*Plant Biotechnology* 35, 177–180 (2018)
DOI: 10.5511/plantbiotechnology.18.0418a

**Note**

# Efficient identification of NLR by using a genome-wide protein domain and motif survey program, Ex-DOMAIN

Mari Narusaka[1], Harunobu Yunokawa[2], Yoshihiro Narusaka[1,*]

[1]Okayama Prefectural Technology Center for Agriculture, Forestry, and Fisheries, Research Institute for Biological Sciences, Okayama 716-1241, Japan; [2]Maze Inc., Tokyo 193-0835, Japan
*E-mail: yo_narusaka@bio-ribs.com Tel: +81-866-56-9450 Fax: +81-866-56-9453

Received January 14, 2018; accepted April 18, 2018 (Edited by T. Mizoguchi)

**Abstract** Genomic and amino acid sequences of organisms are freely available from various public databases. We designed a genome-wide survey program, named "Ex-DOMAIN" (exhaustive domain and motif annotator using InterProScan), of protein domains and motifs to aid in the identification and characterization of proteins by using the InterProScan sequence analysis application, which can display information and annotations of targeted proteins and genes, conserved protein domains and motifs, chromosomal locations, and structural diversities of target proteins. In this study, we indicated the disease resistance genes (proteins) that play an important role in defense against pathogens in *Arabidopsis thaliana* (thale cress) and *Cucumis sativus* (cucumber), by searches based on the conserved protein domains, NB-ARC (a nucleotide-binding adaptor shared by the apoptotic protease-activating factor-1, plant resistance proteins, and *Caenorhabditis elegans* death-4 protein) and C-terminal leucine-rich repeat (LRR), in the nucleotide-binding domain and LRR (NLR) proteins. Our findings suggest that this program will enable searches for various protein domains and motifs in all organisms.

**Key words:** *Arabidopsis*, cucumber, genome, InterProScan, NLR.

Various types of pathogens, i.e., fungi, oomycetes, bacteria, viruses, and nematodes, cause up to 15% reduction in the value of harvests, incurring major economic losses for producers worldwide. Plant innate immune systems respond to these pathogenic infections by subsequently activating sophisticated defense mechanisms against microbial pathogens. Plants recognize conserved pathogen-associated molecular patterns as a part of pattern-triggered immunity, or secreted pathogen effector proteins as a part of effector-triggered immunity (Maekawa et al. 2011). Plant genomes have a significant number of immune receptors, i.e., receptor-like kinases and receptor-like proteins (Jones et al. 2016). In plant disease resistance, typically, the intracellular immune receptors are proteins with a central NB-ARC domain (a nucleotide-binding adaptor shared by apoptotic protease-activating factor-1, plant resistance (R) proteins, and *Caenorhabditis elegans* death-4 protein) and a C-terminal leucine-rich repeat (LRR) domain, consequently termed as the "nucleotide-binding domain and LRR" (NLR) proteins", which recognize pathogen effector proteins (Figure 1). NLRs usually possess either an N-terminal toll/interleukin 1 receptor (TIR) domain or a coiled-coil (CC) domain in their structures (Li et al. 2015; Van der Biezen and Jones 1998), thereby being phylogenetically classified into the subfamilies TIR-domain-containing (TNL) and CC-domain-containing (CNL), respectively. For stable food supply, the use of *R* genes coding for the NLR

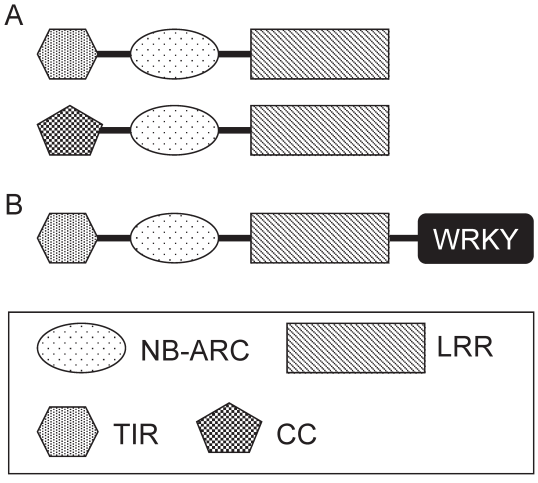


Figure 1. Structures of plant NLRs. (A) Typical plant NLRs. (B) Atypical plant NLRs.

