGIWG
L6	$_{62}$VFLSFRGPDTREQFTDF	$_{125}$YADSKWCLMEL	$_{150}$ILPIFYMVDPSDVRHQTGCYK	$_{257}$VTMVGLYG
M	$_{77}$VFLSFRGPDTRYQITDI	$_{140}$YADSKWCLMEL	$_{165}$IIPIFYMVDPKDVRHQTGPYR	$_{272}$VTMVGLYG
AB012242.7[c]	$_{15}$VFLSFRGEDTRRTIVSH	$_{78}$YTTSRWCLMEL	$_{102}$VLPLFYEVDPSDVRHQRGSFG	$_{207}$VCMVGIWG
AB006706.2[c]	$_{19}$VFVSFRGEDVRKTFVSH	$_{82}$YAASSWCLDEL	$_{103}$IVPIFYEVDPSDVRRQRGSFG	$_{206}$VRMLGIWG
AB012242.8[c]	$_{13}$VFLSFQGLDTRRTFVSH	$_{76}$YASSPLCLDSL	$_{100}$LIPIFYEVDPMDVRKQIGKLY	$_{207}$VRHIKIWG

[a]Predicted proteins from genes *I2, Xa1, RGC2B/Dm3, RPM1, RPS2, RPS5, RPP8, Mi/Meu1, Prf,* 3451069 and AB008271.3 are not shown as no TIR motifs were present.
[b]The consensus is given only as an approximation of the profiles produced from the MEME analysis (complete data are available at http://www.ncgr.org/rgenes).
[c]Hypothetical proteins selected from the phylogenetic tree to represent diverse NBS-encoding sequences.

This suggests that sequences may have moved between chromosomes IV and V. Evidence of more complex duplication events exists: there are clades containing members from different chromosomal locations; for example, sequences at 5.0844 are found paired with 4.0242 in three clades (Figure 1a). An even more complex example is seen in two clades that each contain sequences from locations 5.0963, 4.0848, 2.031 and 4.0613 (Figure 1a). These data suggest that regions containing relatively dissimilar genes have been duplicated within the *Arabidopsis* genome.

NBS-encoding sequences that are related to plant R-genes are rare in EST databases (Table 2). Despite the prevalence of NBS-encoding genes in genomic sequences, only 14 *Arabidopsis* EST sequences were found among 37 000 *Arabidopsis* ESTs in the public databases. ESTs with homology to the LRR (in the 3′ end of R-genes) are more abundant in databases (Botella *et al.*, 1997). However, not all LRR-encoding genes contain NBSs and many are not R-genes (Jones and Jones, 1997). Seven of the 15 ESTs corresponded to NBS sequences on finished BAC or PAC clones, one of which was found three times (from PAC clone MHK7). These five unique ESTs came from chromosomes IV and V, where the majority of the NBS-encoding genes are located. For rice, the public databases contained only two related sequences among 35 000 ESTs. Among 750 000 entries in the DuPont EST database for maize, wheat, sorghum and soybean, 87 NBS-encoding sequences were found, representing 67 non-redundant sequences (Table 2). This low abundance may be the consequence of a combination of low expression levels and several experimental artifacts. Many disease resistance genes are large genes and cDNAs may be prematurely truncated at the 5′ end in libraries that are synthesized using 3′ cDNA priming. Normalization of libraries prior to sequencing will tend to result in over- rather than under-representation. Estimations of expression levels from prevalence in EST libraries are further complicated because the EST collections represent a pool of libraries from a large number of tissues and developmental stages that may express R-genes at different levels.

Several conserved motifs characterize the NBS region in plant resistance genes

Seven major motifs were identified in the NBS region, some of which were specific to the non-TIR class of predicted proteins. To search for patterns adjacent to and including the previously defined P-loop (kinase-1a) and kinase-2 motifs (Traut, 1994), we used a weighted matrix approach based on the MEME algorithm (Bailey and Elkan, 1994). Log-likelihood matrices, derived from the corresponding amino acid frequency matrices, were then used to identify putative motifs within the input sequences (Tables 3 and 4). The 10 most significant motifs identified by the MEME software were considered robust since the least significant of these 10 (TIR1 and an LRR repeat motif) had been described previously (Baker *et al.*, 1997; Jones and Jones, 1997). The next 20 were not found universally in either the TIR or non-TIR sequence sets, although several were specific to single branches within the trees (data not shown). For motif names, previously defined terms were used; an arbitrary descriptor, Resistance Nucleotide Binding Site (RNBS), was used to name novel motifs that may be unique to plant R-genes (Table 4). The RNBS-B motif may play the same functional role as the previously defined kinase-3 motif (interaction with the purine or ribose; Traut, 1994); however, we have assigned it a distinct name because there is positional but no sequence similarity between RNBS-B and kinase-3 motif.

Two domains within the NBS clearly distinguished TIR and non-TIR sequences; we have called these novel motifs

Table 4. Conserved amino acids of R-genes and sequences used to search databases: motifs present in the NBS-encoding region of the aligned sequences

