Frontiers in Plant Science

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fpls.2022.975888

NLRexpress—A bundle of machine learning motif predictors—Reveals motif stability underlying plant Nod-like receptors diversity

Eliza C. Martin[1], Laurentiu Spiridon[1], Aska Goverse[2] and Andrei-José Petrescu[1]*

[1]Department of Bioinformatics and Structural Biochemistry, Institute of Biochemistry of the Romanian Academy, Bucharest, Romania, [2]Laboratory of Nematology, Department of Plant Sciences, Wageningen University, Wageningen, Netherlands

Examination of a collection of over 80,000 Plant Nod-like receptors (NLRs) revealed an overwhelming sequence diversity underlying functional specificity of pathogen detection, signaling and cooperativity. The NLR canonical building blocks—CC/TIR/RPW8, NBS and LRR—contain, however, a number of conserved sequence motifs showing a significant degree of invariance amongst different NLR groups. To identify these motifs we developed NLRexpress—a bundle of 17 machine learning (ML)-based predictors, able to swiftly and precisely detect CC, TIR, NBS, and LRR motifs while minimizing computing time without accuracy losses—aimed as an instrument scalable for screening overall proteomes, transcriptomes or genomes for identifying integral NLRs and discriminating them against incomplete sequences lacking key motifs. These predictors were further used to screen a subset of ∼34,000 regular plant NLR sequences. Motifs were analyzed using unsupervised ML techniques to assess the structural correlations hidden underneath pattern variabilities. Both the NB-ARC switch domain which admittedly is the most conserved region of NLRs and the highly diverse LRR domain with its vastly variable lengths and repeat irregularities—show well-defined relations between motif subclasses, highlighting the importance of structural invariance in shaping NLR sequence diversity. The online NLRexpress webserver can be accessed at https://nlrexpress.biochim.ro.

KEYWORDS

plant resistance, NLR motifs, NB-ARC predictor, LRRpredictor, NLR correlations

Check for updates

OPEN ACCESS

EDITED BY
Volkan Cevik,
University of Bath, United Kingdom

REVIEWED BY
Sung Un Huh,
Kunsan National University,
South Korea
Eui-Hwan Chung,
Korea University, South Korea

*CORRESPONDENCE
Andrei-José Petrescu
andrei.petrescu@biochim.ro

SPECIALTY SECTION
This article was submitted to Plant Pathogen Interactions, a section of the journal Frontiers in Plant Science

RECEIVED 22 June 2022
ACCEPTED 19 August 2022
PUBLISHED 15 September 2022

CITATION
Martin EC, Spiridon L, Goverse A and Petrescu A-J (2022) NLRexpress—A bundle of machine learning motif predictors—Reveals motif stability underlying plant Nod-like receptors diversity.
Front. Plant Sci. 13:975888.
doi: 10.3389/fpls.2022.975888

Introduction

The extensive repertoire of NOD-like receptors (NLR) in plants, which often encode for so-called resistance (R) proteins, have a pivotal role in the plant's innate immune system to counteract a wide range of invading pathogens and pests. The co-evolutionary dynamics between pathogen and host led to a vast expansion of NLR interaction networks and operational mechanisms performed by this class of proteins sharing a more or less similar structural organization (Jones et al., 2016; Schreiber et al., 2016; Shao et al., 2016; Sukarta et al., 2016; Baggs et al., 2017; Wu et al., 2017; Cesari, 2018; Maruta et al., 2022). Canonical NLRs consist of a variable N-terminal domain, a central nucleotide-binding NBS domain and a leucine-rich repeat LRR domain (Sukarta et al., 2016). However, extensions up- and down-stream of this core region are often present and variable domains are even integrated between these core subunits to provide functional specificity to the protein (van der Hoorn and Kamoun, 2008; Cesari et al., 2014; Wu et al., 2015; Khan et al., 2016; Kroj et al., 2016; Andolfo et al., 2019).

Sequence variability of these proteins in the N-ter region led to a broadly accepted classification into several NLR groups. The largest of these are the TNL and CNL classes displaying toward their N-ter end either a toll/interleukin-1 receptor (TIR) or a coiled-coil (CC) domain, respectively. TIR domains are well-conserved structures spanning across the entire tree of life which display the so-called Rossman fold also known as the "ADP-binding βαβ fold" due to the ability to bind the ADP region of dinucleotides (Hanukoglu, 2015). TIR domains are common to several immune system proteins in higher eukaryote organisms. While highly conserved in plant TNLs, the plant TIR group diverges significantly from metazoan TIRs with distinctive, thoroughly analyzed features (Toshchakov and Neuwald, 2020) including specific additions and ablations which make them prone to a precise identification. This high TIR conservation in plant TNLs is in startling contrast to the CNL diversity in which the N-ter CC domains exhibit a very high variability. Most CCs display 4 or 5 predicted helical segments of which the 3rd is the most conserved, embedding the so-called EDVID motif which was proved to have critical functions in several systems (Rairdan et al., 2008; Maekawa et al., 2011; Hao et al., 2013; Casey et al., 2016; Slootweg et al., 2018; Wróblewski et al., 2018) and/or to be involved in CC-LRR interaction as shown by ZAR1 cryo-EM structures (Wang et al., 2019a,b). Among other, less frequently reported extensions in the N-ter region of NLRs were SD, kinase, α/β hydrolase, WRKY (Sarris et al., 2015; De Oliveira et al., 2016; Kroj et al., 2016; Andolfo et al., 2019).

In all NLRs, the most conserved region is the central NB-ARC/NBS domain which acts as an "on/off switch" that changes its configuration upon activation by a matching pathogen effector (Sukarta et al., 2016; Burdett et al., 2019; Wang et al., 2019a; Ma et al., 2020). In plants this domain consists of three subdomains: the Nucleotide Binding Domain (NBD), the Helical Domain, HD1, also known as ARC1 and the Winged Helix Domain, WHD, also termed ARC2. In the inactive state, the three subdomains collectively bind an ADP molecule. During activation, ADP is released and replaced by an ATP. This leads to a drastic change of configuration in which the ARC2 rotates ∼180° with respect to NBD/ARC1. The nucleotide-binding pocket is formed by residues located in seven highly conserved regions spread all over the three NBD, ARC1 and ARC2 subdomains, referred to herein as: VG, P-loop, Walker B—in NBD; RNBS-B, RNBS-C—in ARC1 and GLPL, MHD—in ARC2 (Tameling et al., 2002; Van Ooijen et al., 2008; Qi and Innes, 2013; Wróblewski et al., 2018; Johanndrees et al., 2021; Yu et al., 2021). Along with these seven "pocket" stretches two other regions which are instrumental in providing subdomain stability show a rather high invariance: the RNBS-A in NBD and RNBS-D in ARC2 (Wang et al., 2019b; Ma et al., 2020; Martin et al., 2020b).

Finally, the LRR domain has a solenoidal, horseshoe-like 3D architecture and consists of recurrent stretches of 15–35 amino acids length that start each with a so-called LxxLxL motif, where L is a hydrophobic amino acid, dominantly leucine, and X can be any residue (Kajava, 1998; Kobe and Kajava, 2001). Repeats are held together via a H-bond network formed among the LxxLxL stretches located on the ventral side of the horseshoe. The LRR domain is present not only in NLRs, but in many other immune system proteins, both in plants and metazoans. For instance, in plant Receptor-Like Kinases/Proteins (RLK/RLP) and metazoan TLRs the LRR domain is localized on the extracellular side of the receptor and is responsible for the direct detection of pathogen-associated molecules (Sun et al., 2013; Rajaraman et al., 2016; Eckardt, 2017; Dievart et al., 2020). Some of these RLKs require also the association with other membrane-bound receptors through their own LRR domains such as FLS2-BAK1 (Sun et al., 2013). In cytosolic NLRs, the LRR domains were shown to display various intra- and inter-molecular interacting functions. For example, the cryo-EM structures of ZAR1 CNL and RPP1 TNL (Wang et al., 2019a,b; Ma et al., 2020) indicate that the N-ter region of the LRR interacts with NBS. On the other hand, while in ZAR1 the downstream C-ter region of the LRR binds to the RKS1 kinase which is an adaptor for the recognition of the effector, in other NLRs such as RPP1, Rx or Gpa2, the LRR domain recognizes directly the effector or the pathogen molecule (Slootweg et al., 2017; Ma et al., 2020).

The working environment imposes interesting constraints on the local properties of LRR sequences. Whereas the extracellular LRR domains of RLK/RLP/TLRs display repeats of regular ∼24 aa lengths and almost invariant 3D shape, in cytosolic NLRs LRR repeats are very irregular while the LxxLxL motifs are highly variable and ambiguous, with many local alternative "satellites" (Kobe and Kajava, 2001; Kajava and Kobe, 2002; Sela et al., 2012, 2014; Baudin et al., 2019; Wang et al., 2019b). This imposes a high level of uncertainty in repeat delineation which makes the investigation of structure-function relations problematic and inflicts on model-driven analysis of LRR fold stability, interdomain and protein-protein interactions. Some more accurate machine learning-based tools were recently developed to address this problem (Bej et al., 2014; Martin et al., 2020a; Liu et al., 2022) but these are computationally costly. For instance, the LRRpredictor (Martin et al., 2020a), which was specifically designed to address motif irregularities, relies on the consensus of 8 classifiers of which 4 are based on variability profiles only, built starting from global databases, and 4 rely in addition on secondary structure predictions. Such complex methods clearly require high computational resources which makes them less usable in genome-wide or large proteome datasets scanning.

In this context lighter but similarly precise tools—able to screen overall transcriptomes and proteomes are needed in order to discriminate between complete transcripts and incomplete sequences lacking key functional motifs. To this end we introduce here NLRexpress—a bundle of 17 ML-based predictors designed to identify CC-/TIR-/NBS- and LRR-specific motifs. The main focus in developing NLRexpress was to create an instrument scalable for screening large sets of plant NLRs based on simple but effective NN models able to reduce the computing time with minimal accuracy losses. Motifs identified with NLRexpress in a large NLR atlas (NLRscape, 2022) were then subjected to unsupervised classification in order to identify correlations hidden underneath motif variabilities in individual domains and the overall NLR structure.

The online NLRexpress webserver can be accessed at https://nlrexpress.biochim.ro and the standalone version can be downloaded from https://github.com/eliza-m/NLRexpress.git.

Materials and methods

Training sets and preprocessing

In generating the training sets for NLR motifs we used data found in the NLRscape database which organizes information on over 80,000 plant NLR proteins and fragments (NLRscape, 2022). From the overall set, sequences that contain a single, complete canonical NBS domain with a proper NBD-ARC1-ARC2 subdomain layout were clustered at 90% identity and 90% coverage resulting in a subset of ∼34,000 NLR. This contains around 41% CNLs (CC-NBS-LRR), 15% TNLs (TIR-NBS-LRR), and 1.2% RNLs (RPW8-NBS-LRR) while the remaining 43% sequences are either NLs (NBS-LRR) or display incomplete or a non-canonical domain layout.

Individual CC, TIR, NBS, and LRR domain regions were then extracted from the full protein sequence and further clustered at a minimum of 70% domain coverage and various sequence identity thresholds using MMseqs2 (Steinegger and Söding, 2017).