Abbreviations: Apaf-1, apoptotic protease-activating factor-1; CC, coiled coil; CED-4, *Caenorhabditis elegans* death-4 protein; CNL, CC-domain-containing NLR; ETI, effector-triggered immunity; GFF3, Generic Feature Format version 3; NB-ARC domain, nucleotide-binding adaptor shared by Apaf-1, R proteins, and CED-4; NLR proteins, nucleotide-binding domain and LRR proteins; PAMP, pathogen-associated molecular pattern; PTI, pattern-triggered immunity; R proteins, plant resistance proteins; TIR, Toll/interleukin 1 receptor; TNL, TIR-domain-containing NLR.

This article can be found at http://www.jspcmb.jp/
Published online June 15, 2018

proteins is a major strategy adopted for improving disease resistance in crops. The traditional breeding methods involved identifying and introgressing *R* genes in crops. Therefore, a genome-wide survey of *R* genes is necessary for breeding crops that can resist diseases.

In recent years, various plant genomes have been sequenced and assembled. This genome-related information might rapidly promote plant genomics, genetics, breeding, and basic plant biological research. First, the genome of a model plant (*Arabidopsis thaliana*) used for basic plant research was sequenced in 2000 (Arabidopsis Genome Initiative 2000). The *Arabidopsis* genome provided information and set a foundation for the functional characterization of plant genes and proteins in plant biology.

Cucumber (*Cucumis sativus* L.), belonging to the Cucurbitaceae family, is a major vegetable crop worldwide and is also the first genome-sequenced vegetable crop (Huang et al. 2009; Qi et al. 2013). Several defense-related genes have been identified and cloned in cucumber (Wan et al. 2013); however, the details of the molecular mechanisms of these defensive genes remain unknown.

In the present study, we developed a genome-wide survey program of conserved protein domains and motifs based on the InterProScan sequence analysis application (ver. 5.21–60.0), named "Ex-DOMAIN" (exhaustive domain and motif annotator using InterProScan) ver.1.1, which might help in the identification, information generation, and annotation of proteins and genes, conserved protein domains and motifs, chromosomal locations, and structural diversity of target proteins.

Plants use abundant NLR proteins to detect many kinds of microbial pathogens. The number of NLR genes per plant genome was predicted using computational analysis; for example, 151 such sequences are known to be present in the *A. thaliana* genome, 62 in cucumber, 737 in apple, 438 in rice, 105 in maize, and 49 in moss (Jones et al. 2016).

InterProScan is a freely available sequence analysis tool that matches a protein against the InterPro protein signature database that combines protein signatures from many member databases into a single searchable resource (Jones et al. 2014). The data sets input into this application for this study were collected from public databases (Figure 2), including protein sequences (multi-FASTA files) and gene model information (Generic Feature Format version 3 (GFF3) files) of proteins to specify the chromosomal locations of the genes by using the GFF3 ID of proteins in *A. thaliana* accession Col-0 (sourced from TAIR: https://www.arabidopsis.org/download; Araport11_genes.201606.pep.fasta.gz and Araport11_GFF3_genes_transposons.201606.gff.gz) and cucumber (*C. sativus* L.) (sourced from ftp://www.cucurbitgenomics.org/pub/cucurbit/genome/cucumber/Chinese_long/v2/; cucumber_ChineseLong_v2.gff3.gz, cucumber_ChineseLong_v2_pep.fa.gz). The data sets were scanned using the InterProScan application, and then the databases of protein domains and motifs in *A. thaliana* and cucumber were created. InterProScan allows users to search the databases by using InterProID (InterPro accession or signature accession of the member databases of InterPro) for protein domains and motifs, including advanced searches for order specification of domains and motifs in the targeted proteins and investigation of the existence of gene pairs. In this study, we searched these databases by using the InterProID for NLR, i.e., both IDs of NBS (PF00931) and the subsequent LRRs (PF00560, PF07723, PF07725, PF12799, PF13306, PF13516, PF13855, PF14580, SM00364, SM00365, SM00367, SM00368, SM00369, SM00446, PTHR11017:SF151, PTHR11017:SF160, PTHR24209:SF15, PTHR27004:SF21, PTHR27004:SF37, PS51450, and PR00364) in *A. thaliana* and cucumber. Since each protein domain and motif can have multiple InterProIDs, the input of the InterProID demands a great deal of care.