Gene or Identifier	P-loop/ Kin-1a	RNBS-A - nonTIR	RNBS-A - TIR	Kin-2	RNBS-B	RNBS-C	GLPL	RNBS-D - NonTIR[a]	RNBS-D - TIR[a]
Consensus:	GVGKTT	FDLxAWVCVSQxF	FLENIRExSKKHGLEHLQKKLLSKLL	LLVLDDVW	GSRIIITTRD	YEVxxLSEDEAWELFCKxAF	CGGLPLA	CFLYCALFPED	FLHIACFF
I2	$_{203}$GQGKTT	$_{223}$FDLKAWYCVSEGF		$_{277}$LIVLDDVW	$_{304}$GSKIIVTTRK	$_{325}$IRMGNLSTEASWSLFQRHAF	$_{366}$CKGLPLA	$_{423}$CFSFCAIFPKD	
Xa1	$_{328}$GIGKTT	$_{348}$FNVKIWVYVSDKF		$_{401}$LIVLDDVW	$_{437}$GNMIILTTRI	$_{459}$IKLEALKDDDIWSLFKVHAF	$_{500}$LKGNPLA	$_{559}$CVSYCSLFPKG	
RGC2B/Dm3	$_{182}$GVGKTR	NI		$_{259}$LIVLDDVW	$_{285}$DFKVLLTSRD	$_{309}$INVGLLTEAEAQSLFQQFVE	$_{346}$CCGLPIA	$_{402}$TFLMCGLFPED	
RPM1	$_{202}$GSGKTT	$_{222}$FESYAWVTISKSY		$_{281}$IVVLDDVW	$_{306}$GSRVMMTTRD	$_{330}$HEIELLKEDEAWVLFSNKAF	$_{372}$CQGLPLA	$_{433}$CFLYCSLFPVN	
RPS2	$_{184}$GVGKTT	$_{205}$YDVLIWVQMSREF		$_{257}$LLLLDDVW	$_{282}$KCKVMFTTRS	$_{304}$LRVEFLEKKHAWELFCSKVW	$_{344}$CGGLPLA	$_{406}$CFLYCALFPEE	
RPS5	$_{185}$GVGKTT	$_{206}$FDVVIWVVVSRSS		$_{260}$VLLLDDIW	$_{285}$GCKVAFTTRS	$_{307}$MEVSCLQPEESWDLFQMKVG	$_{347}$CRGLPLA	$_{410}$CFLYCSLFPED	
RPP8	$_{194}$GIGKTT	$_{214}$FDGFAWVCVSQQF		$_{268}$LVVLDDVW	$_{292}$GWKMLLTSRN	$_{315}$FRASILNPEESWKLCERIVF	$_{359}$CGGLPLA	$_{425}$RFLFLAHFPEY	
Mi/Meu1	$_{561}$GSGKTT	$_{571}$FDLRAWCTVDQGC		$_{623}$LIVLDDVW	$_{648}$GSRIILTTRE	$_{671}$LDLRLLRPDESWELLEKRAF	$_{710}$CKGLPLV	$_{776}$CLLYFASFPKD	
3451069[c]	$_{436}$GIGKTE	NI		$_{518}$(FLVVIDNL)[d]	$_{546}$GTHILISTRL	NI	$_{598}$(DKLGRLT)	NI	
AB008271.3	$_{209}$GSGKTT	NI		$_{272}$LVILDDVW	NI	$_{312}$YDVELLNEHEATALFCLSVF	$_{352}$CKGLPLS	$_{413}$CFLVLGAFPED	
N	$_{218}$GVGKTT		$_{247}$FLKDIKENKRGMHSLQNALLSELL	$_{296}$LIVLDDID	$_{322}$GSRIIITTRD	$_{342}$YEVTALPDHESIQLFKQHAF	$_{383}$AKGLPLA		$_{437}$FLDIACFL
RPP5	$_{218}$GIGKST		$_{242}$(FLTYKSTSGSDVSGMKLSWQKELLSEIL)	$_{291}$LILLDDVD	$_{316}$GSRIIVITQD	$_{338}$YEVKLPSQGLALKMISQYAF	$_{377}$(VGSLPLG)		$_{433}$FKCIACFF
L6	$_{267}$GIGKTT		$_{290}$FIDNIRETQEKDGVVVLQKKLVSEIL	$_{344}$LVVLDDVD	$_{369}$QSRFIITSRS	$_{393}$YEVGSMSKPRSLELFSKHAF	$_{432}$TAGLPLT		$_{489}$FLDIACFF
M	$_{282}$GIGKTT		$_{305}$FVDNVRAMQEQKDGIFILQKKLVSEIL	$_{358}$LVVLDDVD	$_{383}$GTRFIITSRN	$_{407}$YEVGSMSEQHSLELFSKHAF	$_{446}$TGGLPLT		$_{503}$FLDIACFF
AB012242.7[c]	$_{217}$GIGKTT		$_{240}$FIEDVGQICKKVDLKCIXXLLCDIL	$_{292}$LFVLDGVD	$_{317}$GSRIIITTRD	$_{339}$YEVKCLQNEDSLKIVKNIAF	$_{379}$AQGLPLA		$_{436}$FIRVACLF
AB006706.2[c]	$_{216}$GVGKTT		$_{239}$FMENVKEVCNRYGVRRLQVEFLCRMF	$_{291}$FIVLDDVD	$_{316}$GSRIIVTTRD	$_{338}$YKVKCLPKKEALQLFCNYAF	$_{378}$ASGLPLA		$_{434}$FLYISCFY
AB012242.8[c]	$_{217}$(DIGKTE)		$_{240}$(VMLKAPQRISRFEEVRLAEYVCLRL)	$_{281}$LLVLDNVN	$_{307}$GSRIITTTRN	$_{329}$YEVLGLEFSEALQLFCLHAF	$_{368}$AGGFPLS		NI

[a]These motifs are found in the region of the protein sequence but been listed separately because the sequences are distinct.
[b]NI = not identified. These motifs were not found using MEME analysis or by visual inspection.
[c]Hypothetical protein – selected from the phylogenetic tree to represent diverse NBS-encoding sequences.
[d]Sequences in parentheses were identified by visual inspection but were below the threshold for recognition using MEME analysis.

