ACADEMIA ROMÂNA
ŞCOALA DE STUDII AVANSATE A ACADEMIEI ROMÂNE
INSTITUTE DE BIOCHIMIE

REZUMATUL TEZEI DE DOCTORAT

Investigarea structurii, dinamicii şi interacţiilor proteinelor prin metode de statistică fizică şi matematică

CONDUCĂTOR DE DOCTORAT:
Dr. Andrei-José PETRESCU, CS I

DOCTORAND:
Eliza Cristina MARTIN

2022

Cuprins

Introducere generală

Evoluțiile tehnologice din domeniul secvențierii genomice și transcriptomice alături de cele din biologia structurală, au condus în ultimul deceniu la o creștere exponențială a datelor biologice colectate. Acest lucru are ca rezultat noi orizonturi pentru analiza mediului biologic complex, a rețelelor de interacție a proteinelor și în înțelegerea variațiilor naturale - cu vaste aplicații biologice și medicale. Cu toate acestea, transformarea datelor biologice brute în cunoștințe, necesită un efort în oglindă în dezvoltarea fluxurilor de lucru de analiză, a platformelor bioinformatice, a modelelor matematice și a instrumentelor de predicție capabile să accelereze ritmul cercetării în biologia modernă.

În acest context mai larg, obiectivul general al acestei teze a fost acela de a dezvolta instrumente bioinformatice aplicate și de a le folosi în abordări experimentale asistate computațional în imunobiologie, menite să ofere o mai bună înțelegere a originii și evoluției proteinelor RAG, mașinăria moleculară cheie a sistemului imun adaptativ, pe de o parte și, pe de altă parte, a interacției secvență-structură din vastul repertoriu de receptori de tip NLR ai plantelor in cadrul imunității înnăscute.

Prima parte a tezei prezintă eforturile de a înțelege originea aparatului RAG (recombination-activating gene) specific vertebratele cu falci ce este responsabil pentru generarea repertoriului extins de receptori imuni unici în celulele B și T. Lucrarea prezentată este rezultatul unei colaborări pluridisciplinare între (i) Prof. Andrei-J. Petrescu, la conducerea Departamentului de Bioinformatică și Biochimie Structurală, IBAR, (ii) Prof. David G. Schatz, președintele Departamentului de Imunobiologie de la Yale School of Medicine, SUA, membru al Academiei Naționale de Științe și al Academiei Naționale de Medicină și (iii) Prof. Pierre Pontaroti, Grupul de biologie evolutivă, Universitatea Aix-Marseille, CNRS SNC5039, Franța. Această parte a tezei descrie eforturile de a identifica noi gene RAG-like în grupuri de organisme mai îndepărtate decât cele raportate anterior, contribuind la extinderea înțelegerii asupra originii acestor gene si apariția acestora

mult mai devreme decât s-a considerat inițial, în clada bilateriană timpurie.

A doua parte a tezei se focalizează pe studiul sistemul imun înnăscut al plantelor, în special al receptorilor intracelulari de tip NLR. Următorul capitol prezintă generarea *in silco* de modele probabilistice ale structurii 3D a ZAR1 - receptor NLR cu spectru larg întâlnit in majoritatea grupurilor taxonomice de plante. Această lucrare a făcut parte dintr-o colaborare interdisciplinară mai amplă cu Prof. Jennifer Lewis, Departamentul de Biologie Vegetală și Microbiană, Universitatea Berkeley din California, SUA. Obiectivul principal al acestui proiect comun este de a extinde înțelegerea structurală a mecanismelor moleculare de activare a receptorilor ZAR1 în recunoașterea unei game largi de patogeni, deoarece astfel de receptori cu spectru larg sunt de interes principal în elaborarea strategiilor de control si preventie a patogenilor ce afectează culturile de plante. Ultimele două capitole prezintă dezvoltarea a două pachete software - LRRpredictor și NLRexpress - care utilizează instrumente de predicție bazate pe învățare automată ce vizează identificarea motivelor de secvență asociate domeniilor NLR și sunt rezultatul colaborării cu Prof. Aska Goverse, Laboratorul de Nematologie, Universitatea Wageningen și Research, Olanda. Pachetul software LRRpredictor a fost conceput pentru a aborda gradul sporit de iregularitate ce caracterizează motivele LRR la receptorii NLR ai plantelor și pentru a oferi o performanță de detecție semnificativ mai bună în comparație cu metodele raportate anterior. Prin utilizarea unei colecții de estimatori de învățare automată ce utilizează informații de secvențe și structură, acest instrument are scopul de a sprijini modelarea structurală și cercetarea în biologie moleculară ce vizează proteinele cu domenii LRR. Ultimul capitol prezintă pachetul software NLRexpress - un flux de lucru de predicție organizat in 3 module, care cuprinde 11 estimatori bazați pe rețele neuronale pentru identificarea motivelor structurale și funcționale cheie ale domeniilor constitutive NLR - CC, NBS și LRR - conceput pentru calcule rapide pentru scanarea bazelor de date de secvențe de dimensiuni mari, precum întregul proteom al unei specii.

