
ROMANIAN ACADEMY
SCHOOL OF ADVANCED STUDIES OF THE ROMANIAN ACADEMY
INSTITUTE OF BIOCHEMISTRY

PH.D. THESIS SUMMARY

# Investigation of protein structure, dynamics and interaction via statistical, physical and mathematical methods

*Scientific coordinator:*
Prof. Dr. Andrei-José PETRESCU

*Ph.D. Student:*
Eliza Cristina MARTIN

2022

# Contents

# General Introduction

The technological developments in the field of genetic and transcriptomic sequencing alongside with those in structural biology have led over the past decade to an exponential growth of the collected biological data. This results in new horizons for the analysis of the complex biological environment and convoluted protein interaction networks and in understanding natural variations - with vital applications in a variety of biological and medical fields. However, transforming raw biological data into knowledge requires a mirrored effort in developing analysis workflows, bioinformatics platforms, mathematical models and prediction tools able to speed up the pace of discovery in modern biology.

In this larger context, the overall objective of this thesis was to develop bioinformatics and applied biocomputing tools and use them in computationally assisted experimental approaches in immunobiology aimed at providing a better understanding of the origin and evolution of RAG proteins the key molecular machinery of the adaptative immunity on one hand, and on the other of the sequence-structural interplay in the vast repertoire of plant NOD-like receptors in innate immunity.

The first part of the thesis presents the efforts in understanding the origin of the RAG (recombination-activating gene) apparatus which in jawed vertebrates is responsible for generating the extensive repertory of unique immune receptors in the B and T cells. The work presented is the result of a pluridisciplinary collaboration between (i) Professor Andrei-J. Petrescu, Head of the Department of Bioinformatics and Structural Biochemistry, IBAR, (ii) with Professor David G. Schatz, Chair of the Department of Immunobiology at Yale School of Medicine, US, member of the National Academy of Sciences and National Academy of Medicine and (iii) with Professor Pierre Pontaroti, Group of Evolutionary Biology, Aix-Marseille Université, CNRS SNC5039, France. This part describes our efforts in identifying new RAG-like genes in more remote branches than previously reported, extending our understanding of the origin of these genes and positing their arrival much

earlier than initially considered, in the early bilaterian clade.

The second part of the thesis focuses on the innate plant immune system, specifically on the intracellular NOD-like receptors (NLRs). The following chapter presents our computational work in generating probabilistic models of the 3D structure of ZAR1 - a broad-spectrum NLR receptor shared by most flowering plant branches. This work was part of a broader interdisciplinary collaboration with Professor Jennifer Lewis, Department of Plant & Microbial Biology, Berkeley University of California, US. The main focus of this joint project is to widen our structural understanding of the activation molecular mechanisms of ZAR1 receptors in recognising a wide variety of pathogen-related molecules, as such broad-spectrum receptors are of primary interest in devising crop pathogen control strategies. The last two chapters present the development of two software packages - LRRpredictor and NLRexpress - which employ machine-learning prediction tools aimed at identifying NLR-associated sequence signatures and are the result of the collaboration with Professor Aska Goverse, Laboratory of Nematology, Wageningen University and Research, the Netherlands. The LRRpredictor software package was designed to address the high pattern irregularity of LRR motifs in plant NLRs and to provide a significantly better detection performance compared to previously existing methods. By utilising a collection of machine learning estimators employing sequence and structural related information, this tool is aimed at assisting structural modelling and molecular biology research involving LRR domains. The last chapter presents NLRexpress software package - a three-module prediction workflow comprising eleven neural network-based estimators aimed at identifying key structural and functional motifs of the NLR constituent domains - CC, NBS and LRR - which is designed for fast computations for screening large sequence databases such as the entire proteome of a specie.

# Chapter 1

# New findings on the origins of the RAG machinery

## 1.1 Introduction & Background

The adaptive immune system, specific only to jawed vertebrates embodies an essential and powerful evolutionary breakthrough, which is believed that provided jawed vertebrates with a major evolutionary advantage (Litman et al., 2010). This extraordinary feature consists of generating a vast repertoire of antigen receptor genes by recombination reactions during lymphocyte development. The V(D)J recombination mechanism is performed by the RAG machinery by operating on the variable, diversity, and joining genes, generating an enormous set of possible antigen genes(Schatz and Swanson, 2011). The origin of the jawed-vertebrate RAG recombinase machinery has been a subject of debate in the last two decades. Given the similarities displayed by the RAG1 catalytic core to DDE transposase, it has been initially assumed that the RAG genes have derived from a class II transposable element. The initially identified element most similar to RAG1 was the Transib transposon (Kapitonov and Jurka, 2005) and in the last 5 years, the discovery of multiple RAG-like (RAGL) genes in invertebrate deuterostomes - several possessing full transposon layout have been reinforcing this hypothesis (Fugmann et al., 2006; Huang et al., 2016; Morales Poole et al., 2017; Zhang et al., 2019).

*In vitro* experiments on cephalochordate *B.belcheri* RAG1&2-like proteins showed their TIR-dependent endonuclease and transposase activity, making it the first known to be

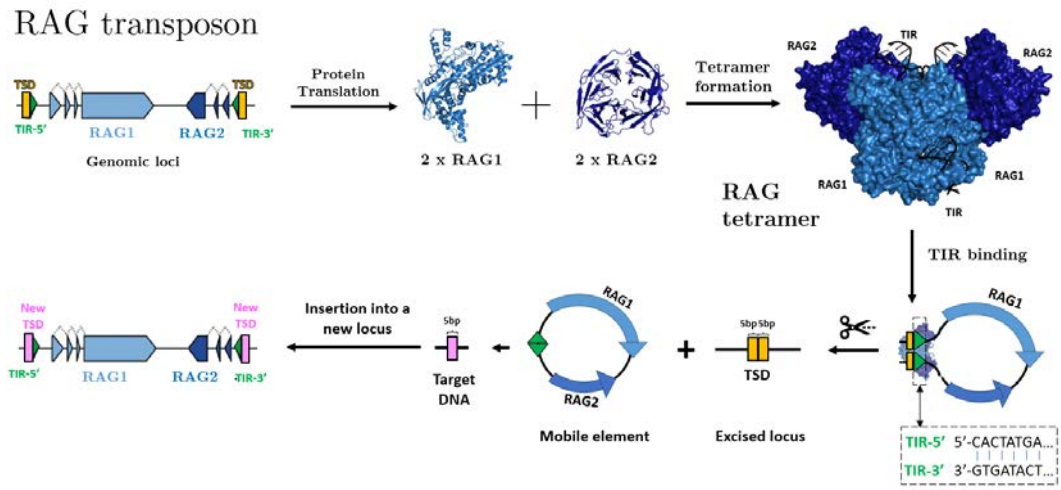



**Figure 1.1:** The amphioxus RAG transposase activity diagram

active *Proto* RAG transposon (Huang et al., 2016). Furthermore, cryo-EM structures of the amphioxus RAG-like machinery in different stages of transposition showed striking similarities with the vertebrate RAG complex cleavage mechanisms, despite their low homology at the protein sequence level (Zhang et al., 2019). Briefly, following the translation of the two RAG1 and RAG2 encoding genes, the RAG tetramer complex is formed and is composed of two RAG1-RAG2 heterodimers (Figure 1.1). The complex recognizes the heptameric region of the TIR margins and brings together the 5' and 3' TIR segments, bending the DNA and forming a circular-like structure (Zhang et al., 2019). The endonuclease complex nicks the DNA at the beginning of the TIR margins, excising the transposon DNA cassette from the genomic locus (Figure 1.1), which will be rejoined by the host repair enzymes. The mobile transposon element is inserted into a new genomic locus, a process in which a new target site duplication (TSD) of 5bp is generated, corresponding to a new insertion locus (Zhang et al., 2019).

Prior to 2019-2020, when this study was conducted, there has been no evidence for RAG transposon activity outside the deuterostome phylum, whereas the Transib transposon has been identified broadly from deuterostomes to the fungi kingdom (Kapitonov and Jurka, 2005; Kapitonov and Koonin, 2015). This chapter presents the identification of several RAG1L-RAG2L gene pairs in the *Protostomia* superphylum inside *Mollusca* and *Nemertea* clades, of which several exhibits the full transposon organisation, with markers of transposase activity (Martin et al., 2020b), preserved CDS that could encode full-length protein products with the complete domain layout. Moreover, less preserved copies were

identified in the *Cnidaria* phylum, outside the *Bilateria* clade, altogether indicating that the RAGL transposon was active outside the deuterostome clade as initially presumed and that it might have arisen in the early metazoan evolution.