For the CC domain, a 30% identity threshold was imposed resulting in 685 representatives. This set was further manually curated to eliminate incomplete CC entries with large deletions of 15–20 aa in helical predicted regions and retain only those entries having CC lengths of over 120 aa including the linker to NBS. By this, a final clean set of 475 sequences was retained for EDVID CC motif training.

A similar approach was used in generating the TIR motifs training data. By contrast to the CC domains which are highly variable, TIRs are significantly more conserved—therefore a 60% identity cutoff was enough to obtain a set of 881 representatives. After the manual inspection and elimination of incomplete motifs, 490 sequences were retained for training the TIR module.

As the NBS region displays not less than nine highly conserved potential motifs (VG; P-loop; RNBS –A, –B, –C, and –D; Walker B, GLPL, and MHD) spread over the three subdomains of NBS, only sequences displaying the complete NBD-ARC1-ARC2 layout were selected. By using a cutoff of 50% sequence identity at 70% coverage, a set of 1,910 representatives was obtained, which was further reduced to 861 by the elimination of incomplete N-/C- termini or missing motifs sequences.

Finally, for LRR motif detection, a curated set of 117 LRR domains containing ∼2,000 core LRR motifs was selected for the overall training set resulting from the following workflow. In a first step, only sequences containing both complete NBS and LRR domains were retained. The LRR domain regions of these sequences were clustered at 50% identity and clusters were ranked according to their dimension. From these, only the top 117 clusters were retained, containing each over 40 sequences while the rest were treated as sparse LRR outliers. Sequences within the top 117 largest clusters were further subjected to LRR motif delineation with LRRpredictor (Martin et al., 2020a) and in-depth sequence profiling and analysis for representative selection. While the level of intra-cluster identity ensures a robust structural homology, the sequence diversity induces local variations of profiles such as the secondary structure state or the LRR motif probability—these were in turn used as selection criteria in choosing the optimal cluster representative.

Feature selection

In most state-of-the-art sequence-based predictors, variability profiles are computed based on remote homologs retrieved from very large databases such as Uniprot20 or Uniclust30/50. However building such HMM/PSSM profiles starting from large databases comes with a significant computational cost, especially when the protein query contains repetitive patterns such as LRR motifs—due to the vast increase in multiple alignment possibilities of such repetitive stretches found in the hits of a given queried sequence. As this is the time-limiting step in generating the NLR profiles, instead of using the Uniprot/Uniclust datasets, we explored the possibility of compiling downscaled custom-generated focused sets in achieving the best tradeoff between the computational cost and the HMM profile relevance. To this end, in building profiles, a set of 2,361 sequences was generated consisting of a collection of 1,361 plant NLRs of at least 500 amino acids filtered at 20% identity and an extra group of 1,000 LRR domains larger than 200 aa filtered at the same identity level, but derived from different protein types and origins—added in order to allow the predictor to detect swiftly in a given genome all kinds of LRRs, not only those NLR specific. This focused dataset was further used by all the 11 predictors included in NLRexpress to generate the training features which consist of match emissions probabilities inferred using the JackHMMER suite (Johnson et al., 2010; Potter et al., 2018) with a 1e-5 expect value cutoff. The first and second iterations of HMM profiles were investigated as features and using them in conjunction proved to yield the best results.

Network architecture—Training and performance evaluation

Supervised classifiers were trained for each of the 17 individual motif types—specific to the CC (extended EDVID), TIR (βA, αA, βC, αC, βD-αD1, αD3), NBS (VG; P-loop; RNBS –A, –B, –C, and –D; Walker B, GLPL, MHD) and LRR (LxxLxL) domains. For each residue in the input sequence, the individual predictors compute the probability to be the starting point of a motif by using as features the negative natural log of HMM emission probabilities to each 20 possible amino acids computed in the first and second iteration over a sliding window of 5 positions upstream and downstream motif borders adding up to a total number of 20 × 2 × (5 + motif length + 5) input features.

For training the individual predictors, the CC, TIR, NBS, and LRR motif datasets were split into 5 equal groups. The first 4 groups were used for parameter optimization using a fourfold cross-validation scheme, while the fifth group was left aside and used as the final test set. In order to tradeoff between class imbalances, non-linear separable problems on small datasets and computational costs, a multilayer perceptron network (MLP) architecture, known to effectively mitigate such issues, was chosen as the supervised algorithm. The hyperparameter tuning stage was conducted using a fourfold cross-validation scheme in which the number of hidden layers, neurons per layer, the solver and regularization parameters were subjected to optimization. A network with three hidden layers of 250, 125, and 100 neurons, respectively, was found to optimally suit all the 17 classifiers, while the Adam stochastic gradient-based optimizer (Kingma and Ba, 2015) with a constant learning rate of 0.001, ReLu activation function and L2 penalty regularization with customized alpha parameters for each predictor were selected as network parameters. The out-of-sample performance was evaluated on the test set using the precision, recall, specificity, F1 and G scores metrics. For training and evaluating the predictors, the scikit-learn library v24.2 (Pedregosa et al., 2011) was used.

Nod-like receptor motifs analysis

NLRexpress was then used to predict NLR motifs on the overall set of ~34,000 NLRs trimmed at 90% identity. As the overall database contains not only regular NLRs but also all sorts of non-canonical sequences with missing, duplicated or shuffled NLR domains, selections were performed to derive subsets fulfilling specific criteria in order to carry out more targeted correlation analysis.

For instance, in order to identify correlations within NBS motifs a subset of only ~20,000 sequences was selected containing at the same time all the nine motifs—only once, in the right order and with a probability estimate higher than 80%. Extended regions conveying each motif and 5 flanking positions on both sides were then extracted and concatenated in the right order. These excised NBS chimeras were subjected to sequence-based clustering using MMseq2 (Steinegger and Söding, 2017) at various identity thresholds using the connected component clustering method as implemented in the MMseq2 suite. Correlation analysis was performed using the CMAT method (Jeong and Kim, 2012) as implemented in VisualCMAT (Suplatov et al., 2018) and MISTIC (Simonetti et al., 2013). Henikoff weights and profile-based pseudocounts were used for join probability adjustments. A minimal Z-score of 3.5 was set as cutoff for both cumulative and proximity mutual information metrics.

For LRR motif distribution analysis domains were selected only if: **(a)** they contain upstream a proper NBS domain with all 9 motifs unambiguously identified in the right order; **(b)** they display the LRR domain downstream NBS with a linker less than 50aa; **(c)** they contain at least 9 LRR repeats of expected regular lengths: 15–50 aa. This last constraint was imposed to avoid ambiguous motif assignments. An exception was made though in cases where the difference in probability among local motif "satellites" found in less than 15 aa surpasses 20%—a situation in which the motif with the highest probability was selected. With these cleaning constraints, a subset of ~6,850 regular NLR sequences were selected containing a total number of ~61,700 LRR motifs.

The LRR motifs were further clustered based on the physico-chemical properties of amino acids in the LxxLxL region. For each of the six positions the hydropathy, charge and volume were used as descriptors summing up to an 18-dimensional clustering embedding. Given its ability to deal with uneven cluster sizes, densities and non-flat geometries the OPTICS algorithm (Ankerst et al., 1999) was selected from several unsupervised clustering methods tested, and ran using its scikit-learn library implementation along with the Xi algorithm as the extraction method and the Minkowski norm 2 metric (Pedregosa et al., 2011).

For the analysis of the N-ter NLR region, TNL/CNL, a sequence-based clustering approach was used similarly to the NBS domain analysis. In the case of TNLs, a subset of 3,774 sequences was gathered to contain a TIR domain with all 6 TIR motifs unambiguously identified in the right order and a downstream proper NBS domain with all 9 NBS motifs. The extracted motif regions with 5 flanking positions on both sides were concatenated and subjected to sequence-based clustering using MMseq2 (Steinegger and Söding, 2017). For CNLs, in addition to the above-mentioned criteria imposed to identify regular NLR proteins, a supplementary condition was used regarding the presence of a single integral CC domain upstream of the NBS. Given that all the potential CNL interdomain interactions were shown to take place within the first five LRR repeats (Qi et al., 2012; Slootweg et al., 2013, 2017; Wang et al., 2019b) CNLs displaying at least five proper repeats, not nine as in LRR domain analysis, were selected resulting in a clean 6,000 CNL subset which was further subjected to a workflow similar to the one described above. As in the NBS analysis—regions covering all the 15 motifs within the CC, NBS and the first five N-ter LRR, each with 5 flanking positions on both sides were extracted and concatenated in the right order to be subjected to sequence-based clustering using MMseq2 (Steinegger and Söding, 2017) with various identity thresholds. Logomaker (Tareen and Kinney, 2020) was used to generate sequence variability plots, while figures illustrating 3D protein structures were edited using (PyMOL, 2018; The PyMOL Molecular Graphics System, Version 2.2.3 Schrödinger, LLC).

Results and discussion

NLRexpress structure, development, and performance analysis

NLRexpress consists of a collection of 17 NN-based predictors trained to identify individual plant NLR domain-specific motifs and it is organized as four prediction modules: CC-, TIR-, NBS-, and LRRexpress that can be also used independently. The CC- and LRRexpress modules contain a single motif predictor for the extended EDVID and LxxLxL pattern respectively, while TIRexpress comprises a bundle of 6 motif predictors for the following most preserved elements: βA, αA, βC, αC, βD-αD1, αD3, and NBSexpress contains a bundle of 9 individual motifs for the following conserved regions: P-loop; RNBS –A, –B, –C, and –D; Walker B, GLPL, and MHD (**Figure 1A**).

The workflow of NLRexpress is shown in **Figure 1**. As input, the predictor accepts amino acid sequences provided by the user in FASTA format. In a first step which is also the most time-consuming, NLRexpress generates the numerical features required by the NN models. Next, the NN-model input is processed and fed to the prediction modules. Users can select either a single CC/TIR/NBS/LRR module or all four together. The output consists of a probability estimate for each residue of the sequence to be the starting position of the given motif (**Figure 1B**).