Capitolul 1

Noi date asupra originii mașinăriei RAG

1.1 Introducere

Sistemul imun adaptativ, specific doar vertebratelor cu fălci reprezintă un mecasim esențial ce a oferit acestora un avantaj evolutiv major (Litman et al., 2010). Această caracteristică extraordinară constă în generarea unui vast repertoriu de gene ale receptorilor de antigeni prin reacții de recombinare în timpul formării limfocitelor. Mecanismul de recombinare V(D)J este realizat de către mașinăria RAG prin operații asupra genelor ”variable” (V), ”diversity” (D) și ”joining” (J), generând un vast set de posibile gene de receptori (Schatz and Swanson, 2011). Originea recombinazei RAG la vertebrate a fost un subiect de dezbatere în ultimele două decenii. Având în vedere asemănările afișate de miezul catalitic al endoucleazei RAG1 cu transpozazele DDE, o ipoteză inițială a fost aceea că genele RAG au derivat dintr-un element mobil de tip II. Primul element mobil identificat ce prezintă omologie ridicată cu RAG1 a fost Transib (Kapitonov and Jurka, 2005), urmat ulterior, în ultimii 5 ani, de descoperirea mai multor gene RAG-like/RAGL în filumul *Deuterostomia* la nevertebrate - dintre care unele elemente prezintă configurația completă a transpozonui, consolidând astfel această ipoteză (Fugmann et al., 2006; Huang et al., 2016; Morales Poole et al., 2017; Zhang et al., 2019).

Experimente *in vitro* utilizând proteinele RAG1&2-like de la cefalocordatul *B.belcheri* au demonstrat activitatea lor endonucleazica și de transpozază, acesta devenind primul transpozon *Proto*-RAG cu activitate probată experimental (Huang et al., 2016). În plus, rezolvarea experimentală a structurilor 3D prin crio-microscopie electronică (cryo-EM)

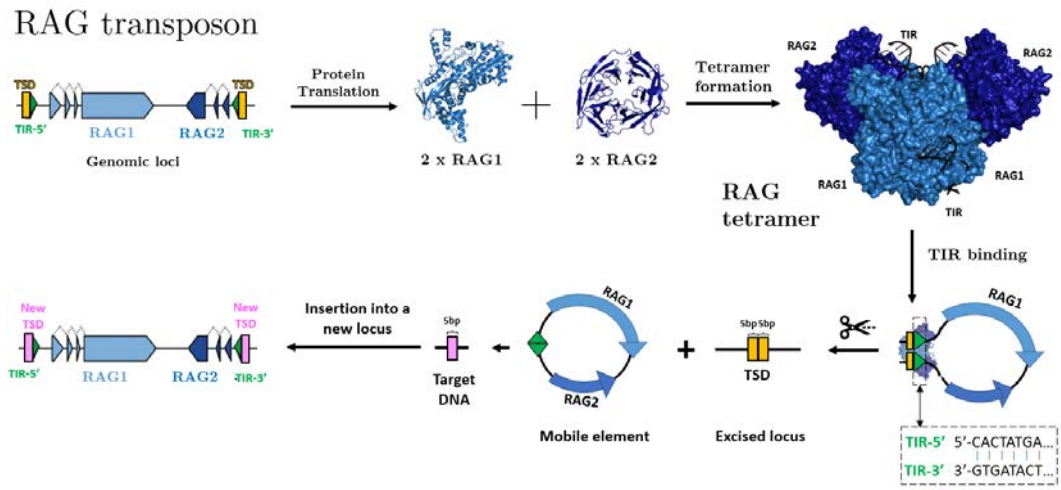


Fig. 1.1: Diagramă a mecanisului de funcționare a transposonului RAG de la amphioxus

ale elementelor RAG de la belcheri, în diferite stadii de funcționare au arătat asemănări semnificative cu mecanismele de clivaj complexe ale RAG de la vertebrate, în ciuda omologiei lor scăzute la nivelul secvenței proteice (Zhang et al., 2019). Succint, în urma translației celor două gene care codifică RAG1 și RAG2, se formează complexul tetramer RAG ce este compus din doi heterodimeri RAG1-RAG2 (Fig. 1.1). Complexul recunoaște regiunea heptamerică a marginilor TIR și reunește segmentele TIR 5' și 3', îndoind ADN-ul și formând o structură circulară (Zhang et al., 2019). Complexul endonucleazei taie ADN-ul la începutul marginilor TIR, excizând caseta transpozonului din locusul genomic (Fig. 1.1), ce ulterior va fi sudat de enzimele de reparare ale gazdei. Elementul mobil este inserat într-un nou locus genomic, proces în care este generată o nouă duplicare a site-ului țintă "Target site duplication" (TSD) de 5pb (Zhang et al., 2019).