## 1.2 Results and Discussions

In order to screen the available genomic and transcriptomic public databases, a pool of previously documented RAG1 and RAG2 sequences was used. As the sequence homology between the vertebrate and invertebrate RAG2 is very low, beyond the detection threshold of blast-based methods, an iterative screening approach was conducted which allowed the detection of novel RAG1/2-like genes in the *Protostomia* and *Cnidaria* clades (Figure 1.2).

In *Protostomia* phylum, RAG1-RAG2 gene pairs were identified in the *Mollusca* clade in oysters (*Crassostrea virginica, Crassostrea gigas, Saccostrea glomerata*), in mussels (*Modiolus philippinarum, Bathymodiolus platifrons*) and in the pearl oyster clade (*Pinctada imbricata*)(Martin et al., 2020b). In *Nemertea* phylum, at the time of the analysis, the only species with available genomic and/or transcriptomic data was the ribbon worm (*Notospermus geniculatus*), where numerous RAG1-RAG2 pairs were identified some of which were supported by mRNA transcriptomic data. In Cnidaria, RAG gene pairs were identified in *Porites rus, Orbicella faveolata* and *Aurelia aurita* (Martin et al., 2020b). Contrasting to protostomes where multiple gene copies were identified in each species, in *Cnidaria* only the *A.aurita* jellyfish exhibits an intact RAG pair was detected, whereas the other identified loci display signs of pseudogenisation.

Specific to the class II DDE transposons is the presence of TIR elements at the boundary of the mobile element. The compatibility region between the TIR segments often spans only at the margins (8-10bp), making the detection of these elements more challenging. In order to discriminate the transposon ends from other such inverted repeats which are frequently found in the genome, a homology variation approach was employed, based on the expectation that different duplicates of the transposons are expected to share a high degree of sequence homology, while the flanking regions should display no homology as these correspond to distinct insertion loci within the genome (Martin et al., 2020b). The presence of TSD duplications flanking the TIR was used as an additional discriminatory

constraint to distinguish between the transposons TIR and premature ends of the cassettes.

Several identified RAG1-RAG2 gene pairs displaying a complete transposon configuration TSD-TIR-RAG1-RAG2-TIR-TSD were found in *C.virginica*, *P.imbricata* and *N.geniculatus*. All the identified TIRs display a heptamer RSS-like region at the beginning of the segment, with perfectly conserved first 3 nucleotides "CAC" - essential for both transposase and recombinase functionality. Similarly to the previously reported TIR elements in deuterostome, the protostome elements do not display a nonamer-like region. Moreover, the length of 5bp of the identified TSD elements in the protostome is consistent with those found in Transib, deuterostome and jawed-vertebrates RAG (Kapitonov and Jurka, 2005; Morales Poole et al., 2017; Zhang et al., 2019).

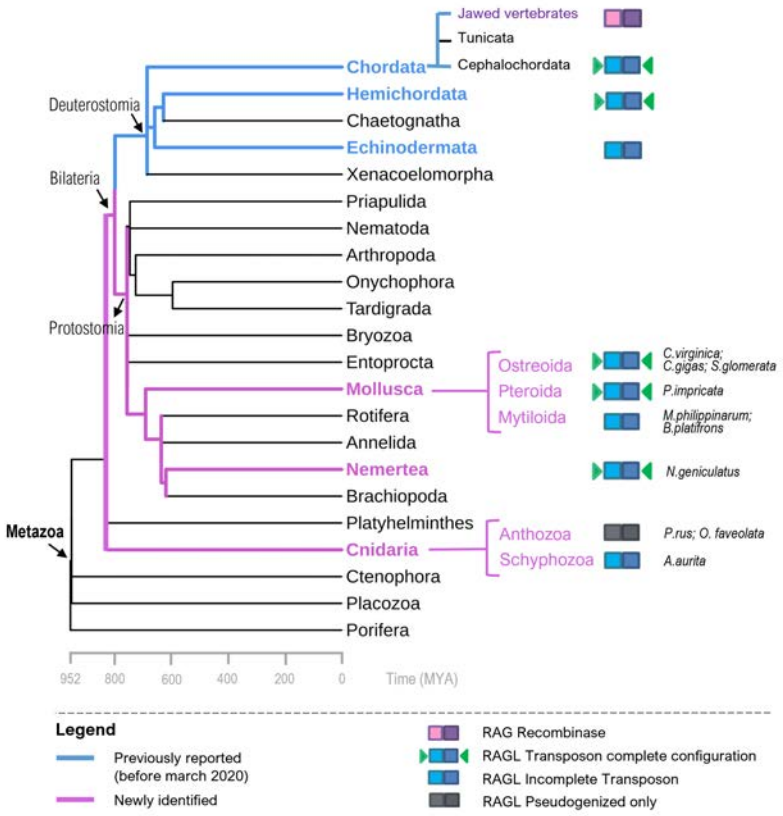



**Figure 1.2:** RAGL distribution across phyla. Blue branches describe the previously reported findings prior to March 2020, while clades identified in this study are shown in magenta. The status of the most preserved loci is illustrated with pictograms.

The phylogeny analysis was performed on the RAG1 catalytic core region of the most preserved identified representatives. The RAG1 tree follows the species phylogeny, indicating vertical evolution within the two bilaterian clades - *Protostomia* and *Deuterostomia* - consistent with the previously reported analyses (Morales Poole et al.,

2017). The updated RAG family partitioning consists in: **(a) RAG-A family** - the closest to the vertebrate RAG; **(b) RAG-B family** - widespread in the deuterostome clade which contains previously reported RAGs in deuterosomes and many of the identified transposons in protostome and cnidarian *A.aurita*; **(c) RAG-C family** - with a single reported member in the hemichordate *P. flava* (Morales Poole et al., 2017); **(d) RAG-D family** - a novel distinct family of RAGs, found only in the nemertean *N.geniculatus*.

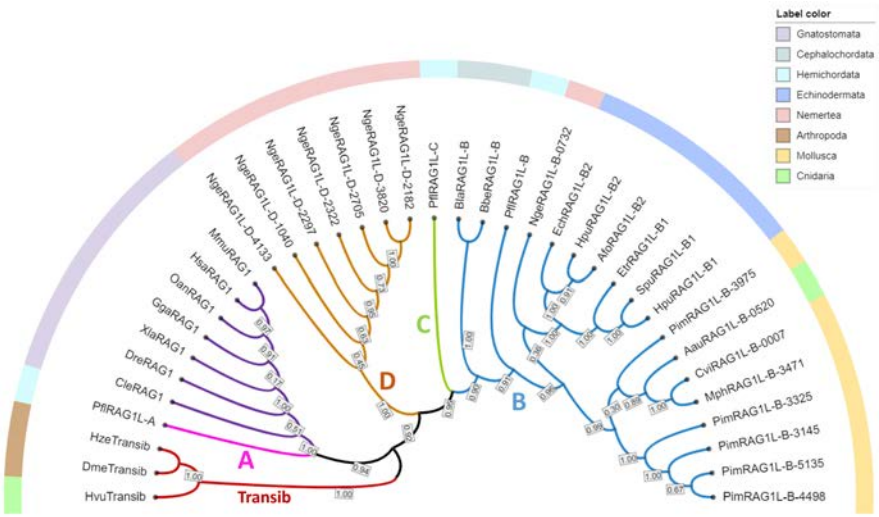
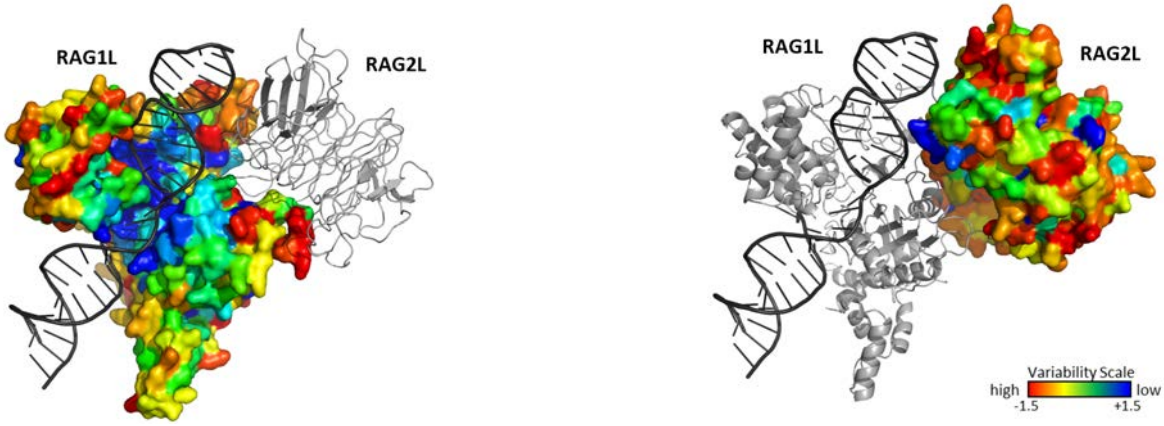



