J Mol Evol (2005) 60:229–237
DOI: 10.1007/s00239-004-0144-2

JOURNAL OF MOLECULAR EVOLUTION

Evolution of NIN-Like Proteins in *Arabidopsis*, Rice, and *Lotus japonicus*

Leif Schauser,[1,2] Wioletta Wieloch,[3] Jens Stougaard[1]

[1] Laboratory of Gene Expression, Department of Molecular Biology, Aarhus University, Gustav Wieds Vej 10, DK-8000 Århus C, Denmark
[2] Bioinformatics Research Center, University of Aarhus, Høegh-Guldbergs Gade, Building 090, DK-8000 Århus C, Denmark
[3] Institute of Parasitology, Polish Academy of Sciences Twarda 51-55, 00-818 Warsaw, Poland

Received: 10 May 2004 / Accepted: 9 September 2004 [*Reviewing Editor:* Professor David Guttman]

Abstract. Genetic studies in *Lotus japonicus* and pea have identified *Nin* as a core symbiotic gene required for establishing symbiosis between legumes and nitrogen fixing bacteria collectively called Rhizobium. Sequencing of additional *Lotus* cDNAs combined with analysis of genome sequences from Arabidopsis and rice reveals that *Nin* homologues in all three species constitute small gene families. In total, the Arabidopsis and rice genomes encode nine and three NIN-like proteins (NLPs), respectively. We present here a bioinformatics analysis and prediction of NLP evolution. On a genome scale we show that in Arabidopsis, this family has evolved through segmental duplication rather than through tandem amplification. Alignment of all predicted NLP protein sequences shows a composition with six conserved modules. In addition, *Lotus* and pea NLPs contain segments that might characterize NIN proteins of legumes and be of importance for their function in symbiosis. The most conserved region in NLPs, the RWP-RK domain, has secondary structure predictions consistent with DNA binding properties. This motif is shared by several other small proteins in both Arabidopsis and rice. In rice, the RWP-RK domain sequences have diversified significantly more than in Arabidopsis. Database searches reveal that, apart from its presence in Arabidopsis and rice, the motif is also found in the algae *Chlamydomonas* and in the slime mold *Dictyostelium discoideum*. Thus, the origin of this putative DNA binding region seems to predate the fungus–plant divide.

Key words: Legume — Rhizobia symbiosis — Nodule inception — Gene family — Gene duplication — NIN-like proteins

Introduction

The genetic constitution allowing legumes to develop root nodules in symbiosis with bacteria belonging to the family Rhizobiaceae, here collectively called *Rhizobium*, is currently being clarified using molecular genetic studies in the model legume *Lotus japonicus* and other legumes. The findings of these studies offer new possibilities for determining the origin of the symbiotic relationship. Development of root nodules is a multistep process mediated by signal exchange between partners. Initially flavones or isoflavones exuded by the plant induce *Rhizobium* to secrete lipochitin-oligosaccharide molecules triggering the compatible host to initiate nodule primordia from already differentiated root cells. Afterwards the microsymbionts invade the nodule primordia and are subsequently endocytosed into plant cells. Several recent reports demonstrate the recruitment of preexisting genes into this specialized organogenic pathway (Schauser et al. 1999; Stracke et al. 2002; Krusell et

Correspondence to: Leif Schauser; *email:* schauser@daimi.au.dk

al. 2002; Madsen et al. 2003). Radutoui et al. 2003). For example, mutations in the *Har1* LRR receptor kinase gene result in deregulated lateral root formation under nonsymbiotic conditions and hypernodulation phenotypes under symbiotic conditions. The Arabidopsis gene most similar to *Har1* is the *Clavata1* gene, participating in the pathway that orchestrates shoot apical meristem growth. During evolution, a *Clavata1* gene has been recruited to serve a function during root nodule development (Krusell et al. 2002). Similarly, the nodulation genes encoding the LysM receptor kinases NRF1 and NRF5 have homologues in non-nodulating Arabidopsis that are likely involved in signal perception and transduction pathways (Madsen et al. 2003; Radutoui et al. 2003). Thus, the evolution of the legume–rhizobium symbiosis seems to build on reprogramming of preexisting pathways.

The initiation of nodule development is dependent on the function of the *Nin* gene (nodule inception. [Schauser et al. 1999]). Mutational inactivation of *Nin* results in an excessive root hair curling response to *Rhizobium* but *nin* mutants do not develop infection threads or initiate cell divisions founding the nodule primordium. This suggests a function downstream of lipochitin-oligosaccharide signal perception and transduction (Schauser et al. 1999). The most prominent feature of the NIN protein is a 60-amino acid (aa)-long sequence that is strongly conserved across a variety of proteins. This region has been named the RWP-RK domain according to invariant amino acids of the consensus sequence and its function has been predicted to be DNA binding and dimerization. *Nin* has nine homologous genes in the Arabidopsis genome predicted to encode NIN-like proteins (NLPs). However, to date no function has been established for any of these genes. In order to decipher the evolutionary history of *Nlp* and *Nin* genes we have compared NLPs *in silico* and determined the phylogenetic relationship in this family consisting of legume, Arabidopsis, and rice sequences. We also present the coding capacities of the Arabidopsis and rice genomes with respect to the RWP-RK domain. Furthermore, we present a model for evolution of the *Nlp* gene family in Arabidopsis.