'RNBS-A' and 'RNBS-D' (Table 4). The RNBS-D motif was especially well-conserved among non-TIR sequences and could be used to design novel degenerate primers for amplification of NBS sequences without N-terminal TIR domains (S. Penuela and N.D. Young, unpublished results). The final residue in the kinase-2 motif can be used to predict the presence of the TIR domain in more than 95% of the cases: a tryptophan residue (W) was found in the non-TIR class, corresponding to an aspartic acid residue (D) in the TIR class (Table 4). Although not the focus of this study, it was also apparent from the MEME analysis that conserved motifs also could be identified in the TIR domain N-terminal to the NBS region. Regions of sequence conservation have been noted previously in the TIR region of cloned resistance genes (Hammond-Kosack and Jones, 1997). In the current analysis, three distinct and highly conserved motifs were uncovered within the TIR domain (TIR-1, TIR-2, and TIR-3; Table 3) and these motifs, compared across all *Arabidopsis* homologs, were found in a surprisingly wide array of combinations and permutations (http://www.ncgr.org/rgenes). Similar motifs are found in Toll-related proteins (Rock *et al.*, 1998). An additional TIR-specific motif was found adjacent to the P-loop (the 'Pre-P-loop' motif listed in Table 3). The motifs in the TIR domain were frequently duplicated and/or reorganized at the N-terminus of NBS sequences; eight different motif arrangements preceding the NBS were observed among known R-genes and full-length NBS sequences derived from the *Arabidopsis* (data not shown).

Several clades were characterized by the appearance of a motif along a unique lineage. Motifs detected for each sequence were converted into a format that could be easily visualized, searched and sorted (http://www.ncgr.org/rgenes). Nearly all NBS sequences (both TIR and non-TIR) had an ordered backbone consisting of P-loop/kinase-2/RNBS-C/RNBS-D/GLPL spanning the NBS domain (http://www.ncgr.org/rgenes). The P-loop (same as 'kinase-1a' or 'G-1') and kinase-2 (or 'G-3') domains have been well characterized in the NBS of many ATP- and GTP-binding proteins (Bourne *et al.*, 1991; Traut, 1994; Walker *et al.*, 1982). These two motifs were expected to have high levels of conservation as they are important domains for the binding of the nucleotide triphosphate (Traut, 1994). The P-loop interacts directly with the phosphate of the bound NTP (Saraste *et al.*, 1990), while the kinase-2 domain contains an aspartate critical for co-ordinating the metal ion (Mg^{2+}) required for phospho-transfer reactions (Traut, 1994). The leucine zipper motif has been reported in RPS2, RPM1, Prf, Mi, RPS5 and RPP8 (Bent *et al.*, 1994; Grant *et al.*, 1995; McDowell *et al.*, 1999; Milligan *et al.*, 1998; Mindrinos *et al.*, 1994; Salmeron *et al.*, 1996; Warren *et al.*, 1998) and has been proposed as characteristic of the non-TIR group of R-gene homologs. We searched the *Arabidopsis* full-length sequences for the pattern of four heptad repeats which define this motif (PROSITE pattern PS00029); only a single match was found. This suggests that the leucine zipper motif is not a reliable characteristic motif of the non-TIR group of R-gene homologs.

Note added in proof: The coiled-coil motif, of which the leucine zipper is a specific example, has recently been identified N-terminal to the NBS of most but not all non-TIR resistance gene products (Pan *et al.*, 1999). Therefore, the coiled-coil motif may be a useful diagnostic motif for the non-TIR group.

Discussion

NBS sequences are common in plant genomes

NBS-encoding sequences are members of one of the largest and most diverse families of genes currently known in plants. We recovered 481 NBS-encoding sequences from the public databases. Data from the *Arabidopsis* genome-sequencing project enabled us to estimate that approximately 200 NBS-encoding genes will eventually be found (although sequence analysis suggests that a significant proportion may be pseudogenes). This implies that nearly ~1% of all *Arabidopsis* genes are committed to just this one component of disease resistance. Physical mapping data revealed that a large number of these sequences are clustered near one another. This suggests that many of these genes have arisen through duplication events. Duplication, divergence and deletions are thought to play important roles in the evolution of this class of genes (Michelmore and Meyers, 1998). In addition to the NBS-encoding genes, plant genomes are likely to contain numerous homologs of other R-genes that do not encode an NBS, such as *Xa21* (Song *et al.*, 1995), *Cf2/5* and *Cf4/9* (Dixon *et al.*, 1996; Hammond-Kosack *et al.*, 1998; Jones and Jones, 1997; Thomas *et al.*, 1997), and *Mlo* (Buschges *et al.*, 1997). In addition, many other gene products are necessary to participate in the signal transduction cascade and mediate the defense response, and therefore a substantial fraction of plant genomes is dedicated to defense.

Two distinct types of R-gene related NBS sequences exist in plants

We used a combination of analytical methods to organize NBS-encoding sequences into biologically meaningful groups and then to characterize them. Valid phylogenetic analyses are dependent upon an alignment of sequences that reflects real biological events. However, for large numbers of diverse sequences there is no methodological consensus on the optimal approach for constructing the most accurate alignment. Alignments are relatively straightforward for sequences that are closely related, such as the plant actin genes (McDowell *et al.*, 1996) or the

developmental MADS-box genes (Theissen *et al.*, 1996). The difficulty in aligning diverse sequences can diminish the accuracy of phylogenetic analysis (Nei, 1996). The NBS sequences identified in this analysis have a high level of diversity; this is most apparent in the long branch lengths of the tree and in the multiple sequence alignment (http://www.ncgr.org/rgenes). This level of diversity suggests that amino acid substitutions have reached saturation at non-conserved residues and that indels (insertions/deletions) have accumulated in these sequences. Advances in alignment algorithms, including the use of hidden Markov models (Krogh *et al.*, 1994) and genetic algorithms (Notredame and Higgins, 1996), complement and may ultimately provide superior alignment results to the common 'progressive alignment' approach (Feng and Doolittle, 1987). In our analysis, CLUSTALW and a hidden Markov model produced essentially similar alignments giving us confidence that we were close to the optimal alignment.