**Figure 1.3:** The updated RAGL phylogeny tree indicates four main RAGL families. The bootstrap branch support is computed using the Maximum Likelihood approach on the catalytic core of RAG1.

The identified RAG1 homologs exhibit preponderantly, with very few exceptions, the complete domain arrangement specific to deuterostome RAG1 transposases and vertebrate RAG1 recombinase. The key positions in the Dimerization and DNA binding domain (DDBD) and the Catalytic RNase H domain (RNH) are unanimously conserved in the identified copies, as well as the Zn-finger motif (C830, C833, H1035, H1040) within the ZnC2 and ZnH2 domains. The identified protostome representatives display a C-terminus tail highly homologous with deuterostome RAG1L containing the cysteine-rich *C**C***GH****C* pattern. All analyzed protostome RAG2L contain a Kelch-like domain followed by a PHD domain. Similarly to the previously reported invertebrate RAG2s, these lack the acidic long hinge connecting the two domains found in vertebrate RAG2. The RAG1L - DNA contact sites, both within the TIR heptamer span, but also in the flanking regions are majorly conserved with respect to the amphioxus RAG1, especially around the catalytic acidic triad, suggesting that protostome RAG1s might exhibit a similar behaviour regarding DNA binding and cleaving as observed in *B.belcheri* (Figure 1.4). On the other

hand, the RAG1L-RAG2L interaction wide and complex surface is poorly conserved when compared to *B.belcheri* RAG, suggesting that the two proteins coevolved together.




(a) RAG1 variability (b) RAG2 variability

**Figure 1.4:** (a) RAG1 and (b) RAG2 variability mapped on the amphyoxus cryo-EM structure (color-code: from red-variable to blue-conserved).

## 1.3 Conclusions

The findings presented herein provide evidence that intact RAG transposons exist in various protostome genomes, several of which are also supported by transcriptomic data. Alongside the fact that they display intact TIR and TSD pairs, complete domain organisation and conservation of the key functional residues, such RAG copies might potentially be currently active in their host organisms. Outside the bilaterian clade, only a few incomplete RAG-like pairs were identified in the *Cnidaria* phylum in various pseudogenisation stages, with only a single gene pair in the *A.aurita* displaying complete protein domain configuration. However, given the scarcity of the available genomic and transcriptomic sequenced data in this taxonomic clade, the current status of RAG transposons cannot be further assessed. Nevertheless, the presence of RAG remnants in *Cnidaria* indicates that the RAG transposon might be older than initially presumed. The phylogeny analysis presented herein is consistent with a vertical evolution trajectory of RAG inside the protostome and deuterostome clades, these could indicate potential stages of domestication of the RAG genes in their host organisms. Further studies of such RAG domestication candidates could be of interest, both for bringing new insights into the vertebrate RAG recombinase domestication phenomenon and also for investigating potentially novel biological functions of RAG in these organisms.

# Chapter 2

# The structure of ZAR1: from in-silico 3D model to experimental validation

## 2.1 Introduction and background

This chapter presents in detail the *in-silico* structural study of ZAR1 NLR receptor from *Arabidopsis thaliana*, which started at the beginning of the doctoral programme in November 2016 and was part of a wider research project in collaboration with Professor Jennifer Lewis, Department of Plant & Microbial Biology, Berkeley University of California. The study aimed at providing a better understanding of the ZAR1 structural determinants in the inter-domain interaction transitions during the activation mechanism.

The ZAR1 receptor is able to mediate the detection of a variety of pathogen-related proteins and effectors via an assortment of adaptor kinases (Lewis et al., 2013; Bastedo et al., 2019), such broad-spectrum NLRs being of great interest for developing pathogen control strategies. Preceding research studies of the group identified a host kinase ZED1 that was required in eliciting an immune response to Hopz1a effector from *Pseudomonas syringae* (Lewis et al., 2008, 2010, 2013), as well as several point mutations and experimental truncations of the ZAR1 sequence that inflict phenotypic changes in the inter-domain interaction profile and/or impact the immune HR response (Baudin et al., 2017).

Developing 3D models of ZAR1 to assist the molecular biology experiments was helpful in formulating hypotheses regarding the inter-domain interactions and putative rationales of

the experimental observations and in further proposing new interventions to experimentally test these hypotheses and provide new data to improve the probabilistic models. In 2019, the cryo-EM structure of ZAR1 was reported and provided us with the opportunity to compare our probabilistic 3D models to the *real* structure, which turned out to be in good agreement in a range of 2-5 deviation when overlapping the domains 3D structures, highlighting the practicality and effectiveness of using probabilistic models in the absence of experimentally-obtained 3D structures.

## 2.2 Results and Discussions

Generating 3D models of ZAR1 domains was challenging due to the extreme low homology with any available 3D structure. At that time no experimentally acquired full protein 3D structures of any plant NLR were available, but only: (i) three CC domain structures with controversial 3D folds discussed below - $\leq$19% identity with ZAR1, (ii) several NBS domains from the metazoan Apaf1 and Ced4 below 21% identity with ZAR1 and (iii) various plant LRR domains all originating from extracellular receptors with significant structural differences. In spite of the low homology with any templates known at that time, probabilistic 3D models of individual ZAR1 models were generated by employing different strategies and further optimised through Molecular Dynamics simulations.

The experimentally solved structures available at that time indicated that the CC domain could adopt two configurations: a 4-helical bundle (4H-CC) as in Rx and Sr33 structures (Hao et al., 2013; Casey et al., 2016) or as a 2-helical bundle (2H-CC) in the MLA10 dimer (Maekawa et al., 2011; Casey et al., 2016), which intriguingly, shares around 85% identity with Sr33. The structural analysis of the two structures revealed that two 4H-CC monomers of Sr33 overlay almost perfectly with the dimer of MLA10 (2 x 2H-CC) (Maekawa et al., 2011; Casey et al., 2016). A possible structural transition consistent with both Sr33 and MLA10 that we hypothesised at the beginning of the study, was that the peripheral first and fourth helical segments unfold from the 4-helical bundle and embrace the other monomer. Such a transition would require that the inter-helical loops to possess certain structural ambivalence, allowing to transit between helical and turn folding.

Based on the generated ZAR1 3D models, several structural inferred hypotheses were

experimentally tested by our collaborators by performing mutagenesis experiments and assessing phenotypic changes in the inter-domain interaction profile via yeast-two-hybrid (Y2H) and bimolecular fluorescence complementation (BiFC), as well as differences in the immune response *in planta*, described in detail in (Baudin et al., 2019). Mutations altering the rich electrostatic charged composition of the inter-helical loops resulted in the partial suppression of the CC dimerisation and the interaction with NBS and LRR domains and in a reduced HR response *in planta* (Baudin et al., 2019), indicating that these loop regions are involved in inter-domain interactions. To further study the involvement of the first helical segment, mutations that reduce the hydrophobicity of the first helix of the CC domain were proposed with the rationale that during activation the first helical segment is the least constraint and can potentially initiate the transition. Experiments indicated a reduced dimerization level and impaired CC-NBS interaction alongside complete suppression of the immune response *in planta*, while it did not impact the CC-LRR interaction, suggesting that the first helix of the CC domain is involved in interaction with NBS (Baudin et al., 2019). Introducing mutations on the conserved EDVID motif yielded impaired dimerisation, total alteration of the CC-NBS and CC-LRR interaction and complete suppression of the HR response, indicating that this region might be involved in both NBS and LRR interface, which was later confirmed by the ZAR1 cryo-EM structure (Baudin et al., 2019).