Anterior acestui studiu realizat în intervalul 2019-2020, nu se cunoșteau dovezi pentru activitatea transpozonului RAG în afara filumului *Deuterostomia*, în timp ce transpozonul Transib a fost identificat pe scară largă de la deuterostomii la regnul fungi (Kapitonov and Jurka, 2005; Kapitonov and Koonin, 2015). Acest capitol al tezei prezintă identificarea mai multor perechi de gene RAG1L-RAG2L în superfilumul *Protostomia* din interiorul filumului *Mollusca* și *Nemertea*, dintre care unele elemente prezintă o configurație completă a transpozonului, cu markeri ai activității transpozazei (Martin et al., 2020b), CDS conservat ce ar putea conduce la produse proteice complete atat ca lungime cat si ca organizare a domeniilor funcționale. Mai mult decât atât, au fost identificate copii mai puțin conservate în filumul *Cnidaria*, în afara încrengăturii *Bilateria*, indicând faptul că transposonul RAGL

a fost activ în afara cladei *Deuterostomia*, așa cum se presupunea inițial și că ar putea avea o origine mai veche in grupul *Bilateria* timpuriu.

1.2 Rezultate și Discuții

Pentru a scana bazele de date genomice și transcriptomice disponibile, a fost utilizată o colecție de secvențe RAG1 și RAG2 documentate anterior. Deoarece omologia de secvență dintre RAG2 este foarte scăzută, sub pragul de detecție al metodelor de tip ”blast”, a fost utilizată o abordare de scanare iterativă ce a permis detectarea de noi gene asemănătoare cu RAG1 și RAG2 în încregăturile *Protostomia* și *Cnidaria* (Fig. 1.2).

În filul *Protostomia*, perechile de gene RAG1-RAG2 au fost identificate în clada *Mollusca* la stridii (*Crassostrea virginica, Crassostrea gigas, Saccostrea glomerata*), la midii (*Modiolus philippinarum, Bathymodiolus platifrons*) și în clada de stridii cu perle (*Pinctada imbricata*)(Martin et al., 2020b). În filul *Nemertea*, la momentul analizei, singura specie cu date genomice și/sau transcriptomice disponibile a fost viermele panglică (*Notospermus geniculatus*), unde au fost identificate numeroase perechi RAG1-RAG2, dintre care unele au fost susținute de date transcriptomice mARN. În *Cnidaria*, perechile de gene RAG au fost identificate în *Porites rus, Orbicella faveolata* și *Aurelia aurita* (Martin et al., 2020b). Spre deosebire de protostomii, unde au fost identificate mai multe copii ale genelor la fiecare specie, la *Cnidaria* doar meduza *A.aurita* prezintă o pereche RAG intactă, în timp ce celelalte loci identificate prezintă semne de pseudogenizare.

Specifica transpozonilor DDE din clasa II este prezența elementelor TIR la marginea elementului mobil. Regiunea de compatibilitate dintre segmentele TIR se întinde adesea doar la margini (8-10 pb), făcând detectarea acestor elemente mai dificilă. Pentru a discrimina capetele transpozonilor de alte astfel de repetiții inversate frecvent prezente în genom, a fost folosită o abordare bazată pe variația omologiei. Astfel, diferitele duplicate ale transpozonilor sunt de așteptat să împărtășească un grad ridicat de omologie de secvență, în timp ce regiunile de flancare nu ar trebui să prezinte omologie, deoarece acestea corespund unor regiuni distincte în genom (Martin et al., 2020b). Prezența elementelor TSD ce flanchează marginile TIR a fost folosită drept constrângere discriminatorie suplimentară pentru a distinge între transpozonii TIR și capetele premature ale casetelor de transpozon.

Mai multe perechi de gene RAG1-RAG2 identificate prezintă o configurație completă de transpozon TSD-TIR-RAG1-RAG2-TIR-TSD si au fost identificate cu precadere în *C.virginica*, *P.imbricata* și *N.geniculatus*. Toate elementele TIR identificate prezintă o regiune asemănătoare heptamerului RSS, cu primele 3 nucleotide "CAC" perfect conservate - esențiale atât pentru funcționalitatea transpozazei, cât și a recombinazei. Similar elementelor TIR raportate anterior la deuterostomii, elementele protostome nu prezintă o regiune asemănătoare nonamerului RSS. Mai mult, lungimea de 5 pb a elementelor TSD identificate la protostomii este în concordanță cu cele găsite în cazul transposonilor Transib sau RAG de la deuterostomii și vertebratele cu falci (Kapitonov and Jurka, 2005; Morales Poole et al., 2017; Zhang et al., 2019).