Materials and Methods

Identification of RWP-RK Sequence Coding Capacity

Previous analysis established the RWP-RK domain as the most conserved region of NIN (Schauser et al. 1999). We used this motif (NIN amino acid positions, 559 to 649) as a query in order to identify homologous sequences. For this purpose, we used Blastp for searching NCBIs Genbank and tblastn against the rice genome. The rice sequences (*Oryza sativa* L. ssp. *Indica* [Yu et al.]) were downloaded from the NCBI Web site (August 2003). The *Oryza sativa* L. ssp. *Japonica* genome (Goffer et al. 2002) was searched using the Syngentas Web site (http://portal.tmri.org/rice/).

Alignment of one sequence, At4g35270, with the other NLPs revealed a mispredicted intron–exon boundary due to erroneous in silico splicing. One Arabidopsis EST (λ-PRL2 107G21T7) encoding parts of the same gene was sequenced in its entire length (designated *AtNlp2*; EMBL accession no. AJ579912.1). This resulted in a corrected gene sequence where the sequence GAAAGTGATGAT TCATTCACGCAGTTTCATTTTCATGTTGCA was removed. Thus the sequence confirmed the suspicion of erroneous in silico splicing of the genomic sequence and revealed yet another annotation error of the genomic sequence at the 3′ end of the sequence.

The pea *NIN* orthologue was identified experimentally (Borisov et al. 2003).

Two full-length Lotus cDNAs, designated *LjNlp1* (EMBL accession no. AJ579910.1) and *LjNlp2* (EMBL accession no. AJ579912.1), were isolated from a nodule cDNA library by hybridization with the RWP-RK domain encoding region of *LjNin* as a probe.

Alignment and Phylogenetic Tree Construction

Alignment of the sequences was performed using ClustalX and the following parameters; gap open, 0.2; gap elongation, 0.05; and the Gonnet 250 substitution matrix. Phylogenetic and molecular evolutionary analyses of this alignment were conducted using MEGA version 2.1 (Kumar et al. 2001). The phylogeny tree was constructed using the minimal evolution method with Poisson correction for amino acid distance and handling gap/missing data by pairwise deletion. Confidence values were obtained by 1000-fold bootstrap tests. Maximum likelihood estimation of this tree by PHYLIP (dnaml, Felsenstein, 1995) using a manually curated (i.e., removal of gaps and ambiguities) multiple alignment of codons (DNA sequences) resulted in an identical topology.

Analysis of NLP Evolution

Eleven protein sequences encoded by genes flanking each of the nine Arabidopsis NLPs (five on each side and the NLP itself) were extracted from TAIRs SeqViewer (http://www.arabidopsis.org/servlets/sv) and concatenated. The NLP sequences were masked in order to allow us to focus on the surrounding regions. These nine blocks were subsequently compared to each other using blastp. The Blast output was processed using Python scripts to create a visual output similar to Fig. 5. Individual proteins with paralogues in other blocks were then searched against the entire protein content of the Arabidopsis genome in order to identify those sequences that are reciprocal best hits.

Secondary Structure Predictions

Secondary structure prediction of the RWP domain was carried out using the BMERCs PSA server (White et al. 1994; http://bmerc-www.bu.edu/psa/).

Results

RWP-RK Domain Containing Proteins and NIN-like Proteins in Arabidopsis and Rice

A genomewide database search using the highly conserved RWP-RK domain of NIN (aa 559 to 649) was performed in order to identify conceptual NLP and RWP-RK domain containing proteins in Ara-