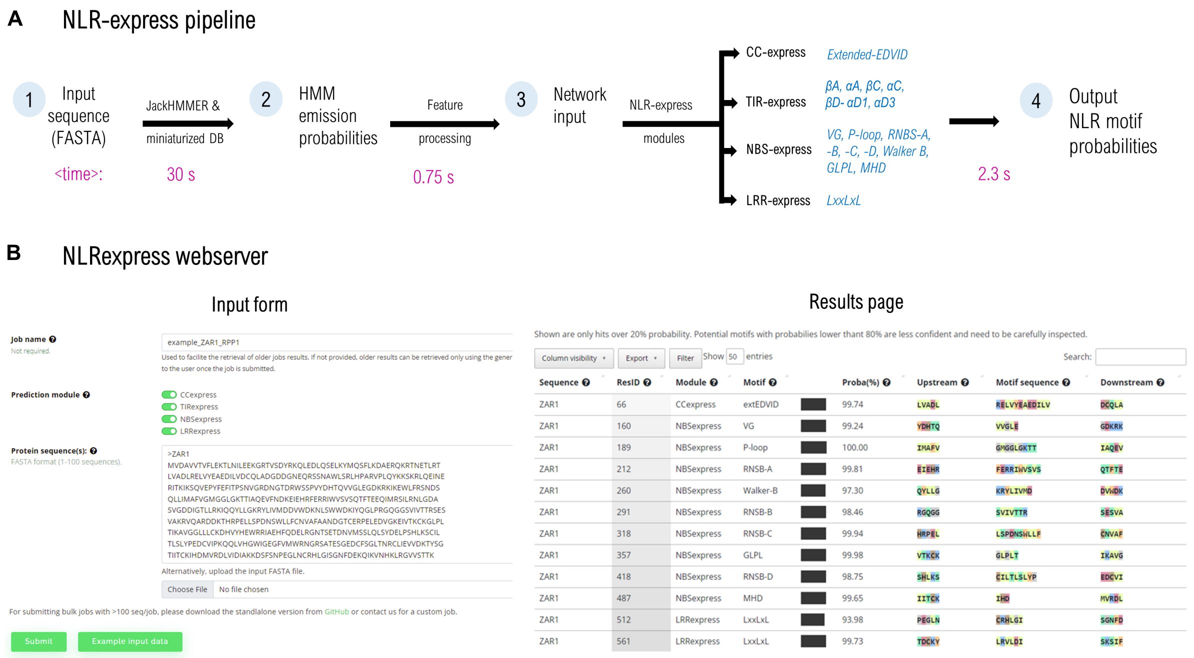
For generating the features, HMM profiles are usually learnt from a set of homologs queried over a protein sequence database. Today state of the art predictors generally rely in this step on global protein DBs such as Uniprot20 (∼30 GB; ∼8.3 million sequences), and Uniclust30 (∼9 GB; ∼10 million sequences), the scanning of which results in high computational costs. In order to reduce the execution time of this stage, NLRexpress alternatively relies on a downscaled, manually curated dataset. This ensures an optimal tradeoff between performance and speed. For instance, using the downscaled dataset reduces eightfold the time required to build the HMM profile for ZAR1 using Uniclust30, at the same accuracy. However, even when starting from such small datasets of only ∼2,400 seq, the longest duration of feature generation is 15-fold higher than the rest of the workflow including predictions—which take on average less than 3.5 s per protein (**Figure 1A**).

As described under methods, curated sets of CC, TIR, NBS, and LRR domains were used to train and test the 17 individual NN models. The four domain sets were divided into five subsets of which the first 4 were used in a fourfold cross-validation scheme for parameter tuning while the fifth was left aside for a final out-of-sample predictor performance testing. The workflow results in four predictor modules for the CC, TIR, NBS, and LRR motifs, respectively.

The CC predictor focuses on the EDVID motif found in many CC domain classes. The initial training trials were performed using the minimum EDVID motif span and proved less effective. However, by using an extended version of the motif spanning over 12 positions covering the upstream region involved in the CC-LRR interfacing in the ZAR1 cryoEM structure (Wang et al., 2019a,b)—with *RDbbbDb<u>EDbbD</u>* as consensus (b-hydrophobic residues)—significantly improved the F1 score to a proper 96% value (**Figure 2**).

The TIR module identifies six highly conserved plant TNLs regions (Toshchakov and Neuwald, 2020): βA, αA, βC, αC, βD-αD1, αD3 (where α and β refer to helical and beta-strand elements and A-D correspond to structural modules of the TIR domain). Given their high conservation, predictors attain almost perfect performance with F/G scores above 99% (**Figure 2**).

The NBS predictor targets 9 motifs: *P-loop; RNBS –A, –B, –C, and –D; Walker B, GLPL,* and *MHD*. The highest scores were reached by the most conserved NBS motifs with precision, recall and specificity and F/G scores surpassing 96% (**Figure 2**). A somehow lower sensitivity/recall of only 88% was



Job name ❓
Not required.

example_ZAR1_RPP1

Used to facilite the retrieval of older jobs results. If not provided, older results can be retrieved only using the gener
to the user once the job is submitted.

Prediction module ❓

- ● CCexpress
- ● TIRexpress
- ● NBSexpress
- ● LRRexpress

Protein sequence(s): ❓
FASTA format (1-100 sequences).

```
>ZAR1
MVDAVVTVFLEKTLNILEEKGRTVSDYRKQLEDLQSELKYMQSFLKDAERQKRTNETLRT
LVADLRELVYEAEDILVDCQLADGDDGNEQRSSNAWLSRLHPARVPLQYKKSKRLQEINE
RITKIKSQVEPYFEFITPSNVGRDNGTDRWSSPVYDHTQVVGLEGDKRKIKEWLFRSNDS
QLLIMAFVGMGGLGKTTIAQEVFNDKEIEHRFERRIWVSVSQTFTEEQIMRSILRNLGDA
SVGDDIGTLLRKIQQYLLGKRYLIVMDDVWDKNLSWWDKIYQGLPRGQGGSVIVTTRSES
VAKRVQARDDKTHRPELLSPDNSWLLFCNVAFAANDGTCERPELEDVGKEIVTKCKGLPL
TIKAVGGLLLCKDHVYHEWRRIAEHFQDELRGNTSETDNVMSSLQLSYDELPSHLKSCIL
TLSLYPEDCVIPKQQLVHGWIGEGFVMWRNGRSATESGEDCFSGLTNRCLIEVVDKTYSG
TIITCKIHDMVRDLVIDIAKKDSFSNPEGLNCRHLGISGNFDEKQIKVNHKLRGVVSTTK
```

Alternatively, upload the input FASTA file.

Choose File No file chosen

For submitting bulk jobs with >100 seq/job, please download the standlalone version from GitHub or contact us for a custom job.

Submit Example input data

Shown are only hits over 20% probability. Potential motifs with probabilies lower thant 80% are less confident and need to be carefully inspected.

Column visibility ▾ Export ▾ Filter Show 50 entries Search:

Sequence ❓	ResID ❓	Module ❓	Motif ❓		Proba(%) ❓	Upstream ❓	Motif sequence ❓	Downstream ❓
ZAR1	66	CCexpress	extEDVID	■	99.74	LVADL	RELVYEAEDILV	DCQLA
ZAR1	160	NBSexpress	VG	■	99.24	YDHTQ	VVGLE	GDKRK
ZAR1	189	NBSexpress	P-loop	■	100.00	IMAFV	GMGGLGKTT	IAQEV
ZAR1	212	NBSexpress	RNSB-A	■	99.81	EIEHR	FERRIWVSVS	QTFTE
ZAR1	260	NBSexpress	Walker-B	■	97.30	QYLLG	KRYLIVMD	DVWDK
ZAR1	291	NBSexpress	RNSB-B	■	98.46	RGQGG	SVIVTTR	SESVA
ZAR1	318	NBSexpress	RNSB-C	■	99.94	HRPEL	LSPDNSWLLF	CNVAF
ZAR1	357	NBSexpress	GLPL	■	99.98	VTKCK	GLPLT	IKAVG
ZAR1	418	NBSexpress	RNSB-D	■	98.75	SHLKS	CILTLSLYP	EDCVI
ZAR1	487	NBSexpress	MHD	■	99.65	IITCK	IHD	MVRDL
ZAR1	512	LRRexpress	LxxLxL	■	93.98	PEGLN	CRHLGI	SGNFD
ZAR1	561	LRRexpress	LxxLxL	■	99.73	TDCKY	LRVLDI	SKSIF

FIGURE 1
(A) NLRexpress pipeline diagram with typical computation times shown below for a queried NLR sequence of average length. **(B)** Screen capture of the NLRexpress webserver: input form and the results page containing the prediction output. Average times may vary depending on hardware and sequence properties.

NLR motifs prediction – performance evaluation

Module	NLR motif	Consensus	TN	FP	FN	TP	Precision	Recall	Specificity	F-score	G-score
CC-express	ext-EDVID	RDbbbDbEDbbD	12739	2	6	85	0.98	0.93	1.00	0.96	0.96
	βA	FbSFRGEDbR	17078	0	1	96	1.00	0.99	1.00	0.99	0.99
	αA	xxFbSHL	17371	0	1	97	1.00	0.99	1.00	0.99	0.99
TIR-express	βC	SRbxbbbF	17273	0	0	98	1.00	1.00	1.00	1.00	1.00
	αC	WCLxEL	17469	0	0	98	1.00	1.00	1.00	1.00	1.00
	βD-αD1	VbPbFYxbDPSDVRxQ	16489	0	0	98	1.00	1.00	1.00	1.00	1.00
	αD3	WRxALxxbAxbxG	16783	0	0	98	1.00	1.00	1.00	1.00	1.00
	VG	bbGRE	62294	1	20	152	0.99	0.88	1.00	0.94	0.94
	P-loop	GbGGbGKTT	61607	0	0	172	1.00	1.00	1.00	1.00	1.00
	RNBS-A	FDbRbWbxVS	61429	6	2	170	0.97	0.99	1.00	0.98	0.98
	Walker B	KRFbbbbDD	61774	5	5	167	0.97	0.97	1.00	0.97	0.97
NBS-express	RNBS-B	KbbbTTR	61949	2	3	169	0.99	0.98	1.00	0.99	0.99
	RNBS-C	LxxxExWxLF	61435	0	1	171	1.00	0.99	1.00	1.00	1.00
	GLPL	GLPLA	62295	0	0	172	1.00	1.00	1.00	1.00	1.00
	RNBS-D	CFbYCxLFP	61607	0	1	171	1.00	0.99	1.00	1.00	1.00
	MHD	bHD	62636	3	0	172	0.98	1.00	1.00	0.99	0.99
LRR-express	LRR	LxxLxL	21435	33	29	367	0.92	0.93	1.00	0.92	0.92

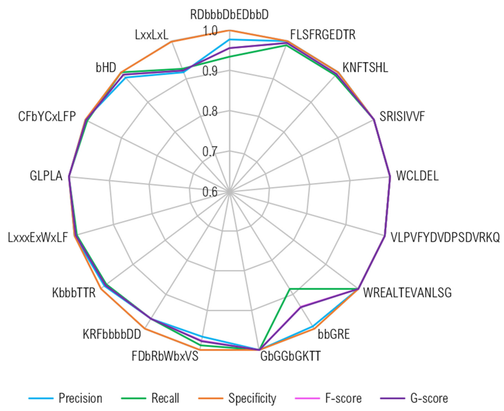


FIGURE 2
Out-of-sample performance of the NLRexpress predictors on the test set (TP, true positive; TN, true negative; FN, false negative; TP, true positive). Cross-validation out- and in-sample performance results are shown in Supplementary Table 1.

shown by the VG motif which is expected given its higher variability. Detailed stats on the cross-validation *out-of-/in-sample* performance can be examined from Supplementary Table 1.

Finally, LRRexpress focuses on the detection of *LxLxxL* motifs *in LRR domains in general*. This yields a total F-score of 92% on the test set, with a balanced precision to recall ratio. As an additional provision ZAR1, ROQ1 and RPP1 clusters were set aside from the training data and used only in the out-of-sample predictor performance testing as these are so far the only plant CNLs/TNLs with known full-protein 3D structure. LRRexpress is able to accurately identify 12/13 repeats in the case of ZAR1, while 19/20 and 21/21 in the case of RPP1 and ROQ1, respectively, which is consistent with the ∼90% sensitivity estimated on the test set.

While the out-of-sample results assessed above involve only plant NLRs, the extrapolation ability and performance of LRRexpress were next investigated on other protein classes containing LRR domains.

Using a set of 178 structures from PDB containing LRR domains at 90% sequence identity, consisting of ∼2,000 manually curated and delineated LRR repeats from Martin et al. (2020a), LRRexpress reaches an F1 score of 92% when considering only the core LRR motifs and 88% when the border, more diverse N-ter and C-ter motifs are included (**Supplementary Figure 1A**). Compared to the plant NLR set, the precision/recall ratio is much more tilted toward higher precision rates (98%) rather than sensitivity (88% on core motifs and 80% when border motifs are included).