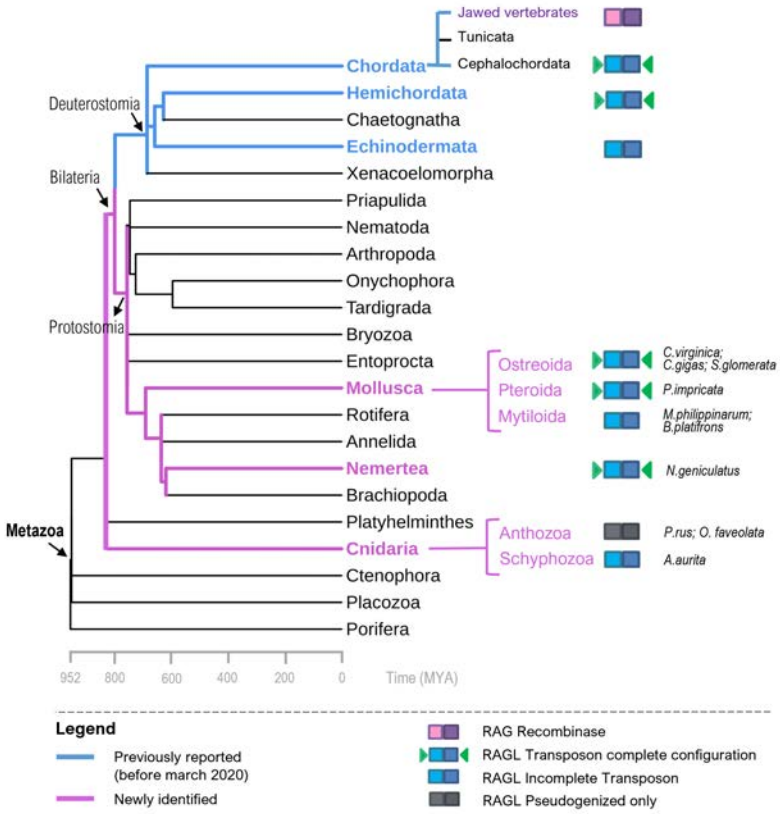


Fig. 1.2: Distribuția taxonomică a elementelor RAGL. Ramurile albastre descriu elementele RAG raportate înainte de martie 2020, în timp ce ramurile magenta descriu cladele în care a fost identificat RAGL în acest studiu. Starea celor mai conservate loci identificate este ilustrată cu pictograme descrise in legenda figurii

Analiza de filogenie a fost efectuată pe regiunea nucleului catalitic RAG1 a celor mai conservați reprezentanți identificați. Arborele RAG1 urmează filogenia speciei, indicând evoluția verticală în cadrul celor două clade bilateriane - *Protostomia* și *Deuterostomia* -

în concordanță cu analizele raportate anterior (Morales Poole et al., 2017). Partiționarea actualizată a familiei RAG constă în: **(a) familia RAG-A** - cea mai apropiată de RAG de vertebrate; **(b) familia RAG-B** - răspândită în clada deuterostomelor care conține elemente RAG raportate anterior în deuterostomii și mulți dintre transpozonii identificați în *Protostomia* și *A.aurita*; **(c) familia RAG-C** - cu un singur membru raportat în hemicordatul *P. flava* (Morales Poole et al., 2017); **(d) familia RAG-D** - o nouă familie distinctă, identificata numai în clada *Nemertea* la *N.geniculatus*.

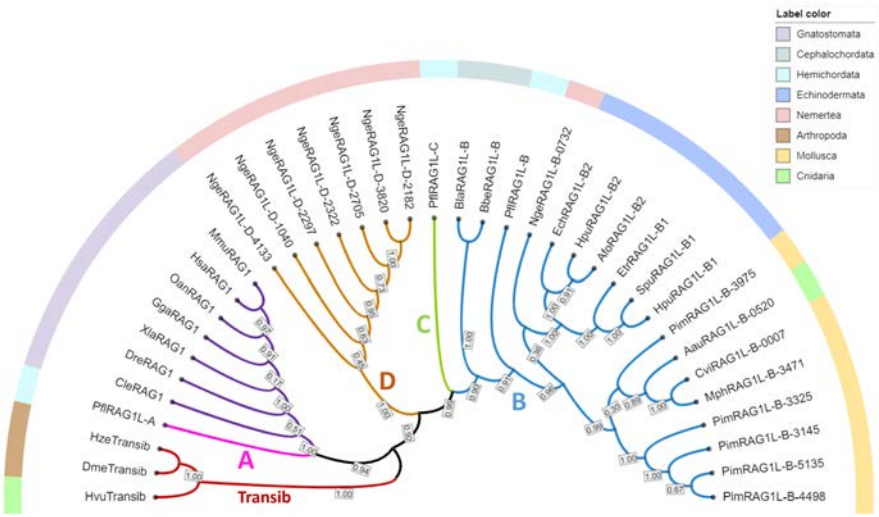


Fig. 1.3: Arborele de filogenie RAGL actualizat. Suportul bootstrap este calculat folosind Maximum Likelihood pe regiunea catalitică a RAG1.