In 2019, cryo-EM structures of ZAR1 were reported in three stages of the activation mechanism: inactive monomeric ADP-binding state, activated monomeric state with absent nucleotide and oligomeric activated ATP-binding state (Wang et al., 2019b,a). This provided us with the opportunity to evaluate our probabilistic 3D models generated before the experimentally obtained structure. At the level of the CC domain, the cryo-EM structures reveal drastic conformational changes - from a 4H-CC to a 3H-CC conformation during activation. In the 4H-CC conformation, only the first half of H4a is part of the 4-helical bundle while in our initial model, the entire H4 helix was modelled as part of the helical bundle, whereas the rest of the 4H-CC model is in good overall agreement, with RMSD values of 3.9 between the model and the cryo-EM structure (Figure 2.2). The proposed NBS model shows a high structural agreement both for the inactive and activated conformations (RMSD of 4.0 and 5.1 ). The LRR model also displays a good superposition at of 4.7 RMSD with the cryo-EM, in conformity with the overall shape, curvature pitch and radius (Figure 2.2). Moreover, the model correctly represented the structural environment

surrounding the catalytic positions involved in binding the ADP/ATP ligand.

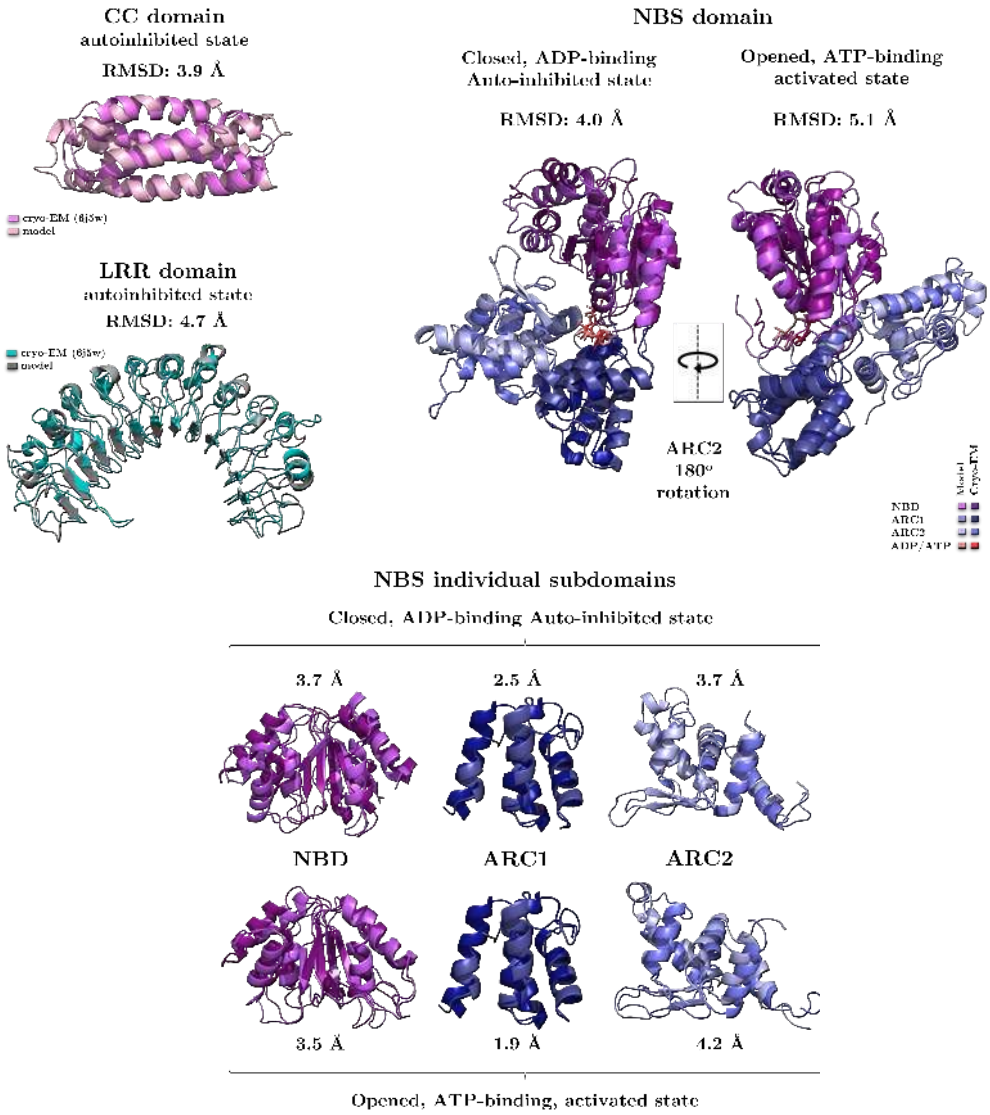



**Figure 2.2:** ZAR1 model vs cryo-EM in inactive/activated conformations (6j5w, 6j5t).

## 2.3 Conclusions

The chapter herein describes the structural analysis and generation of probabilistic 3D models for ZAR1 domains prior to the appearance of the cryoEM structure. This provided us with the opportunity to compare the initially proposed models with the experimental structure and also to scrutinise the hypotheses formulated based on the probabilistic model in the light of the new data. The model displays quite a good agreement with the 3D structure, highlighting the practicality and efficacy of employing computational analysis and probabilistic models in the absence of experimentally-driven 3D data.

# Chapter 3

# LRRpredictor - the challenge of irregular LRR pattern detection in plant NLRs

## 3.1 Introduction & Background

The leucine-rich repeat (LRR) architecture is central to the immune system, in pathogen detection and signal transduction upon recognition across the entire life tree, from *Archaea* to *Mammalia* (Enkhbayar et al., 2004). LRR domains adopt a 'horseshoe'-like solenoid 3D structure composed of spires generated by tandem LRR repeats of ∼15-30 amino acid length. These repeats are held together throughout a beta-sheet network formed on the inward side of the domain and characterised by the presence of a conserved sequence pattern, termed the LRR motif (Kajava and Kobe, 2002). The LRR motif consensus shows significant variations across protein classes and phyla, the most minimalist LRR pattern shared by all classes having 'LxxLxL' as a consensus ('L' - any hydrophobic amino acid, most frequently leucine; 'x' - any amino acid ). Furthermore, studies on plant NLRs domains showed a far more increased frequency of motif irregularities when compared to their metazoan counterparts or compared to plant extracellular receptors (Sela et al., 2014; Wang et al., 2019b).

Understanding the structural factors of the binding specificity of LRR domains unfolds the prospect of receptor engineering for pathogen control, with vast implications both in medicine and agricultural fields. The lack of sensitivity of current approaches in properly detecting LRR motifs by their amino acid sequence is a huge drawback in bioinformatic analysis, reliable 3D modelling and mapping between sequence/structural particularities and

biological behaviour. Difficulty in detecting individual motifs resides in the fact that the minimalistic motif is extremely trivial and such patterns are often expected to randomly occur in non-LRR proteins.

Presented in this chapter is the development of LRRpredictor - a new LRR pattern detection method composed of an ensemble of estimators that aim to bring increased versatility to pattern irregularities than existing methods by using resampling techniques. Further on, the performance and behaviour of the LRRpredictor are evaluated in comparison with the existing methods on a dataset of annotated domains from different classes (plant NLRs, RLKs and RLPs and animal NLRs and TLRs).

## 3.2 Results and Discussion

Available annotated domain databases were used to gather LRR domains with known 3D structure. After applying a redundancy filter at 90% identity, 178 protein chains (PDB-LRR-90) comprising ∼2100 LRR repeats were further used in analysis and subjected to LRR repeat delineation based on their structural data and beta-sheet network. A second redundancy filter of 50% identity was applied for obtaining the training set data, at the level of individual LRR repeats, obtaining a set of ∼850 highly divergent repeats (PDB-LRR-50).