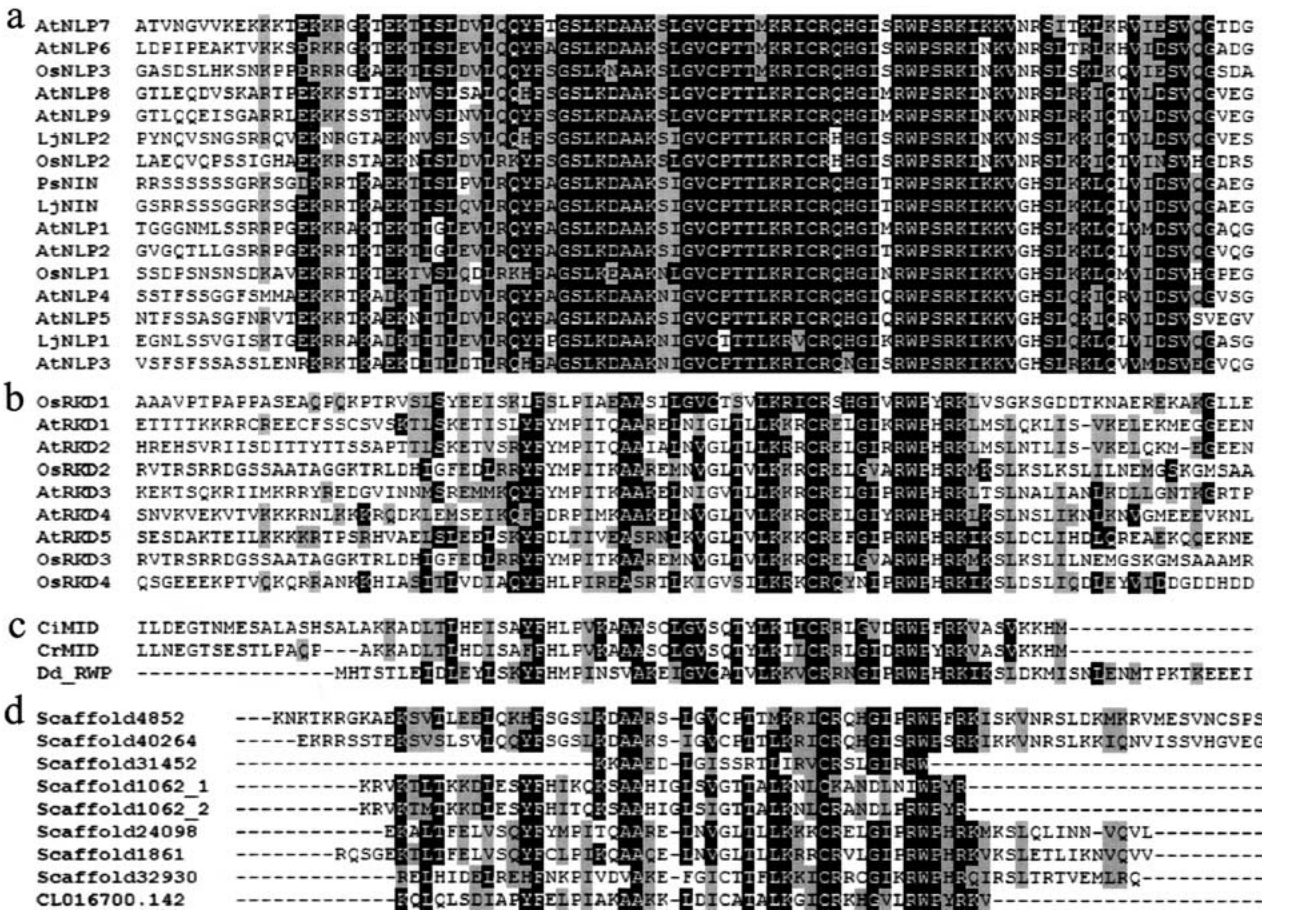

Fig. 1. Alignment of the RWP-RK domain of 40 proteins. Black and gray shading represents identity and similarity in at least 50% of the sequences aligned, respectively. Domains shown are as follows. **a.** NIN and NLP proteins sequenced in this study or found in protein databases, **b.** RKD proteins from protein databases. **c.** Dd_RWP, from the N terminus of the hypothetical protein from *Dictyostelium discoideum* (gb|AAL99316.1|); CiMID, C terminus of the Mid protein from Chlamydomonas incerta; CrMID, C terminus from the Mid protein from *Chlamydomonas reinhardtii*. **d**. Not annotated hits in the rice genome. The rice sequences Scaffold31452, Scaffold1062_1, and Scaffold1062_2 are only distantly related.

bidopsis and rice. In total, the Arabidopsis genome encodes 14 proteins carrying a RWP-RK domain. These proteins can be divided into two classes: NLPs with similarity to NIN along their entire length (800–900 aa) and smaller proteins sharing only the RWP-RK domain with NIN. We propose that the former class, with nine Arabidopsis members, be named AtNLP1–AtNLP9 according to evolutionary distance to NIN (see Fig. 1 and 2), the closest relative being AtNLP1 (Table 1). For the five smaller proteins we propose AtRKD proteins (*R*WP-R*K* *d*omain). The rice proteome contains three annotated OsNLPs and four annotated OsRKDs (Table 1). Searching for further coding capacity with regard to the RWP-RK domain, this time in the rice genomes deposited at NCBI and Syngenta, identified another nine DNA sequences with a high similarity at the amino acid level when translated. Potentially, one of these sequences represents a fourth OsNLP; the rest, OsRKDs. Apart from the NLP and RKD classes already defined in Arabidopsis, a new class seems to have evolved in rice. Notably, two putative genes found in the vicinity of each other on the same scaffold (assembly of sequences at Syngenta) have sequences that, when translated, contain an extra amino acid in the RWP-RK domain (Scaffold 1062 BGI nos. 1 and 2). Whether these are genes or pseudogenes awaits the completion of the sequencing of the genomes of these rice varieties.

Secondary Structure Predictions

All RWP-RK sequences are predicted to fold into highly ordered secondary structures. Predictions by the PSA server indicate that an α-helical basic region is followed by a helix–turn–helix and a helical leucine zipper, spaced by loops. This overall structure prediction, reminiscent of the ubiquitous basic leucine-zipper and helix–turn–helix classes of DNA-binding domains, led to the speculation that the RWP-RK domain might be involved in DNA binding and protein dimerization (Schauser et al. 1999).

