




Article

LRRpredictor—A New LRR Motif Detection Method for Irregular Motifs of Plant NLR Proteins Using an Ensemble of Classifiers

Eliza C. Martin ¹, Octavina C. A. Sukarta ², Laurentiu Spiridon ¹, Laurentiu G. Grigore ³, Vlad Constantinescu ¹, Robi Tacutu ¹, Aska Goverse ^{2,*} and Andrei-Jose Petrescu ^{1,*}

¹ Department of Bioinformatics and Structural Biochemistry, Institute of Biochemistry of the Romanian Academy, Splaiul Independentei 296, 060031 Bucharest, Romania; eliza.martin@biochim.ro (E.C.M.); laurentiu.spiridon@biochim.ro (L.S.); vlad.ion.constantinescu@gmail.com (V.C.); robi.tacutu@gmail.com (R.T.)

² Laboratory of Nematology, Wageningen University and Research, 6700ES Wageningen, The Netherlands; octavina.sukarta@wur.nl

³ Space Comp SRL, 041512 Bucharest, Romania; laur@itprod.eu

* Correspondence: aska.goverse@wur.nl (A.G.); andrei.petrescu@biochim.ro (A.-J.P.); Tel.: +31-31-748-5086 (A.G.); +40-21-233-9069 (A.-J.P.); Fax: +40-21-233-9068 (A.-J.P.)

Received: 7 February 2020; Accepted: 4 March 2020; Published: 8 March 2020



Abstract: Leucine-rich-repeats (LRRs) belong to an archaic procaryal protein architecture that is widely involved in protein–protein interactions. In eukaryotes, LRR domains developed into key recognition modules in many innate immune receptor classes. Due to the high sequence variability imposed by recognition specificity, precise repeat delineation is often difficult especially in plant NOD-like Receptors (NLRs) notorious for showing far larger irregularities. To address this problem, we introduce here LRRpredictor, a method based on an ensemble of estimators designed to better identify LRR motifs in general but particularly adapted for handling more irregular LRR environments, thus allowing to compensate for the scarcity of structural data on NLR proteins. The extrapolation capacity tested on a set of annotated LRR domains from six immune receptor classes shows the ability of LRRpredictor to recover all previously defined specific motif consensuses and to extend the LRR motif coverage over annotated LRR domains. This analysis confirms the increased variability of LRR motifs in plant and vertebrate NLRs when compared to extracellular receptors, consistent with previous studies. Hence, LRRpredictor is able to provide novel insights into the diversification of LRR domains and a robust support for structure-informed analyses of LRRs in immune receptor functioning.

Keywords: leucine-rich repeat prediction; supervised learning; LRR motif; LRR structure; NOD-like receptors; R proteins

1. Introduction

The leucine-rich-repeat (LRR) domains are present in all of the tree of life branches. As they are involved in protein–protein interactions, LRR domains are found in receptors having a vast number of functions such as pathogen detection, immune response propagation, hormone perception, enzyme inhibition, or cell adhesion [1]. In both plants and mammals, a number of studies have detailed adverse effects associated with mutations in the LRR domains such as that reported for various immune-related receptors, resulting in compromised functions and enhanced disease progression [2]. For example, mutating a single residue in the LRR domain of the rice Pita receptor results in complete loss of recognition against the fungus *Magnaporthe grisea* [3] while mutations in the metazoan NLRC4-LRR contributes to autoinflammatory disease phenotypes [4]. Additionally, mutations in the

LRRK2 kinase enzyme, lead to Parkinson's disease and other associated inflammatory diseases [5,6], whereas mutations in leucine-rich proteoglycans have been previously shown to be involved in osteoarthritis [7], and last but not least PRELP mutations might have a role in Hutchinson–Gilford, an accelerated progeroid syndrome characterized by premature aging [8]. Hence, understanding the structural aspects of binding properties and specificities of LRR domains opens wide possibilities for receptor engineering with vast implications not only for improved crop resistance to plant diseases, but also for a wide range of medical applications.

In innate immunity, LRR modules are found in various domain organizations in many receptor classes such as plant receptor-like kinases (RLK), receptor-like proteins (RLP), NOD-like receptors (NLR), or metazoan NLR and Toll-like receptors (TLR). In plant basal immunity, LRR N-terminal domains face the extracellular environment and are found in either receptor-like kinases (RLK) or receptor-like proteins (RLPs) depending on the presence or absence of a C-terminal kinase domain on the cytosolic side of the receptor. By contrast, LRRs constitute the C-terminal domains of intracellular NOD-like receptors (NLR), also known as resistance (R) proteins, and face the cytosolic environment to mediate resistance against specific pathogens. Depending on their N-terminal domain, which is either a coiled-coil (CC) or a toll-like receptor domain (TIR), R proteins fall into two main NLR classes: the CNL and TNL receptors, respectively [9]. Both these classes contain however a central nucleotide binding domain (NBS) which acts as a 'switch' that changes its conformation upon ADP/ATP binding [9,10]. Metazoan NLRs show a similar organization with plant NLRs. They encode a variety of N-terminal 'sensors' (caspase activation and recruitment domains—CARD, baculovirus inhibitor of apoptosis repeat—BIR, etc.), the central 'switch' STAND domain (signal transduction ATPases with numerous domains) - NBS/NACHT domain (NAIP (neuronal apoptosis inhibitory protein), CIITA (MHC class II transcription activator), HET-E (incompatibility locus protein from *Podospora anserina*) and TP1 (telomerase-associated protein)) and the LRR domain at the C-terminal end. Last but not least, we mention here the metazoan toll-like receptors (TLRs) that have an extracellular LRR sensor domain as seen in the RLK/RLP case and a TIR domain on the cytosolic side involved in signal transduction [11].

From a structural point of view LRR domains have a solenoidal 'horseshoe' like 3D architecture composed of a variable number of repeats varying each from ≈ 15 to ≈ 30 amino acids in length. Repeats are held together through a network of hydrogen bonds which forms a beta sheet located on the ventral side of the 'horseshoe'. This is generated by a conserved sequence pattern named the LRR motif that in its minimal form is of the type 'LxxLxL' where L is generally leucine and to a lesser degree other hydrophobic amino acids [12]. Comprehensive sequence analysis of LRR immune receptors resulted in several classifications of LRR domains showing preferred amino acid conservation outside the minimal motif such as the two type classification proposed by Matsushima et al. [13] for TLR receptors or the seven type classification proposed by Kobe and Kajava [14] for all known LRR domains across all Kingdoms. However, exceptions to such rules are frequent as revealed by the Hidden Markov Model approach carried out by Ng et al. [15]. This highlighted the fact that most of the analyzed classes of human proteins containing LRR domains also display many irregular motifs alongside repeats showing the well-defined class specific motif [15].

While the above mentioned receptor classes were shown to present LRR irregularities [15], studies on plant NLR proteins such as Lr10 and Pm3 from wheat, Rx1 and Gpa2 from potato, or ZAR1 from *Arabidopsis* show that their LRR domains have a far more variable and irregular structure than their extracellular counterparts [16–22]. These factors combined contribute to the challenge for the accurate prediction of LRR motifs in plant NLRs.

A proper annotation of each LRR motif in a given LRR domain is instrumental in generating an accurate 3D model [12,23] and by this in properly defining the domain surface and identifying potential protein–protein interaction interfaces. An illustrative example is the conservation mapping performed by Helft et al. in 2011, which was used to identify new interaction partners of plant RLPs and RLKs by studying conserved 3D relationships among amino acids inferred from annotation of LRR repeats [24].

Based on our previous work, identifying the individual true motifs in a LRR domain is hindered by the following: (a) in its minimal form, a 'LxxLxL' pattern is trivial and frequently occurs randomly in any protein; (b) in many cases several 'LxxLxL' patterns do overlap in less than 15 aa range in NLR-LRRs making the precise delineation difficult; (c) the number of 3D experimental structures from which to learn is low; and (d) this small 3D learning set is class and phyla biased—as around half of the structures are of mammalian origin while plant NLRs only have one recently documented structure [21,22].

Thus, given the above described indeterminacies the precise LRR motif identification becomes the most problematic step in the correct repeat delineation within a LRR domain. This also explains why LRR domains and their individual repeats are poorly annotated in genomes or protein databases in contrast to the better annotated, relatively more conserved NBS domain, which has therefore been used in phylogenetic analyses [10,25]. Hence, these major limitations hamper the study of NLRs at various levels such as in the context of plant innate immunity. To address these challenges, in this paper we propose a new LRR motif detection method: LRRpredictor, designed to be more sensitive to motif irregularities than the existing methods like LRRfinder [26] or LRRsearch [27] and to detect irregular and short LRR signatures as are often found in plant NLRs, but not limited to this class.

We assessed how LRRpredictor behaves within different classes of immune-related receptors that contain LRR domains, such as plant NLRs, RLPs, and RLKs and vertebrate NLRs and TLRs with the aim to provide novel insights into the diversification of LRR domains and their role in the functioning of immune receptors.

2. Materials and Methods

2.1. Assembly and Analysis of the LRR Structural Dataset

Various protein domain databases, such as CATH [28], Pfam [29], and Interpro collection [30] were used to obtain a dataset of 611 structure files of proteins annotated to contain LRR domains. These files were processed and filtered out to extract a clean set of LRR chains sharing less than 90% sequence identity using Pisces server [31]. This set containing 178 LRR chains were visually inspected and subjected to LRR repeat delineation based on the distinctive LRR ventral beta-sheet secondary structure pattern. Annotated LRR domains consisting in less than five LRR repeats, as well as incomplete repeats not covering at least five amino acids upstream and downstream of the "LxxLxL" minimal motif were further eliminated.

Using this procedure, we generated the 90% identity data set, ID90, consisting of 172 N-ter LRR 'entry' repeats (N), 1792 LRR 'core' repeats (L), and 154 C-ter LRR 'exit' repeats (C) (File S1). To avoid redundancy in the training data the level of identity has to be further significantly reduced. However, given the small size of ID90 (<180 chains), a trade-off between increase in entropy and loss of data had to be reached. As seen from Figure A1a, a proper inflection point shapes up at around 50% identity and was considered the best compromise in generating a nonredundant set of repeats. In practical terms, the nonredundant ID50 set was generated from ID90 by selecting repeats showing less than eight identical amino acids on a 16 amino acid window centered on the 'LxxLxL' minimal LRR motif, i.e., the window comprising five amino acids upstream and downstream 'LxxLxL'. This nonredundant ID50 set was comprised of 106 N-ter 'entry' repeats (N), 659 'core' repeats (L), and 88 C-ter 'exit' repeats (C), i.e., ~40% of the 90ID set (Figure 1, File S1).

Jensen–Shannon divergence (JSD) scores (Figure 1e) were computed using Capra et al. implementation [32], using the BLOSUM62 matrix for background probabilities and a window parameter 0. The phyla distribution shown in Figure 1c was computed using the Environment for Tree Exploration (ETE3) library v3.1.1 [33].

2.2. Training and Testing Datasets Construction

In order to provide a representative collection of non-LRR examples, we selected a representative example of each CATH [28] domains' topology (except LRR) from a nonredundant dataset provided by CATH where all proteins share less than 20% identity or have a less than 60% overlap (cath-dataset-nonredundant-S20 set-09.12.2019). Given potential synchronization problems between various databases used to build the overall learning set comprising (a) the nonredundant 50ID LRRs, containing the 'entry'-, 'core'-, 'exit'-repeats and the flanking nonLRR domains when present and (b) the CATH nonLRR domains—the data was subjected to a third redundancy filter performed with a similar CATH methodology, aimed at eliminating sequences that fail one of the below bounds:

- the length of the alignment is over 100 and the identity is over 20%.
- length of the alignment is between 40 and 100 with an identity over 20% and the overlap with respect to both sequences is more than 60%.
- LRR repeats with alignments lengths ≥ 16 aa and $\geq 50\%$ identical (equivalent of at most 8/16 aa constraint imposed initially on the motifs).

The final dataset built as above and used herein for training and testing classifiers, contains 648 LRR core repeats, 100 N-ter entry, and 67 C-ter exit nonredundant repeats (including the LRR domain flanking regions) and 875 non-LRR domains from CATH.