The 3D superposition of different LRR repeats showed high topological resemblance extending both upstream and downstream the minimalist 6 aa long $L_0XXL_3XL_5$ motif with at least 5 residues in both directions (Figure 3.1b). Therefore, a 16 aa interval from position -5 to +10 around the $L_0$ position was further referred to as the 'extended' motif. An important facet that limits the analysis is the extremely high taxonomic bias of the available 3D data on LRR proteins in comparison with the equivalent baseline distribution of sequence databases such as Uniref-50. Approximately half of the LRR repeats within PDB-LRR-50 originate from mammalian species, while in UniRef-50 mammalian proteins share is less than 3% in all proteins as well as in annotated LRR domains (Figure 3.1d). By contrast, plant R proteins are extremely poorly represented, with a single experimental structure (Wang et al., 2019b,a) reported prior to 2020, while the majority of plant LRR repeats belong to extracellular RLP and RLK proteins.

LRRpredictor was trained on a protein set comprising the curated collection of 850 highly

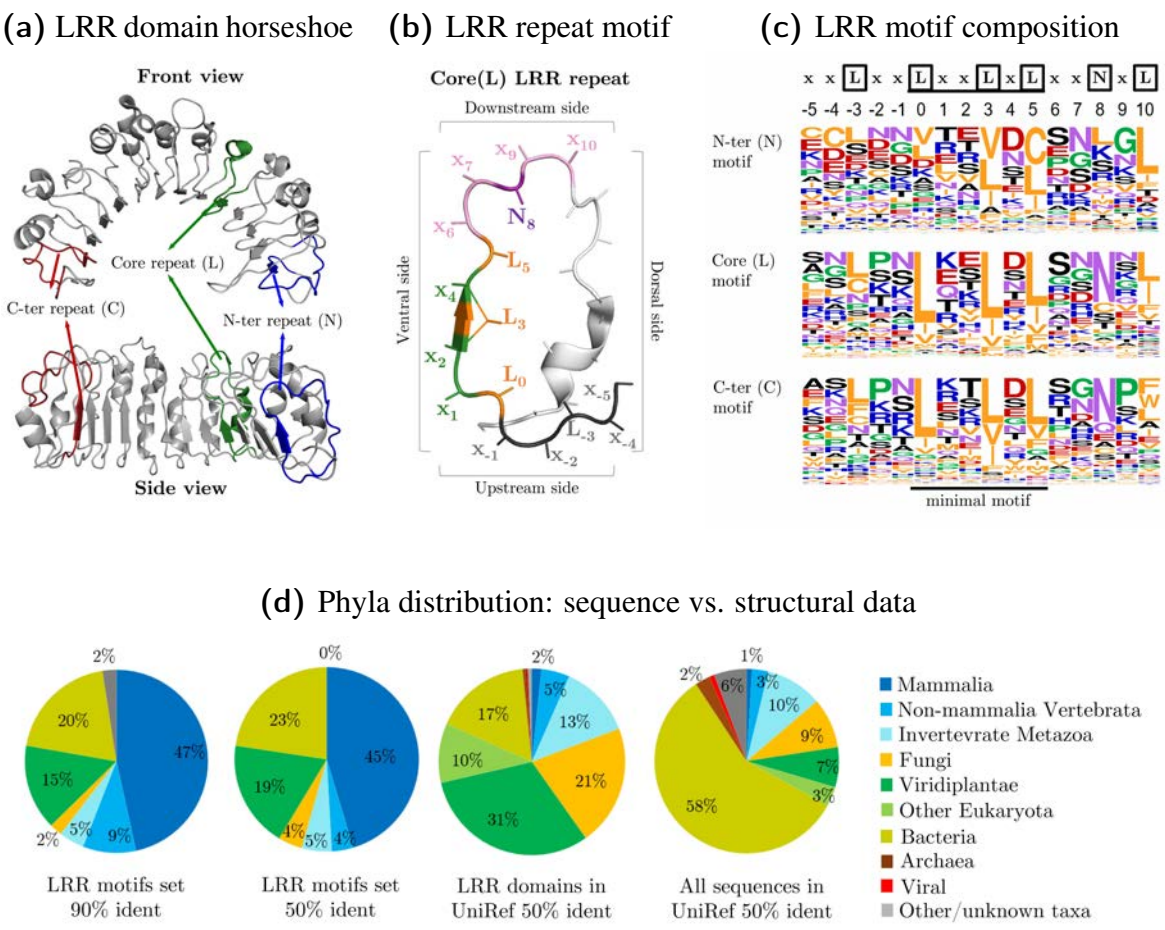



**Figure 3.1:** **(a)** LRR architecture exemplified on ZAR1 structure (PDB: 6j5w). **(b)** Zoom-in perspective of a LRR repeat. **(c)** Residue composition of the N-ter, core and C-ter motifs (PDB-LRR-50 set). **(d)** Taxonomic distribution of the LRR motif sets versus UniRef-50 sequence database. Figured derived from (Martin et al., 2020a).

diverse LRR motifs (at 50% identity) and a set of non-LRR proteins from each CATH 3D fold topology. To supplement the predictor with broader sequence-related context information, position-specific scoring matrices (PSSMs) profiles were used instead of the raw protein sequence. These profiles are derived from residue transition probabilities conditioned by the protein group they belong to and are expected to supply wider context information and underline the key conserved positions and relationships between residues. Besides sequence-related features, structural-based features were also explored - such as secondary structure, solvent accessibility and disorder intrinsic predictions. Given the low number of LRR repeat samples in the structural dataset, different artificial sample generation methods were employed.

The optimised final predictor - LRRpredictor - consists of an ensemble of eight classifiers

that employ different supervised learning techniques and class-imbalance treatments which are aggregated together via a soft voter approach. Half of the constituent estimators rely only on sequence-related information, whereas the other half uses both sequence and structural features. The training of the LRRpredictor was conducted using a 4-fold cross-validation approach on 80% of the dataset, and the out-of-sample tests were employed on the remaining 20% of the data.

Across the different cross-validation and test sets, the recall, precision and F1 scores of the ensemble predictor vary in the range of 85-97% for all LRR motif types and in the 89-98% range when only the core (L) motifs are considered. LRRpredictor, as an ensemble predictor, performs better overall when compared to the individual constituent estimators, both on the test set, but also on each cross-validation set. LRRpredictor outperforms other LRR motif predictors such as LRRsearch (Bej et al., 2014) and LRRfinder (Offord and Werling, 2013) (**??**). The LRRpredictor pipeline was tested on four classes of solenoid proteins - trimeric, pectate lyase, ankyrin and armadillo (50 seq/set) - which show the closest resemblance to the LRR fold. LRRpredictor is capable to distinguish between *true* LRR motifs and other LRR-like patterns, as no estimates over 50% LRR motif probability are obtained on all four sets.

Given the high taxonomic bias observed in the structural database on which the LRR predictor relies compared to the sequence database and more specific to the scarce structural data on plant LRR domains, next evaluated was the ability of LRRpredictor to extrapolate on different immune-related LRR-containing proteins. Datasets of the most representative immune-related protein classes that contain LRR domains were gathered: 4 sets from plant NLRs proteins (CNL and TNL) and extracellular LRR-containing receptors (RLK and RLP) and 2 sets from vertebrates - cytosolic NLRs and extracellular TLR - as described in the methods section. The identified LRR repeats using LRRpredictor display a good coverage of the LRR domain span annotated in the Interpro database (Mitchell et al., 2019) in all six datasets case. Approx. 75% of CNLs and 50% of TNLs LRR domains lack any repeat annotation in Interpro, whereas the extracellular receptors annotations cover between 30-80% of the LRR domain. The coverage attained by LRRpredictor is significantly higher with values over 90% coverage per LRR domain in two-thirds of each protein class data set. The analysis of the detected LRR motifs in each class, reveals apparent distinctions between the six immune receptor classes (Figure 3.2). While the minimum motif span - $L_0XXL_3XL_5$

- is invariant across sets, outside this region the CNL and TNL groups display increased variability, whereas the extracellular receptors display a prolonged pattern.