The output data were exported to tab-delimited text files. We performed genome-wide analyses for the NLR genes (proteins) in *A. thaliana* and cucumber and subsequently identified a total of 158 and 53 NLR genes in these two species, respectively. The NLR proteins usually include, but not always, two major types, phylogenetically classified as TNL and CNL. By running the programs using both IDs of TIR-domain (SSF52200, PS50104, PF13676, SM00255, G3DSA:3.40.50.10140, and PF01582) or CC-domain (cd14798, and Coil) and the subsequent NBS-LRR, we identified 99 TNLs and 52 CNLs from *A. thaliana*, and 18 TNLs and 16 CNLs from cucumber (Table 1, Supplementary Tables S1, S2).

In addition, some NLRs consist of domain and/or motif combinations different from those in regular NLR structures, i.e., the WRKY DNA-binding domain, the LIM domain, the Solanaceae domain, a protein kinase domain, a zinc finger domain, and an MAPKKK domain (Li et al. 2015). For example, the atypical TNL *Arabidopsis* RRS1, which cooperates with RPS4 to confer resistance to different pathogens in plants (Narusaka et al. 2009), has an extra WRKY transcription factor domain. By running the programs by using the IDs of WRKY (PF03106, G3DSA:2.20.25.80, SSF118290, PS50811, and SM00774), we identified 73 and 63 proteins with the WRKY domain in *A. thaliana* and cucumber, respectively. Based on the findings of this study, we identified three of the 158 NLR genes in *A. thaliana* to encode proteins containing the WRKY domain; in contrast, in cucumber, none of the NLR genes encoded proteins that contained the WRKY domain. These findings might help us understand the structure, protein–protein interaction, and pathogen recognition of NLR

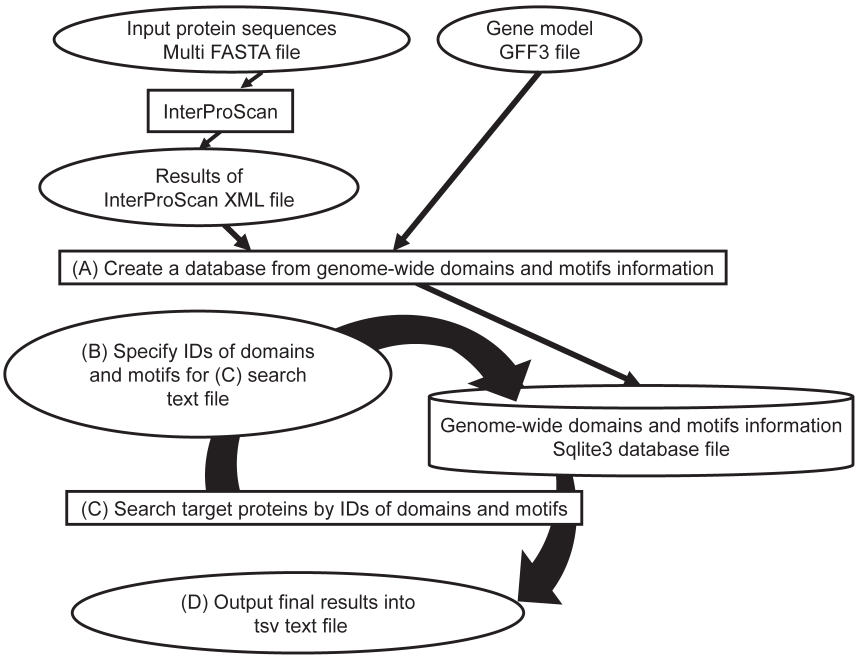


Figure 2. Flowchart of the procedure. (A) This program runs InterProScan, and then creates a database that stores the IDs of the domains and motifs with the genomic positions and annotations for every protein in an organism. (B) This file specifies the IDs of the domains and motifs for the target protein. In addition, it allows the specification of the order of the domains and motifs on the target protein. (C) This program searches the database through the text files created by procedure (B), and then provides the list of proteins. In addition, this program can designate the search condition, distance between genes, and gene orientation on the DNA strands. (D) The results file contains protein information (input protein sequence ID, gene name, annotation comment, genome position, etc.) that matches the search condition, as well as domain and motif information in the protein. For paired protein searches, this file contains paired information.

Table 1. The number of NLR genes identified by genome-wide analyses.

| Species | Common name | NLRs | TNLs | CNLs | Paired NLR genes[a] |
|---|---|---|---|---|---|
| *Arabidopsis thaliana* | Thale cress | 158 | 99 | 52 | 46 |
| *Cucumis sativus* L. | Cucumber | 53 | 18 | 16 | 10 |

[a]The number of genes defined by head-to-head (inverted) tandem arrangement within 10 kb in the genome of *A. thaliana* and *C. sativus* genomes.

function.