In order to further evaluate the LRRexpress performance a thorough comparison with a former more laborious method implemented in LRRpredictor (Martin et al., 2020a) was undertaken. For this LRRexpress was run on the same set used previously for testing the LRRpredictor out-of-sample performance. While this yields a somehow lower F-score when compared to LRRpredictor, ∼85% vs. 90%, it displays on the other hand a higher precision, 99% vs. 92%. Results indicate that LRRexpress shows an overall 88–99% agreement rate with LRRpredictor in pinpointing the same LRR motifs on various LRR-containing protein classes: cytosolic plant resistance proteins (CNL and TNL), extracellular plant receptors with or without kinase domains (RLK/RLP) and their vertebrate corresponding classes—cytosolic NLRs and extracellular TLRs (**Supplementary Figure 1C**). However, the cross-diagonal cases where the two predictors disagree have a large proportion of either atypical motif patterns or result from ambiguities related to motif selection in regions showing multiple local alternatives.

To assess its discriminative capacity, LRRexpress was also tested using as inputs other solenoidal architectures containing high levels of LxxLxL patterns that might result in false-positive predictions. On five benchmark sets each comprising 50 sequences from ankyrin, armadillo, trimeric, and pectate lyases classes from Martin et al. (2020a), LRRexpress is able to correctly classify the LxxLxL patterns as non-LRR motifs, with almost no false positives in 1,000–2,700 LxxLxL patterns per set (**Supplementary Figure 1B**).

NLRexpress was further subjected to speed benchmarks on two sets of sequences: (–1) the entire *A. thaliana* proteome consisting of ∼137.000 sequences from UniprotKB; and (–2) A random set of 1,000 plant NLRs retrieved from the overall set. These were subjected to the NLRexpress pipeline for identifying NBS and LRR motifs, on an Intel(R) Core(TM) i9-9900K CPU machine running in three parallel jobs of 4 CPU threads each. The whole Arabidopsis proteome was screened in 31.5 h yielding 930 sequences with complete nine-motif NBS and 4,461 sequences with LRR domains (at least 5 repeats). Given that variability profiles are generated in NLRexpress starting from a reduced and targeted set of clean NLRs supplemented by LRRs from diverse sources—during a genome/proteome/transcriptome scanning the program swiftly discards, in less than 2 s/non-NLR-sequence the "non-NLR/LRR" regions spending most of the time in building profiles of the regions embedding the NLR-specific domains—arr. 32 s/NLR-sequence on average over the 1,000 NLR set, which was scanned entirely in ∼9 h by a 4 CPU thread job.

These results indicate that the LRR module of NLRexpress and LRRpredictor complement each other—given that LRRpredictor is structure-oriented while LRRexpress is sequence oriented, based on a "clean" set of examples—and in conjunction are able to generate an accurate description of LRR domains for structural modeling. On the other hand, this comes with a clear speed advantage for LRRexpress which in addition is able to scan entire genomes/proteomes/transcriptomes for identifying all LRR domains in a reasonable time.

Motif analysis

NLRexpress was then next used to analyze the overall set of ∼34,000 NLR and NLR fragments at 90% identity in order to identify the main patterns and correlations shaping up in NLR proteins at both individual-domain and global NLR levels.

As the 17 motifs are the main invariants of an NLR sequence these were subjected to clustering, based both on amino acid sequence with Blosum-derived metrics and on physio-chemical properties such as hydropathy, side-chain charge and size.

While distant in sequence, some of these motifs are or might be brought close in space by folding or during the NLR functioning cycle. For instance, seven of the NBS motifs are brought together to form the ADP/ATP binding site of the domain, moreover, the CC/TIR, NBS, and LRR might be permanently or transiently in contact with each other in given functional states of an NLR. Therefore the motif analysis was extended to investigate potential correlations between groups of motifs according to the functional context.

The coiled-coil extended EDVID motif

Within the overall set of ∼34,000 NLR and fragment sequences, the CCexpress predictor was able to identify with over 80% probability ∼22,800 extended-EDVID motifs in 20,000 sequences. This excess is due to the fact that many sequences in our database have a more complex domain layout than that of a canonical CC-NBS-LRR layout, with repeated, multiple and/or shuffled arrangements of CC, NBS, LRR, TIR. By focusing now exclusively on the canonical CNL (X-CC-NBS-LRR-X) subset consisting of ∼13,150 sequences 97% display such extended-EDVID motifs (12,815).

The six toll/interleukin-1 receptor motifs

A set of 3,774 TNL sequences which simultaneously contain all 6 TIR and 9 NBS motifs was gathered, and the excised motif sequences with a margin frame of five residues were concatenated and subjected to clustering as described in methods. Using a threshold of 55% identity, around 75% of the TNL sequences cluster in top six largest clusters (**Figure 3**).

The cryo-EM structures of RPP1 and ROQ TNLs (Ma et al., 2020; Martin et al., 2020b) posit the TIR domains in an asymmetrical tetrameric configuration, binding two ATP molecules between TIRs type 1 and 2 conformation. The nucleotide-binding site 1 consists of βA strand and αC helix and are highly conserved among all TIR classes—particularly the catalytic glutamate and neighboring cysteine in positions 1&4 of αC motif residue experimentally proved essential for NAD* hydrolysis and/or cNMP synthetase activities (Ma et al., 2020; Lapin et al., 2022; Yu et al., 2022). The second ATP binding site is located on the neighboring TIR and consists in (i) the βD-αD1 beta-loop-helix, particularly with conserved Phe and Pro in positions 4 and 9 respectively and (ii) the helical element αD3 and the downstream loop (positions 8–14), more variable across various TNL classes. Besides binding ATP, the regions αA, αC, βD-αD1, and αD3 also participate in TIR-TIR interfacing in the tetrameric structure, and display a higher variability between the identified clusters, suggesting a coevolution process.

The nine NBS motifs

The above-mentioned sequence complexity also inflicts upon the NBS subdomain integrity and organization—as some of the sequences show missing or duplicated regions containing one or multiple NBS motifs. Therefore, the overall set was cleaned to retain only those sequences that contain simultaneously all of the 9 NBS motifs, predicted by NBSexpress, in the right order, with a probability higher than 80%. This reduces the initial set to around 20,000 sequences, with a similar rejection rate of ∼60% as that seen when curating the NBS training set. The 9 motifs were then simultaneously clustered based on their amino acid sequence at different identity thresholds. At 55% identity, 85% of the NBS motifs are conveyed into the top 9 clusters, the largest one comprising ∼28% of all sequences (**Figure 4**). Interestingly this clustering method, based exclusively on the sequence information confined only within the 9 NBS motifs also adequately separates between NLR types suggesting that differences in the molecular environment in which NBS is confined are a driving force in diversification.

As expected, the most invariant motifs are those containing amino acids forming the ADP/ATP binding sites—especially the P-loop region, Walker-B β-strand, RNBS-B, GLPL and MHD—which is consistent for instance with mutagenesis experiments results on NRC1, SW5, I-2, Mi-1, etc., indicating that many critical single amino acid substitutions are located in the NBS highly conserved motifs (Tameling et al., 2006; Van Ooijen et al., 2008; Sueldo et al., 2015; de Oliveira et al., 2018; Wang et al., 2020). By contrast, the VG, RNBS-A, –C, –D motifs are more diverse, with cluster-specific characteristics.

Of these, the most diverse is the VG motif (*bbGRE*), located right at the border between the CC and NBS domains. In all clusters, the consensus extends with two more hydrophobic positions, denoted by *b*, upstream VG: *bb*VG. The cryo-EM structure of ZAR1 shows that these *bb*V are in contact with the nucleotide in both resting and activated states. The next two positions downstream of the conserved glycine (positions 3 and 4 in **Figure 4**) are cluster-specific: *VGRE/D* in the largest CNL cluster (NBS #1), while *VGIE/D* (NBS #2,#5), *VGEx* (NBS-#7) in the other CNL clusters. The most atypical is cluster NBS-#4 which displays a significantly divergent motif *FESR*, in which the glycine is replaced by a serine. Oppositely, in the case of TNLs and RNLs the motif consensus is robust—*VGIE/D*—with a hydrophobic and acidic residue downstream of the VG (positions 3, 4; **Figure 4**).

Next in sequence is the P-loop motif (also termed Walker-A with consensus *GbGGbGKTT*) which is highly invariant across the identified clusters, with modest variations at given positions. Position —1 is either populated by hydrophobic aliphatic (mostly valine) or aromatic residues (tyrosine—in several CNL groups or tryptophan in TNLs). Position 1, which is majorly composed of aliphatic hydrophobic residues, usually methionine, is frequently replaced with proline in a small cluster of TNLs (NBS-#6) found only in malvids clade. The most notable differences are spotted in the RNLs group, where the P-loop motif is less stringent, especially in positions 0–2 (**Figure 4**).

The highly hydrophobic motif RNBS-A is located in the central, second beta-strand of the NBD and has a structural role in stabilizing strands 1 and 3 which both contribute to the formation of the nucleotide binding site. While quite variable across clusters, a strong preference shapes up for Phe in position 0 of all clusters, with the CNLs Trp in position 5 replaced by Phe in TNLs and RNLs (**Figure 4**).

Walker-B is located in the next beta-strand. As being involved in nucleotide binding, this motif—*KRFbbbbDDbW*—is highly invariant. Here, letter *b* denotes hydrophobic residues which are mostly aliphatic with the exception of the last one, which in CNLs is frequently a tryptophan (**Figure 4**).

As also being involved in ligand binding, the RNBS-B motif—K*bbb*TTR—is also highly conserved regardless of the NLR type, with variability occurring only in the flanking regions. Notably, in the small cluster of malvid TNLs (NBS-#6) the arginine in position 6 is less conserved, which is interesting as this position is expected to actively participate in binding the nucleotide (**Figure 4**).

Within the RNBS-C motif (consensus L*xxx*ExW*x*LF), only the leucines from positions 0 and 8 are highly conserved as they are part of the binding site, while the surrounding positions of the motif are subjected to increased variability. This can be explained by the location of the motif, which covers the

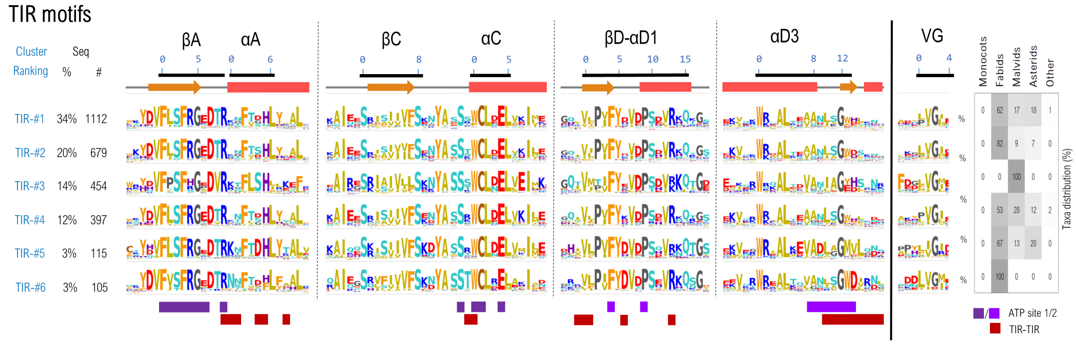


FIGURE 3
Sequence clustering of the TIR domains based on the confined motif regions. Motif variability is expressed as relative entropy displayed as letter height, while the taxonomic distributions of each cluster is depicted in the right panel. The ATP contacts within 5Å are mapped below the motif consensus (light/dark purple—ATP binding sites 1 and 2 PDB: 7crc). Amino acids are colored according to their properties as follows: yellow, hydrophobic; orange, aromatic; brown, cysteine; cyan, polar; red, acidic; blue, basic; gray, proline, and glycine.