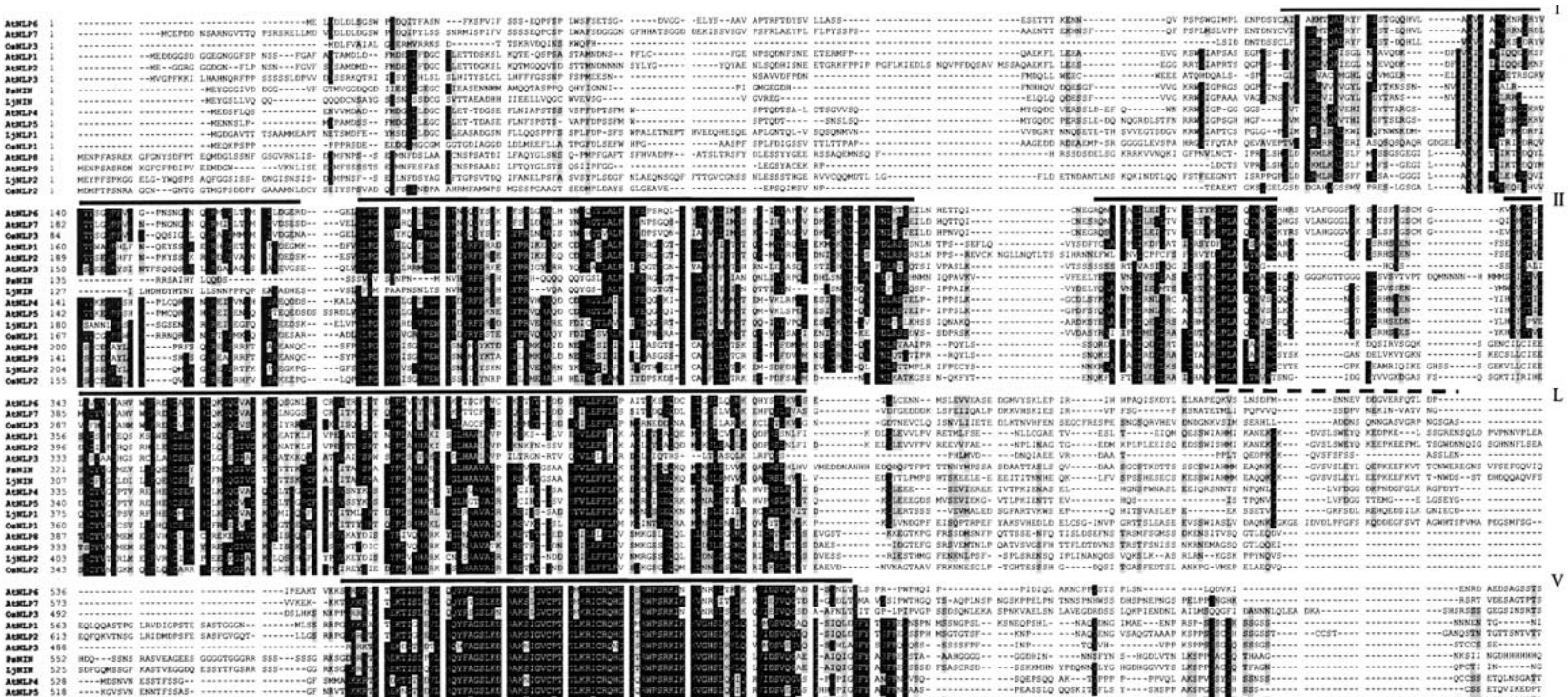

Fig. 2. Alignment of NLPs from legumes, Arabidopsis and rice. The legume sequences were obtained by translating cDNA sequences from *Lotus* and pea, whereas the Arabidopsis and rice NLPs were identified through annotation of the respective genomic sequences. One artifact produced by in silico splicing has been removed by sequencing Arabidopsis ESTs (see Materials and Methods), but others may still exist. Six blocks of conservation are identified, indicated by the thick solid lines at the top of the alignment and numbered as indicated to the right of the alignment. The RWP-RKP domain makes up block V, whereas the PB1 domain (Ponting et al. 2002) is located in block VI. A block conserved in NINs, NLP1 and NLP2 is indicated by dashes and denoted (Borisov et al. 2003). Identity and similarity in at least 50% of the sequences aligned are represented by black and gray shading, respectively.

Table 1. Nomenclature of Arabidopsis and rice NLPs and RKDs

NLP/RKD	Gene product name
AtNLP1	At2g17150
AtNLP2	At4g35270
AtNLP3	At4g38340
AtNLP4	At1g20640
AtNLP5	At1g76350
AtNLP6	At1g64530
AtNLP7	At4g24020
AtNLP8	At2g43500
AtNLP9	At3g59580
OsNLP1	OJ1004C08.16 Chr 3
OsNLP2	OSJNBa0067K08.21 Chr 4
OsNLP3	BAA92920 Chr 1
AtRKD1	At1g18790
AtRKD2	At1g74480
AtRKD3	At5g66990
AtRKD4	At5g53040
AtRKD5	At4g35590
OsRKD1	OSJNBa0004G10.17
OsRKD2	OSJNBa0066C06.7
OsRKD3	OSJNBa0024F24.30
OSRKD4	OSJNBa0017B10.3

NIN-like Proteins Contain Several Conserved Domains

From legumes, only the *Lotus* and the pea *Nin* sequences were available at the outset of this study. To provide further insight into the legume gene family and broaden the basis for comparative study, we isolated two full-length cDNAs *LjNlp1* and *LjNlp2*, from a *Lotus* root nodule library.

Full alignment of the Arabidopsis NLPs together with the annotated rice NLPs and the experimentally determined legume NIN and NLPs delimits six blocks of conservation (Fig. 2). Blocks I to IV are unique to this protein family, as confirmed by BLAST searches with this region as a query. In the region encoding blocks I and II, an ~210-bp deletion has occurred in the ancestor of the *Lotus* and pea *Nin* genes, indicating that this region might be crucial for the specific function of these proteins in nodule development. Computational prediction of the secondary structure by the PSA server suggests the presence of an amphipathic α-helix in the region specific for NLPs. It is possible that this domain constitutes a helix–turn–helix motif since some weaker predictions of a turn and a second helix follow the strong first helix prediction (not shown). The corresponding region in *Lotus* and pea NIN proteins lacks homology and no regular structure is predicted. This suggests that a special feature of legume NIN proteins is the lack of a domain, rather than the gain of structural elements.