Phylogenetic analysis indicates that plant NBS domains can be categorized into two major types, the TIR and non-TIR containing groups. The TIR branch of the tree consists exclusively of sequences from dicot species, including a large number of *Arabidopsis* sequences. The high proportion of *Arabidopsis* genes reflects the relative abundance of *Arabidopsis* genomic sequences in the databases. However, our analysis demonstrates that within *Arabidopsis*, TIR sequences outnumber the non-TIR sequences by a factor of three. In contrast, all the NBS sequences available from monocots are all non-TIR and largely found in one of two non-TIR subgroups. Further analysis of the *Oryza* BAC-end sequences at Clemson University and the DuPont EST databases provided further evidence for the absence of detectable TIR-encoding sequences; sequencing of the rice genome will provide comprehensive data on the absence of TIR-containing sequences in this grass species. It is now important to analyze a variety of other monocot species outside the *Poaceae* to determine whether the absence of non-TIR encoding genes is a characteristic of all monocot species. It is also interesting that a *Pinus* sequence was found among the TIR-containing sequences. Assuming that this PCR product is not an experimental artifact, this sequence suggests that TIR-NBS-encoding genes pre-date the divergence of angiosperms and gymnosperms. Furthermore, sequences with homology to TIR motifs are abundant in both *Drosophila* and mammalian genomes (Gay and Keith, 1991; Rock *et al.*, 1998); therefore, the TIR motif is ancient. Although many of these mammalian proteins also contain LRRs, outside of plants, only one gene encoding an NBS has been found with weak homology to a TIR motif (Aravind *et al.*, 1999). If TIR sequences are truly absent from monocots, this family of genes must have either evolved rapidly beyond recognition or have been lost from monocots early in the evolution of this taxon. Furthermore, the TIR and NBS domains must have been linked either before or early in the evolution of angiosperms.

All NBS sequences, both TIR and non-TIR, have a highly conserved backbone of amino acid motifs. TIR-containing sequences, however, lack two motifs within the NBS domain found in non-TIR sequences. The motifs in common between TIR and non-TIR sequences also differ in specific residues, most notably the kinase-2 motif. Therefore, these residues are predictive of the N-terminal features of the protein. With a similar approach, we characterized three conserved motifs within the TIR region, although the order of these motifs within the TIR region is not highly conserved (http://www.ncgr.org/rgenes). Sequences that are related to the *Drosophila* Toll protein and active in host defense have also been isolated from insects and mammals (Qureshi *et al.*, 1999; Rock *et al.*, 1998). The interaction of specific ligands with the TIR regions of these proteins activates intracellular signaling pathways leading to a cascade of cytoplasmic events (Hoffmann *et al.*, 1999). The relative conservation of the NBS may be critical to the activation of the signal transduction pathway involving hydrolysis of NTPs, while the variation in the TIR organisation could lead to diverse specificities.

NBS sequences are being increasingly isolated from a variety of plant species using PCR primers designed to the conserved motifs of the NBS region. In *Arabidopsis*, this approach produced largely non-TIR sequences related to *RPS2* (Aarts *et al.*, 1998a; Speulman *et al.*, 1998), while in soybean (*Glycine max*) most sequences were TIR-related (Kanazin *et al.*, 1996; Yu *et al.*, 1996). In these experiments, only five to 11 different groups of sequences were isolated; the TIR/non-TIR bias may be due to low sample sizes or may result from the slightly different primers used by each laboratory. With the large number of sequences available and grouped into several clades, it is now possible to redesign degenerate primers to isolate specific classes of R-gene homologs. It should be possible to design primers to the motifs defined in Tables 3 and 4 for the specific amplification of particular groups of NBS-encoding sequences. Comprehensive attempts to isolate the full range of NBS-encoding sequences from a particular genome will require the use of a variety of primer pairs and PCR conditions.

Plant NBS-LRR R-genes encode a novel class of nucleotide binding proteins

The P-loop and other NBS motifs are highly conserved within protein families of related function. Consequently, it is possible to use a sequence consensus from ATP- and GTP-binding proteins to categorize and predict the func-

tion of a novel NBS protein based solely on its sequence (Saraste *et al.*, 1990; Traut, 1994). A search of the SwissProt databases using an example set of seven P-loop consensus sequences identified distinct families of ATP- and GTP-binding proteins (Saraste *et al.*, 1990). All of these families were based on conserved residues in a 15 amino acid sequence around the P-loop, and the conservation of the sequence signatures was strict enough to identify and correctly predict the function of over 160 protein sequences based on the partial sequence of the NBS contained therein (Saraste *et al.*, 1990). A more specific classification of members of the GTPase superfamily could similarly identify members of functional groups from a diverse set of genes based on conservation within the four conserved motifs of the GTPase nucleotide binding site (Bourne *et al.*, 1991). The sequence signatures in the conserved motifs are probably no less specific to the various classes of NBS-encoding plant genes. Our sequence analysis of diverse R-gene homologs identified novel consensus sequences (Table 3), indicating that this is a new class of nucleotide binding proteins. The G-4 region that is characteristic of the GTPase superfamily (Bourne *et al.*, 1991) and critical for specific recognition of the guanine base (Hopp, 1995) is absent in the NBS of sequences used in this study. This suggests that plant R-gene homologs do not belong to the GTPase superfamily; therefore, these gene products may bind ATP rather than GTP. Interestingly, biochemical data for the related mammalian *Apaf-1* gene product demonstrate that the NBS forms oligomers through self-association (Srinivasula *et al.*, 1998), indicating that the NBS may be important for both nucleotide binding and oligomerization. However, this has yet to be demonstrated for any plant R-genes.