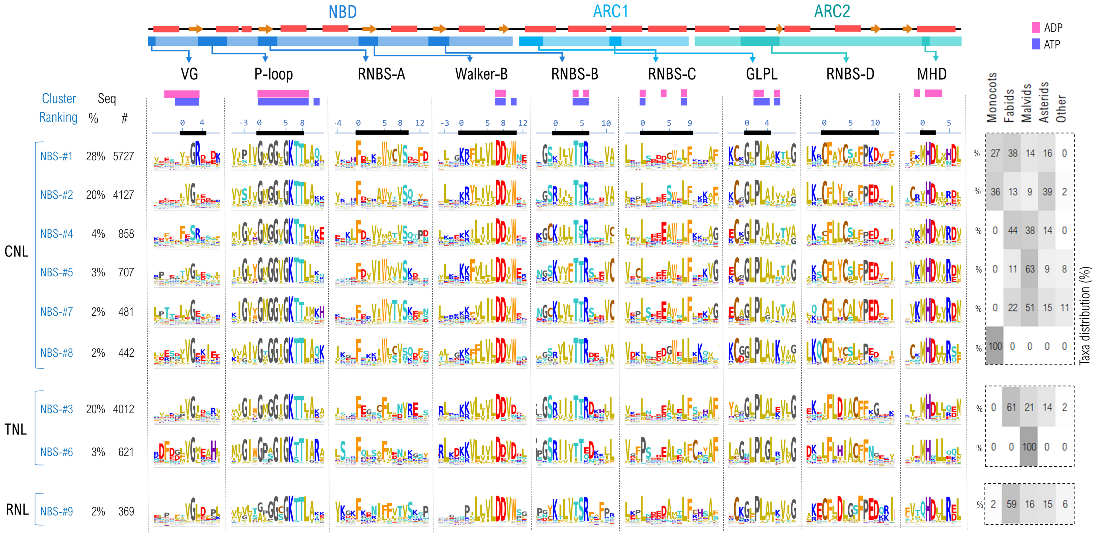


FIGURE 4
Sequence clustering of NBS domains based on the confined 9 NBS motifs. Motif variability is expressed as relative entropy displayed as letter height—the larger the more conserved. Top clusters are grouped according to the preponderant NLR class: CNL, TNL, or RNL. The taxonomic distributions of each cluster is depicted in the right panel. The ADP/ATP contacts within 5Å are mapped above the motif consensus (magenta—the inactive ADP-bound ZAR1 PDB: 6j5w; purple—the activated ATP-binding ZAR1 PDB: 6j5t). Amino acids are colored as described in Figure 3 caption.

NBD-ARC1 linker and the beginning of the first helical segment of ARC1 subdomain (Figure 4).

The GLPL motif is very well conserved across NLR classes, as being a key region in forming the binding site. Notable particularities that can be mentioned are the increased preference for cysteine at position −2 in CNLs and RNLs compared to TNLs in which this cysteine is less conserved.

The RNBS-D motif—CFbYCxLFP—is not involved in binding the ligand but in NBS-LRR interface and shows the highest divergences across the NLR classes. Common to all three NLR classes seems only the high occurrence of Phe in positions 1 and 7. Interestingly positions 0 and 4 appear to show class-specific preference—with two Cys in CNLs; Cys and Leu,respectively, in RNLs; and Leu and Ile in TNLs. In addition the triad Pro-Glu-Asp shapes up in CNLs and RNLs but not in TNLs (Figure 4).

Finally, the MHD motif is well conserved across all NLR classes, whereas variations occur in regions next to the motif. Interestingly, only the NBS-#6 small cluster on TNLs does not have the aspartic acid conserved in position 2, while in RNLs, the methionine in position 0 is often substituted by glutamine (Figure 4).

As seven out of nine NBS motifs are brought close in space in forming the nucleotide binding sites, a correlation

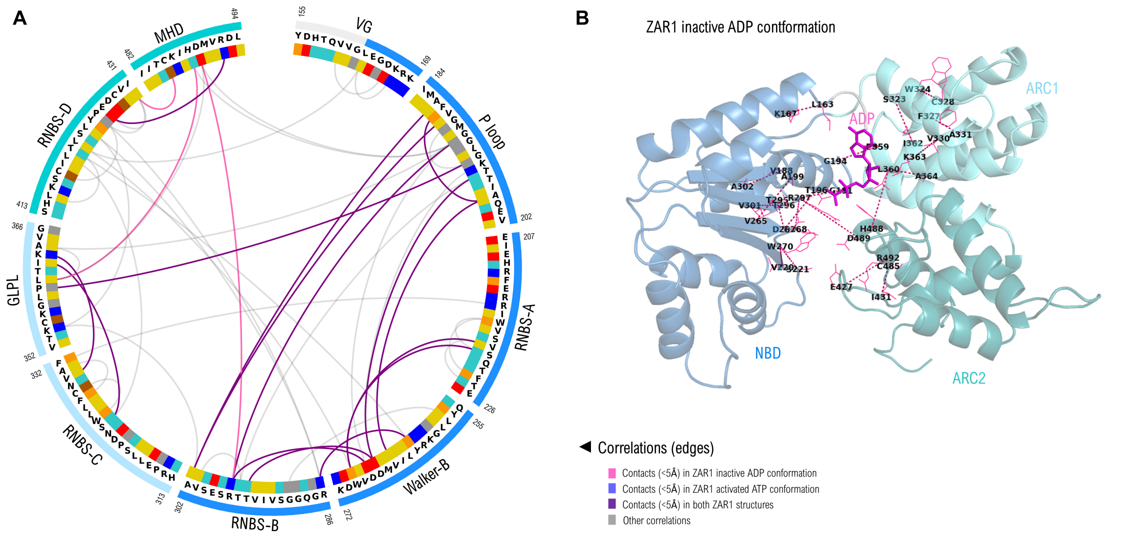


FIGURE 5
(A) Correlated amino acid pairs within the nine NBS motifs. Correlations endorsed by 3D space proximity (<5Å) in ZAR1 inactive ADP cryo-EM structure (6j5w) are shown in violet, while contacts found also in pentameric ATP structure (6j5t) are shown in purple. Correlations not supported by 3D proximity are depicted in gray and were computed using VisualCMAT (Suplatov et al., 2018). Amino acid color code is detailed in **Figure 2** caption. **(B)** Correlated interaction pairs mapped onto the cryo-EM inactive ZAR1 structure (PDB: 6j5w). Amino acid pairs are depicted with dotted lines. Domain color code is as in **(A)**: NBD, blue; ARC1, light blue; ARC2, teal.

analysis was performed in order to identify co-evolving residues. Only the close-in-space correlated pairs were further analyzed and shown in **Figure 5**, which are numbered according to their positions in ZAR1. As can be seen from **Figure 5** most of these close-in-space, correlated positions are found in the NBD subdomain. This can be explained by its large size but also by the beta-sheet architecture that brings close in space residues found far apart in sequence in a more intricate way. Most significant inter-subdomain correlations shape up between the P-loop (NBD) and GLPL (ARC1); GLPL (ARC1) and MHD (ARC2) and between RNBS-B (NBD) and MHD (ARC2). Notably, the correlation identified between residue 297 in RNBS-B "*bbbTTR*" and the ARC2-MHD-489, forms in the ZAR1 structure a saline bridge (R297-D489).

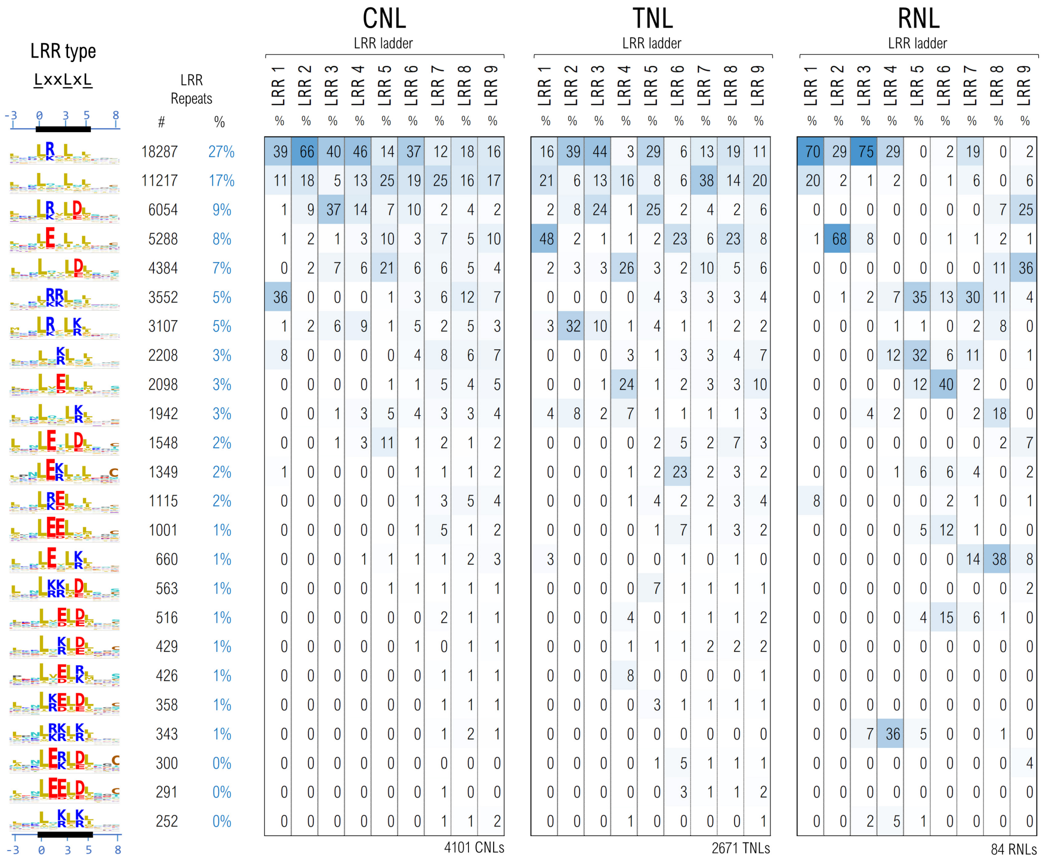
The leucine-rich repeat motifs—Clustering by physico-chemical properties

Screening the overall NLR dataset with LRRexpress yields a collection of ∼468,000 motifs in ∼31,100 sequences. As LRR domains may be found as standalone or part of many other protein classes only sequences containing regular NBS domains along with LRR domains larger than 9 repeats with no delineation ambiguities were retained for analysis, as described under methods.