Blocks V and VI are more abundant domains in the protein universe. Block V is the RWP-RK domain, predicted to be involved in DNA binding and dimerization. Block VI shows strong homology to the PBI domain, a protein–protein interaction domain enabling heterodimerization between PB1 domain containing proteins (Ponting et al. 2002).

Analysis of NLP Evolution

The full alignment shown in Fig. 2 was used to infer a phylogenetic tree (Fig. 3). This NLP phylogenetic tree, encompassing three plant families, suggests that at least three paralogous NLPs have existed in the common ancestor of mono-and eudicotyledons. Under this hypothesis, none of the three ancestral NLPs have been duplicated or deleted in rice. In Arabidopsis, two ancestral NLPs have since undergone one round of duplication, whereas the third ancestral NLP has duplicated several times since the divide of eudicots from monocotyledons. Some of these duplicated NLPs have since been deleted in Arabidopsis. In an alternative scenario, the ancestral species contained four NLPs, one of which has since been lost in rice (see below). This fourth NLP would, under this hypothesis, be the ancestor of NIN proteins.

Gene families can arise through segmental duplications of chromosomal regions, resulting in a scattered occurrence, or through tandem amplification, resulting in a clustered occurrence of family members. As observed for the rice sequences, no two NLPs are physically located near each other (i.e., on the same scaffold). This indicates that the rice NLP family has arisen through segmental duplication rather than single gene duplications that would result in clustered occurrence of members. To test whether this holds true in Arabidopsis, we mapped the locations of the nine AtNLPs together with the five AtRKDs onto Arabidopsis chromosomes (Fig. 4). The distribution of occurrences is correlated, indicating segmental duplication of large chromosomal regions as the underlying force for creating this family in Arabidosis. In order to identify the evolutionary relationship between duplicated segments, we compared the protein content of regions surrounding individual NLPs (Fig. 5). For each of the eleven segments, we concatenated the protein sequences encoded by the *Nlp* gene with those originating from five neighboring genes on either side, resulting in nine long sequences, consisting of 11 proteins each. These nine sequences were then compared to each other using BLAST. If two homologous proteins found in different NLP containing blocks are reciprocal best hits in the Arabidopsis proteome, this indicates a common history of these blocks. Four pairs of duplicated blocks can be identified in this way (Fig. 5). The relationship of orthologues matches the phylogenetic relationship inferred from the NLP sequence alignment (Fig. 3).

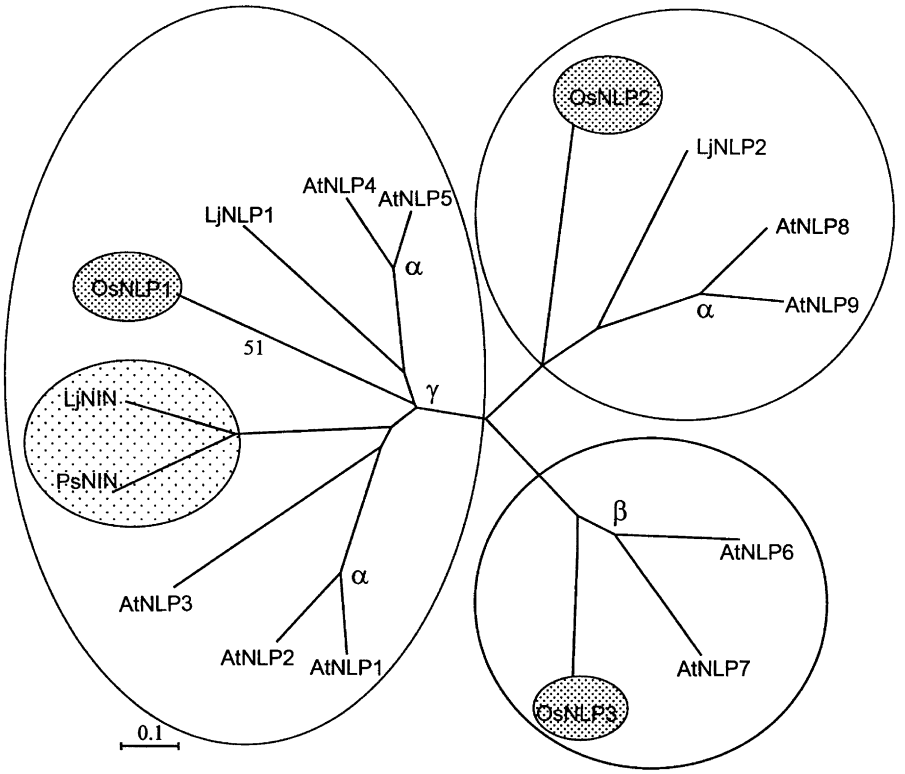

Fig. 3. Phylogenetic analysis of NLPs. The sequence alignment from Fig. 2 was used to calculate the phylogenetic relationship between the proteins using the program MEGA. A recent whole-genome duplication event specific to Arabidopsis has given rise to four AtNLP pairs. The origin of three clades, exemplified by the rice sequences OsNLPl, OsNLP2, and OsNLP3, predates the monocot/eudicot divide. Duplicated NLP pairs located in syntenic regions detected by Bowers et al. (2003) are indicated with their nomenclature (α, β, and γ). Bootstrap values for all branches were above 96 except for OsNLP1, as indicated (51).