The consistent structural arrangement of motifs suggests that the NBS domains of R-gene products have similar or identical biochemical functions. However, the multiple sequence alignment demonstrates that segments between the conserved motifs are highly divergent (http://www.ncgr.org/rgenes). One possible explanation for the variability among plant NBS sequences is that regions between conserved motifs are not critical to biochemical function and therefore diverse sequences can function in similar ways. Previous alignments of diverse NBS-encoding sequences from a variety of organisms identified conserved residues critical to nucleotide binding (Bourne *et al.*, 1991; Saraste *et al.*, 1990; Traut, 1994; Walker *et al.*, 1982). The topology of these domains is quite similar over a variety of proteins, although the conserved regions are often separated by highly diverse sequences (Bourne *et al.*, 1991). Alterations in the size and composition of the residues that intervene between the conserved motifs may subtly alter the position of the bound ATP (Traut, 1994). If the spacing and relative position of the motifs is conserved, the underlying activity of the NBS is maintained and insertions of varying sizes are tolerated in the loops that connect the critical residues (Milner-White *et al.*, 1991).

It is possible that some of the differences between the NBS sequences encoded by R-genes influence the specificity of interactions either with a pathogen elicitor or with proteins downstream in the signal transduction cascade. Diversity among NBS sequences could be critical to directing participation in the distinct signaling pathways that have been identified by genetic analysis (Aarts *et al.*, 1998b). Receptor-like functions, such as protein binding and recognition, are probably determined elsewhere in the protein, possibly by the LRR. The LRR is known to be involved in protein–protein interactions and ligand-binding in diverse proteins (Kobe and Deisenhofer, 1994). Biochemical evidence of this binding by the LRR has yet to be reported for R-gene products. Sequence comparisons in cloned R-genes suggest that the LRR region rather than the NBS domain determines specificity to pathogen elicitors (Botella *et al.*, 1998; Dangl and Holub, 1997; Hammond-Kosack and Jones, 1997; McDowell *et al.*, 1998; Meyers *et al.*, 1998; Warren *et al.*, 1998). The TIR domain may also be responsible for some aspects of pathogen recognition; the allelic *L6* and *L7* R-genes of flax differ only in their N-terminal (TIR) region, yet have different specificities (Ellis *et al.*, 1999). Further biochemical and genetic analyses as well as studies utilizing site-directed mutagenesis and X-ray crystallography are necessary to better define the structural and functional roles of particular motifs and of diversity within the NBS domain.

With this and future characterizations of NBS sequences, it is becoming possible to predict the class and function of new sequences. The profiles derived from the sequences used in this study are available through the World Wide Web (http://www.ncgr.org/rgenes). Newly identified sequences can be compared and classified using these profiles and by analysis of regions flanking the NBS, such as the 5′ end that may encode a TIR domain. Further sequence analysis in *Arabidopsis* and other plant species may reveal additional families of closely related NBS sequences. However, our study utilized half of the *Arabidopsis* genome plus hundreds of NBS sequences from other sources, so it seems unlikely than any major classes will have escaped our analysis. Phenotypic characterization of plant NBS-encoding genes will determine if these sequences play a functional role outside of disease resistance.

Experimental procedures

Isolation of sequences using degenerate primers

DNA templates were prepared using genomic DNA isolated from each plant species using a modified CTAB extraction protocol (Bernatzky and Tanksley, 1986). PCR conditions were as previously described using the degenerate oligonucleotide primers from Shen *et al.* (1998).

Similarity searches for sequences encoding resistance gene-like NBS motifs

BLAST version 2.0.3 (Altschul *et al.*, 1997) was used to search the GenBank non-redundant (NR) database and the Genome Sequence Database (GSDB) at the National Center for Genome Resources (http://www.ncgr.org). TBLASTN searches were performed on dbEST (expressed sequence tags), dbGSS (genome sequence survey, a database comprised of BAC end sequences), and HTGS (unfinished genome sequences) at NCGR, as well as the *Oryza* GSS database (http://www.genome.clemson.edu) and the EST database at DuPont (unpublished). Nine known NBS-encoding R-gene protein sequences were used to query the databases: N (A54810), L6 (LUU27081), M (LUU73916), RPP5 (ATU97106), Prf (LEU65391), RPM1 (A57072), RPS2 (A54809), I2C-1 (AF004878) and Mi/Meu1 (AF039682). In addition, several divergent NBS sequences selected on the basis of their outlying phylogenetic positions were also used: P25941; AF113948, AB008271 (Kazusa identifier MUK11.3, [http://www.kazusa.or.jp/arabi/]), AB012242 (Kazusa identifier K24G6.8) and AL031326. Searches were conducted using the N-terminal portion of the sequence as defined by the amino acid sequence up to 120 amino acids C-terminal to the GLPL motif in the NBS. This avoided the identification of sequences with similarity to only the leucine-rich repeat (LRR) region of resistance genes. The searches were performed during the last week of March 1999. The threshold expectation value was set to 0.0001, a value empirically determined to filter out most irrelevant hits. Other numerical options were left at default values.

The BLAST output was parsed and organized by a series of Perl scripts. Individual target sequences found multiple times in the output were coalesced; these occurred because of introns and single DNA sequences, such as BACs, containing multiple distinct R-genes. Regions of similarity separated by less than 1.2 kb were interpreted as parts of a single gene. These parsed data can be obtained as a tab-delimited spreadsheet from http://www.ncgr.org/rgenes.