From this set, 1/5th was used to generate the test dataset, while the remaining 4/5 were used to build the training datasets, preserving the class ratio between the sets. The test dataset contains 40,241 amino acid samples of which only 150, i.e., less than 0.4%, are initiating LRR motifs. Similarly, over the training set less than 0.5% of the samples are LRR initiators. The training set was further split into four cross-validation sets that were used for parameter optimization. All these sets are provided in File S2.

2.3. Feature Selection and Data Pre-Processing

In developing LRRpredictor we tested sequence-based (SeqB) features: solely or combined with structural based (StrB) features. The SeqB features comprise position-specific scoring matrices PSSM over the above discussed 16 amino acids interval summing up to 320 features corresponding to 20 amino acid types over the 16 positions. The StrB features comprise: (a) the three state (H—helix, E—extended, C—coil) secondary structure probabilities, (b) the three class (B—buried, M—medium and E—exposed) residue relative solvent accessibility, RSA probabilities and (c) intrinsic disorder probability—summing up to seven extra structural features per residue, resulting in a total of 432 features per 16 aa window. The structural based predictions were performed with RaptorX-Property software [34–37]. Sequence PSSMs were computed on Uniprot20 protein sequence database, using HHblits [38,39] that is based on HMM-HMM alignments shown to improve accuracy of alignments at low sequence homology levels.

In the pre-processing stage, feature variables were normalized, centered, and rescaled, as standard procedure involves. Data whitening using principal component analysis (PCA) decomposition was not used as it did not provide better performance on the tested classifiers.

2.4. Machine Learning Model Selection

Several classifiers such as support vector classification (SVC) [40], multi-layer perceptron (MLP) [41,42], and AdaBoost [43] as well as several oversampling techniques such as Adasyn [44] and SMOTE-based varieties [45–47], or over- and under-sampling combined approaches SmoteTomek [48] and SmoteEEN [49], were tested and parameter optimized via cross-validation using Scikit-learn library v.0.22.1 [50]. Multiclass estimators for N-entry (N), core (L), and C-exit (C) motif types that use either one-vs.-one or one-vs.-rest approaches were also investigated, but they performed worse than when treating all LRR motifs as a single class.

The best performing classifiers with tuned parameters were further studied in the context of a soft voter (that averages predicted probabilities of the ensemble constituents), and a final predictor,

further referred to as LRRpredictor, was chosen based on its out-of-sample performance on test set and overfitting behavior on the training data. LRRpredictor is composed of a set of eight classifiers (C1–C8) that use different strategies and consider all N, L, C motif types as a single class, aggregated within an ensemble based on the soft voting scheme, as shown in Figure 2d.

Classifiers C1–C4 use solely sequence-based features while C5–C8 use both sequence and structural-based features. Classifiers C1 and C5 use the support vector classification (SVC) algorithm [40], with a radial basis function (RBF) kernel, one-vs.-rest ('ovr') decision function. The margin penalty and the RBF scale (gamma) parameters were optimized through grid search to 1 and 0.01 for C1 and 1 and 0.001 for C4, respectively. Class imbalance was treated by adjusting the SVM weights inversely proportional to class frequency and class probabilities were inferred using sigmoid probability calibration.

Classifiers C2, C3, C6, C7 use multi-layer perceptron (MLP) [41,42]. A depth of three hidden layers was sufficient to describe the system, as adding additional hidden layers provided little to no difference in out-of-sample performance. The number of hidden nodes for each hidden layer was selected via grid search as follows: C2 (300-250-100), C3 (250-150-100), C5 (250-150-100), C6 (125-100-10). Classifiers C2, C3, C7 use the Limited-Memory BFGS [51] solver, while C6 uses Adam [52] optimizer for stochastic gradient descent [53] with early-stopping over a validation fraction of 0.2. All four classifiers use rectified linear unit (ReLU) activation function [54].

Classifiers C3 and C7 approach the imbalance problem through synthetic resampling using the combined over- and under-sampling method SMOTETomek [48], as implemented in imbalanced-learn library v 0.6.1 [55].

Classifiers C4 and C8 use an ensemble boosting approach—AdaBoost [43]—using tree classifiers of depth 1, as base estimators, SAMME.R real boosting algorithm, and sigmoid probability calibration. A maximum number of 50 base estimators was selected to maximize performance while avoiding overfitting.

2.5. Assembly of Protein Family Sets Containing LRR Domains

In order to investigate LRRpredictor behavior on previously annotated LRR domains from various functional protein groups, we generated a collection of randomly selected 500 representatives from Uniprot50 database (i.e., below 50% identity between themselves at a given minimum overlap—version available at 20.11.2019-release-2019_10) which were annotated by Interpro to contain a LRR domain (IPR032675 and Interpro v77.0 protein2ipr database).

A total of six groups were generated: four groups of sequences of CNLs, TNLs, RLKs, and RLPs protein classes from flowering plants and two groups of TLRs and NLRs from vertebrates. Given the high conservation of vertebrate TLRs this set gathered only ≈ 350 sequences (File S3).

Within the CNL group, there were included only proteins annotated by Interpro to contain a single coiled-coil (CC) domain, a single NBS domain, and a LRR domain in this order, and sequences that contained a different domain organization, such as two annotated NBS domains or a different domain order were not included in the analysis. Similarly, for the TNL group we selected only sequences that contain a TIR-NBS-LRR domain organization. The RLK group was built with sequences displaying a "LRR-TM predicted region-kinase" domain organization, while the RLP group contained sequences with "LRR-TM" organization and did not contain other annotated domains by Interpro. In generating the vertebrate NLR group we included any annotated NACHT or NBS domains followed by a LRR domain annotation without discriminating on the N-terminal domain, as animal NLRs can have upstream of the NACHT/NBS domain a multitude of N-terminal domain types, while vertebrate TLRs group contains sequences with a "LRR-TM-TIR" configuration. Transmembrane predictions were performed using Phobius [56].

In analyzing the length of the LRR domains covered by individual repeat annotations, we used all Interpro annotation codes associated with LRR repeat types. We considered as having the status of 'annotated as domain' LRRs with the IPR032675 label and 'annotated as repeats' any amino acids

that had attached by at least one predictor part of Interpro collection one of the following tags: leucine-rich repeat (IPR001611), leucine-rich repeat, typical subtype (IPR003591), leucine-rich repeat, cysteine-containing subtype (IPR006553), leucine-rich repeat 2 (IPR013101), leucine-rich repeat 3 (IPR011713), leucine rich repeat 4 (IPR025875), BspA type leucine rich repeat region (IPR026906), CD180 leucine-rich repeat (IPR041281), DUF4458 domain-containing protein, leucine-rich repeat (IPR041403). Annotations referring to the N-ter cap of the LRR domain (IPR000372, IPR041302) were not considered as these are not LRR repeats.

2.6. Assessment of LRR Motif Conservation Across Protein Groups

Intra- and inter-group sequence variability was also analyzed using a subset of 1000 predicted 16 aa extended motifs from each group. In order to avoid a potential bias induced by false ‘entry’ (N) or ‘exit’ (C) repeats, only ‘core’ (L) repeats were used in this analysis. The similarity measure used here is the distance mapping defined by Halperin et al. [57]. This consists of the inner product of BLOSUM scores between each pair of amino acids summed up over the motif span, as this function can be used as a metric distance for several BLOSUM matrices. Considering d to be the distance between a pair of amino acids i and j , that have the $s(i, j)$ BLOSUM score:

$$d(i, j) = s(i, i) + s(j, j) - 2 \cdot s(i, j) \quad (1)$$

The distance between two sequences a and b of equal length l , would be the sum of distances of each pair of amino acids a_i and b_i across the length of the sequence:

$$D_{a,b} = \sum_{i=0}^l d(a_i, b_i), \quad (2)$$

This definition of distance is expected to reflect amino acids compatibilities, as BLOSUM scores are inferred from amino acid mutation probabilities observed on large datasets. As a BLOSUM matrix we selected an updated version of the original BLOSUM matrix, which was recently recalculated on a large dataset and satisfies the triangle inequality. (RBLOSUM59_14.3) [58,59].

Starting from the above described distance function, we calculated Silhouette coefficients [60] between each pair of groups, and precomputed distances were used for manifold learning using metric multi-dimensional scaling (MDS) [61] as implemented in Scikit-learn library [50].

Sequences logos were generated using Weblogo [62], figures showing protein structures were obtained using PyMol [63], while other plots were generated in Microsoft Office or by using Matplotlib library [64].

3. Results

3.1. Available LRR Domains in Structural Data

A collection of 611 PDB structures previously annotated by several protein domain databases, such as CATH [28], Pfam [29], and Interpro-collection [30] to contain LRR horseshoe architectures was obtained. This collection was used to derive a clean set, ID90, of 178 LRR chains displaying 90% identity that was structurally analyzed in order to structurally delineate the LRR repeats based on the beta-sheet network. By this, a dataset of ≈ 2100 LRR motifs was obtained, as shown in Figure 1a. It is interesting to note here that less than 20% of these are annotated as LRR motifs in Pfam even though the 178 sequences were derived from known 3D structures.

The LRR motif annotation of each repeat was performed starting with the first position (L_0) of the minimal motif ‘ $L_0XXL_3XL_5$ ’, position that marks the beginning of the ventral side of the horseshoe domain (Figure 1b). Superposition of the 2100 repeats indicates that the structural similarity extends in most of the cases over five positions upstream and downstream of the minimal motif defining a 16 positions region which is referred herein as the ‘extended’ motif (Figure A1d). Due to this, the structural

LRR diversity concentrates mainly onto the dorsal side of the horseshoe which imposes onto the curvature and the overall geometry of the domain (Figure 1b).

As duplications of highly similar LRR repeats within the same LRR domain is abundant in the ID90 set, we opted to perform a second redundancy filter at the level of LRR repeats as described (M&M). This results in the ID50 nonredundant set consisting of ≈ 850 LRR repeats, that approximates well the ID90 distribution of lengths (Figure A1c), phyla (Figure 1c), and the ratio between marginal N-terminal (N) and C-terminal (C) versus interior motifs (L) (Figure A1b).

The ‘entry’ N-ter LRR motifs are less regular than the ‘core’ motifs, especially at the first hydrophobic position (L_0) that is often found solvent exposed, as this position marks the end of the inter-domain linker and the beginning of the LRR domain. By contrast, the ‘exit’ C-ter LRR motifs better resemble the ‘core’ motifs (L) amino acid composition and the conventional LRR motif ‘ $L_0xxL_3xL_5xx(N/C)_8xL_{10}$ ’ (Figure 1d). Interestingly, the stringency for leucine occurrence sequentially decreases from L_0 to L_3 and L_5 in core repeats, allowing other amino acids to be present in L_3 and L_5 more frequently (Figure 1d). This structurally correlates with a larger accessible space of the protein core structure around L_3 and L_5 positions, as can be seen from Figures 1b and A1d. It is also worth noting that the third L_3 position upstream of $LxxLxL$ has a significant hydrophobic propensity presumably allowing the solenoid to form (Figure 1d).

Another important facet that has to be carefully pondered is the high phyla bias of the structural data when compared to the baseline phyla distribution of the UniRef50 database. As can be seen from Figure 1c, around 50% of the repeats in ID50 are of mammalian origin while the UniRef50 baseline is of less than 3% in both annotated LRR proteins or any protein. Moreover, the $\approx 20\%$ plant LRR motifs present in ID50 originate overwhelmingly from RLP and RLK proteins while plant NLRs are poorly represented in this set, with only a single 3D structure recently reported for the ZAR1 NLR protein from *Arabidopsis thaliana* [21,22].