Omologii RAG1 identificați prezintă în mod preponderent, cu foarte puține excepții, aranjamentul complet al domeniilor funcționale specific transpozazelor RAG1 de la deuterostome și al recombinazei RAG1 de la vertebrate. Pozițiile cheie din domeniul de dimerizare și legare la ADN (DDBD) si din domeniul catalitic RNH sunt conservate în unanimitate în copiile identificate, precum și motivul Zn-finger în domeniile ZnC2 și ZnH2. Reprezentanții identificați la protostomii prezintă o coadă C-ter foarte similară cu RAGL de la deuterostomii ce prezintă motivul *C**C***GH****C*. Toate secvențele de proteine RAG2L analizate de la protostomii conțin un domeniu de tip Kelch, urmat de un domeniu PHD bogat în cisteină. Similar secvențelor de RAG2 la deutorostomii, acestora le lipsește regiunea linker acidă specifică vertebratelor. Siturile de contact dintre RAG1 si ADN, atât în regiunea de contact cu heptamerul, cât și în regiunile de flancare sunt conservate în mare parte în special în preajma triadei catalitice, sugerând că elemente RAG1 de la protostomii ar putea prezenta un comportament similar în ceea ce privește legarea și scindarea ADN-ului

cu *B.belcheri* (Fig. 1.4). Pe de altă parte, suprafața extinsă și complexă de interacție dintre RAG1L-RAG2L este slab conservată, sugerând o posibilă co-evoluție intre cele două proteine RAG1L și RAG2L.

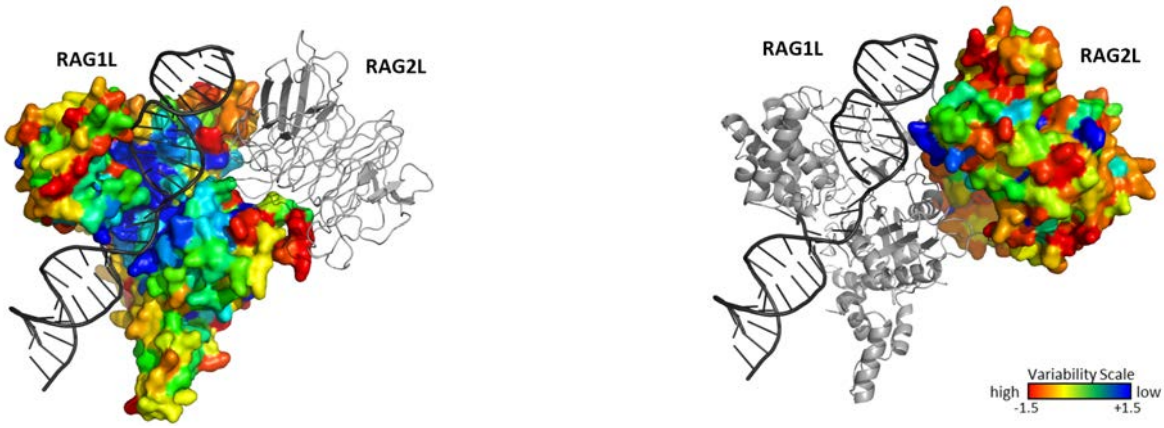


Fig. 1.4: Analiză de variabilitatea pentru RAG1 (a) și RAG2 (b) mapată pe structura cryo-EM (roșu-variabil, albastru-conservat).

1.3 Concluzii

Rezultatele prezentate in cadrul tezei indică dovezi pentru activitatea transpozonilor RAG în diverse specii de protostome, o parte din loci find susținute de date transcriptomice. Acestea prezintă perechi intacte de elemente TIR și TSD, o configurație completă a domeniilor funcționale și conservarea reziduurilor cataitice cheie, indicând faptul ca astfel de transposoni RAG ar putea fi active în prezent în organismele lor gazdă. În afara cladei bilateriane, doar câteva perechi incomplete asemănătoare RAG au fost identificate în filul *Cnidaria* în diferite stadii de pseudogenizare, cu o singură pereche completă de gene în *A.aurita*. Cu toate acestea, prezența fragmentelor RAG în *Cnidaria* indică faptul că transposonul RAG ar putea avea origini mai vechi decât se presupunea inițial. Analiza de filogenie prezentată este în concordanță cu o traiectorie de evoluție verticală a RAG în interiorul cladelor protostome și deuterostome, fapt ce ar putea indica posibile stagii de domesticire a genelor RAG în organismele lor gazdă. Studii ulterioare ale unor astfel de candidați ar putea fi de interes, atât pentru a aduce noi perspective asupra fenomenului de domesticire a recombinazei RAG la vertebrate, cât și pentru investigarea unor funcții biologice potențial noi ale RAG în aceste organisme.