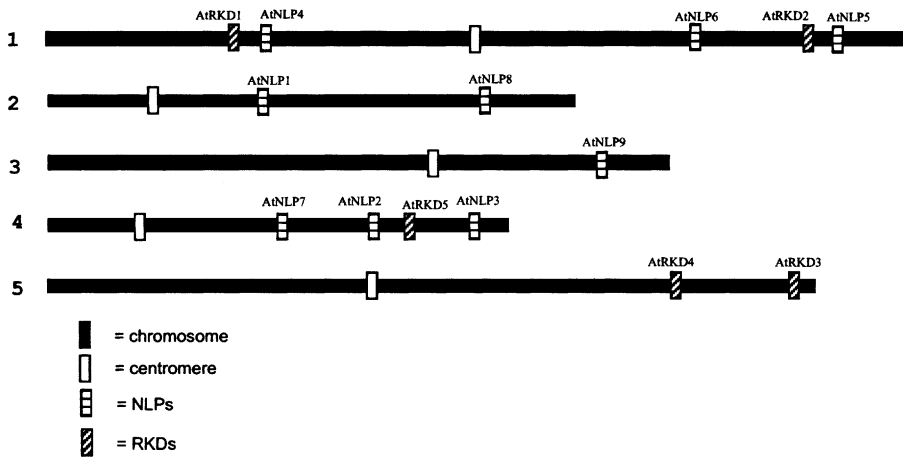

Fig. 4. Chromosomal location of genes encoding NLPs and other RWP-RK domain containing proteins in the Arabidopsis genome. Chromosome numbers are indicated on the left. Gene names are indicated. AtNLP and AtRKD numbering is according to Table 1.

Thus, the pairs NLP1 and 2, NLP4 and 5, NLP6 and 7, and NLP8 and 9 have arisen as a result of a recent duplication as indicated by the presence of additional paralogous protein pairs in their neighborhoods. In addition, this neighborhood analysis reveals that NLP3 and NLP4 share a common ancestor. This duplication is presumably of a more ancient date.

Dating the Duplication Events

In order to put our findings in the larger context of plant genome evolution, we compared the relationship of the Arabidopsis NLPs to lists of genes defining historical duplication events compiled by Bowers et al. (2003).

Using a comparative approach involving seven plant species representing major taxonomic families, Bowers et al. (2003) date three whole-genome duplication events, termed α, β, and γ detectable in the present-day Arabidopsis genome. The most recent whole-genome duplication occurred after the divergence of Arabidopsis from most other eudicots and is termed the α-duplication. It is estimated that 30% of *Arabidopsis* genes remained syntenic in the ≤ 86 million years since this α-duplication, the rest of the genes being reshuffled or lost. Only 13% of the genes remained syntenic in the much older γ-duplication.

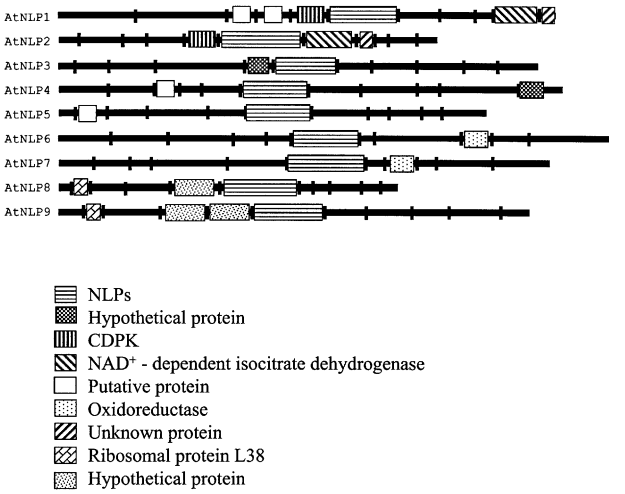

Fig. 5. Detection of segmental duplications in regions of the Arabidopsis genome encompassing NLPs. The sequences of 11 proteins surrounding each NLP (5 on each side) were concatenated to form one block. A vertical black bar indicates the concatenation of two protein sequences. This was done for all nine Arabidopsis NLPs, resulting in nine blocks, which were then searched against each other using Blast. Reciprocal best hits are indicated.

By comparing Figs. 3 and 5 with the results from Bowers et al. (2003) and Paterson et al. (2004), a detailed view of the duplication history of NLPs in Arabidopsis emerges. An ancient "γ"-duplication, dating back about 300 MYA, predates the divide of eucotyledons and monocotyledons. This and other nondetectable early duplication events have given rise to four NLP clades, three of which are common to rice and Arabidopsis (Fig. 3). Bowers et al. (2003) trace one NLP duplication back to the γ-event. This finding argues in favor of the hypothesis of an ancestral state of four NLPs. A second round of duplications, postdating divergence, has occurred in both the eucotyledon and the monocotyledon lineages. Arabidopsis NLP6 and NLP7 are located in syntenic regions originating from this "β"-duplication dating to about 150–200 MYA. A third round of duplication occurred in a common ancestor to the Brassicaceae. Results of this "α"-duplication are the Arabidopsis gene pairs NLP1 and 2, NLP4 and 5, and NLP8 and 9. The relationship between NLP3 and NLP4 uncovered by block analysis (Fig. 4) is not detected by the method of Bowers et al. (2003), presumably because the duplicated region is too short or because synteny is too degraded.