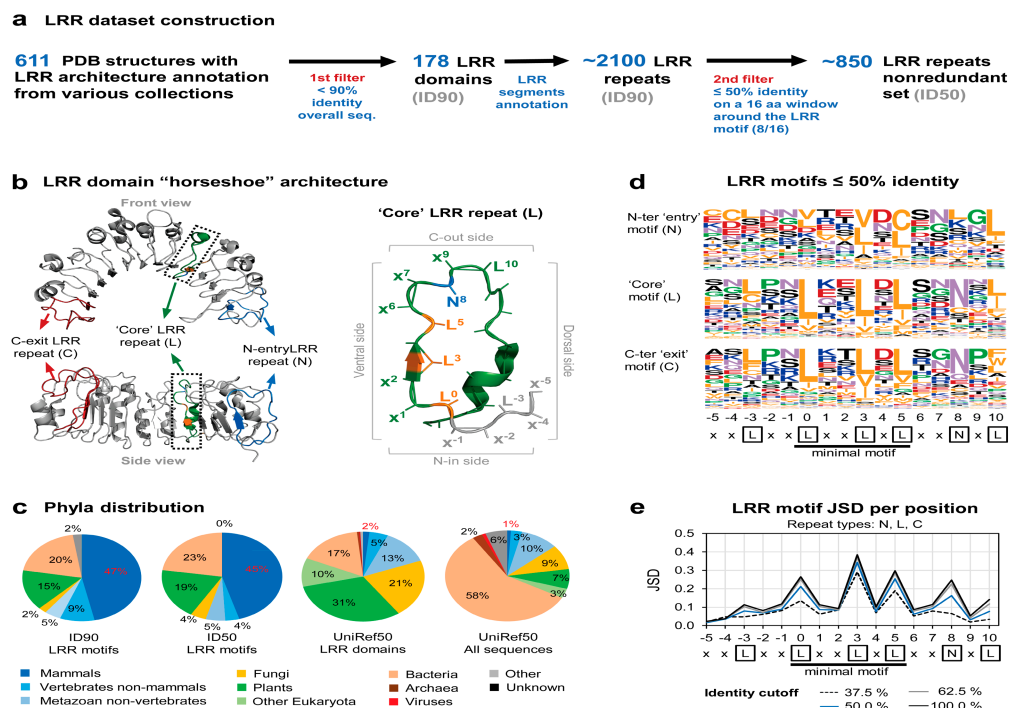


Figure 1. Available leucine-rich-repeat (LRR) domains in structural data. (a) LRR structural dataset construction. (b) LRR domain horseshoe architecture illustrated on the only plant NLR cryo-EM structure available—ZAR1—from *Arabidopsis thaliana* (left) and zoom-in view of a LRR repeat (right)

(PDB: 6J5W). The hydrophobic positions in the minimal 'L₀xxL₃xL₅' motif are shown in orange. The first N-entry repeat (blue) and the last C-exit repeat (red) are also mapped on the structure. (c) Phyla distribution of the initial LRR motif set ID90, the 50% identity trimmed LRR motifs set (ID50), annotated LRR proteins and all proteins from the UniRef50 database (from left to right). Percent values corresponding to the mammals group are shown in red. (d) Frequency plot of amino acid composition of the N-entry, core and C-exit motifs on the 50% identity trimmed set. Amino acids are colored according to their properties as follows: hydrophobic (yellow), acidic (red), basic (blue), asparagine and glutamine (purple), proline and glycine (green), others (black). (e) Jensen–Shannon divergence (JSD) score for each position of the LRR motif at different identity thresholds. Higher values show increased conservation.

3.2. Development of the LRRpredictor Method

In order to train a machine learning (ML) estimator for detecting LRR motifs we used an overall dataset comprising the filtered LRR ID50 dataset and a collection of 875 non-LRR domains composed of one representative of each CATH topology (Figure 2a).

As discussed in Section 1, the sequence patterns corresponding to the ≈ 850 actual *true structural LRR motifs* identified in ID50 are quite common in any protein. We will name here such sequence patterns as *potential motifs*. As expected, Table 1 shows that *potential motifs* occur with more or less equal probability in both the LRR and non-LRR domains of the overall dataset. Moreover, even when taking into account only LRR domains the number of *potential motifs* is larger than the number of *true structural LRR motifs* (Table 1). This allows the ML estimators to learn to detect *true motifs* from the far larger set of *potential motifs* by taking into account the larger 16 amino acid sequence context in which the *true motifs* are embedded. In this way the method developed herein can be used not only to delineate repeats in a given LRR domain but also to discriminate between protein products that do not have LRR domains from those hosting such domains.

Table 1. Occurrence of LRR sequence patterns in the overall dataset used to train the machine learning (ML) estimators.

LRR-Like Pattern	Total Number	Full Training & Testing Dataset (CV 1-4 and Test Sets)			
		NonLRR Proteins		LRR Proteins	
		False Motifs	True Motifs	False Motifs	True Motifs
LxxLxL (3L)	296	114	0	27	155
LxxLxL/LxxLxL/LxxLxL (2L)	1,060	773	0	147	140
LxxLxL/LxxLxL/LxxLxL (1L)	7,239	5,875	0	1,192	172
LxxLxL	12,247	10,149	0	1,417	681
LxxLxLxxN	811	438	0	76	297
LxxLxLxxC	273	163	0	41	69
LxxLxLxx(N/C)xL	618	269	0	39	310
Number of predicted positions (16 aa sliding windows):		148,540		25,658	
				815 LRR motifs	

L—strictly leucine; L—hydrophobic without leucine (I, V, M, F, W, Y, C, A); L—hydrophobic (L, I, V, M, F, W, Y, C, A); x—any amino acid.

In developing LRRpredictor we tested 'sequence-based' features based on position-specific scoring matrices-PSSMs either solely or combined with 'structural-based' features as described in Section 2 (Figure 2c). PSSM profiles are expected to provide context information on the overall sequence, to highlight the key amino acids position that are conserved, as the amino acids scores are derived from amino acid substitution probabilities conditioned by the homologues family they belong to. Therefore, it is expected that irregular LRR motifs would be more detectable when using sequence profiles, rather than amino acid sequence alone.

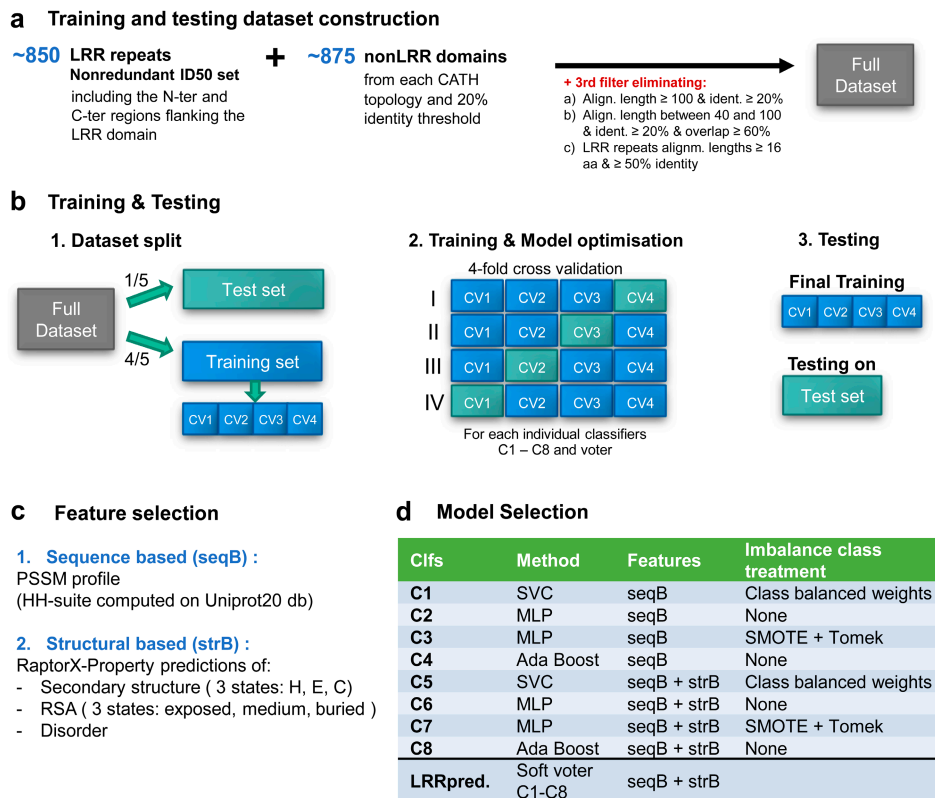


Figure 2. LRRpredictor training and testing workflow: (a) training and testing dataset construction. (b) schematic representation of the training and testing procedure, (c) selected features, and (d) selected classifiers aggregated into LRRpredictor.

The dataset was split into five parts: one part was initially separated as test set and the other four were used as a training set in parameter tuning using a four-fold cross-validation (CV) approach, where models were iteratively trained on three of the CV sets and tested on the remaining fourth (Figure 2b). A pool of estimators (representing algorithms for classification) that used either (1) sequence-based or (2) both sequence and structural features and (3) various imbalance class treatments were optimized via cross-validation. Finally, best performing estimators were studied in the context of an ensemble estimator. The selected ensemble classifier, further referred to as LRRpredictor is a soft voter aggregating eight classifiers C1–C8 (Figure 2d) which were trained to detect the LRR motif starting position—i.e., L_0 position from the minimalistic LRR motif ' $L_0xxL_3xL_5$ '.

Finally, LRRpredictor was trained on the entire training set (all four CV sets) and tested on the test set which had been set aside.

3.3. Assessment of LRRpredictor Performance

The precision of LRRpredictor given by the fraction of true-positives (TP) predicted results over the sum of true-positives (TP) and false-positives (FP) varies between 89% and 97% on the test set and within cross-validation sets (Figure 3a). Similarly, the recall (also known as sensitivity), given by the fraction of TP over TP + false-negatives (FN) varies between 85% and 93%, while the F1-score (representing the harmonic mean between precision and recall) varies between 87% and 95% on the test set and cross-validation sets (Figure 3a,b).

a LRRpredictor performance across datasets

Dataset	Classifier	All motif types (N+L+C)			Only 'core' motifs (L)			TN	FP		FN			TP			
		Precision	Recall	F1 score	Precision	Recall	F1 score		nonLRR prot.	LRR prot.	N-entry (N)	Core (L)	C-exit (C)	N-entry (N)	Core (L)	C-exit (C)	
CV	CV1	LRRpredictor	0.950	0.923	0.936	0.940	0.982	0.961	33007	0	7	8	2	1	11	110	11
	CV2	LRRpredictor	0.924	0.897	0.910	0.911	0.957	0.933	35829	0	13	9	6	3	12	133	12
	CV3	LRRpredictor	0.893	0.854	0.873	0.879	0.928	0.903	36030	0	16	7	9	7	12	116	6
	CV4	LRRpredictor	0.967	0.916	0.941	0.961	0.961	0.961	33349	0	6	8	6	2	14	147	13
Test	LRRpredictor	0.928	0.860	0.893	0.915	0.899	0.907	35116	0	10	7	12	2	12	107	10	

b F1 scores of LRRpredictor and its classifiers

Classifier	Method	Features	Imbalance treatment	Cross-validation				Test
				CV1	CV2	CV3	CV4	
C1	SVC	seqB	Class balanced w eights	0.930	0.891	0.864	0.926	0.877
C2	MLP	seqB	None	0.930	0.906	0.864	0.929	0.891
C3	MLP	seqB	SMOTE + Tomek	0.914	0.895	0.856	0.890	0.850
C4	Ada Boost	seqB	None	0.928	0.892	0.841	0.917	0.872
C5	SVC	seqB + strB	Class balanced w eights	0.931	0.914	0.866	0.911	0.868
C6	MLP	seqB + strB	None	0.937	0.920	0.862	0.939	0.893
C7	MLP	seqB + strB	SMOTE + Tomek	0.944	0.903	0.864	0.924	0.827
C8	Ada Boost	seqB + strB	None	0.927	0.905	0.823	0.911	0.878
LRRpredictor	voter C1 - C8	seqB + strB	None	0.936	0.910	0.873	0.941	0.893

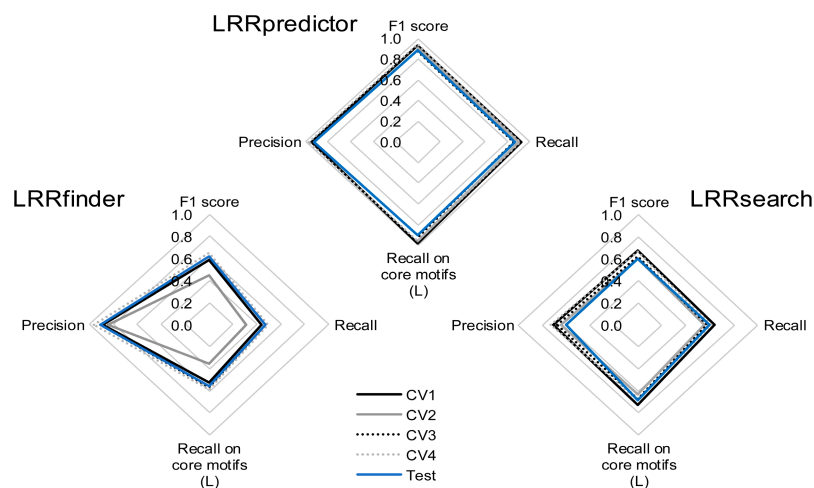
c Comparison between LRRpredictor and other LRR motif predictors

Figure 3. LRRpredictor performance analysis: (a) LRRpredictor performance across datasets: precision, recall, and F1 scores are shown either considering all the LRR motif types (N-entry, core, and C-exit types), either solely core motifs (L); also shown are the true negative (TN), false positive (FP), false negative (FN) and true positive (TP) counts. (b) F1 scores of LRRpredictor and its individual classifiers. (c) Comparison between LRRpredictor and other LRR motif predictors: LRRfinder [26] and LRRsearch [27] (computed on their webservers using default parameters).