Recent studies also indicate that some NLRs function in pairs, such as the TNL pair RPS4 and RRS1 in *A. thaliana*, both of which are required to confer resistance to bacterial and fungal pathogens (Cesari et al. 2014; Narusaka et al. 2009), and the CNL pair RGA4 and RGA5 in rice (Cesari et al. 2013, 2014; Okuyama et al. 2011). *RPS4*/*RRS1* and *RGA4*/*RGA5* are localized near each other in a head-to-head orientation in each genome. By using the program developed in this study, we surveyed the databases of the protein domains and motifs in *A. thaliana* and cucumber for paired NLRs defined by head-to-head (inverted) tandem arrangement within 10 kb in the genome. We identified at least 46 and 10 NLR genes in the *A. thaliana* and cucumber genomes, respectively. The discovery of paired NLRs might contribute to the breeding of crops specifically for disease resistance, but not for the molecular understanding of NLR activation.

The InterProScan application is used for protein sequence analysis worldwide. Protein function prediction from genomic sequences is the ultimate goal in bioinformatics, which can provide useful information about target proteins. We designed a genome-wide survey program for protein domains and motifs to aid in the identification and characterization of proteins by using the InterProScan sequence analysis application and *A. thaliana* and *C. sativus* as model organisms. We conclude that this program can be used for conducting searches for various protein domains and motifs in other organisms as well.

## Acknowledgements

We would like to thank Aya Okada, Shoko Nieda, Yasuyo Katayama, and Masami Miyamoto of RIBS for their excellent technical assistance. This work was supported by the Science and Technology Research Promotion Program for Agriculture, Forestry, Fisheries, and Food Industry awarded to Y.N., by Research program on development of innovative technology (funding agency: Bio-oriented Technology Research Advancement Institution, NARO) to Y.N., and by Grants-in-Aid for Scientific Research (KAKENHI) to Y.N. (15K07321) and M.N. (16K08152).

## References


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

Cesari S, Bernoux M, Moncuquet P, Kroj T, Dodds PN (2014) A novel conserved mechanism for plant NLR protein pairs: The "integrated decoy" hypothesis. *Front Plant Sci* 5: 606

Cesari S, Thilliez G, Ribot C, Chalvon V, Michel C, Jauneau A, Rivas S, Alaux L, Kanzaki H, Okuyama Y, et al. (2013) The rice resistance protein pair RGA4/RGA5 recognizes the *Magnaporthe oryzae* effectors AVR-Pia and AVR1-CO39 by direct binding. *Plant Cell* 25: 1463–1481

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* 41: 1275–1281

Jones JD, Vance RE, Dangl JL (2016) Intracellular innate immune surveillance devices in plants and animals. *Science* 354: aaf6395

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30: 1236–1240

Li X, Kapos P, Zhang Y (2015) NLRs in plants. *Curr Opin Immunol* 32: 114–121

Maekawa T, Kufer TA, Schulze-Lefert P (2011) NLR functions in plant and animal immune systems: So far and yet so close. *Nat Immunol* 12: 817–826

Narusaka M, Shirasu K, Noutoshi Y, Kubo Y, Shiraishi T, Iwabuchi M, Narusaka Y (2009) *RRS1* and *RPS4* provide a dual *Resistance*-gene system against fungal and bacterial pathogens. *Plant J* 60: 218–226

Okuyama Y, Kanzaki H, Abe A, Yoshida K, Tamiru M, Saitoh H, Fujibe T, Matsumura H, Shenton M, Galam DC, et al. (2011) A multifaceted genomics approach allows the isolation of the rice *Pia*-blast resistance gene consisting of two adjacent NBS-LRR protein genes. *Plant J* 66: 467–479

Qi J, Liu X, Shen D, Miao H, Xie B, Li X, Zeng P, Wang S, Shang Y, Gu X, et al. (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat Genet* 45: 1510–1515

Van der Biezen EA, Jones JD (1998) Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem Sci* 23: 454–456

Wan H, Yuan W, Bo K, Shen J, Pang X, Chen J (2013) Genome-wide analysis of NBS-encoding disease resistance genes in *Cucumis sativus* and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops. *BMC Genomics* 14: 109