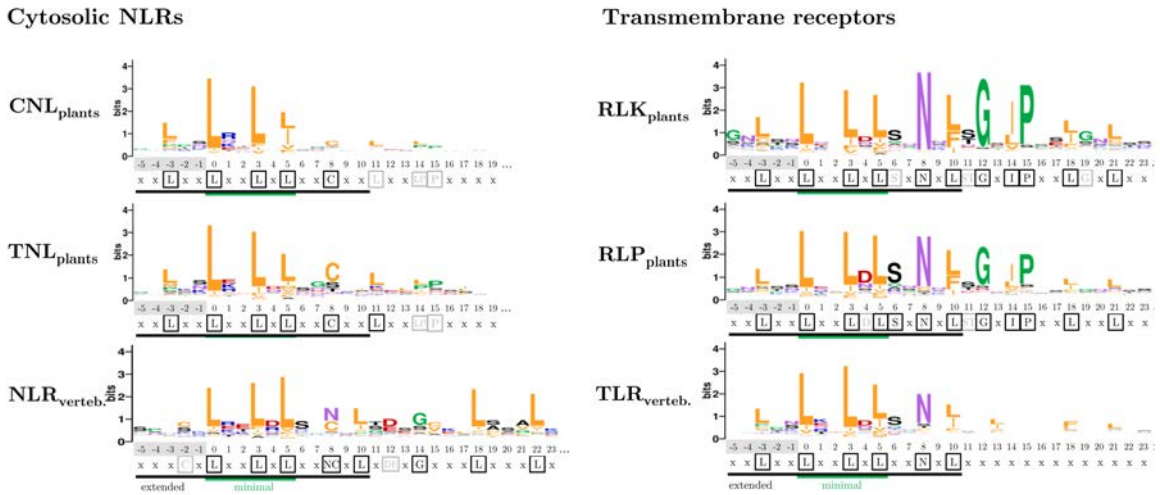



**Figure 3.2:** LRR motif consensuses across different immune-related protein classes. Sequence variability is expressed as relative entropy and depicted as letter height (higher means higher conservation). Figured derived from (Martin et al., 2020a).

## 3.3 Conclusion

The results presented in this chapter show that LRRpredictor displays a good performance on the structural data available at the moment. Moreover, the behaviour of the LRRpredictor on different immune-related LRR-containing protein classes indicates a good extrapolation ability, especially on the CNL and TNL resistance plant proteins, which are characterised by increased pattern irregularity and are poorly represented in the structural training dataset. The predictor is able to cover annotated LRR domain spans in the Interpro database with significantly higher coverage rates compared to other LRR repeat annotations methods. Moreover, the identified LRR motifs follow the previously reported consensus signatures of each investigated immune receptors classes.

In conclusion, LRRpredictor is a tool that aims to assist structural-informed research in understanding the sequence-structure-function interplay of individual immune receptors, which is essential in receptor engineering and pathogen detection control.

# Chapter 4

# NLRexpress - a collection of plant NLR motif predictors

## 4.1 Introduction & context

The previous chapter describes the development of LRRpredictor (Martin et al., 2020a) - an LRR motif predictor designed to address high motif irregularities such in the case of plant NLRs, which consists of a collection of eight individual classifiers employing different machine learning, artificial sampling strategies. The LRRpredictor classifiers compute variability PSSM profiles, built using the global Uniprot-20 protein databases and structural properties predictions - which both requires high computational resources, which makes LRRpredictor less feasible for screening large datasets. A fast tool able to scan entire organism proteomes or transcriptomes and annotate key functional motifs is valuable in comparative sequence analysis, discriminating complete NLR transcripts from ones lacking a specific motif, generating accurate 3D models and analysing changes in protein-protein interaction surfaces.

Next efforts were made to bypass these limitations and investigate light-weight neural network models able to reduce the execution time and computational resources required with minimum performance loss. Besides focusing on the *LxxLxL* motifs describing each repeat LRR element, the analysis was extended to include also sequence motifs found in the other NLR-specific domains such as the NBS and CC domains - as these conserved positions have vital roles such as ADP/ATP binding, inter-domain interaction or for the structural stability

of the 3D fold (Wang et al., 2019a; Ma et al., 2020).

This chapter presents NLRexpress - a collection of ML-based predictors designed to identify sequence motifs specific to resistance proteins in the CC, NBS and LRR domains and to be scalable to screening large data sets. The pipeline was used to scan a ∼34,300 plant NLRs and the detected motifs were clustered and analysed to identify inter-motif correlations using unsupervised learning techniques.

## 4.2 Results and Discussions

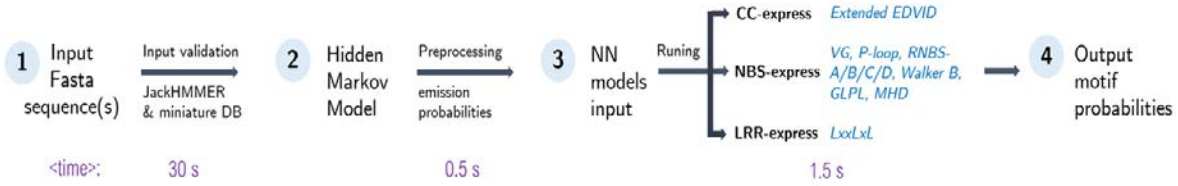
NLRexpress comprises a collection of 11 neural network classifiers trained each to detect individual sequence motifs specific to plant NLRs. NLRexpress is set up as 3 modular prediction units as follows: (i) CCexpress - extended EDVID motif; (ii) NBSexpress - *VG/hhGRE*, *P-loop / Walker-A*, *Walker-B*, *A/B/C/D-RNBS*, *GLP*L motifs and *MHD*; (iii) LRRexpress - LxxLxL pattern.

The NLR-express workflow is displayed in Figure 4.1 and begins with the user input protein sequence(s) in FASTA format. The first step consists in generating the input features required by the prediction models, including building the variability HMM profiles. The individual predictors part of each module is run independently, each of them returning as output the probability estimate of each position of the input sequence(s) to start the given motif.

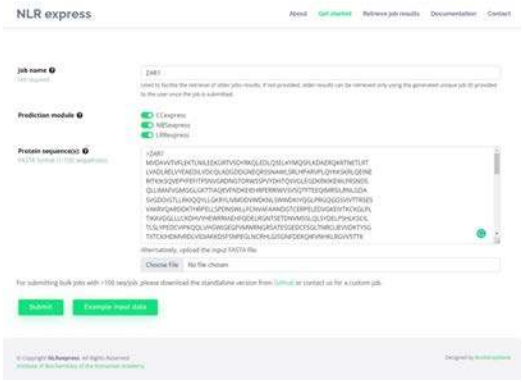
Many state-of-the-art prediction methods designed for protein sequence data rely on using features inferred from HMM models, which are computationally expensive due to the large size of the commonly used protein target databases such as Uniprot-20 or Uniclust-30. To drastically reduce the time spent during this stage, custom NLR-oriented miniaturized search databases were investigated in order to achieve the best trade-off between performance and execution time. The training of the eleven NN models corresponding to each individual motif was performed starting from a curated set of CC, NBS and LRR motifs from plant NLRs as described in detail in the thesis, using a 4-fold cross-validation schema for hyper-parameter optimisation.

In the case of the NBSexpress module, the eleven motifs show significantly higher conservation, which reflects in the performance of the individual predictors with precision and recall scores above 96%. The LRRexpress module yields balanced precision and

(a) NLRexpress workflow




(b) User input data form (screen capture)




(c) Job results page (screen capture)

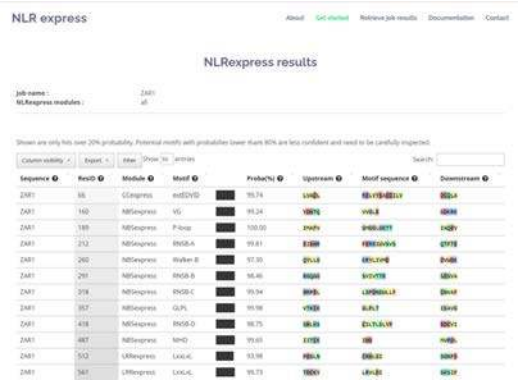



**Figure 4.1:** **(a)** NLRexpress workflow with average execution time of each stage. **(b)** NLRexpress webserver - screen captures (https://nlrexpress.biochim.ro).

sensitivity with F and G scores of 92% on the test set. Next, LRRexpress was run on the curated 3D structural set of ∼2000 LRR motifs trimmed at 90% identity, described in the previous chapter and in (Martin et al., 2020a). On this set, LRR-express attains an overall F1-score of approx. 92% when taking into account only the core LRR motifs alone, while ∼88% when including the more irregular marginal repeats.