3.4. LRRpredictor Behavior on Protein Families Containing LRR Domains

As the available structural data is scarce, we further evaluated the extrapolation capabilities of LRRpredictor on a set of LRR domains annotated in Interpro collection. Groups of the most representative protein functional classes containing LRR domains were generated as follows: four groups from flowering plants—resistance proteins (CNL_{plants} and TNL_{plants}) and extracellular receptors (RLK_{plants} and RLP_{plants}) and two groups from vertebrates—NLR_{vert} and TLR_{vert} as described in Section 2.

Selected sequences from each group were subjected to LRRpredictor motif detection. The repeat length distribution of the predicted LRR repeats (Figure 4a), is consistent with previously reported lengths within all protein groups of the seven type Kobe–Kajava (KK) classification [14,65]. The repeat length distribution of extracellular LRR domains (RLK_{plants}, RLP_{plants}, and TLR_{vert}) show a sharp peak at 24 amino acids, in agreement with the most frequent repeat length within plant-specific (PS) from KK classification [14,65]. As they often contain large helices over the dorsal side of the LRR horseshoe, vertebrate NLRs repeats have longer lengths (25–30 aa) as previously shown by the same classification,

while plant NLRs (CNL_{plants} and TNL_{plants}) have a larger distribution with a lower peak shaping up toward lower value side (20–24 aa) of repeat lengths range.

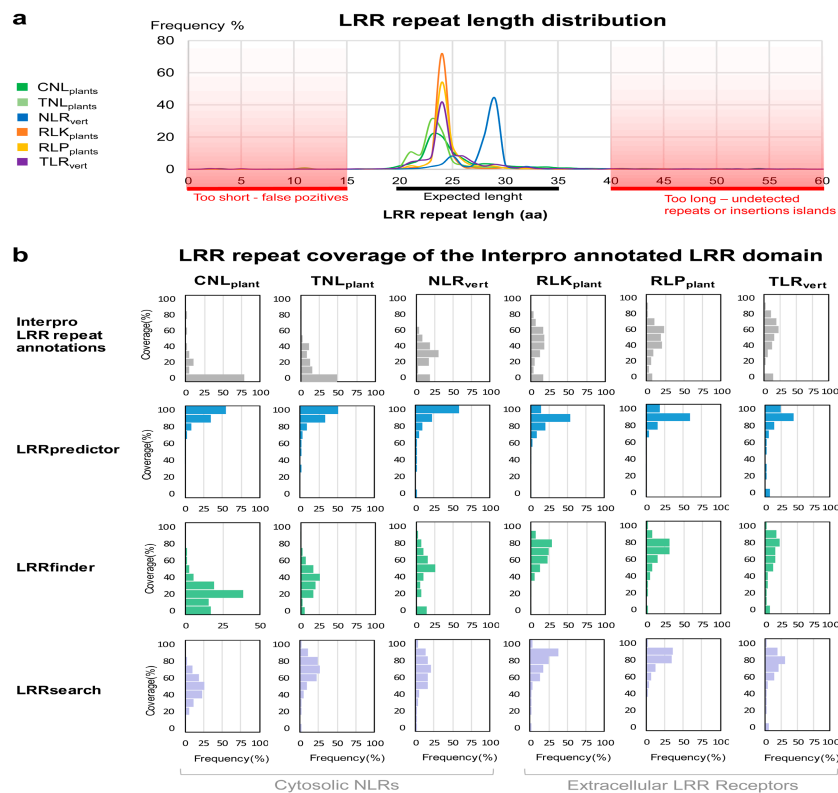


Figure 4. LRRpredictor behavior on Interpro annotated LRR domains from different classes. (a) Length distribution of the predicted repeats using LRRpredictor within each protein group. C-terminal motifs were not used in computing the distribution. Repeat lengths size prone to ambiguity—i.e., either too short (potential FP) or too long (potential FN)—are shaded in red. (b) Distributions of the Interpro annotated LRR domain length that is covered by Interpro LRR repeat annotations (grey) or by predicted repeats using LRRpredictor (blue), LRRfinder (green), and LRRsearch (purple). Coverage percent distributions are shown within each protein group.

As seen in Figure 4a, repeat lengths are rarely found outside the 19–35aa range, cases in which prediction becomes ambiguous. Too short repeats are improbable due to structural constraints and might indicate false positive predictions. Similarly, too long repeats—over 40 amino acids—could indicate either the presence of undetected repeats (false negatives) or cases in which an insertion or ‘island’ shapes up protruding the horseshoe structure (Figure 4a). Very large gaps between LRR motifs (more than 100 aa) were not included in computing the length distribution as these are rather indicating the presence of an inserted domain flanked by two LRR domains.

We further analyzed the percent of the annotated LRR domain span that is covered by LRRpredictor and compared the predicted LRR motifs to LRRfinder [26] and LRRsearch [27] predictions and to the existing motif annotations from Interpro collection. In doing so we defined as predicted repeats motifs separated by 15–35 amino acids. Predicted motifs that superpose or cluster within 15 amino acids were counted only once, while when the distance between two motifs was higher than 35, the repeat was considered to be a potential terminal repeat or contain a domain break and the first 24 aa of such a stretch was assigned as a predicted repeat, given that this is the most frequent repeat length over the structural data.

The repeat coverage of analyzed LRR domains predicted by LRRpredictor, LRRfinder and LRRsearch were compared to the extent Interpro repeat annotations using a coverage percentage (CP)

defined as the ratio between the sum of predicted/annotated repeat length vs the overall LRR domain length (Figure 4b).

Plant NLRs from flowering plants show the lowest level of repeat annotation as 75% and 50% of CNL and TNL LRRs in Interpro lack *any* repeat annotation resulting in CP = 0% (Figure 4b). In comparison, repeats in vertebrate NLRs are better annotated in a CP ranging within 20–60% of the LRR domain. Even higher Interpro repeat annotations are shown by the extracellular plant and vertebrate receptors with CP ranging most frequently between 30% and 80% of the LRR domain size (Figure 4b). For all six receptor classes analyzed herein, both LRRfinder and LRRsearch slightly increase the LRR coverage as compared to Interpro annotations especially in the case of extracellular receptors, with LRRsearch surpassing LRRfinder in the case of plant NLRs (Figure 4b).

As also can be seen from Figure 4b, in comparison to Interpro and the two predictors mentioned above, LRRpredictor covers far larger regions of LRR domains with coverage percentages (CP) exceeding 60% and almost complete coverage in over 50% in all six groups (Figure 4b). It is interesting to note that Interpro annotation of extracellular LRR domains also include the N-terminal cap region, that is *not* formally a LRR repeat. This results in the fact that LRRpredictor covers in most cases only $\approx 90\%$ of this domain, instead of 100% as in NLR groups (Figure 4b).

3.5. Predicted Repeats Consensus in Each Class

Further, the amino acid composition of the predicted LRR motifs was investigated solely on the ‘core’ predicted LRR repeats, i.e., repeats that are flanked by other predicted repeats within a 15–35 aa range. In short, the results presented below clearly indicate that LRRpredictor is able to detect and reproduce all the consensus motifs previously defined for well-studied classes of RLKs, NLRs, and TLRs (Figure 5).

This is especially the case for vertebrate NLRs. The consensus follows the ribosomal inhibitor (RI) type - ‘ $x_3xxL_0xxL_3xL_5xx(N/C)_8xL_{10}xxxg_0xxLxxoLxx$ ’ [14,65], with position ‘-3’ being less relevant for this class of repeats. Additionally, the vertebrate TLRs predicted motif consensus matches the

“T” type motif: $L_3xxL_0xxL_3xL_5xxN_8xL_{10}xxL_{13}xxx(F/L)_{18}xxL_{21}xx$

defined in Matushima et al. classification [13] rather than the less encountered

“S” type motif: $L_3xxL_0xxL_3xL_5xxN_8xL_{10}xxL_{13}Px(x)LPxx$.

In the case of plant extracellular receptors, the predicted motifs from RLK_{plant} and RLP_{plant} groups show a prolonged pattern that is in perfect agreement with the plant-specific (PS) type from Kobe and Kajava classification [14,65]— $L_3xxL_0xxL_3xL_5xxN_8xL_{10}(S/T)_{11}GxIPxxLxxLGx$. Interestingly, the kinase containing receptors (RLK) have a more prominent consensus (Figure 5).

On the other hand, the predicted motifs in plant NLRs comprising the CNL_{plant} and TNL_{plant} groups display a remote similarity with the cysteine-containing (CC) type as defined by Kobe and Kajava classification [14,65]:

“(C/L) $_3xxL_0xxL_3xL_5xxC_8xxITDxxOxxL(A/G)xx$ ”—where O is any nonpolar residue.

While the extended motif is satisfied (16 aa), a difference worth noting is that in both CNL and TNL groups cysteine is rare in position ‘-3’ and outside this region any similarity with CC-type ends. Both plant NLR groups mainly confine their consensus to only the minimal $L_0xxL_3xL_5$ motif, with TNL extending it a little bit with C_8 position. By contrast, in plant extracellular receptors the consensus expands beyond the 16 amino acids of the ‘extended’ region covering all the four sides of the LRR solenoid. Despite being analogous in composition, the TNL_{plant} group consensus is more pronounced, especially at positions C_8 and L_{11} (Figure 5).

In all six classes L_0 , L_3 , and L_5 of minimal motif are as expected overwhelmingly hydrophobic, with all three positions occupied by leucine in around 50% of the cases, except CNLs where leucine occurrence seems less stringent (Figure 6). When compared to the Kobe and Kajava classification, the majority of the motifs fall under the expected class and very few cross terms are seen between them (Figure 6).

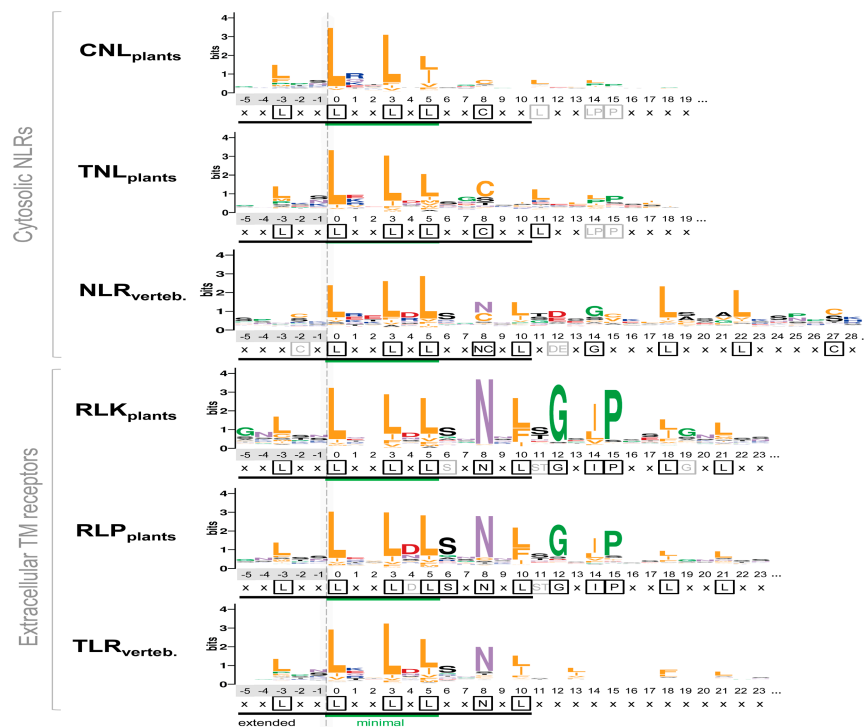


Figure 5. Consensuses of the LRR motifs predicted by LRRpredictor across different classes. Logo heights correspond to amino acid relative entropy (in bits), higher heights implying higher conservation. A consensus for each class is displayed below each logo, highly conserved positions being shown in black boxes, while less conserved in gray. Minimal motif ‘L₀xL₃xL₅’ (green line) and the extended motif (black line) are indicated below each logo. Amino acids are colored according to their properties as in Figure 1d.