Discussion

Domains

The high degree of conservation of the RWP-RK domain indicates purifying selection due to heavy constraints on its three-dimensional structure. Secondary structure predictions of this domain indicate the presence a basic helix followed by a helix–turn–helix motif and an amphipathic leuzine zipper. These structural elements are consistent with a function in DNA-binding and protein dimerization. The RWP-RK domain is found in a variety of proteins, indicating a functional module with conserved structure and function. It can be located in different regions of a given protein. In the *Chlamydomas* mating type-determining protein Mid, the RWP-RK domain terminates the protein and lacks the leucine zipper extension seen in many other proteins of this family. The *Dictyostelium* protein is unique in that the RWP-RK domain makes up the N-terminal region. These proteins, together with a family of other small proteins termed RKDs, do not seem to contain other conserved regions (apart from the RWP-RK domain) which might shed light on their function(s). The larger NIN-like proteins, in contrast, are multidomain proteins with a high degree of conservation. Apart from the six domains identified from the multiple alignment of the family member sequences, it is apparent that their overall length as well as the relative order of the domains is conserved. The RWP-RK domain is situated in domain V. The PB1 domain found in all NLPs (domainVI) is involved in the heterodimerization with other PB1 domain containing proteins (Ponting et al. 2002). PB1 domains predominantly occur in eukaryotic signaling molecules, such as kinases. One can envisage a scenario where NLPs receive signals through their PB1 domain and mediate responses through their DNA binding abilities, a situation reminiscent to the two-component system, where a phosphorelay cascade results in transcriptional changes of target genes by activation of a transcription factor (Inoue et al. 2001).

Close inspection of the NLP multiple alignment (Fig. 2) reveals that parts of domains I and II are deleted in *Lotus* and pea NIN proteins. Secondary structure prediction of this region in NLPs revealed a well-defined helical structure of this region lacking in NIN proteins. This might indicate the loss of a specific function in the recruitment of NINs to the rhizobial symbiosis.

NLP evolution

The phylogenetic tree inferred from the NLP alignment suggests that at least three copies of this protein existed in the common ancestor to mono- and eudicotyledons. Two of these copies have since undergone one round of duplication in Arabidopsis, whereas the third copy has duplicated several times since the divide of eudicotyledons from monocotyledons. None of the three ancestral copies has duplicated in rice.

There are no close relatives to the legume NIN proteins in rice or Arabidopsis. Arabidopsis AtNLP1, AtNLP2, and AtNLP3 and rice OsNLP1 are the closest relatives of legume NINs.

Legumes and Arabidopsis belong to the Rosids and diverged 90 million years ago (MYA) (Yang et al. 1999). During this period, the nitrogen-fixing rhizobacteria–legume symbiosis has evolved (Kistner and Parniske, 2002). The common ancestor had the ability of mycorrhizal symbiosis, an ability Arabidopsis has since lost. A number of shared genetic components of the two symbioses have been identified in legumes (Schauser et al. 1998; Stracke et al. 2002; Stougaard 2001), indicating their common ancestry. It is thought that the rhizobial symbiosis evolved by building on the mycorrhizal symbiosis. The recruitment of additional components from the plant genome enabled the more elaborate rhizobial symbiosis. One of the genetic components unique to this symbiosis is *Nin*. Mutant *nin* legumes are specifically deficient in the rhizobial symbiosis, with the mycorrhizal symbiosis unaffected. Since all plants contain genes encoding NLPs, *Nin* provides evidence for the hypothesis of recruitment of preexisting genes to the specialized function of rhizobial symbiosis. Another prominent recent example of such recruitment is the regulation of nodule number allowed to develop on legume roots upon inoculation with rhizobia by the *Har1* gene. The Arabidopsis homologue most similar to *Har1* is *Clv1*, involved in the regulation of shoot apical meristem size (Krusell et al. 2002).

Involvement of Segmental Duplication in NLP Evolution in Arabidopsis

The eudicot/monocot divide dates to about 200 MYA (Wikstrom et al. 2001). Much of the Arabidopsis genome has been scrambled and duplicated since (Vision et al. 2000; Simillion et al. 2002; Bowers et al. 2003), explaining why so little synteny to rice exists (Goff et al. 2002). Our analysis of conserved blocks surrounding NLPs was able to identify homeologous segments in Arabidopsis but failed to identify orthologous segments in rice (data not shown).

Duplications of single regulatory genes are usually not of selective advantage, due to disequilibrium in their expression (Ohno 1970). Therefore the fate of recently duplicated single genes usually is to accumulate mutations and rapidly degenerate to pseudogenes (Lynch and Conery 2000). Whole-genome duplication, on the other hand, is proposed to be the major force behind speciation. These duplications have a larger chance of survival because the production of all proteins increases proportionally. Redundant genes are then free to participate in the evolution of novel traits, such as, in the current example, the ability to participate in new types of symbiosis. Often, however, redundancy persists as seen for the three Arabidopsis SEPALLATA MADS-box genes (Pelaz et al. 2000). Single and double mutants do not have a phenotype, whereas the triple mutant has. Segmental duplication results in a dispersed pattern of paralogue distribution, as observed for NLPs in Arabidopsis (Fig. 4). The occurrence of homeologous genes in the vicinity of NLP paralogues also suggests their origin by segmental duplication (Fig. 5).

Several pairs of duplicated blocks can be identified in this way. Most of these blocks have previously been identified and dated by Bowers et al. (2003), allowing a detailed view on the evolution of Arabidopsis NLPs. All major duplication events detected by Bowers et al. (2003) are identifiable in the NLP phylogeny (superimposed on Fig. 3).