A further auxiliary test was assessing the behaviour of LRRexpress on other non-LRR solenoid architectures containing the *LxxLxL* pattern which might provide a source of confusion/false positive predictions. For this, were used the five benchmark sets of 50 sequences each from protein classes - ankyrin, pectate lyases, trimeric and armadillo - previously described within the preceding chapter and in (Martin et al., 2020a). On these sets, LRRexpress is capable of correctly classifying the *L**L*L* patterns when occurring outside the LRR architecture context - with almost no false positives in sets containing each between 1000-2700 LRR-like LxxLxL patterns.

The NLR-express pipeline was further used on the larger set of 34314 plant NLRs trimmed at 90% identity. The predicted individual motifs were subjected to clustering either based on residue similarity metrices or based on the overall physiochemical properties

(hydrophobicity, volume and electrostatic charge).

Considering the proximity of the eleven NBS motifs in the 3D space - seven of them actively participating in the formation of the ADP/ATP binding pocket - the analysis of the individual regions taken separately would conceal relevant relationships shaping up at long distances in sequence. Therefore, the NBS motifs of the cleaned set of $\sim$20000 were extracted, concatenated and collectively clustered based on an amino acid similarity measure at various identity thresholds as described in the methods section. At a cutoff of 55% identity, around 85% of the sequences are conveyed in the top ten largest clusters. The most invariant motif areas are, as expected, the ones directly involved in ADP/ATP binding, notably within the P-loop, Walker-B, B-RNBS motif in the NBD subdomain, and GLPL and MHD motifs within the ARC1 and ARC2 subdomains, whereas cluster-specific characteristics are shaping up within the more diverse motifs: the hhGRE and the A/C/D-RNBS motifs.

As the LRR motif is the most variable among the other NLR-related motifs, the next analysed was how the motif variability distributes depending on the position in the LRR domain of NLRs. A set of excised LxxLxL pattern regions (approx. 65000) were subjected to clustering using unsupervised machine-learning approaches in an embedding describing their physio chemical properties (hydrophobicity, charge and size). A strong predilection for positively charged motifs is seen in all NLR classes within the first four repeats, the most frequently occurring LRR pattern being of type *LRxLxL*. Contrarily to CNLs, the first LRR repeat in TNLs shows a preference for an acidic environment in position 1 of the motif, while in RNLs, a strong preference for *LRxLxL* type is noticed in the case of the first and third LRR repeats.

## 4.3 Conclusion

The results presented within this chapter indicate that NLRexpress - a prediction pipeline gathering a collection of predictors designed to detect CC, NBS and LRR motifs specific to plant NLRs - displays a good performance on the benchmark tests employed herein and might be of use in assisting different types of NLR-related investigations. Besides applications in structural modelling, due to its computational speed improvements, it is feasible for large-scale sequence analysis, such as screening an entire proteome of a species or in comparative analyses of large sets of orthologs.

# List of personal contributions

## Publications

### Publication as primary author

1. Martin EC*, Vicari C,* Tsakou-Ngouafo L, Pontarotti P, Petrescu AJ, Schatz DG. *"Identification of RAG-like transposons in protostomes suggests their ancient bilaterian origin."* **Mobile DNA**. 11, 12 (2020). [PMID: 32399063]

   **IF**: 4.06; **AI**: 2.6; **Citations (WoS)**: 11

2. Martin EC, Sukarta OCA, Spiridon L, Grigore LG, Constantinescu V, Tacutu R, Goverse A, Petrescu A-J, *"LRRpredictor - A New LRR Motif Detection Method for Irregular Motifs of Plant NLR Proteins Using an Ensemble of Classifiers"*, **Genes** 11(3), 286-300 (2020). [PMID: 32182725]

   **IF**: 3.69; **AI**: 1.2; **Citations (WoS)** 12

3. Manoliu LCE* ,Martin EC*, Milac AL, Spiridon L, *"Effective Use of Empirical Data for Virtual Screening against APJR GPCR Receptor."*, **Molecules**; 26(16):4894, 2021. [PMID: 34443478]

   **IF**: 4.41; **AI**: 0.7; **Citations (WoS)** -

4. Mernea M*, Martin EC*, Petrescu AJ, Avram S., *"Deep Learning in the Quest for Compound Nomination for Fighting COVID-19."*, **Curr.Med.Chem** 28(28), 5699-5732 (2021) [PMID: 33441063]

   **IF**: 4.53; **AI**: 0.8; **Citations (WoS)** -

* shared first co-autorship

### Publication as co-author

5. Manica G, Ghenea S, Munteanu CVA, Martin EC, Butnaru C, Surleac M, Chiritoiu GN, Alexandru PR, Petrescu AJ, Petrescu SM, *"EDEM3 Domains Cooperate to Perform Its Overall Cell Functioning."*, **Int.J.Mol.Sci**; 22(4):2172 (2021). [PMID: 33671632]

   **IF**: 5.92; **AI**: 1.2; **Citations (WoS)** 1

6. Baudin M, Martin EC, Sass C, Hassan JA, Bendix C, Sauceda R, Diplock N, Specht CD, Petrescu AJ, Lewis JD, *"A natural diversity screen in Arabidopsis thaliana*

*reveals determinants for HopZ1a recognition in the ZAR1-ZED1 immune complex.*", **Plant Cell Environ.**; 44(2):629-644, 2021. [PMID: 33103794]

**IF**: 7.33; **AI**: 1.9; **Citations (WoS)** 1

7. Baudin M, Schreiber KJ, Martin EC, Petrescu AJ, Lewis JD, *"Structure-function analysis of ZAR1 immune receptor reveals key molecular interactions for activity."*, **Plant J.**; 101(2):352-370, 2020. [PMID: 31557357]

   **IF**: 6.41; **AI**: 2.2; **Citations (WoS)** 10

8. Ionescu AE, Mentel M, Munteanu CVA, Sima LE, Martin EC, Necula-Petrareanu G, Szedlacsek SE., *"Analysis of EYA3 Phosphorylation by Src Kinase Identifies Residues Involved in Cell Proliferation."*, **Int.J.Mol.Sci.**; 20(24):6307, 2019. [PMID: 31847183]

   **IF**: 5.92; **AI**: 1.2; **Citations (WoS)** 5

9. Wróblewski T, Spiridon L, Martin EC, Petrescu AJ, Cavanaugh K, Truco MJ, Xu H, Gozdowski D, Pawłowski K, Michelmore RW, Takken FLW., *"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 Biology**.; 16(12): e2005821 (2018). [PMID: 30540748]

   **IF**: 8.38; **AI**: 4.0; **Citations (WoS)** 26

10. Slootweg EJ, Spiridon LN, Martin EC, Tameling WIL, Townsend PD, Pomp R, Roosien J, Drawska O, Sukarta OCA, Schots A, Borst JW, Joosten MHAJ, Bakker J, Smant G, Cann MJ, Petrescu AJ, Goverse A., *"Distinct Roles of Non-Overlapping Surface Regions of the Coiled-Coil Domain in the Potato Immune Receptor Rx1."*, **Plant Physiol.**; 178(3): 13010-1331 (2018). [PMID: 30194238]

    **IF**: 6.30; **AI**: 2.4; **Citations (WoS)** 7

# Software Packages

- **LRRpredictor**

  GitHub: https://github.com/eliza-m/LRRpredictor_v1

  Webserver: https://lrrpredictor.biochim.ro/

- **NLRexpress**

  GitHub: https://github.com/eliza-m/NLRexpress

  Webserver: https://nlrexpress.biochim.ro/

# Conference presentations

## Oral presentations

- Martin EC, Ifrimescu F, Spiridon L, Goverse A, Petrescu AJ. "An atlas of plant NLR proteins" Jul 2021 | Molecular Plant-Microbe Interactions 34 (7)

## Poster presentations

- Vicari C, Martin EC, Tsakou L, Morales Poole JR, Zhang Y, Petrescu AJ, Pontarotti P, Schatz DG. Update on the distribution of the RAG transposon through the deuterostomes. Poster: 33 P; **23rd Evolutionary Biology Meeting at Marseilles**, France, 24-27 September. (2019)

- Wróblewski T, Spiridon L, Martin EC, Petrescu AJ, Cavanaugh K, Truco MJ, Xu H, Gozdowski D, Pawłowski K, Michelmore RW, Takken FLW. The CC domains of NLR-type pathogen receptors play essential roles in oligomerization, network formation and immune signalling. IS-MPMI XVIII Congress, Glasgow, Scotland, July 14-18 2019.