LRR motif classes	LRR pattern expression	intracellular			extracellular		
		CNL _{plant}	TNL _{plant}	NLR _{vert}	RLK _{plant}	RLP _{plant}	TLR _{vert}
Minimal	W6	29	44	51	55	55	46
	W6	92	95	84	95	95	94
*Cysteine containing (CC) (intracell.)	W11	31	49	24	6	6	4
	W16	26	40	6	4	4	3
	W16+	23	37	0	1	1	1
*Ribonuclease inhibitor like (RI) (intracell.)	W11	7	13	54	80	80	62
	W16	7	13	54	80	80	62
	W16+	1	1	51	15	15	9
*Plant specific (extracell.)	W11	2	4	31	74	74	59
	W16	2	3	8	60	60	45
	W16+	0	0	0	39	39	0
*Typical (extracell.)	W11	2	4	31	74	74	59
	W16	2	3	8	60	60	45
	W16+	0	1	2	8	8	27

* Adapted from KK classification
 L - strictly leucine
 L - any hydrophobic amino acid (leu, val, ile, met, phe, tyr, trp, cys or ala)
 x - any aminoacid

Figure 6. Distribution of LRR motif types defined by Kobe and Kajava (KK) [14] predicted with LRRpredictor across the six receptor classes. As the motif consensuses from KK classification were very strict, we adapted these consensuses to different sequences windows (W6, W11, W16, or more) centered around the minimal motif as shown in the table. Percentages of the predicted motifs compatible with each consensus are shown with grey bars.

CNLs and TNLs seem more dispersed even on a shorter 11 amino acid window consensus (W11), while the extracellular receptors obey in over 60% of the cases the corresponding W11 pattern that is shared simultaneous by all three classes (Figure 6).

3.6. LRR Motifs Variability Across Classes

Sequence variability is of critical importance for LRR domain function and in contrast to their common structural pattern, a wide spread in the sequence space is expected. To assess this, we analyzed the extended motifs, predicted by LRRpredictor, both the intra- and inter- group sequence similarity. This was performed over subsets of randomly selected 1000 examples of ‘core’ (L) motifs from each group. We selected as similarity measure a metric distance function [57] derived from BLOSUM scores which reflect the structural compatibility between amino acids, as described in Section 2. Using this metric, we calculated the distance between each predicted LRR motif from all groups and analyzed how these distances behave intra- and inter-groups.

Intra-group all-vs.-all distances distribution shows that the extracellular groups RLK, RLP from plants and TLRs from vertebrates form a denser group in terms of conservation, than plant and vertebrate NLRs (Figure 7a left). Figure 7b shows the silhouette coefficients. These scores show how separated two given clusters are, based on the distance between samples from each group, the maximal value of 1 corresponding to perfectly separated clusters, value 0 corresponds to clusters that coincide, while negative values with a minimum of -1 correspond to the case where samples from one group actually cluster better with the opposite group that is being compared. Silhouette coefficient of all versus all analyzed groups indicate that the NLR groups form a rather overlapping cluster, that has an increased variability among its sample motifs (i.e., expanded cluster) (Figure 7b left). Extracellular plant receptors RLK and RLP clusters are overlapping and have a more reduced span in terms of variability (i.e., more conserved motifs), while vertebrate TLR overlap plant RLK and RLP receptors have a slightly increased variability (Figure 7a,b left). Interestingly, within the minimal LRR motif region ‘L₀XXL₃XL₅’ there are no significant differences between groups (Figure 7a,b right).

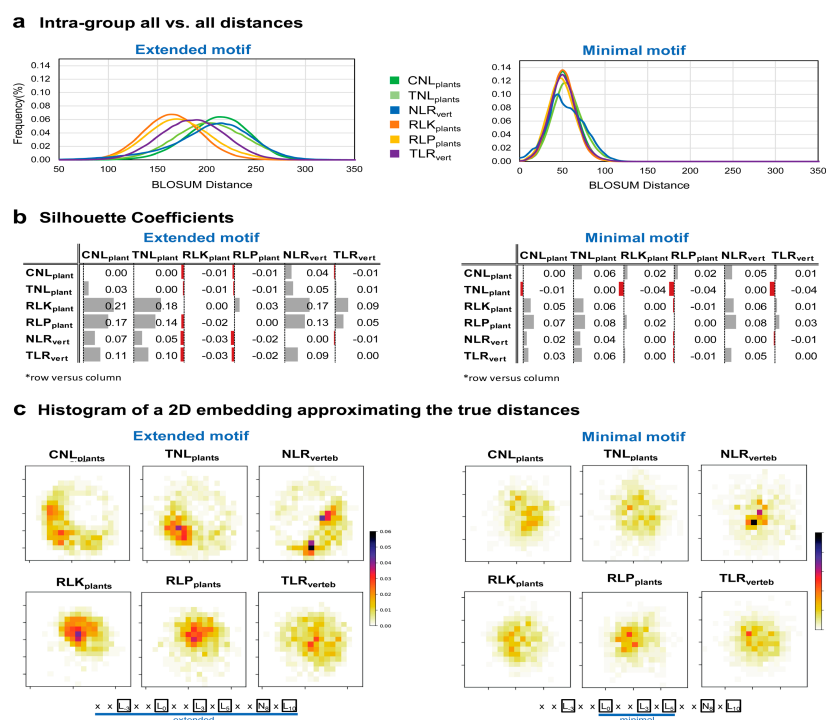


Figure 7. LRR motifs variability in different protein classes. (a) Intra-group all-vs.-all distances on the extended (left) and minimal (right) motif (b) Silhouette coefficients inter-groups extended (left) and minimal (right) motif. (c) Histogram of a 2D embedding approximating the true distances between points for the extended (left) and minimal (right) motif. Histograms were computed using a 20 × 20 bins grid. Extended and minimal motif histograms cannot be compared as they refer to different sequence spaces.

To have an overall view on the sequence dispersion in each protein class containing LRR domains Figure 7c shows the 2D embedding of the high dimensional sequence space of both the extended and minimal motifs of each class. Nonetheless such a reduction gives only a rough representation of distance relations between clusters in the original space as the normalized stress parameter (*stress-1*) of this 2D embedding is 0.25 and 0.21 for the extended and minimal motif space, respectively [66].

3.7. LRRpredictor Specificity Tested on Solenoid Architectures

From a structural point of view the LRR protein architecture belongs to the larger class of solenoidal architectures which are defined by specific repeated structural patterns. Given the repetitiveness of such structures we asked if LRRpredictor is able to discriminate between LRR motifs and other repetitive sequence patterns. The main candidates considered for possible misclassifications are two classes of beta sheet repeat proteins—which are the closest structural relatives of LRR domains: pectate lyases (PeLs) and trimeric LpxA architectures and two helical repetitive classes: armadillo and ankyrin architecture (Figure A2b). To this end, 50 sequences from each of the above four classes annotated as such by Interpro were randomly selected from UniRef50. Figure A2a shows the probabilities returned by LRRpredictor that the potential motifs occurring in the 200 sequences are true LRR structural motifs. As can be seen in all four classes taken into account, the vast majority of potential motifs have a probability lower than 10% to be true motifs. Only 0.1% of such sites show a probability between 10% and 20% to be true motifs and none of these sites reaches a threshold of 40% for being a true LRR motif (Figure A2a). From a technical point of view this result shows that LRRpredictor is highly specific for LRR domains. On the other hand this result is even more interesting from a structural and biological point of view indicating that even if LRRs and PeLs were considered to be members of the same LRR superfamily [67] the structural principles upon which they are built are different and presumably the two classes have diverged very early in evolution.

4. Discussion

Given the high number of indeterminacies generated by sequence variability, a proper annotation of LRR motifs and the correct delineation of repeats is critical in identifying potential protein–protein interaction sites of LRR domains.

Here, we show that LRRpredictor is able to address this problem and by this, can be of use as a new tool in the analysis of especially plant NLR sequences that display a larger variability and irregularity as compared to other LRR domains [9,68]. This often results in the superposition or presence in less than a minimal repeat distance of potential alternative LRR motifs, as can be seen from Figure 8 illustrating such indeterminacies found on a 150 amino acid stretch from the potato CNL Gpa2 LRR domain.

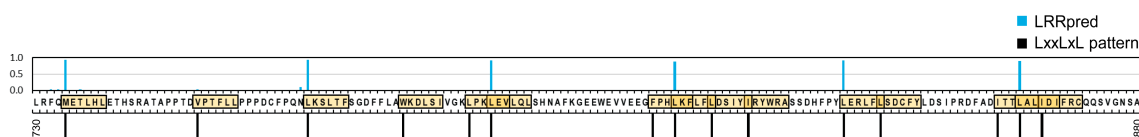


Figure 8. LRR motif and repeat indeterminacies onto a 150 aa stretch in Gpa2 potato NLR. Potential motifs that follow the minimal ‘LxxLxL’ pattern (where L is any hydrophobic amino acid) are illustrated above the sequence with black bars and yellow highlight, while LRRpredictor results are shown above with blue bars.

Given the scarcity of structural learning data consisting of less than 180 LRR structures with lower than 90% identity and only ≈ 850 motifs at hand in ID50 (<50% identity), in order to maximize LRRpredictor extrapolation abilities, the method was set to rely on aggregating a collection of eight classifiers based on different strategies, two of them designed to perform a massive oversampling of the real data (Figure 2).

In this context, LRRpredictor shows to perform well, with overall *precision*, *recall*, and *F1* scores ranging between 85% and 97% on both test and cross validation sets (Figure 3a). In addition, LRRpredictor increases its performances when taking into account only the ‘core’ repeats (L), as the main prediction problems relate only to the N-‘entry’ repeats (N)—i.e., the first repeat of the LRR domain (Figure 3a, Table A1). This can be explained in part by the increased irregularity of the sequence in this region, but also by the small sample size of the N-‘entry’ (N) motifs when compared to the ‘core’ (L) motifs.

It is also important to note here the fact that false positives are almost never found in nonLRR domains but always in proteins containing LRR domains (Table A1). Here, such false positives shape up in close vicinity to the marginal repeats—where the LRR motif characteristics are more diffuse, or in linkers or different domains neighboring the LRR, but found in a ‘one repeat range’ to the N-entry motif.

Other false predictions are caused by alignment artefacts. These yield to an offset of 1–3 amino acids in the predicted LRR motif starting position. Alignment artefacts are also frequently seen in regions with high beta structure propensity of insertion loops or ‘islands’ protruding from the LRR domain structures. This is mainly due to the fact that the multiple alignment on which PSSM relies forces the protruding loop in the queried sequence to align to regular repeats in the template LRRs of the database.

Unfortunately, the number of such insertion loops or ‘islands’ is so small in ID90/ID50 that estimators cannot learn from the existing data to discriminate such false positives. Thus, only careful structural analysis performed in later modelling stages can handle such cases.

Results on both cross-validation and test sets show that estimators using structural features in addition to the sequence based features (C5–C8) perform on average only slightly better compared to those sequence based only (C1–C4), with some interesting improvements on *F1* scores (Figure 3b). This only marginal improvement may indicate that RaptorX-Property [34] training on the overall structural database that might have marginally overlapped with our testing dataset did not affect the results. Nevertheless, C5–C8 are expected to be better extrapolators (Table A1), while the structural predictions on which (C5–C8) are based, and that are present in the output file can prove instrumental in further dealing with ambiguous cases where two LRR motif signatures partially superpose or are within the limit of a repeat.