Arabidopsis NLPs might be functionally redundant genes, which could explain the fact that none of these genes have been assigned a function yet in mutational screens. Double mutants with defects in both members of a duplicated pair might reveal a function.

Origin of the RWP-RK Domain

Until recently, the RWP-RK domain has been thought to be plant specific (Riechman et al. 2000). The finding of a protein with this domain in *Dictyostelium discoideum* expands its presence to a second kingdom (Amoebozoa). The recent sequencing of chromosome 2 of *D. discoideum* (Gloeckner et al. 2002) revealed that its genome exhibits greater similarity to metazoans than to plants or fungi. Systematically, amoebazoans are placed at a position before the branching of the metazoa and fungi, but after the divergence of the plant kingdom (Baldauf et al. 2003). Unless the RWP-RK domain has been acquired by *Dictyostelium* through horizontal gene transfer later in evolution, its presence here implies that the common ancestor to metazoans and plants already contained genes encoding this motif. The absence of this domain in the proteome of metazoans and fungi might reflect gene loss in these phyla rather than novel evolution in plants. The RWP-RK is thus likely to be an ancient motif, predating the fungus–plant divide.

Acknowledgments. We would like to thank Lene Heedgaard Madsen for critical comments on the manuscript. L.S. is supported by Danish Research Council Grant SNF 21-01-0329.

References

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

Baldauf SL (2003) The deep roots of eukaryotes. Science 300:1703–1706

Borisov AY, Madsen LH, Tsyganov VE, Umehara Y, Voroshilova VA, Batagov AO, Sandal N, Frederiksen A, Schauser L, Ellis N, Tikhonovich IA, Stougaard J (2003) The Sym35 gene required for root nodule development in *Pisum sativum* is an orthologue of *Nin* from *Lotus japonicus*. Plant Physiol 131:1009–1017

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:383–384

Felsenstein J (1995) PHYLIP. Phylogeny inference package, version 3.5. Department of Genetics, University of Washington, Seattle,

Goff SA, et al. (2002) A draft dequence of the rice genome (Oryza sativa L. ssp. Japonica). Science 296:92–100

Inoue T, Higuchi M, Hashimoto Y, Seki M, Kobayashi M, Kato T, Tabata S, Shinozaki K, Kakimoto T (2001) Identification of CRE1 as a cytokinin receptor from Arabidopsis. Nature 409:1060–1063

Kistner C, Parniske M (2002) Evolution of signal transduction in intracellular symbiosis. Trends Plant Sci 7:511–518

Krusell L, Madsen LH, Sato S, Aubert G, Genua A, Szczyglowski K, Duc G, Kaneko T, Tabata S, de Bruijn F, Pajuelo E, Sandal N, Stougaard J (2002) Shoot control of root development and nodulation is mediated by a receptor-like kinase. Nature 420:422–425

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software) Arizona State University, Tempe, Vol. 1, pp 1–2

Lynch M, Conery J (2000) The evolutionary fate and consequences of duplicated genes. Science 290:1151–1155

Madsen EB, Madsen LH, Radutoiu S, Olbryt M, Rakwalska M, Szczyglowski K, Sato S, Kaneko T, Tabata S, Sandal N, Stougaard J (2003) A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. Nature 425:637–640

Ohno S (1970) Evolution by gene duplication. Springer Verlag, Heidelberg, Germany

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polypoidization predating divergence of cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA 101:9903–9908

Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF (2000) B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature 405:200–203

Ponting CP, Ito T, Moscat J, Diaz-Meco MT, Inagaki F, Sumimoto H (2002) OPR, PC and AID: all in the PB1 family. Trends Biochem Sci 27:10

Radutoiu S, Madsen LH, Madsen EB, Felle HH, Umehara Y, Grønlund M, Sato S, Nakamura Y, Tabata S, Sandal N, Stougaard J (2003) Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. Nature 425:585–592

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C-Z, Keddie J, Adam L, Pineda O., Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G-L (2000) Arabidopsis transcription factors: Genome-wide comparative analysis among eukaryotes. Science 290:2105–2110

Schauser L, Handberg K, Sandal N, Stiller J, Thykjaer T, Pajuelo E, Nielsen A, Stougaard J (1998) Symbiotic mutants deficient in nodule establishment identified after T-DNA transformation of Lotus japonicus. Mol Gen Genet 259:414–423

Schauser L, Roussis A, Stiller J, Stougaard J (1999) A symbiotic regulator controlling development of symbiotic root nodules. Nature 402:191–195

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y (2002) The hidden duplication past of Arabidopsis thaliana. Proc Natl Acad Sci USA. 99:13627–13632

Stougaard (2001) Genetics and genomics of root symbiosis. Curr Opin Plant Biol 4:328–333

Stracke S, Kistner C, Yoshida S, Mulder L, Sato S, Kaneko T, Tabata S, Sandal N, Stougaard J, Szczyglowski K, Parniske M (2002) A plant receptor-like kinase required for both bacterial and fungal symbiosis. Nature 417:959–962

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 24:4876–4882

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science 290:2114–2117

White JV, Stultz CM, Smith TF (1994) Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. Math Biosci 119:35–75

Wikstrom N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond B Biol Sci, 268:2211–2220

Yang Y-W, Lai K-N, Tai P-Y, Li W-H (1999) Rates of nucleotide substitution in angiosperm mitichondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. J Mol Evol 48:597–604

Yu J, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Indica*). Science 296:79–92