Alignment and phylogenetic analysis of sequences

For the purpose of alignment and analysis, predicted protein sequences were trimmed to include the amino acids from the conserved P-loop to the 'GLPL' motif (defined in Table 3). A Markov model was produced from this region using Hmmer 2.0 (S. Eddy, http://hmmer.wustl.edu/). Only sequences that matched at least 138 amino acids against the Markov model were considered. At this stage, sequences were filtered to remove exact duplicates, which resulted from searching multiple databases. Sequences were then aligned by five iterations of alignment using CLUSTALW (Thompson *et al.*, 1994). In each iteration, a neighbor-joining tree was generated which was then used as the guide tree for the next cycle of alignment. The reiterative process was necessary because of the mutual dependency between the alignment and the tree. The initial alignment used the CLUSTALW default options. For subsequent alignments, a distance matrix was constructed with the Protdist program of the PHYLIP package (Felsenstein, 1993) using the PAM distance measure. In each round, the matrix was used to build a neighbor-joining tree (Saitou and Nei, 1987) using the Neighbor program from PHYLIP. The tree topology was used as the guide tree in the next round of alignment by CLUSTALW.

The final alignment was also used to derive trees by parsimony analyses using the Protpars program of PHYLIP. The optimality criterion of Protpars was included to minimize the number of replacement substitutions needed to place the observed sequences in a tree, taking the genetic code into account (Felsenstein, 1993). Because both distance and parsimony trees supported separate TIR and non-TIR groups, these two sets of sequences were segregated for subsequent analyses.

Bootstrapping was performed by resampling with replacement of alignment positions using the Seqboot program of PHYLIP. This yielded an arbitrary number of pseudo-replicate data matrices that reflected the variability of the original data set. One hundred such pseudo-replicates were analyzed by both distance and parsimony analyses. The multiple trees resulting from each tree-building algorithm were summarized by the Consense program of PHYLIP.

Analysis of conserved motif structures

Multiple Expectation Maximization for Motif Elicitation or 'MEME' (Bailey and Elkan, 1994) was used to analyze conserved motif structures among NBS sequences. The output of MEME consists of a profile, a mathematical description of a gapless conserved sequence pattern. Each position in the profile corresponds to a numerical vector that describes the probability of observing each amino acid at that position. Using MEME, a matrix was constructed containing the ratio of logarithms of an amino acid frequency at a given position over its overall frequency in the database. Matches between the profile and any sequence was scored by summing the log-ratios for each amino acid along the width of the profile. Regions with substantial conservation were identified based on their information content; those that exceeded a minimum criterion were declared a putative motif. A program was written to read and search for MEME motifs using MEME threshold values to identify significant matches in our data set.

MEME version 2.2 (Bailey and Elkan, 1995a; Bailey and Elkan, 1995b) was used to find conserved regions in a training set of 418 NBS-encoding sequences, trimmed as described above. The training set was selected from sequences found in the NR database by BLASTP using a range of query sequences. After an initial exploration, we specified a limit of 30 motifs with other options set to default values. The biological significance of the motifs was determined based on both the distribution in the phylogenetic trees and on previous descriptions of such motifs (Baker *et al.*, 1997; Ori *et al.*, 1997). An individual profile describing amino acid frequencies was generated for each motif.

Acknowledgements

We gratefully acknowledge the laboratories of Drs Frederick Bliss and Steven Knapp for contributing unpublished sequences of *Prunus* and sunflower, respectively, for use in this analysis. We thank the anonymous reviewers for helpful comments. This work was supported in part by USDA NRI grant no. 95–37300–1571 to R.W.M. and USDA NRI grant no. 98–35300–6168 to N.D.Y. Partial support for B.C.M. was provided by an NSF graduate research fellowship.

References

Aarts, M.G.M., Hekkert, B.L., Holub, E.B., Beynon, J.L., Stiekema, W.J. and Pereira, A. (1998a) Identification of R-gene homologous DNA fragments genetically linked to disease resistance loci in *Arabidopsis thaliana*. *Molec. Plant-Microbe Interact.* **11**, 251–258.

Aarts, N., Metz, M., Holub, E., Staskawicz, B.J., Daniels, M.J. and Parker, J.E. (1998b) Different requirements for EDSI and *NDR1* by disease resistance genes define at least two *R* gene-

mediated signaling pathways in *Arabidopsis*. *Proc. Natl Acad. Sci. USA*, **95**, 10306–10311.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Anderson, P.A., Lawrence, G.J., Morrish, B.C., Ayliffe, M.A., Finnegan, E.J. and Ellis, J.G. (1997) Inactivation of the flax rust resistance gene *M* associated with loss of a repeated unit within the leucine-rich repeat coding region. *Plant Cell*, **9**, 641–651.

Aravind, L., Dixit, V.M. and Koonin, E.V. (1999) The domains of death: evolution of the apoptosis machinery. *Trends Biochem. Sci.* **24**, 47–53.

Arumuganathan, K. and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Pl. Mol. Biol. Reporter*, **9**, 208–218.

Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (ISMB '94).* Menlo Park, CA: AAAI Press, pp. 28–36.

Bailey, T.L. and Elkan, C. (1995a) Unsupervised learning of multiple motifs in biopolymers using EM. *Machine Learning*, **21**, 1–2.

Bailey, T.L. and Elkan, C. (1995b) The Value of Prior Knowledge in Discovering Motifs with MEME. *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (ISMB '95)*. Menlo Park, CA: AAAI Press, pp. 21–29.

Baker, B., Zambryski, P., Staskawicz, B. and Dinesh-Kumar, S.P. (1997) Signaling in plant–microbe interactions. *Science*, **276**, 726–733.

Bent, A.F. (1996) Plant disease resistance genes: function meets structure. *Plant Cell*, **8**, 1757–1771.

Bent, A.F., Kunkel, B.N., Dahlbeck, D., Brown, K.L., Schmidt, R.L., Giraudat, J., Leung, J.L. and Staskawicz, B.J. (1994) *RPS2* of *Arabidopsis thaliana*: a leucine-rich repeat class of plant disease resistance genes. *Science*, **265**, 1856–1859.

Bernatzky, R. and Tanksley, S.D. (1986) Genetics of actin-related sequences in tomato. *Theor. Appl. Genet.* **72**, 314–321.