Figures 3c and 4b compare predictions of three existing engines. LRRpredictor outperforms LRRsearch [27] and LRRfinder [26]. This is expected as the two previous methods were designed to focus mainly on specific LRR classes such as vertebrate TLRs or NLRs, respectively, while LRRpredictor relies on a newer larger dataset and was designed to identify LRR motifs in general. However, despite focusing on specific protein classes, both LRRsearch and LRRfinder show comparable efficiency in covering annotated LRR domains in plant extracellular receptors but decreased capabilities on plant NLRs (CNLs and TNLs) (Figure 4b). Furthermore, both LRRsearch and LRRfinder were intended for fast computation and they use a predefined PSSM matrix computed on a curated collection of LRR domains, instead of performing case by case basis sequence profiles as our method does.

However, the increased performance of LRRpredictor comes with an attached computational cost and is not easily scalable for scanning large protein sequences databases such UniprotKB. The main reason for this is that generating case by case sequence profiles and performing predictions for each estimator aggregated in LRRpredictor is more computationally demanding than LRRfinder and LRRsearch workflow.

Another matter of concern was related to the phyla bias of the database on which LRRpredictor relies, as $\approx 50\%$ of ID90 have mammalian origin while the share of mammalian—from total—annotated LRRs in UniRef50 is only 2% (Figure 1c). Moreover, groups such as plant NLRs are extremely poorly represented, as only very recently the first plant NLR structure was reported [21,22].

In this context, in order to investigate the extrapolation capabilities of LRRpredictor we used a set of LRR domains annotated in Interpro collection from the six most representative immune receptor classes: R-proteins and extracellular receptors from flowering plants (CNL_{plants}, TNL_{plants},

RLK_{plants}, RLP_{plants}) and their vertebrate counterparts (NLR_{vert}, TLR_{vert}). The LRR motifs predicted by LRRpredictor show a good coverage of the LRR domains annotated by Interpro and follow the expected repeat length distribution for all these six classes [14,65] (Figure 4a,b). Moreover, the predicted motifs reproduce the expected LRR motif consensus of each protein class (Figure 5) from Kobe and Kajava classification [14,65]. Combined, these indicate that LRRpredictor is able to extrapolate well in different LRR motif classes which is especially important for plant NLRs.

Analysis of LRRpredictor detected motifs showed clear differences between the six classes within the extended 16 aa motif. Whether variation in these extended motifs directly relate to the functional diversification of the different receptor classes still remains to be addressed. By contrast, within the minimal 6 aa LRR motif region—L₀XXL₃XL₅—there are no significant differences between the six groups (Figure 7). This might suggest a common root of minimal structural criteria imposed by the solenoidal architecture from which the six classes have diverged to fulfil specific tasks in specific environments. For receptor function, such a solenoidal domain organization in which only three positions over a ≈ 25 repeat length are loosely conserved has two-fold evolutionary advantages: first the solenoid architecture ensures a large solvent exposed surface area [10] and second a high sequence variability can be achieved without disturbing the tertiary structure.

The increased conservation seen at the level of the extended motif among all three extracellular LRR classes—plant RLK, RLP and vertebrate TLRs—when compared to plant and vertebrate NLRs could be related to N-glycosylation and the constraints imposed by the extracellular environment. On the one hand, plant NLRs recognize directly or indirectly a suite of pathogen effectors or (perturbations) of their host targets conferring host specific immunity. Single amino acid changes in the effector can already be detected or are sufficient to evade recognition by a NLR, resulting in a co-evolutionary arms race between pathogen effectors and host immune receptor [69,70]. In contrast, extracellular LRRs recognize often conserved microbial patterns to confer basal immunity thus lacking such a strong driver for diversification [71]. Vertebrate NLRs act more like basal immune receptors in innate immunity, recognizing conserved microbe-associated molecular patterns (MAMPs). The greater diversifying selection imposed by fast-evolving effectors may therefore account for the co-evolution of structurally highly variable LRR motifs in plant NLRs. In the future, it will be interesting to relate our LRR structural annotations to specific functional NLR sub-classes. This is relevant, for instance, as some plant NLRs-types are described to have a downstream ‘helper’ function rather than a role as a canonical ‘sensor’.

Another aspect is that LRR domains in plant NLRs have a dual role. They not only contribute to pathogen recognition, but also negatively regulate the switch function [72]. Hence, it will be interesting to link LRR structural annotations to specific intramolecular domain interactions between LRRs and other NLR subdomains to better understand the co-evolution of protein domains in NLRs. It is shown that subtle mutations in the interface between LRR and NB-ARC can have a major effect on NLR functioning, often resulting in constitutive immune activation or a complete loss-of-function [17,72]. This shows the tight link between structural and functional constraints underlying the shaping of NLRs in plants. Additionally, the link between LRR structural annotations and complex formation with other host proteins will be interesting to assess. LRR domains are known to interact with other components like chaperones (e.g., SGT1) which are required for proper NLR folding and functioning [73], or kinases (e.g., ZED1, RKS1) [21,22,74] but also NLR hetero- and homodimers are often formed [9,75,76] which could impose additional structural constraints on the shape and irregularity of LRR domains in plant NLRs.

5. Conclusions

The results presented herein indicate that LRRpredictor shows a good performance on the available 3D data and good extrapolation capabilities on plant NLRs (CNL/TNL), which are poorly represented in the training dataset. Predicted LRR repeats using LRRpredictor significantly increase the coverage of Interpro annotated LRR domains from main immune receptors groups. In addition, these predicted

repeats are consistent with previously defined motif consensus from all studied groups and also follow the repeat length range specific to each class. In conclusion, LRRpredictor is a tool worth using in research topics related to understanding immune receptors functions and structure-informed strategies for pathogen control technologies.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2073-4425/11/3/286/s1>, File S1: LRR structural datasets, File S2: LRRpredictor cross-validation and testing raw data, File S3: Predicted LRR motifs within each protein class.

Availability: LRRpredictor webserver can be accessed at <https://lrrpredictor.biochim.ro> and on GitHub repository at https://github.com/eliza-m/LRRpredictor_v1 alongside with installation and usage documentation.

Author Contributions: Conceptualization, A.-J.P., A.G., E.C.M.; Methodology, E.C.M., L.S., V.C., R.T., A.-J.P., A.G.; Software, E.C.M.; Validation, V.C., L.S., L.G.G.; Formal analysis, E.C.M., O.C.A.S., R.T., A.-J.P., A.G.; Investigation, E.C.M., O.C.A.S.; Resources, L.G.G.; Data curation, E.C.M., O.C.A.S., L.S.; Writing—original draft preparation, E.C.M., A.-J.P., A.G.; Writing—review and editing, E.C.M., A.-J.P., A.G., R.T., V.C., O.C.A.S.; Visualization, E.C.M., A.-J.P.; Supervision, A.G., A.-J.P.; Project administration, A.-J.P., A.G.; Funding acquisition, A.-J.P., L.S., R.T., A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Research and Innovation, CNCS - UEFISCDI grants PN-III-ID-PCE-2016-0650 (E.C.M. and A.J.P) and PN-III-P1-1.1-TE-2016-1852 (E.C.M. and L.S.), by the National Authority for Scientific Research and Innovation, and Ministry of European Funds, through the Competitiveness Operational Programme 2014-2020, POC-A.1-A.1.1.4-E-2015 [Grant number: 40/02.09.2016, ID: P_37_778] (R.T., V.C., L.S.) and Romanian Academy programs 1 & 2 of IBAR (E.C.M., L.S. and A.J.P) and by the Dutch Technology Foundation STW (O.C.A.S. and A.G.)

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

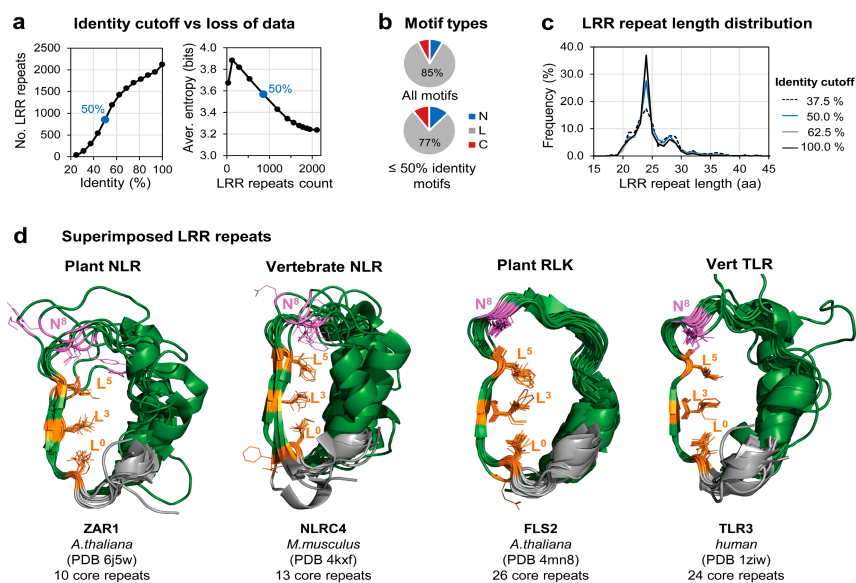


Figure A1. Available LRR domains structural data. (a) Identity cut-off versus loss of data: plots of the loss of samples (left) and increase in entropy (right) at different identity thresholds. Displayed is the Shannon entropy averaged over the 16 amino acid extended motif. (b) Composition of LRR motif types: N-entry (N), core (L), and C-exit (C) LRR motifs in the initial set (ID90) and in the 50% identity trimmed dataset (ID50). (c) LRR repeat length distribution at different identity thresholds. C-exit motifs were not used. (d) Structural superposition of the LRR repeats from a plant NLR, vertebrate NLR, plant RLK and vertebrate TLR structures [22,77–79] (from left to right). Hydrophobic positions of the minimal 'L₀xxL₃xL₅' motif are shown in orange and position N₈ in purple.

a LRRpredictor behaviour on other solenoidal repetitive 3D folds

	Pectate lyases	Trimeric LpxA	Armadillo	Ankyrin
LRR-like patterns				
LxxLxxL (3L)	16	10	54	51
LxxLxxL / LxxLxxL / LxxLxxL (2L)	130	53	192	246
LxxLxxL / LxxLxxL / LxxLxxL (1L)	1080	469	1511	1643
LxxLxxL	2163	1028	2130	2672
LxxLxxLxxN	232	40	116	219
LxxLxxLxxC	38	23	29	62
LxxLxxLxx xL	138	27	60	100
LRRpredictor	37784	14484	28152	32614
probability range				
0 - 10%	46	8	6	6
10 - 50%	0	0	0	0
50 - 100%				
Number of predicted positions (16 aa sliding windows)	37830	14492	28158	32620

L - strictly leucine
 E - hydrophobic without leucine (I, V, M, F, W, Y, C, A)
 L - hydrophobic (L, I, V, M, F, W, Y, C, A)
 x - any amino acid

b Other Solenoid repetitive architectures

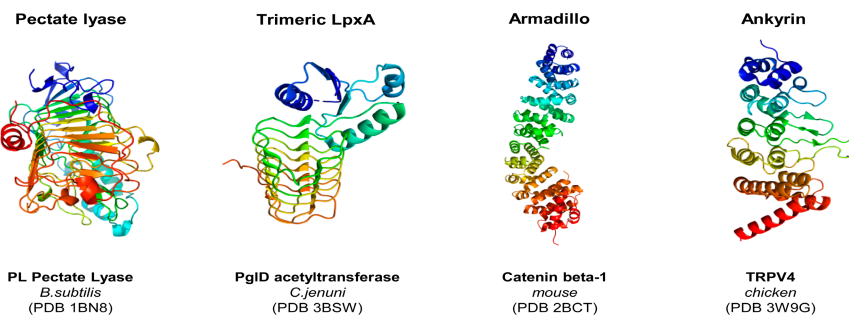


Figure A2. (a) LRRpredictor behavior on other solenoidal architectures. Shown are occurrence counts of LRR-like patterns versus LRRpredictor probabilities counts histogram. (b) Overall 3D structure of the four examined classes of solenoidal proteins [80–83].