- Baudin M, Schreiber KJ, Martin EC, Petrescu AJ, Lewis JD. Structure-function analysis of ZAR1 immune receptor reveals key molecular interactions for activity. Poster: 358-P1 IS-MPMI XVIII Congress, Glasgow, Scotland, July 14-18 2019..

- Martin EC, Spiridon L, Caldararu O, Petrescu AJ. Plant R-Protein Structure-Model Vs. Cryo-EM Comparison.Annual International Conference of RSBMB, Iasi, September 24-27, 2019. Poster abstract published in J.Exp.Mol.Biol..; 20(3), 19 (2019)

- Martin EC, Caldararu O, Ruta LL, Ghenea S, Surleac MD, Spiridon L, Milac AL, Farcasanu IC, Petrescu AJ. De novo Peptide Design for Enhanced Heavy Metal Accumulation. Annual International Conference of RSBMB, Timisoara, June 8-9, 2017. Poster abstract published in New Frontiers in Chemistry.; 26 (2). S4_P4 (2017)

Results presented in this thesis were financially supported by UEFISCDI grants PN-III-ID-PCE-2016-0650, PN-III-P1-1.1-TE2016-1852, PN-III-P3-3.5-EUK-2017-02-0030/Nr. 63/2018 and PN-III-P4-IDPCE-2020-2444 and by the Romanian Academy Programs of IBAR: Structural and systemic research in immunobiology and gerontomics.

# Acknowledgements

Foremost, I would like to convey my immense gratitude to my advisor Professor Andrei-José Petrescu for the continuous research support and mentoring during the Ph.D. programme, for including me in a multidisciplinary international team, as well for his inspiration, enthusiasm and encouragement, which allowed me to grow and learn plenty of invaluable things during this journey.

I wish to extend my thanks to Professor David G. Schatz from the Yale University School of Medicine for his insightful guidance and encouragement during the last two years and for leading me to work on diverse exciting projects.

I also want to thank Professor Pierre Pontaroti from Marseille University, Professor Aska Goverse and her team from Wageningen University, as well as Professor Jennifer Lewis and Dr. Maël Baudin from the Berkeley University of California and Dr. Tadeusz Wroblewski from the University of California, Davis for their support, fruitful advice and discussions which facilitated me to widen my research and skills on the topic of plant immunity.

Moreover, I wish to show my appreciation to the members of the Department of Bioinformatics and Structural Biology of IBAR, specifically to Dr. Laurentiu Spiridon, Dr. Adina Milac, Dr. Marius Surleac and Teodor Sulea for their fruitful advice and help, inspiring discussions (especially at late hours when we were working together before deadlines), and for all the fun we have had in the last years.

Not least of all, many thanks to my family, especially Viviana and Cornel who have supported and been there for me for all these years, as well to my grandmother Viorica and my uncle Mircea.

# Bibliography


Bastedo DP, Khan M, Martel A, Seto D, et al. PLOS Pathog., 15(7):e1007900, 2019. doi: 10.1371/journal.ppat.1007900.

Baudin M, Hassan JA, Schreiber KJ, and Lewis JD. Plant Physiol., 174(4):2038–2053, 2017. doi: 10.1104/pp.17.00441.

Baudin M, Schreiber KJ, Martin EC, Petrescu AJ, and Lewis JD. Plant J., 1:352–370, 2019. doi: 10.1111/tpj.14547.

Bej A, Sahoo BR, Swain B, Basu M, et al. Comput. Biol. Med., 53:164–170, 2014. doi: 10.1016/j.compbiomed.2014.07.016.

Casey LW, Lavrencic P, Bentham AR, Cesari S, et al. Proc. National Acad. Sci. United States America, 113(45):12856–12861, 2016. doi: 10.1073/pnas.1609922113.

Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, and Matsushima N. Proteins: Struct. Function Genet., 54(3):394–403, 2004. doi: 10.1002/prot.10605.

Fugmann SD, Messier C, Novack LA, Andrew Cameron R, and Rast JP. Proc. National Acad. Sci. United States America, 103(10):3728–3733, 2006. doi: 10.1073/pnas.0509720103.

Hao W, Collier SM, Moffett P, and Chai J. J. Biol. Chem., 288(50):35868–35876, 2013. doi: 10.1074/jbc.M113.517417.

Huang S, Tao X, Yuan S, Zhang Y, et al. Cell, 166(1):102–114, 2016. doi: 10.1016/j.cell.2016.05.032.

Kajava AV and Kobe B. Protein Sci., 11(5):1082–1090, 2002. doi: 10.1110/ps.4010102.

Kapitonov VV and Jurka J. PLoS Biol., 3(6):0998–1011, 2005. doi: 10.1371/journal.pbio.0030181.

Kapitonov VV and Koonin EV. Biol. Direct, 10(1):20, 2015. doi: 10.1186/s13062-015-0055-8.

Lewis JD, Abada W, Ma W, Guttman DS, and Desveaux D. J. bacteriology, 190(8):2880–2891, 2008. doi: 10.1128/JB.01702-07.

Lewis JD, Wu R, Guttman DS, and Desveaux D. PLoS Genet., 6(4):1–13, 2010. doi: 10.1371/journal.pgen.1000894.

Lewis JD, Lee AHY, Hassan JA, Wan J, et al. Proc. National Acad. Sci., 110(46):18722–18727, 2013. doi: 10.1073/pnas.1315520110.

Litman GW, Rast JP, and Fugmann SD. Nat. Rev. Immunol., 10(8):543–553, 2010.

Ma S, Lapin D, Liu L, Sun Y, et al. Sci. (New York, N.Y.), 370(6521), 2020. doi: 10.1126/SCIENCE.ABE3069.

Maekawa T, Cheng W, Spiridon LN, Töller A, et al. Cell Host Microbe, 9(3):187–199, 2011. doi: 10.1016/j.chom.2011.02.008.

Martin EC, Sukarta OCA, Spiridon L, Grigore LG, et al. Genes, 11(3):286, 2020a. doi: 10.3390/genes11030286.

Martin EC, Vicari C, Tsakou-Ngouafo L, Pontarotti P, et al. Mob. DNA 2020 11:1, 11(1): 1–20, 2020b. doi: 10.1186/S13100-020-00214-Y.

Mitchell AL, Attwood TK, Babbitt PC, Blum M, et al. Nucleic Acids Research, 47(D1): D351–D360, 2019. doi: 10.1093/nar/gky1100.

Morales Poole JR, Huang SF, Xu A, Bayet J, and Pontarotti P. Immunogenetics, 69(6): 391–400, 2017. doi: 10.1007/s00251-017-0979-5.

Offord V and Werling D. Innate Immun., 19(4):398–402, 2013. doi: 10.1177/1753425912465661.

Schatz DG and Swanson PC. Annu. Review Genet., 2011. doi: 10.1146/annurev-genet-110410-132552.

Sela H, Spiridon LN, Ashkenazi H, Bhullar NK, et al. Mol. Plant-Microbe Interactions, 27 (8):835–845, 2014. doi: 10.1094/MPMI-01-14-0009-R.

Wang J, Hu M, Wang J, Qi J, et al. Science, 364(6435):eaav5870, 2019a. doi: 10.1126/science.aav5870.

Wang J, Wang J, Hu M, Wu S, et al. Science, 364(6435), 2019b. doi: 10.1126/science.aav5868.

Zhang Y, Cheng TC, Huang G, Lu Q, et al. Nature, 569(7754):79–84, 2019. doi: 10.1038/s41586-019-1093-7.