Bevan, M., Bancroft, I., Bent, E. *et al.* (1998) Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. *Nature*, **391**, 485–488.

Botella, M.A., Coleman, M.J., Hughes, D.E., Nishimura, M.T., Jones, J.D.G. and Sommerville, S.C. (1997) Map positions of 47 *Arabidopsis* sequences with sequence similarity to disease resistance genes. *Plant J.* **12**, 1197–1211.

Botella, M.A., Parker, J.E., Frost, L.N., Bittner-Eddy, P.D., Beynon, J.L., Daniels, M.J., Holub, E.B. and Jones, J.D. (1998) Three genes of the *Arabidopsis* *R*PP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *Plant Cell*, **10**, 1847–1860.

Bourne, H.R., Sanders, D.A. and McCormick, F. (1991) The GTPase superfamily: conserved structure and molecular mechanism. *Nature*, **349**, 117–127.

Buschges, R., Hollricher, K., Panstruga, R. *et al.* (1997) The barley *Mlo* gene: a novel control element of plant pathogen resistance. *Cell*, **88**, 695–705.

Collins, N.C., Webb, C.A., Seah, S., Ellis, J.G., Hulbert, S.H. and Pryor, A. (1998) The isolation and mapping of disease resistance gene analogs in maize. *Molec. Plant-Microbe Interact.* **11**, 968–978.

Dangl, J. and Holub, E. (1997) La dolce vita: a molecular feast in plant–pathogen interactions. *Cell*, **91**, 17–24.

Dixon, M.S., Jones, D.A., Keddie, J.S., Thomas, C.M., Harrison, K. and Jones, J.D.G. (1996) The tomato *Cf-2* disease resistance locus comprises two functional genes encoding leucine-rich repeat proteins. *Cell*, **84**, 451–459.

Downward, J. (1990) The ras superfamily of small GTP-binding proteins. *Trends Biochem. Sci.* **15**, 469–472.

Ellis, J.G., Lawrence, G.J., Luck, J.E. and Dodds, P.N. (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell*, **11**, 495–506.

Felsenstein. J. (1993) PHYLIP (Phylogeny Inference Package). Seattle: University of Washington.

Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360.

Gay, N.J. and Keith, F.J. (1991) *Drosophila Toll* and IL-1 receptor. *Nature*, **351**, 355–356.

Grant, M.R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., Innes, R.W. and Dangl, J.L. (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. *Science*, **269**, 843–846.

Hammond-Kosack, K.E., Tang, S., Harrison, K. and Jones, J.D.G. (1998) The tomato *Cf-9* disease resistance gene functions in tobacco and potato to confer responsiveness to the fungal avirulence gene product Avr 9. *Plant Cell*, **10**, 1251–1266.

Hammond-Kosack, K.E. and Jones, J.D.G. (1997) Plant disease resistance genes. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 575–607.

Hoffmann, J.A., Kafatos, F.C., Janeway, C.A., Jr and Ezekowitz, R.A.B. (1999) Phylogenetic perspectives in innate immunity. *Science*, **284**, 1313–1318.

Holub, E.B. (1997) Organization of resistance genes in *Arabidopsis*. In *The Gene-for-Gene Relationship in Host–Parasite Interactions* (Crute, I., Holub, E. and Burdon, J., eds). Wallingford, UK: CAB International.

Hopp, T.P. (1995) Evidence from sequence information that the interleukin-1 receptor is a transmembrane GTPase. *Prot. Sci.* **4**, 1851–1859.

Jones. D.A. and Jones, J.D.G. (1997) The role of leucine-rich repeat proteins in plant defences. *Adv. Bot. Res.* **24**, 90–167.

Kanazin, V., Marek, L.F. and Shoemaker, R.C. (1996) Resistance gene analogs are conserved and clustered in soybean. *Proc. Natl Acad. Sci. USA*, **93**, 11746–11750.

Kaziro, Y., Itoh, H., Kozasa, T., Nakafuku, M. and Satoh, T. (1991) Structure and function of signal-transducing GTP-binding proteins. *Annu. Rev. Biochem.* **60**, 349–400.

Kobe, B. and Deisenhofer, J. (1994) The leucine-rich repeat: a versatile binding motif. *Trends Bio. Sci.* **19**, 415–421.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.

Kuhlman, P. and Palmer, J.D. (1995) Isolation, expression, and evolution of the gene encoding mitochondrial elongation factor *Tu* in *Arabidopsis thaliana*. *Plant Mol. Biol.* **29**, 1057–1070.

Kunkel, B.N. (1996) A useful weed put to work – genetic analysis of disease resistance in *Arabidopsis thaliana*. *Tr. Genetics*, **12**, 63–69.

Lawrence, G.J., Finnegan, E.J., Ayliffe, M.A. and Ellis, J.G. (1995) The *L6* gene for flax rust resistance is related to the *Arabidopsis* bacterial resistance gene *RPS2* and the tobacco viral resistance gene. *N. Plant Cell*, **7**, 1195–1206.

Leister, D., Ballvora, A., Salamini, F. and Gebhardt, C. (1996) A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants. *Nature Genet*, **14**, 421–429.

McDowell, J.M., Dhandaydham, M., Long, T.A., Aarts, M.G., Goff, S., Holub, E.B. and Dangl, J.L. (1998) Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *R*PP. 8 locus of *Arabidopsis*. *Plant Cell*, **10**, 1861–1874.

McDowell, J.M., Huang, S., McKinney, E.C., An, Y.Q. and Meagher, R.B. (1996) Structure and evolution of the actin gene family in *Arabidopsis thaliana*. *Genetics*, **142**, 587–602.

Meyers, B.C., Shen, K.A., Rohani, P., Gaut, B.S. and Michelmore, R.W. (1998) Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. *Plant Cell*, **10**, 1833–1846.

Michelmore, R.W. and Meyers, B.C. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8**, 1113–1130.