Table A1. Detailed performance analysis on LRRpredictor and its classifiers. Precision recall and F1 scores are shown for in-sample (i.e., training data) and out-of-sample data (i.e., test data that was not used in training), for both cross-validation and test phase. In the cross-validation stage, classifiers were trained on three of the cross-validation (CV) sets and tested on the fourth set in an iterative manner, while in testing stage, classifiers were trained on all four CV sets and evaluated on the test set left aside from the beginning. The counts for true-negatives (TN), false-positives (FP), false-negative (FN), and true-positive (TP) within each set are also shown. As marginal repeats (N-entry and C-exit types) have a lower detection rate, also included is the recall calculated only with respect to ‘core’ repeats (L), indicated with blue font. Performance scores shading is according to a value based colormap from yellow (0.75) to blue (1.00).

Dataset	Classifier	In-Sample			Out-Of Sample													
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	TN	FP		FN			TP			Recall on Core(L) Only	
								-	Non-LRR Proteins	LRR Proteins	N-Entry (N)	Core (L)	C-Exit (C)	N-Entry (N)	Core (L)	C-Exit (C)		
Cross validation	CV 1	C1	0.897	0.998	0.945	0.930	0.930	0.930	33004	0	10	7	2	1	12	110	11	0.982
		C2	0.938	0.904	0.921	0.936	0.923	0.930	33005	0	9	8	2	1	11	110	11	0.982
		C3	0.877	0.887	0.882	0.904	0.923	0.914	33000	3	11	8	2	1	11	110	11	0.982
		C4	0.942	0.866	0.902	0.956	0.902	0.928	33008	0	6	9	3	2	10	109	10	0.973
		C5	0.894	0.925	0.910	0.924	0.937	0.931	33003	1	10	6	2	1	13	110	11	0.982
		C6	0.919	0.912	0.915	0.943	0.930	0.937	33006	0	8	7	2	1	12	110	11	0.982
		C7	0.976	1.000	0.988	0.938	0.951	0.944	33005	0	9	2	4	1	17	108	11	0.964
		C8	0.943	0.853	0.895	0.970	0.888	0.927	33010	0	4	10	3	3	9	109	9	0.973
	LRRpredictor	0.930	0.918	0.924	0.950	0.923	0.936	33007	0	7	8	2	1	11	110	11	0.982	
	CV 2	C1	0.904	1.000	0.950	0.874	0.909	0.891	35819	0	23	9	4	3	12	135	12	0.971
		C2	0.930	0.900	0.915	0.903	0.909	0.906	35825	0	17	8	5	3	13	134	12	0.964
		C3	0.901	0.890	0.895	0.892	0.897	0.895	35823	0	19	10	5	3	11	134	12	0.964
		C4	0.954	0.855	0.902	0.937	0.851	0.892	35832	0	10	13	8	5	8	131	10	0.942
		C5	0.904	0.937	0.920	0.914	0.914	0.914	35827	1	14	8	4	3	13	135	12	0.971
		C6	0.947	0.904	0.925	0.925	0.914	0.920	35829	0	13	8	4	3	13	135	12	0.971
		C7	0.974	1.000	0.987	0.898	0.909	0.903	35824	2	16	7	4	5	14	135	10	0.971
		C8	0.953	0.878	0.914	0.944	0.869	0.905	35833	0	9	10	9	4	11	130	11	0.935
	LRRpredictor	0.941	0.908	0.924	0.924	0.897	0.910	35829	0	13	9	6	3	12	133	12	0.957	
	CV 3	C1	0.925	0.998	0.960	0.881	0.847	0.864	36028	0	18	7	10	7	12	115	6	0.920
		C2	0.949	0.943	0.946	0.903	0.828	0.864	36032	0	14	8	12	7	11	113	6	0.904
		C3	0.922	0.904	0.913	0.879	0.834	0.856	36028	0	18	7	11	8	12	114	5	0.912
		C4	0.972	0.898	0.934	0.899	0.790	0.841	36032	0	14	7	18	8	12	107	5	0.856
		C5	0.932	0.945	0.938	0.887	0.847	0.866	36029	0	17	8	9	7	11	116	6	0.928
		C6	0.944	0.931	0.938	0.891	0.834	0.862	36030	0	16	9	10	7	10	115	6	0.920
C7		0.979	1.000	0.989	0.861	0.866	0.864	36024	2	20	6	8	7	13	117	6	0.936	
C8		0.968	0.906	0.936	0.883	0.771	0.823	36030	0	16	10	18	8	9	107	5	0.856	
LRRpredictor	0.950	0.939	0.945	0.893	0.854	0.873	36030	0	16	7	9	7	12	116	6	0.928		
CV 4	C1	0.885	1.000	0.939	0.936	0.916	0.926	33343	0	12	8	6	2	14	147	13	0.961	
	C2	0.923	0.884	0.903	0.961	0.900	0.929	33348	0	7	9	7	3	13	146	12	0.954	
	C3	0.818	0.901	0.858	0.851	0.932	0.890	33324	2	29	7	4	2	15	149	13	0.974	
	C4	0.931	0.848	0.888	0.971	0.868	0.917	33350	0	5	12	10	3	10	143	12	0.935	
	C5	0.907	0.920	0.913	0.906	0.916	0.911	33337	0	18	8	5	3	14	148	12	0.967	
	C6	0.921	0.907	0.914	0.951	0.926	0.939	33346	0	9	7	5	2	15	148	13	0.967	
	C7	0.983	1.000	0.992	0.922	0.926	0.924	33340	4	11	7	5	2	15	148	13	0.967	
	C8	0.944	0.846	0.892	0.970	0.858	0.911	33350	0	5	15	9	3	7	144	12	0.941	
LRRpredictor	0.923	0.912	0.917	0.967	0.916	0.941	33349	0	6	8	6	2	14	147	13	0.961		

Table A1. Cont.

Dataset	Classifier	In-Sample			Out-Of Sample												Recall on Core(L) Only
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	TN	FP		FN			TP			
								-	Non-LRR Proteins	LRR Proteins	N-Entry (N)	Core (L)	C-Exit (C)	N-Entry (N)	Core (L)	C-Exit (C)	
Test	C1	0.900	0.997	0.946	0.874	0.880	0.877	35107	0	19	6	10	2	13	109	10	0.916
	C2	0.962	0.956	0.959	0.941	0.847	0.891	35118	0	8	8	12	3	11	107	9	0.899
	C3	0.882	0.896	0.889	0.852	0.847	0.850	35104	1	21	8	13	2	11	106	10	0.891
	C4	0.940	0.874	0.906	0.907	0.840	0.872	35113	1	12	8	14	2	11	105	10	0.882
	C5	0.895	0.934	0.914	0.862	0.873	0.868	35105	1	20	7	10	2	12	109	10	0.916
	C6	0.940	0.901	0.920	0.928	0.860	0.893	35116	0	10	7	12	2	12	107	10	0.899
	C7	0.990	1.000	0.995	0.827	0.827	0.827	35100	4	22	7	16	3	12	103	9	0.866
	C8	0.942	0.874	0.906	0.920	0.840	0.878	35115	0	11	9	13	2	10	106	10	0.891
	LRRpredictor	0.943	0.928	0.936	0.928	0.860	0.893	35116	0	10	7	12	2	12	107	10	0.899

References

1. Enkhbayar, P.; Kamiya, M.; Osaki, M.; Matsumoto, T.; Matsushima, N. Structural Principles of Leucine-Rich Repeat (LRR) Proteins. *Proteins Struct. Funct. Bioinform.* **2004**, *54*, 394–403. [[CrossRef](#)] [[PubMed](#)]
2. Warren, R.F.; Henk, A.; Mowery, P.; Holub, E.; Innes, R.W. A mutation within the leucine-rich repeat domain of the arabidopsis disease resistance gene RPS5 partially suppresses multiple bacterial and downy mildew resistance genes. *Plant Cell* **1998**, *10*, 1439–1452. [[CrossRef](#)] [[PubMed](#)]
3. Jia, Y.; McAdams, S.A.; Bryan, G.T.; Hershey, H.P.; Valent, B. Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *EMBO J.* **2000**, *19*, 4004–4014. [[CrossRef](#)] [[PubMed](#)]
4. Moghaddas, F.; Zeng, P.; Zhang, Y.; Schützle, H.; Brenner, S.; Hofmann, S.R.; Berner, R.; Zhao, Y.; Lu, B.; Chen, X.; et al. Autoinflammatory mutation in NLRC4 reveals a leucine-rich repeat (LRR)–LRR oligomerization interface. *J. Allergy Clin. Immunol.* **2018**, *142*, 1956–1967.e6. [[CrossRef](#)] [[PubMed](#)]
5. Paisán-Ruiz, C.; Jain, S.; Evans, E.W.; Gilks, W.P.; Simón, J.; Van Der Brug, M.; De Munain, A.L.; Aparicio, S.; Gil, A.M.; Khan, N.; et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson’s disease. *Neuron* **2004**, *44*, 595–600. [[CrossRef](#)] [[PubMed](#)]
6. Zimprich, A.; Biskup, S.; Leitner, P.; Lichtner, P.; Farrer, M.; Lincoln, S.; Kachergus, J.; Hulihan, M.; Uitti, R.J.; Calne, D.B.; et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **2004**, *44*, 601–607. [[CrossRef](#)] [[PubMed](#)]
7. Ni, G.X.; Li, Z.; Zhou, Y.Z. The role of small leucine-rich proteoglycans in osteoarthritis pathogenesis. *Osteoarthritis Cartilage.* **2014**, *22*, 896–903. [[CrossRef](#)]
8. Lewis, M. PRELP, collagen, and a theory of Hutchinson-Gilford progeria. *Ageing Res. Rev.* **2003**, *2*, 95–105. [[CrossRef](#)]
9. Sukarta, O.C.A.; Sloopweg, E.J.; Goverse, A. Structure-informed insights for NLR functioning in plant immunity. *Semin. Cell Dev. Biol.* **2016**, *56*, 134–149. [[CrossRef](#)]
10. Urbach, J.M.; Ausubel, F.M. The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 1063–1068. [[CrossRef](#)]
11. Matsushima, N.; Miyashita, H.; Enkhbayar, P.; Kretsinger, R.H. Comparative Geometrical Analysis of Leucine-Rich Repeat Structures in the Nod-Like and Toll-Like Receptors in Vertebrate Innate Immunity. *Biomolecules* **2015**, *5*, 1955–1978. [[CrossRef](#)] [[PubMed](#)]
12. Kajava, A.V.; Kobe, B. Assessment of the ability to model proteins with leucine-rich repeats in light of the latest structural information. *Protein Sci.* **2002**, *11*, 1082–1090. [[CrossRef](#)]
13. Matsushima, N.; Tanaka, T.; Enkhbayar, P.; Mikami, T.; Taga, M.; Yamada, K.; Kuroki, Y. Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. *BMC Genom.* **2007**, *8*, 124. [[CrossRef](#)] [[PubMed](#)]
14. Kobe, B.; Kajava, A.V. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* **2001**, *11*, 725–732. [[CrossRef](#)]
15. Ng, A.C.Y.; Eisenberg, J.M.; Heath, R.J.W.; Huett, A.; Robinson, C.M.; Nau, G.J.; Xavier, R.J. Human leucine-rich repeat proteins: A genome-wide bioinformatic categorization and functional analysis in innate immunity. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 4631–4638. [[CrossRef](#)] [[PubMed](#)]
16. Sela, H.; Spiridon, L.N.; Petrescu, A.J.; Akerman, M.; Mandel-Gutfreund, Y.; Nevo, E.; Loutre, C.; Keller, B.; Schulman, A.H.; Fahima, T. 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.* **2012**, *13*, 276–287. [[CrossRef](#)]
17. Sloopweg, E.J.; Spiridon, L.N.; Roosien, J.; Butterbach, P.; Pomp, R.; Westerhof, L.; Wilbers, R.; Bakker, E.; Bakker, J.; Petrescu, A.J.; et al. 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.* **2013**, *162*, 1510–1528. [[CrossRef](#)]
18. Sela, H.; Spiridon, L.N.; Ashkenazi, H.; Bhullar, N.K.; Brunner, S.; Petrescu, A.-J.; Fahima, T.; Keller, B.; Jordan, T. 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.* **2014**, *27*, 835–845. [[CrossRef](#)]