Capitolul 2

Structura ZAR1: de la modelul 3D in-silico la validarea experimentală

2.1 Introducere

Acest capitol prezintă în detaliu studiul structural *in-silico* al receptorului ZAR1 NLR de la *Arabidopsis thaliana*, care a început la începutul programului de doctorat în noiembrie 2016 și a făcut parte dintr-un proiect de cercetare mai amplu în colaborare cu Prof. Dr. Jennifer Lewis, Departamentul de Biologie Microbiană a Plantelor, Universitatea Berkeley din California. Studiul și-a propus să ofere o mai bună înțelegere a determinanților structurali în tranzițiile de interacție inter-domeniu în timpul mecanismului de activare.

Receptorul ZAR1 are rolul de a media detectarea unei varietăți de proteine indicatori ai prezenței patogenilor prin intermediul unor kinaze adaptor (Lewis et al., 2013; Bastedo et al., 2019), astfel de receptori NLR cu spectru larg fiind de mare interes pentru dezvoltarea strategiilor de control al patogenilor. Studiile anterioare ale grupului au identificat kinaza ZED1, indispensabilă pentru generarea răspunsului imun in urma detectiei proteinei Hopz1a de la *Pseudomonas syringae* (Lewis et al., 2008, 2010, 2013), precum și mai multe mutații și trunchieri experimentale ale ZAR1 ce induc modificări fenotipice în profilul de interacție inter-domenii și/sau impactează răspunsul imun (Baudin et al., 2017).

Dezvoltarea modelelor 3D ale proteinei ZAR1 pentru a asista experimentele de biologie moleculară a fost utilă în formularea ipotezelor privind interacțiile inter-domenii și în

propunerea de noi intervenții in vederea probării experimentale a acestora și pentru a oferi noi date esentiale pentru optimizarea modelelor probabilistice. În 2019, a fost raportată structura crio-EM a proteinei ZAR1, fapt ce ne-a oferit oportunitatea de a compara modelele 3D probabilistice cu structura *reala* - ce s-a dovedit a fi într-un bun acord într-un interval de 2-5 $\mathring{A}$ deviatie RMSD la suprapunerea structurilor 3D ale domeniilor individuale, evidențiind caracterul practic și eficacitatea utilizării modelelor probabilistice în absența structurilor 3D obținute experimental.

2.2 Rezultate și Discuții

Generarea de modele 3D ale domeniilor ZAR1 a fost ingreunată din cauza omologiei extrem de scăzute cu orice structură 3D disponibilă la acel moment. In absenta unei structuri 3D complete obținute experimental pentru receptori NLR de la plante, raportate erau doar: (i) 3 structuri de domeniu CC cu arhitecturi 3D controversate discutate mai jos - la $\leq$19% identitate cu ZAR1, (ii) domenii NBS provenind de la metazoarele Apaf1 și Ced4 sub 21% identitate cu ZAR1 și (iii) diferite domenii LRR de la plante, toate provenind din receptori extracelulari cu diferențe structurale semnificative. În ciuda omologiei scăzute, au fost generate modelele 3D probabilistice corespunzatoare domeniilor individuale ale receptorului ZAR1, si au fost optimizate prin simulări de dinamică moleculară.

Structurile rezolvate experimental disponibile la acel moment au indicat faptul că domeniul CC ar putea adopta două configurații: o arhitectura cu 4 elici (4H-CC) corform structurilor 3D ale proteinelor Rx și Sr33 (Hao et al., 2013; Casey et al., 2016) sau o arhitectura cu 2 elici (2H-CC) asemenea dimerul MLA10 (Maekawa et al., 2011; Casey et al., 2016), care in ciuda diferentelor structurale, prezinta $\sim$85% identitate cu Sr33. Analiza structurală a celor două arhitecturi a indicat faptul că cei doi monomeri 4H-CC ai Sr33 se suprapun aproape perfect cu dimerul MLA10 (2 x 2H-CC) (Maekawa et al., 2011; Casey et al., 2016). O posibilă tranziție structurală compatibilă cu ambele arhitecturi 3D, propusă la începutul studuiului, a fost aceea că primul și al patrulea segment elicoidal periferic se desprind și îmbrățișează celălalt monomer. O astfel de tranziție ar necesita ca regiunile linker care conectează elicele periferice să posede o anumită ambivalență structurală, permițând rotații si tranzitii de structura secundara.