The cleaned set of ∼61,700 LRR motifs from ∼6,850 seq were further clustered using unsupervised ML techniques in a space describing their hydropathy, charge and size, as described under methods. As various parts of the LRR horseshoe were shown to be involved in different types of internal and external interactions (Wang et al., 2019a; Ma et al., 2020) motifs were further grouped according to their position in the repeat ladder for 4,101 CNLs, 2,671 TNLs, and 84 RNLs.

Figure 6 shows the top 24 largest LRR motif clusters and their distribution in the LRR ladder for the three NLR classes. The most frequent pattern displays lysine and arginine in position 1 which is mainly found within the first 4–6 repeats of the analyzed NLRs. In general, the first 4–6 repeats of all three NLR classes show a high preference for positively charged LRR motifs, especially arginine. As can be seen from **Figure 6**, the cluster separation is mainly driven by the distribution of charge features in various solvent-exposed positions of the LxxLxL pattern.

In general more bulky charged amino acids, R&E, are preferred in positions 1 and 2 of the LxxLxL motif, while the lighter ones K&D are more frequent in position 4 of the motif. Remarkably Arginine in position 1 of the motif is dominant in the first four CNL repeats, in repeats 2 and 3 in TNLs and in repeats 1, 3 and 4 in RNLs. By contrast, aspartic acid in position 4 starts shaping up above repeat 3. Interestingly clusters with glutamic acid in position 1 are dominant in the first TNL and second RNL repeat in contrast with CNLs. Aside from the presence of Glu in the first TNL/RNL repeats, **Figure 6** also indicates other interesting differences among the three NLR classes – for instance, while the second uncharged cluster shapes up quite frequently (>15%) and evenly in all CNL/TNL repeats, this is almost absent in RNLs with the exception of the first repeat (<4%).



LRR Repeats		CNL									TNL									RNL								
#	%	LRR 1 %	LRR 2 %	LRR 3 %	LRR 4 %	LRR 5 %	LRR 6 %	LRR 7 %	LRR 8 %	LRR 9 %	LRR 1 %	LRR 2 %	LRR 3 %	LRR 4 %	LRR 5 %	LRR 6 %	LRR 7 %	LRR 8 %	LRR 9 %	LRR 1 %	LRR 2 %	LRR 3 %	LRR 4 %	LRR 5 %	LRR 6 %	LRR 7 %	LRR 8 %	LRR 9 %
18287	27%	39	66	40	46	14	37	12	18	16	16	39	44	3	29	6	13	19	11	70	29	75	29	0	2	19	0	2
11217	17%	11	18	5	13	25	19	25	16	17	21	6	13	16	8	6	38	14	20	20	2	1	2	0	1	6	0	6
6054	9%	1	9	37	14	7	10	2	4	2	2	8	24	1	25	2	4	2	6	0	0	0	0	0	0	0	7	25
5288	8%	1	2	1	3	10	3	7	5	10	48	2	1	1	2	23	6	23	8	1	68	8	0	0	1	1	2	1
4384	7%	0	2	7	6	21	6	6	5	4	2	3	3	26	3	2	10	5	6	0	0	0	0	0	0	0	11	36
3552	5%	36	0	0	0	1	3	6	12	7	0	0	0	0	4	3	3	3	4	0	1	2	7	35	13	30	11	4
3107	5%	1	2	6	9	1	5	2	5	3	3	32	10	1	4	1	1	2	2	0	0	0	1	1	0	2	8	0
2208	3%	8	0	0	0	0	4	8	6	7	0	0	0	3	1	3	3	4	7	0	0	0	12	32	6	11	0	1
2098	3%	0	0	0	0	1	1	5	4	5	0	0	1	24	1	2	3	3	10	0	0	0	0	12	40	2	0	0
1942	3%	0	0	1	3	5	4	3	3	4	4	8	2	7	1	1	1	1	3	0	0	4	2	0	0	2	18	0
1548	2%	0	0	1	3	11	1	2	1	2	0	0	0	0	2	5	2	7	3	0	0	0	0	0	0	0	2	7
1349	2%	1	0	0	0	0	1	1	1	2	0	0	0	1	2	23	2	3	2	0	0	0	1	6	6	4	0	2
1115	2%	0	0	0	0	0	1	3	5	4	0	0	0	1	4	2	2	3	4	8	0	0	0	0	2	1	0	1
1001	1%	0	0	0	0	0	1	5	1	2	0	0	0	0	1	7	1	3	2	0	0	0	0	5	12	1	0	0
660	1%	0	0	0	1	1	1	1	2	3	3	0	0	0	0	1	0	1	0	0	0	0	0	0	0	14	38	8
563	1%	0	0	0	0	1	1	1	1	1	0	0	0	0	7	1	1	1	1	0	0	0	0	0	0	0	0	2
516	1%	0	0	0	0	0	0	2	1	1	0	0	0	4	0	1	1	1	2	0	0	0	0	4	15	6	1	0
429	1%	0	0	0	0	0	1	0	1	1	0	0	0	1	1	1	2	2	2	0	0	0	0	0	0	0	0	0
426	1%	0	0	0	0	0	0	1	1	1	0	0	0	8	0	0	0	0	1	0	0	0	0	0	0	0	0	0
358	1%	0	0	0	0	0	0	1	1	1	0	0	0	0	3	1	1	1	1	0	0	0	0	0	0	0	0	0
343	1%	0	0	0	0	0	0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	7	36	5	0	0	1	0
300	0%	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	1	1	1	0	0	0	0	0	0	0	0	4
291	0%	0	0	0	0	0	0	1	0	0	0	0	0	0	0	3	1	1	2	0	0	0	0	0	0	0	0	0
252	0%	0	0	0	0	0	0	1	1	2	0	0	0	1	0	0	0	0	1	0	0	2	5	1	0	0	0	0

FIGURE 6
Clustering of the LRR motifs in a physico-chemical properties embedding—top 24 clusters and their distribution among position in the LRR domain ladder in the three NLR classes. Displayed values are percentages normalized vertically, with respect to the total number of repeats in the given position of the ladder. Amino acids are colored as described in Figure 3 caption.

The CNL clustering data are consistent with the experimental reports indicating that at least in some systems such as ZAR1, Rx, SW5 charge complementarity between the first 4-6 LRR repeats and the CC and NBS domains is involved in CC-LRR and NBS-LRR interfacing and has a critical role in correct CNL functioning (Slootweg et al., 2013; de Oliveira et al., 2018; Wang et al., 2019b).

Interdomain correlations in CNL receptors

Due to their overall fold or to the conformational changes taking place during NLR functioning, the individual domains might be brought close in space, permanently or transiently. Inter-domain contacts could leave traces in correlations between motifs which are the most invariant regions of this highly diverse class of proteins. To identify such correlations we simultaneously clustered all the 11 NLR motifs in the case of CNLs, mapping results onto the recently released structural data (Wang et al., 2019a,b). As CNLs represent the largest group of NLR proteins (~14,100 CNLs vs. ~5,100 TNLs and ~400 RNLs in the analyzed set) and were shown to host a relatively conserved region readily detectable with ML techniques—the "extended EDVID" motif—we separated the main groups of this class of NLRs using clustering techniques described under the methods. The best cluster separation is reached at ~37% identity where around 80% of the sequences do gather into the top seven CNL clusters. As can be seen from Figure 7A, five out of the seven main clusters: CNL#1, 2, 4, 5, and 7 display a strong charge complementarity between the EDVID and the first 5 LRR motifs, while clusters CNL#3 and 6 completely or partially lack strict EDVID pattern and charge match, with motifs being more diverse and neutral.

When mapping the motifs onto the ZAR1 cryo-EM structure the charge complementarity locates at the CC-LRR interface suggesting that sequences in CNL#1, 2, 4, 5, and 7 might adopt CC-LRR configurations similar to that seen in ZAR1. Some marked differences are noticeable though between clusters in LRR repeats 1–4 and the EDVID region which displays only two main conserved positions 9 and 12 in all these clusters.

Interestingly also the first two LRR motifs are quite irregular returning low probabilities by both LRRexpress and LRRpredictor, especially due to the unusual usage of amino acids

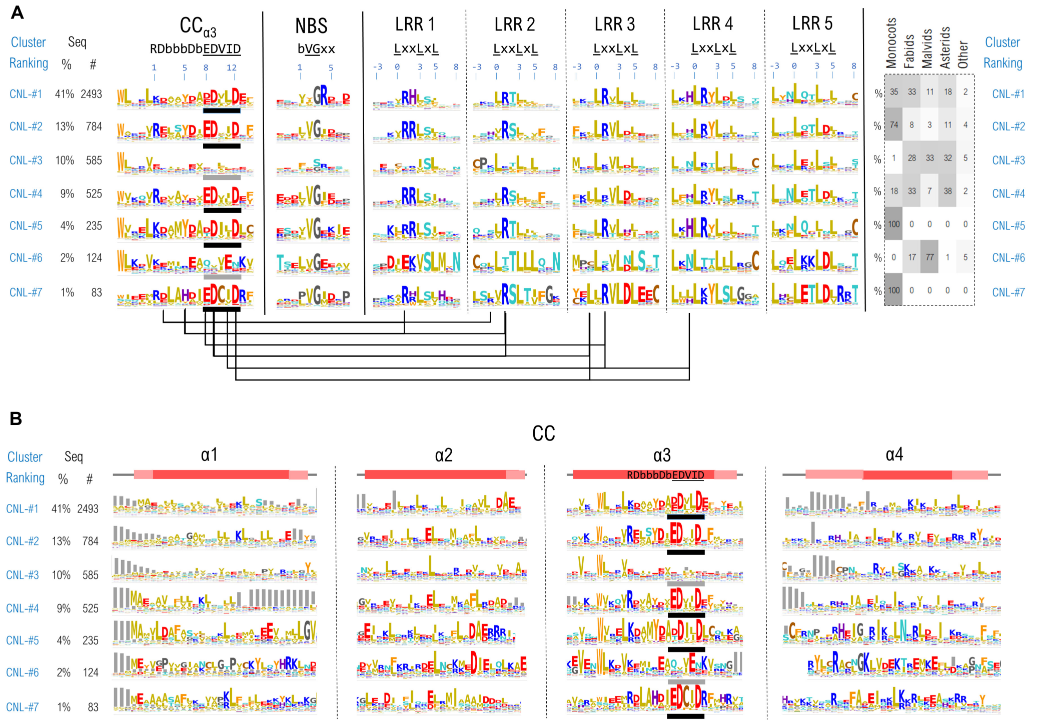


FIGURE 7
(A) CNL proteins clustering based on the regions confining the CC extended EDVID (α3 helix) and the first 5 LRR motifs. Motif variability is expressed as relative entropy displayed as letter height (the higher the more conserved). The taxonomic distributions of each cluster is depicted in the right panel. The ZAR1 CC-LRR contacts within 5Å are mapped below the motif consensus (PDB: 6j5w). **(B)** Zoom in on the conservation profile of the CC domain of the top CNL clusters from **(A)**.

in positions −3 (upstream) and L_0 of the motif. This is related most likely to the involvement of these regions in CC-LRR interfacing which constitutes a stabilizing factor, while further repeats require a structural reinforcement provided by canonical leucines in these two positions.