Milligan, S.B., Bodeau, J., Yaghoobi, J., Kaloshian, I., Zabel, P. and Williamson, V.M. (1998) The root knot nematode resistance gene *Mi* from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. *Plant Cell*, **10**, 1307–1319.

Milner-White, E.J., Coggins, J.R. and Anton, I.A. (1991) Evidence for an ancestral core structure in nucleotide-binding proteins with the type A motif. *J. Mol. Biol.* **221**, 751–754.

Mindrinos, M., Katagiri, F., Yu, G.-L. and Ausubel, F.M. (1994) The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell*, **78**, 1089–1099.

Nei, M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* **30**, 371–403.

Notredame, C. and Higgins, D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucl. Acids Res.* **24**, 1515–1524.

Ori, N., Eshed, Y., Paran, I., Presting, G., Aviv, D., Tanksley, S., Zamir, D. and Fluhr, R. (1997) The *I2C* family from the wilt disease resistance locus *I2* belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. *Plant Cell*, **9**, 521–532.

Pan, Q., Wendel, J. and Fluhr, R. (1999) Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J. Mol. Evol.* In press.

Parker, J.E., Coleman, M.J., Szabo, V., Frost, L.N., Schmidt, R., van der Biezen, E.A., Moores, T., Dean, C., Daniels, M.J. and Jones, J.D.G. (1997) The *Arabidopsis* downy mildew resistance gene *R*PP. 5 shares similarity to the toll and interleukin-1 receptors with N and L6. *Plant Cell*, **9**, 879–894.

Qureshi, S.T., Gros, P. and Malo, D. (1999) Genetic control of lipopolysaccharide responsiveness by Toll-like receptor genes. *Tr. Genetics.* **15**, 291–294.

Rock, F.L., Hardiman, G., Timans, J.C., Kastelein, R.A. and Bazan, J.F. (1998) A family of human receptors structurally related to *Drosophila* Toll. *Proc. Natl Acad. Sci. USA*, **95**, 588–593.

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

Salmeron, J.M., Oldroyd, G.E.M., Rommens, C.M.T., Scofield, S.R., Kim, H.S., Lavelle, D.T., Dahlbeck, D. and Staskawicz, B.J. (1996) Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. *Cell*, **86**, 123–133.

Saraste, M., Sibbald, P.R. and Wittinghofer, A. (1990) The P-loop – a common motif in ATP- and GTP-binding proteins. *Tr. Biochem. Sci.* **15**, 430–434.

Shen, K.A., Meyers, B.C., Islam-Faridi, N., Stelly, D.M. and Michelmore, R.W. (1998) Resistance gene candidates identified using PCR with degenerate oligonucleotide primers map to resistance gene clusters in lettuce. *Mol. Plant-Microbe Interact.* **11**, 815–823.

Simons, G., Groenendijk, J., Wijbrandi, J. *et al.* (1998) Dissection of the fusarium *I2* gene cluster in tomato reveals six homologs and one active gene copy. *Plant Cell*, **10**, 1055–1068.

Song, W.Y., Wang, L.L., Kim, H.S. *et al.* (1995) A receptor kinase-like protein encoded by the rice disease-resistance gene, *Xa21. Science*, **270**, 1804–1806.

Speulman, E., Bouchez, D., Holub, E.B. and Beynon, J.L. (1998) Disease resistance gene homologs correlate with disease resistance loci of *Arabidopsis thaliana. Plant J.* **14**, 467–474.

Srinivasula, S.M., Ahmad, M., Fernandes-Alnemri, T. and Alnemri, E.S. (1998) Autoactivation of procaspase-9 by Apaf-1-mediated oligomerization. *Mol. Cell*, **1**, 949–957.

Theissen, G., Kim, J.T. and Saedler, H. (1996) Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. *J. Mol. Evol.* **43**, 484–516.

Thomas, C.M., Jones, D.A., Parniske, M., Harrison, K., Balint-Kurti, P.J., Hatzixanthis, K. and Jones, J.D.G. (1997) Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognitional specificity in *Cf-4* and *Cf-9. Plant Cell*, **9**, 2209–2224.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.

Traut, T.W. (1994) The functions and consensus motifs of nine types of peptide segments that form different types of nucleotide binding-sites. *Eur. J. Biochem.* **222**, 9–19.

Van der Biezen, E.A. and Jones, J.D.G. (1998) The NB-ARC domain: a novel signaling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr. Biol.* **8**, 226–227.

Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) Distantly related sequences in the α- and β-subunits of ATP synthetase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951.

Wang, Z.X., Yano, M., Yamanouchi, U., Iwamoto, M., Monna, L., Hayasaka, H., Katayose, Y. and Sasaki, T. (1999) The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. *Plant J.* **19**, 55–64.

Warren, R.F., Henk, A., Mowery, P., Holub, E. and Innes, R.W. (1998) A mutation within the leucine-rich repeat domain of the *Arabidopsis* disease resistance gene *RPS5* partially suppresses multiple bacterial and downy mildew resistance genes. *Plant Cell*, **10**, 1439–1452.

Whitham, S., Dinesh-Kumar, S.P., Choi, D., Hehl, R., Corr, C. and Baker, B. (1994) The product of the tobacco mosaic virus resistance gene *N*: similarity to *Toll* and the interleukin-1 receptor. *Cell*, **78**, 1101–1115.

Yoshimura, S., Yamanouchi, U., Katayose, Y., Toki, S., Wang, Z.X., Kono, I., Kurata, N., Yano, M., Iwata, N. and Sasaki, T. (1998) Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl Acad. Sci. USA*, **95**, 1663–1668.

Yu, Y.G., Buss, G.R. and Saghai Maroof, M.A. (1996) Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site. *Proc. Natl Acad. Sci. USA*, **93**, 11751–11756.