19. Rajaraman, J.; Douchkov, D.; Hensel, G.; Stefanato, F.L.; Gordon, A.; Ereful, N.; Caldararu, O.F.; Petrescu, A.J.; Kumlehn, J.; Boyd, L.A.; et al. An LRR/Malectin receptor-like kinase mediates resistance to non-adapted and adapted powdery mildew fungi in barley and wheat. *Front. Plant Sci.* **2016**, *7*, 1836. [[CrossRef](#)]
20. Baudin, M.; Schreiber, K.J.; Martin, E.C.; Petrescu, A.J.; Lewis, J.D. Structure–function analysis of ZAR1 immune receptor reveals key molecular interactions for activity. *Plant J* **2020**, *101*, 352–370. [[CrossRef](#)]
21. Wang, J.; Hu, M.; Wang, J.; Qi, J.; Han, Z.; Wang, G.; Qi, Y.; Wang, H.-W.; Zhou, J.-M.; Chai, J. Reconstitution and structure of a plant NLR resistosome conferring immunity. *Science* **2019**, *364*, eaav5870. [[CrossRef](#)]
22. Wang, J.; Wang, J.; Hu, M.; Wu, S.; Qi, J.; Wang, G.; Han, Z.; Qi, Y.; Gao, N.; Wang, H.W.; et al. Ligand-triggered allosteric ADP release primes a plant NLR complex. *Science* **2019**, *364*, eaav5868. [[CrossRef](#)]
23. Kajava, A.V.; Vassart, G.; Wodak, S.J. Modeling of the three-dimensional structure of proteins with the typical leucine-rich repeats. *Structure* **1995**, *3*, 867–877. [[CrossRef](#)]
24. Helft, L.; Reddy, V.; Chen, X.; Koller, T.; Federici, L.; Fernández-Recio, J.; Gupta, R.; Bent, A. LRR conservation mapping to predict functional sites within protein leucine-rich repeat domains. *PLoS ONE* **2011**, *6*, e21614. [[CrossRef](#)]
25. Gao, Y.; Wang, W.; Zhang, T.; Gong, Z.; Zhao, H.; Han, G.Z. Out of Water: The Origin and Early Diversification of Plant R-Genes. *Plant Physiol.* **2018**, *177*, 82–89. [[CrossRef](#)]
26. Offord, V.; Werling, D. LRRfinder2.0: A webserver for the prediction of leucine-rich repeats. *Innate Immun.* **2013**, *19*, 398–402. [[CrossRef](#)]
27. Bej, A.; Sahoo, B.R.; Swain, B.; Basu, M.; Jayasankar, P.; Samanta, M. 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.* **2014**, *53*, 164–170. [[CrossRef](#)]
28. Dawson, N.L.; Lewis, T.E.; Das, S.; Lees, J.G.; Lee, D.; Ashford, P.; Orengo, C.A.; Sillitoe, I. CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **2017**, *45*, D289–D295. [[CrossRef](#)]
29. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2019**, *47*, D427–D432. [[CrossRef](#)]
30. Mitchell, A.L.; Attwood, T.K.; Babbitt, P.C.; Blum, M.; Bork, P.; Bridge, A.; Brown, S.D.; Chang, H.Y.; El-Gebali, S.; Fraser, M.I.; et al. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **2019**, *47*, D351–D360. [[CrossRef](#)]
31. Wang, G.; Dunbrack, R.L. PISCES: A protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589–1591. [[CrossRef](#)]
32. Capra, J.A.; Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **2007**, *23*, 1875–1882. [[CrossRef](#)]
33. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638. [[CrossRef](#)]
34. Wang, S.; Li, W.; Liu, S.; Xu, J. RaptorX-Property: A web server for protein structure property prediction. *Nucleic Acids Res.* **2016**, *44*, W430–W435. [[CrossRef](#)]
35. Wang, S.; Sun, S.; Xu, J. AUC-maximized deep convolutional neural fields for protein sequence labeling. In *Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin, Germany, 2016; Volume 9852 LNAI, pp. 1–16.
36. Wang, S.; Ma, J.; Xu, J. AUCpreD: Proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. In *Proceedings of the Bioinformatics*; Oxford University Press: Oxford, UK, 2016; Volume 32, pp. i672–i679.
37. Wang, S.; Peng, J.; Ma, J.; Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **2016**, *6*, 18962. [[CrossRef](#)]
38. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173–175. [[CrossRef](#)]
39. Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S.J.; Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **2019**, *20*, 473. [[CrossRef](#)]
40. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
41. Pal, S.K.; Mitra, S. Multilayer Perceptron, Fuzzy Sets, and Classification. *IEEE Trans. Neural Netw.* **1992**, *3*, 683–697. [[CrossRef](#)]

42. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [[CrossRef](#)]
43. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [[CrossRef](#)]
44. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks*; IEEE: Piscataway, NJ, USA, 2008; pp. 1322–1328.
45. Han, H.; Wang, W.-Y.; Mao, B.-H. *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3644.
46. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
47. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline Over-sampling for Imbalanced Data Classification. In *Proceedings: Fifth International Workshop on Computational Intelligence & Applications*; IEEE: Piscataway, NJ, USA, 2009; Volume 2009, pp. 24–29.
48. Batista, G.E.; Bazzan, A.L.C.; Monard, M.C. *Balancing Training Data for Automated Annotation of Keywords: A Case Study*. In *WOB; UFRGS: Porto Alegre, Brazil, 2003*; pp. 10–18.
49. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [[CrossRef](#)]
50. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
51. Liu, D.C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503–528. [[CrossRef](#)]
52. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings*; International Conference on Learning Representations; ICLR: San Diego, CA, USA, 2015.
53. Robbins, H.; Monro, S. A Stochastic Approximation Method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [[CrossRef](#)]
54. Hahnloser, R.H.R.; Sarpeshkar, R.; Mahowald, M.A.; Douglas, R.J.; Seung, H.S. Digital selection and analogue amplification coexist in a cortex- inspired silicon circuit. *Nature* **2000**, *405*, 947–951. [[CrossRef](#)]
55. Lemaitre, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
56. Käll, L.; Krogh, A.; Sonnhammer, E.L.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **2004**, *338*, 1027–1036. [[CrossRef](#)]
57. Halperin, E.; Buhler, J.; Karp, R.; Krauthgamer, R.; Westover, B. Detecting protein sequence conservation via metric embeddings. *Bioinformatics* **2003**, *19*, 122–129. [[CrossRef](#)]
58. Govindarajan, R.; Leela, B.C.; Nair, A.S. RBLOSUM performs better than CorBLOSUM with lesser error per query. *BMC Res. Notes* **2018**, *11*, 328. [[CrossRef](#)] [[PubMed](#)]
59. Styczynski, M.P.; Jensen, K.L.; Rigoutsos, I.; Stephanopoulos, G. BLOSUM62 miscalculations improve search performance. *Nat. Biotechnol.* **2008**, *26*, 274–275. [[CrossRef](#)] [[PubMed](#)]
60. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [[CrossRef](#)]
61. Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 1–27. [[CrossRef](#)]
62. Crooks, G.E.; Hon, G.; Chandonia, J.-M.; Brenner, S.E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14*, 1188–1190. [[CrossRef](#)]
63. Schrödinger, L.L.C. The PyMOL Molecular Graphics System, Version 2.2.3. Available online: <https://pymol.org/2/> (accessed on 10 November 2019).
64. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 99–104. [[CrossRef](#)]
65. Kajava, A.V. Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.* **1998**, *277*, 519–527. [[CrossRef](#)]
66. Borg, I.; Groenen, P.J.F. *Modern Multidimensional Scaling*, 2nd ed.; Chapter: MDS Models and Measures of Fit; Springer: Berlin, Germany, 2005; pp. 37–61.
67. Ward, C.W.; Garrett, T.P.J. The relationship between the L1 and L2 domains of the insulin and epidermal growth factor receptors and leucine-rich repeat modules. *BMC Bioinform.* **2001**, *2*, 4. [[CrossRef](#)]

68. Padmanabhan, M.; Cournoyer, P.; Dinesh-Kumar, S.P. The leucine-rich repeat domain in plant innate immunity: A wealth of possibilities. *Cell. Microbiol.* **2009**, *11*, 191–198. [[CrossRef](#)]
69. Dodds, P.N.; Rathjen, J.P. Plant immunity: Towards an integrated view of plant–pathogen interactions. *Nat. Rev. Genet.* **2010**, *11*, 539–548. [[CrossRef](#)]
70. Ravensdale, M.; Nemri, A.; Thrall, P.H.; Ellis, J.G.; Dodds, P.N. Co-evolutionary interactions between host resistance and pathogen effector genes in flax rust disease. *Mol. Plant Pathol.* **2011**, *12*, 93–102. [[CrossRef](#)]
71. Franchi, L.; Warner, N.; Viani, K.; Nuñez, G. Function of Nod-like receptors in microbial recognition and host defense. *Immunol. Rev.* **2009**, *227*, 106–128. [[CrossRef](#)]
72. Borrelli, G.M.; Mazzucotelli, E.; Marone, D.; Crosatti, C.; Michelotti, V.; Valè, G.; Mastrangelo, A.M. Regulation and evolution of NLR genes: A close interconnection for plant immunity. *Int. J. Mol. Sci.* **2018**, *19*, 1662. [[CrossRef](#)]
73. Leister, R.T.; Dahlbeck, D.; Day, B.; Li, Y.; Chesnokova, O.; Staskawicz, B.J. Molecular genetic evidence for the role of SGT1 in the intramolecular complementation of Bs2 protein activity in *Nicotiana benthamiana*. *Plant Cell* **2005**, *17*, 1268–1278. [[CrossRef](#)]
74. Lewis, J.D.; Lee, A.H.-Y.; Hassan, J.A.; Wan, J.; Hurley, B.; Jhingree, J.R.; Wang, P.W.; Lo, T.; Youn, J.-Y.; Guttman, D.S.; et al. The Arabidopsis ZED1 pseudokinase is required for ZAR1-mediated immunity induced by the *Pseudomonas syringae* type III effector HopZ1a. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 18722–18727. [[CrossRef](#)]
75. Mestre, P.; Baulcombe, D.C. Elicitor-mediated oligomerization of the tobacco N disease resistance protein. *Plant Cell* **2006**, *18*, 491–501. [[CrossRef](#)]
76. Huh, S.U.; Cevik, V.; Ding, P.; Duxbury, Z.; Ma, Y.; Tomlinson, L.; Sarris, P.F.; Jones, J.D.G. Protein-protein interactions in the RPS4/RRS1 immune receptor complex. *PLoS Pathog.* **2017**, *13*, e1006376. [[CrossRef](#)]
77. Hu, Z.; Yan, C.; Liu, P.; Huang, Z.; Ma, R.; Zhang, C.; Wang, R.; Zhang, Y.; Martinon, F.; Miao, D.; et al. Crystal structure of NLRC4 reveals its autoinhibition mechanism. *Science* **2013**, *341*, 172–175. [[CrossRef](#)]
78. Sun, Y.; Li, L.; Macho, A.P.; Han, Z.; Hu, Z.; Zipfel, C.; Zhou, J.-M.; Chai, J. Structural basis for flg22-induced activation of the Arabidopsis FLS2-BAK1 immune complex. *Science* **2013**, *342*, 624–628. [[CrossRef](#)]
79. Choe, J.; Kelker, M.S.; Wilson, I.A. Crystal structure of human toll-like receptor 3. *Science* **2005**, *22*, 581–585. [[CrossRef](#)]
80. Pickersgill, R.; Jenkins, J.; Harris, G.; Nasser, W.; Robert-Baudouy, J. The structure of Bacillus subtilis pectate lyase in complex with calcium. *Nat. Struct. Biol.* **1994**, *1*, 717–723. [[CrossRef](#)]
81. Olivier, N.B.; Imperiali, B. Crystal structure and catalytic mechanism of PglD from Campylobacter jejuni. *J. Biol. Chem.* **2008**, *283*, 27937–27946. [[CrossRef](#)]
82. Huber, A.H.; Nelson, W.J.; Weis, W.I. Three-dimensional structure of the armadillo repeat region of β -catenin. *Cell* **1997**, *90*, 871–882. [[CrossRef](#)]
83. Takahashi, N.; Hamada-Nakahara, S.; Itoh, Y.; Takemura, K.; Shimada, A.; Ueda, Y.; Kitamata, M.; Matsuoka, R.; Hanawa-Suetsugu, K.; Senju, Y.; et al. TRPV4 channel activity is modulated by direct interaction of the ankyrin domain to PI(4,5)P₂. *Nat. Commun.* **2014**, *5*, 4994. [[CrossRef](#)]