The top largest CNL groups, clustered based on the most conserved 11 motifs (CC-EDVID, NBS, and LRR motifs) were next inspected with respect to the entire span of the CC domain (**Figure 7B**). Strong consensus signatures are observed solely within the third helix (α3)—with tryptophan upstream the extended EDVID motif highly conserved in most CNL groups. By contrast, the remaining three helical segments display a high sequence and length variability, preserving though the hydrophobic positions involved in coil-coiled zipping.

Among these, the first helix (α1) displays the highest variability which may be linked to both NLR localization and functioning. For instance, a general more hydrophobic profile is particularly seen in the CNL groups displaying canonical EDVID signatures. By contrast, groups lacking a strict EDVID pattern and CC-LRR charge match (CNL-#3&#6), display also a significantly higher sequence and length variability in α1.

The "MADA" motif in its extended form "MA(D/E)AxVSFxVxKLxxLLxxEx" (Adachi et al., 2019), also present in ZAR1 is seen mainly in CNL groups #1&#4 which is consistent with Adachi et al. (2019) observation that this motif is found in around one fifth of plant CNLs. While certain positions in this extended motif were experimentally shown to be essential in the activation and/or the protruding of the plasma membrane in ZAR1 (Baudin et al., 2019; Wang et al., 2019a; Bi et al., 2021) this is by no means an universal feature in CNLs. Other NLRs that display a N-ter α1 CC region highly non-homologous to ZAR1 attach to the membrane via N-myristoylation and/or palmitoylation such as RPS5 (Qi et al., 2012), while other NLRs such as Rx1 and RPS4, have a nucleo-cytoplasmic localization during their activation cycle (Tameling et al., 2012; Qi and Innes, 2013).

Conclusion

Results presented herein indicate that NLRexpress—the bundle of 17 ML based predictors designed to identify CC, TIR, NBS and LRR specific sequence motifs—performs fast and accurately in identifying NLRs on large datasets and might be of use in overall genome, transcriptome or proteome screening. In addition, LRRexpress—the NLRexpress module for LRR motif prediction was tuned for general LRR detection and may be used as a fast and accurate standalone tool in identifying LRR domains and delineating repeats in other protein classes such as extracellular plant RLK/RLP or metazoan LRRs, making it a complementary tool to the existing DeepLRR (Liu et al., 2022), LRRpredictor (Martin et al., 2020a), LRRsearch (Bej et al., 2014), LRRfinder (Offord and Werling, 2013), etc.

NLRexpress was then use to detect motifs in a large set of ∼34,000 plant NLRs and cluster these proteins based on their motif properties using unsupervised methods in order to identify correlations within and between the canonical domains shaping up underneath the NLR significant diversity.

All in all, NLRexpress is designed as a flexible bundle for the fast prediction of main NLR domains in large sequence datasets. Using LRRexpress alone will pinpoint all proteins displaying LRR domains, most of which are immune system receptors. Using it in conjunction with NBSexpress will identify all NLRs irrespective to their group, while adding the CCexpress and/or TIRexpress modules will bring further information related to the main NLR groups—TNL and CNL.

Data availability statement

The datasets and source code presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: all training models and datasets can be found as tar.gz archives at https://nlrexpress.biochim.ro/documentation.php within the download section. The source code of NLRexpress can be found at: https://github.com/eliza-m/NLRexpress.

Author contributions

A-JP and AG conceived the research framework. A-JP, AG, and LS supervised the work. EM carried out software implementation, validation, and data analysis. All authors contributed to analysis and interpretation of the data and writing of the manuscript, read, and approved the final manuscript.

Funding

EM, LS, and A-JP acknowledged financial support from the Romanian Academy Programs 1 and 2 of IBAR.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.975888/full#supplementary-material

References

Adachi, H., Contreras, M., Harant, A., Wu, C. H., Derevnina, L., Sakai, T., et al. (2019). An N-terminal motif in NLR immune receptors is functionally conserved across distantly related plant species. *eLife* 8:e49956. doi: 10.7554/ELIFE.49956

Andolfo, G., Di Donato, A., Chiaiese, P., De Natale, A., Pollio, A., Jones, J. D. G., et al. (2019). Alien Domains Shaped the Modular Structure of Plant NLR Proteins. *Genome Biol. Evol.* 11, 3466–3477. doi: 10.1093/GBE/EVZ248

Ankerst, M., Breunig, M. M., Kriegel, H. P., and Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. *ACM SIGMOD Rec.* 28, 49–60. doi: 10.1145/304181.304187

Baggs, E., Dagdas, G., and Krasileva, K. V. K. (2017). NLR diversity, helpers and integrated domains: Making sense of the NLR IDentity. *Curr. Opin. Plant Biol.* 38, 59–67. doi: 10.1016/j.pbi.2017.04.012

Baudin, M., Schreiber, K. J., Martin, E. C., Petrescu, A. J., and Lewis, J. D. (2019). Structure–function analysis of ZAR1 immune receptor reveals key molecular interactions for activity. *Plant J.* 1, 352–370. doi: 10.1111/tpj.14547

Bej, A., Sahoo, B. R., Swain, B., Basu, M., Jayasankar, P., and Samanta, M. (2014). LRRsearch: An asynchronous server-based application for the prediction of leucine-rich repeat motifs and an integrative database of NOD-like receptors. *Comput. Biol. Med.* 53, 164–170. doi: 10.1016/j.compbiomed.2014.07.016

Bi, G., Su, M., Li, N., Liang, Y., Dang, S., Xu, J., et al. (2021). The ZAR1 resistosome is a calcium-permeable channel triggering plant immune signaling. *Cell* 184:3528–3541.e12. doi: 10.1016/j.cell.2021.05.003

Burdett, H., Bentham, A. R., Williams, S. J., Dodds, P. N., Anderson, P. A., Banfield, M. J., et al. (2019). The Plant "Resistosome": Structural Insights into Immune Signaling. *Cell Host Microbe* 26, 193–201. doi: 10.1016/j.chom.2019.07.020

Casey, L. W., Lavrencic, P., Bentham, A. R., Cesari, S., Ericsson, D. J., Croll, T., et al. (2016). The CC domain structure from the wheat stem rust resistance protein Sr33 challenges paradigms for dimerization in plant NLR proteins. *Proc. Natl. Acad. Sci. U.S.A.* 113, 12856–12861. doi: 10.1073/pnas.1609922113

Cesari, S. (2018). Multiple strategies for pathogen perception by plant immune receptors. *New Phytol.* 219, 17–24. doi: 10.1111/nph.14877

Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T., and Dodds, P. N. (2014). A novel conserved mechanism for plant NLR protein pairs: The "integrated decoy" hypothesis. *Front. Plant Sci.* 5:606. doi: 10.3389/FPLS.2014.00606/ABSTRACT

de Oliveira, A. S., Boiteux, L. S., Kormelink, R., and Resende, R. O. (2018). The Sw-5 Gene Cluster: Tomato Breeding and Research Toward Orthotospovirus Disease Control. *Front. Plant Sci.* 9:1055. doi: 10.3389/FPLS.2018.01055

De Oliveira, A. S., Koolhaas, I., Boiteux, L. S., Caldararu, O. F., Petrescu, A. J., Resende, R. O., et al. (2016). Cell death triggering and effector recognition by Sw-5 SD-CNL proteins from resistant and susceptible tomato isolines to Tomato spotted wilt virus. *Mol. Plant Pathol.* 17, 1442–1454. doi: 10.1111/MPP.12439

Dievart, A., Gottin, C., Peacuterin, C., Ranwez, V., and Chantret, N. (2020). Origin and Diversity of Plant Receptor-Like Kinases. *Annu. Rev. Plant Biol.* 71, 131–156. doi: 10.1146/ANNUREV-ARPLANT-073019-025927

Eckardt, N. A. (2017). The Plant Cell Reviews Plant Immunity: Receptor-Like Kinases, ROS-RLK Crosstalk, Quantitative Resistance, and the Growth/Defense Trade-Off. *Plant Cell* 29, 601–602. doi: 10.1105/TPC.17.00289

Hanukoglu, I. (2015). Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. *Biochem. Mol. Biol. Educ.* 43, 206–209. doi: 10.1002/BMB.20849

Hao, W., Collier, S. M., Moffett, P., and Chai, J. (2013). Structural basis for the interaction between the potato virus X resistance protein (Rx) and its cofactor ran GTPase-activating protein 2 (RanGAP2). *J. Biol. Chem.* 288, 35868–35876. doi: 10.1074/jbc.M113.517417

Jeong, C. S., and Kim, D. (2012). Reliable and robust detection of coevolving protein residues. *Protein Eng. Des. Sel.* 25, 705–713. doi: 10.1093/PROTEIN/GZS081

Johanndrees, O., Baggs, E. L., Uhlmann, C., Locci, F., Läßle, H. L., Melkonian, K., et al. (2021). Differential EDS1 requirement for cell death activities of plant TIR-domain proteins. *bioRxiv* [Preprint]. . doi: 10.1101/2021.11.29.470438

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* 11:431. doi: 10.1186/1471-2105-11-431

Jones, J. D. G., Vance, R. E., and Dangl, J. L. (2016). Intracellular innate immune surveillance devices in plants and animals. *Science* 354, aaf6395–aaf6395. doi: 10.1126/science.aaf6395

Kajava, A. V. (1998). Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.* 277, 519–527. doi: 10.1006/jmbi.1998.1643

Kajava, A. V., and Kobe, B. (2002). Assessment of the ability to model proteins with leucine-rich repeats in light of the latest structural information. *Protein Sci.* 11, 1082–1090. doi: 10.1110/ps.4010102

Khan, M., Subramaniam, R., and Desveaux, D. (2016). Of guards, decoys, baits and traps: Pathogen perception in plants by type III effector sensors. *Curr. Opin. Microbiol.* 29, 49–55. doi: 10.1016/j.mib.2015.10.006

Kingma, D. P., and Ba, J. L. (2015). "Adam: A method for stochastic optimization. in 3rd International Conference on Learning Representations," in *ICLR 2015 – Conference Track Proceedings (International Conference on Learning Representations, ICLR)*, (San Diego).

Kobe, B., and Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* 11, 725–732. doi: 10.1016/S0959-440X(01)00266-4

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X., and Morel, J. B. (2016). Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol.* 210, 618–626. doi: 10.1111/nph.13869

Lapin, D., Johanndrees, O., Wu, Z., Li, X., and Parker, J. E. (2022). Molecular innovations in plant TIR-based immunity signaling. *Plant Cell* 34, 1479–1496. doi: 10.1093/PLCELL/KOAC035

Liu, Z., Ren, Z., Yan, L., and Li, F. (2022). DeepLRR: An Online Webserver for Leucine-Rich-Repeat Containing Protein Characterization Based on Deep Learning. *Plants* 11:136. doi: 10.3390/PLANTS11010136

Ma, S., Lapin, D., Liu, L., Sun, Y., Song, W., Zhang, X., et al. (2020). Direct pathogen-induced assembly of an NLR immune receptor complex to form a holoenzyme. *Science* 370:eabe3069. doi: 10.1126/SCIENCE.ABE3069

Maekawa, T., Cheng, W., Spiridon, L. N., Töller, A., Lukasik, E., Saijo, Y., et al. (2011). Coiled-coil domain-dependent homodimerization of intracellular barley immune receptors defines a minimal functional module for triggering cell death. *Cell Host Microbe* 9, 187–199. doi: 10.1016/j.chom.2011.02.008

Martin, R., Qi, T., Zhang, H., Liu, F., King, M., Toth, C., et al. (2020b). Structure of the activated ROQ1 resistosome directly recognizing the pathogen effector XopQ. *Science* 370:eabd9993. doi: 10.1126/SCIENCE.ABD9993

Martin, E. C., Sukarta, O. C. A., Spiridon, L., Grigore, L. G., Constantinescu, V., Tacutu, R., et al. (2020a). LRRpredictor—A New LRR Motif Detection Method for Irregular Motifs of Plant NLR Proteins Using an Ensemble of Classifiers. *Genes* 11:286. doi: 10.3390/genes11030286

Maruta, N., Burdett, H., Lim, B. Y. J., Hu, X., Desa, S., Manik, M. K., et al. (2022). Structural basis of NLR activation and innate immune signalling in plants. *Immunogenet* 74, 5–26. doi: 10.1007/S00251-021-01242-5

NLRscape, (2022). *NLRscape - A Plant NLR Atlas*. Available online at: https://nlrscape.biochim.ro/. [Accessed Mar 15, 2022].