Pe baza modelelor 3D generate pentru ZAR1, mai multe ipoteze structurale au fost testate experimental de colaboratorii noștri prin efectuarea de experimente de mutageneză și evaluarea modificărilor fenotipice în profilul de interacțiune inter-domeniu prin *yeast-two-hibrid* (Y2H) și *bimolecular fluorescence complementation*(BiFC), precum și diferențe în răspunsul imun *in planta*, descrise în detaliu în (Baudin et al., 2019). Mutații la nivelul buclelor linker ce modifică compoziția lor bogat încărcată electrostatic au condus la suprimarea parțială a dimerizării și a interacției CC-NBS si CC-LRR, precum și la un răspuns HR redus *in planta* (Baudin et al., 2019), indicând faptul că aceste regiuni sunt implicate în interacțiile dintre domenii. Pentru a studia rolul primului segment helical în activare, au fost propuse mutații care reduc hidrofobicitatea primei elici din domeniul CC cu rațiunea că în timpul activării primul segment helical prezintă o mai mică constrângere și poate iniția tranzițiile conformaționale (Baudin et al., 2019). Experimentele prin Y2H au indicat un nivel redus de dimerizare și interacție CC-NBS alături de suprimarea completă a răspunsului imun *in planta*, dar nu a afectat interacția CC-LRR, sugerând faptul că primul helix al domeniului CC este implicat în interacția CC-NBS (Baudin et al., 2019). Introducerea de mutații la nivelul motivul conservat EDVID, a condus la impactarea procesului de dimerizare, alterarea totală a interacției CC-NBS și CC-LRR și suprimarea completă a răspunsul HR, indicând faptul că această regiune ar putea fi implicată în interfața CC-NBS și CC-LRR, ceea ce a fost confirmat ulterior de structura crio-EM a ZAR1 (Baudin et al., 2019).

În 2019, structurile crioEM ale proteinei ZAR1 au fost raportate în 3 etape ale mecanismului de activare: stare monomerică inactivă ce leagă ADP, stare monomerică activată cu nucleotidă absentă și în stare oligomerică activată, ce leagă ATP (Wang et al., 2019b,a). La nivelul domeniului CC, structurile crioEM ZAR1 relevă modificări conformaționale drastice in timpul procesului de activare - de la o conformație 4H-CC în starea inactivă la o conformație 3H-CC în starea activată. În conformația 4H-CC, doar prima jumătate a segmentului H4 face parte din domeniul CC, pe când în modelul inițial, întregul helix H4 a fost modelat ca parte a domeniului CC, iar restul modelului este într-un bun acord, cu valori RMSD de 3,9 $\mathring{A}$. față de structura crioEM (Fig. 2.2). Modelele propuse pentru domeniul NBS prezintă un acord structural ridicat cu structura crio-EM, atât pentru conformația inactivă, cât și pentru ca activată, cu o valoare RMSD de 4,0$\mathring{A}$ și respectiv 5,1$\mathring{A}$. Modelul domeniului LRR prezintă o bună suprapunere cu structura crio-EM cu valori RMSD de 4,7$\mathring{A}$ și o bună conformitate a curburii si razei generale a domeniului. Mai mult, regiunea adiacentă situsului

catalitic de legarea a ADP/ATP a fost reprezentat corect de model.