Offord, V., and Werling, D. (2013). LRRfinder2.0: A webserver for the prediction of leucine-rich repeats. *Innate Immun.* 19, 398–402. doi: 10.1177/1753425912465661

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46:W200–W204. doi: 10.1093/nar/gky448

PyMOL,. (2018). *The PyMOL Molecular Graphics System, Version 2.2.3*. New York, NY: Schrödinger, LLC.

Qi, D., DeYoung, B. J., and Innes, R. W. (2012). Structure-function analysis of the coiled-coil and leucine-rich repeat domains of the RPS5 disease resistance protein. *Plant Physiol.* 158, 1819–1832. doi: 10.1104/pp.112.194035

Qi, D., and Innes, R. W. (2013). Recent advances in plant NLR structure, function, localization, and signaling. *Front. Immunol.* 4:348. doi: 10.3389/fimmu.2013.00348

Rairdan, G. J., Collier, S. M., Sacco, M. A., Baldwin, T. T., Boettrich, T., and Moffett, P. (2008). The coiled-coil and nucleotide binding domains of the potato Rx disease resistance protein function in pathogen recognition and signaling. *Plant Cell* 20, 739–751. doi: 10.1105/tpc.107.056036

Rajaraman, J., Douchkov, D., Hensel, G., Stefanato, F. L., Gordon, A., Ereful, N., et al. (2016). An LRR/Malectin receptor-like kinase mediates resistance to non-adapted and adapted powdery mildew fungi in barley and wheat. *Front. Plant Sci.* 7:1836. doi: 10.3389/fpls.2016.01836

Sarris, P. F., Duxbury, Z., Huh, S. U., Ma, Y., Segonzac, C., Sklenar, J., et al. (2015). A Plant Immune Receptor Detects Pathogen Effectors that Target WRKY Transcription Factors. *Cell* 161, 1089–1100. doi: 10.1016/J.CELL.2015.04.024

Schreiber, K. J., Baudin, M., Hassan, J. A., and Lewis, J. D. (2016). Die another day: Molecular mechanisms of effector-triggered immunity elicited by type III secreted effector proteins. *Semin. Cell Dev. Biol.* 56, 124–133. doi: 10.1016/j.semcdb.2016.05.001

Sela, H., Spiridon, L. N., Ashkenazi, H., Bhullar, N. K., Brunner, S., Petrescu, A. J., et al. (2014). Three-dimensional modeling and diversity analysis reveals distinct AVR recognition sites and evolutionary pathways in wild and domesticated wheat Pm3 R genes. *Mol. Plant Microbe Interact.* 27, 835–845. doi: 10.1094/MPMI-01-14-0009-R

Sela, H., Spiridon, L. N., Petrescu, A. J., Akerman, M., Mandel-Gutfreund, Y., Nevo, E., et al. (2012). Ancient diversity of splicing motifs and protein surfaces in the wild emmer wheat (Triticum dicoccoides) LR10 coiled coil (CC) and leucine-rich repeat (LRR) domains. *Mol. Plant Pathol.* 13, 276–287. doi: 10.1111/j.1364-3703.2011.00744.x

Shao, Z. Q., Xue, J. Y., Wu, P., Zhang, Y. M., Wu, Y., Hang, Y. Y., et al. (2016). Large-Scale Analyses of Angiosperm Nucleotide-Binding Site-Leucine-Rich Repeat Genes Reveal Three Anciently Diverged Classes with Distinct Evolutionary Patterns. *Plant Physiol.* 170, 2095–2109. doi: 10.1104/PP.15.01487

Simonetti, F. L., Teppa, E., Chernomoretz, A., Nielsen, M., and Marino Buslje, C. (2013). MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Res.* 41:W8–W14. doi: 10.1093/NAR/GKT427

Slootweg, E., Koropacka, K., Roosien, J., Dees, R., Overmars, H., Lankhorst, R. K., et al. (2017). Sequence Exchange between Homologous NB-LRR Genes Converts Virus Resistance into Nematode Resistance, and Vice Versa. *Plant Physiol.* 175, 498–510. doi: 10.1104/pp.17.00485

Slootweg, E. J., Spiridon, L. N., Martin, E. C., Tameling, W. I. L., Townsend, P. D., Pomp, R., et al. (2018). Distinct roles of non-overlapping surface regions of the coiled-coil domain in the potato immune receptor Rx1. *Plant Physiol.* 178, 1310–1331. doi: 10.1104/pp.18.00603

Slootweg, E. J., Spiridon, L. N., Roosien, J., Butterbach, P., Pomp, R., Westerhof, L., et al. (2013). Structural determinants at the interface of the ARC2 and leucine-rich repeat domains control the activation of the plant immune receptors Rx1 and Gpa2. *Plant Physiol.* 162, 1510–1528. doi: 10.1104/pp.113.218842

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi: 10.1038/nbt.3988

Sueldo, D. J., Shimels, M., Spiridon, L. N., Caldararu, O., Petrescu, A.-J., Joosten, M. H. A. J., et al. (2015). Random mutagenesis of the nucleotide-binding domain of NRC1 (NB-LRR Required for Hypersensitive Response-Associated Cell Death-1), a downstream signalling nucleotide-binding, leucine-rich repeat (NB-LRR) protein, identifies gain-of-function mutations in t. *New Phytol.* 208, 210–223. doi: 10.1111/nph.13459

Sukarta, O. C. A., Slootweg, E. J., and Goverse, A. (2016). Structure-informed insights for NLR functioning in plant immunity. *Semin. Cell Dev. Biol.* 56, 134–149. doi: 10.1016/j.semcdb.2016.05.012

Sun, Y., Li, L., Macho, A. P., Han, Z., Hu, Z., Zipfel, C., et al. (2013). Structural basis for flg22-induced activation of the Arabidopsis FLS2-BAK1 immune complex. *Science* 342, 624–628. doi: 10.1126/science.1243825

Suplatov, D., Sharapova, Y., Timonina, D., Kopylov, K., and Švedas, V. (2018). The visualCMAT: A web-server to select and interpret correlated mutations/co-evolving residues in protein families. *J. Bioinform Comput. Biol.* 16:1840005. doi: 10.1142/S021972001840005X

Tameling, W. I. L., Elzinga, S. D. J., Darmin, P. S., Vossen, J. H., Takken, F. L. W., Haring, M. A., et al. (2002). The Tomato R Gene Products I-2 and Mi-1 Are Functional ATP Binding Proteins with ATPase Activity. *Plant Cell* 14, 2929–2939. doi: 10.1105/TPC.005793

Tameling, W. I. L., Nooijen, C., Ludwig, N., Boter, M., Slootweg, E., Goverse, A., et al. (2012). RanGAP2 Mediates Nucleocytoplasmic Partitioning of the NB-LRR Immune Receptor Rx in the Solanaceae, Thereby Dictating Rx Function. *Plant Cell* 22, 4176–4194. doi: 10.1105/TPC.110.077461

Tameling, W. I. L., Vossen, J. H., Albrecht, M., Lengauer, T., Berden, J. A., Haring, M. A., et al. (2006). Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation. *Plant Physiol.* 140, 1233–1245. doi: 10.1104/pp.105.073510

Tareen, A., and Kinney, J. B. (2020). Logomaker: Beautiful sequence logos in Python. *Bioinformatics* 36, 2272–2274. doi: 10.1093/BIOINFORMATICS/BTZ921

Toshchakov, V. Y., and Neuwald, A. F. (2020). A survey of TIR domain sequence and structure divergence. *Immunogenetics* 72, 181–203. doi: 10.1007/s00251-020-01157-7

van der Hoorn, R. A. L., and Kamoun, S. (2008). From Guard to Decoy: A new model for perception of plant pathogen effectors. *Plant Cell* 20, 2009–2017. doi: 10.1105/tpc.108.060194

Van Ooijen, G., Mayr, G., Kasiem, M. M. A., Albrecht, M., Cornelissen, B. J. C., and Takken, F. L. W. (2008). Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *J. Exp. Bot.* 59, 1383–1397. doi: 10.1093/jxb/ern045

Wang, J., Chen, T., Han, M., Qian, L., Li, J., Wu, M., et al. (2020). Plant NLR immune receptor Tm-22 activation requires NB-ARC domain-mediated self-association of CC domain. *PLoS Pathog.* 16:e1008475. doi: 10.1371/JOURNAL.PPAT.1008475

Wang, J., Hu, M., Wang, J., Qi, J., Han, Z., Wang, G., et al. (2019a). Reconstitution and structure of a plant NLR resistosome conferring immunity. *Science* 364:eaav5870. doi: 10.1126/science.aav5870

Wang, J., Wang, J., Hu, M., Wu, S., Qi, J., Wang, G., et al. (2019b). Ligand-triggered allosteric ADP release primes a plant NLR complex. *Sience* 364:eaav5868. doi: 10.1126/science.aav5868

Wróblewski, T., Spiridon, L., Martin, E. C., Petrescu, A. J., Cavanaugh, K., Truco, M. J., et al. (2018). Genome-wide functional analyses of plant coiled–coil NLR-type pathogen receptors reveal essential roles of their N-terminal domain in oligomerization, networking, and immunity. *PLoS Biol.* 16:e2005821. doi: 10.1371/journal.pbio.2005821

Wu, C. H., Abd-El-Haliem, A., Bozkurt, T. O., Belhaj, K., Terauchi, R., Vossen, J. H., et al. (2017). NLR network mediates immunity to diverse plant pathogens. *Proc. Natl. Acad. Sci. U.S.A.* 114, 8113–8118. doi: 10.1073/pnas.1702041114

Wu, C. H., Krasileva, K. V., Banfield, M. J., Terauchi, R., and Kamoun, S. (2015). The"sensor domains"of plant NLR proteins: More than decoys? *Front. Plant Sci.* 6:134. doi: 10.3389/FPLS.2015.00134

Yu, D., Song, W., Tan, E. Y. J., Liu, L., Cao, Y., Jirschitzka, J., et al. (2022). TIR domains of plant immune receptors are 2',3'-cAMP/cGMP synthetases mediating cell death. *Cell* 185:2370-2386.e18. doi: 10.1016/J.CELL.2022.04.032

Yu, X., Zhong, S., Yang, H., Chen, C., Chen, W., Yang, H., et al. (2021). Identification and Characterization of NBS Resistance Genes in Akebia trifoliata. *Front. Plant Sci.* 12:2459. doi: 10.3389/FPLS.2021.758559