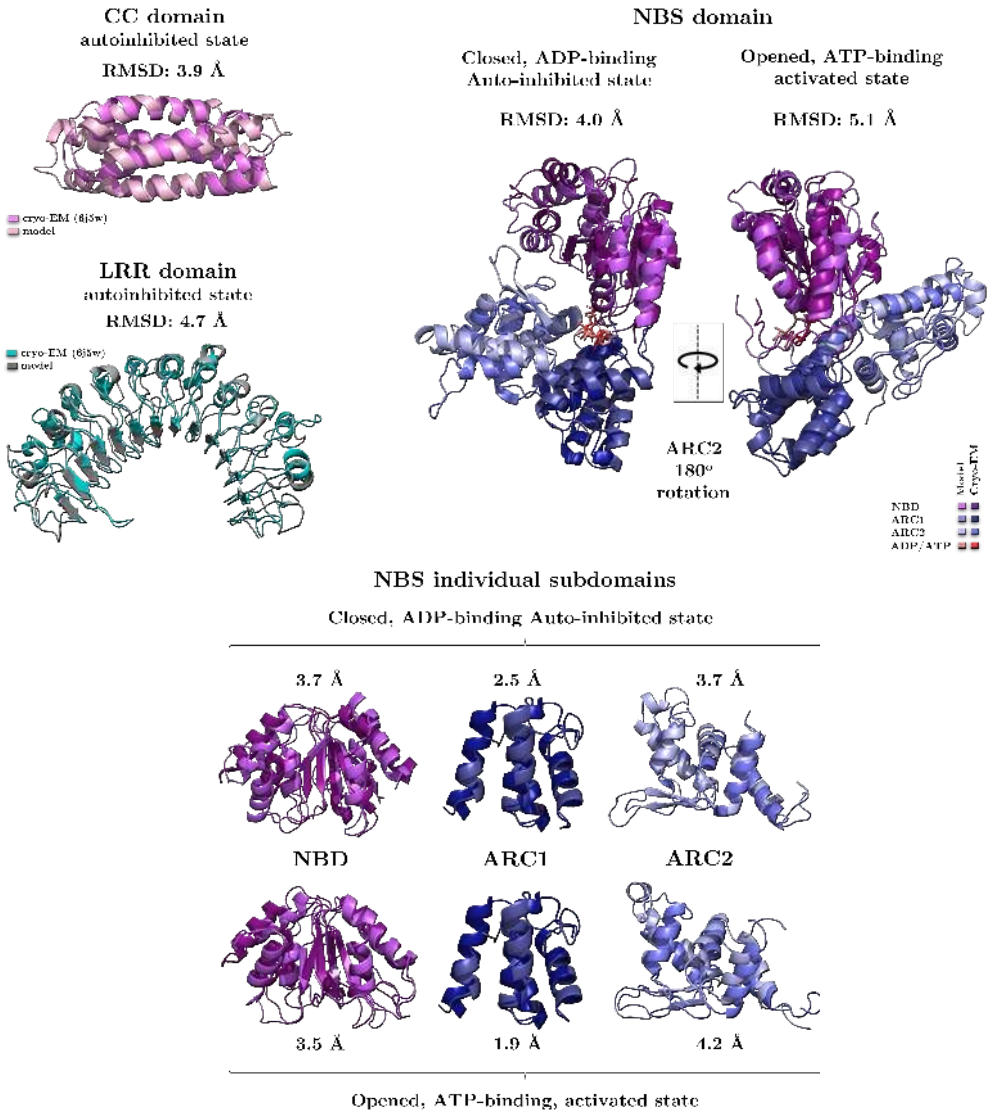


Fig. 2.2: Modelul ZAR1 vs structura cryo-EM în conformația inactivă/activată (6j5w, 6j5t).

2.3 Concluzii

Acest capitol descrie analiza structurală și generarea de modele 3D probabilistice pentru domeniile receptorului NLR ZAR1, realizate anterior apariției structurii crioEM. Acest lucru ne-a oferit posibilitatea de a compara modelele propuse inițial cu structura experimentală și, de asemenea, de a analiza ipotezele formulate pe baza modelului în lumina noilor date. Modelul prezintă un acord destul de bun cu structura 3D, evidențiind caracterul practic și eficacitatea utilizării analizei computaționale și modelelor probabilistice în absența datelor 3D obținute experimental.

Capitolul 3

LRRpredictor - detecția motivelor LRR iregulare în receptorii NLR de la plante

3.1 Introducere

Arhitectura de repetiții bogate în leucină LRR (*Leucine-rich repeat*) este esențială pentru sistemul imun, în detectarea agenților patogeni și transducția semnalului și este întâlnită în întregul arbore al speciilor, de la *Archaea* la *Mammalia* (Enkhbayar et al., 2004). Domeniile LRR adoptă o structură 3D de solenoid asemănătoare cu o ”potcoavă” compusă din segmente generate de regiuni repetitive cu lungimi de ∼15-30 aminoacizi. Aceste segmente sunt consolidate structural de existența unei rețele de foi beta pe partea interioară a domeniului și este caracterizată de prezența unui motiv de secvență conservat, denumit motiv LRR (Kajava and Kobe, 2002). Consensusul motivului LRR arată variații semnificative între clasele de proteine și grupuri taxonomice, motivul LRR minimalist împărtășit de toate clasele având *LxxLxL* drept consensus (L-orice aminoacid hidrofob, cel mai frecvent leucină; x-orice aminoacid). În plus, studiile asupra domeniilor NLR la plante, au arătat o frecvență mult mai crescută a motivelor iregulare în comparație cu omologii lor de la metazoare sau în comparație cu receptorii extracelulari ai plantelor (Sela et al., 2014; Wang et al., 2019b).

Înțelegerea factorilor structurali ai specificității de legare a domeniilor LRR este de interes din perspectiva ingineriei receptorilor pentru controlul patogenilor, cu implicații vaste atât în domeniul medicinei, cât și al agriculturii. Lipsa de sensibilitate a abordărilor actuale în detectarea corectă a motivelor LRR doar pornind de la secvența sa de aminoacizi

este un dezavantaj semnificativ în analiza bioinformatică, modelarea 3D acurată și în studiul relațiilor dintre secvență, particularitățile structurale și comportamentul biologic. Dificultatea în detectarea motivelor individuale constă în faptul că motivul minimalist este extrem de trivial și adesea astfel de motive apar in mod aleatoriu în proteine non-LRR.

În acest capitol este prezentată dezvoltarea LRRpredictor - o nouă metodă de detectare a motivelor LRR ce constă într-un ansamblu de estimatori ce își propun să aducă o versatilitate sporită iregularității motivelor față de metodele existente, prin utilizarea tehnicilor de reeșantionare. Performanța și comportamentul metodei sunt evaluate în comparație cu metodele existente pe un set de date de domenii adnotate din diferite clase (receptori NLR, RLK și RLP de la plante și respectiv NLR și TLR de la animale).

3.2 Rezultate și discuții

Bazele de date de domenii adnotate disponibile au fost folosite pentru a identifica domenii LRR cu structură 3D cunoscută. După aplicarea unui filtru de redundanță de 90% identitate, 178 de structuri (PDB-LRR-90) cuprinzând ∼2100 segmente LRR au fost utilizate în continuare în analiză și au fost supuse delimitării segmentelor LRR pe baza datelor lor structurale și a rețelei de foi beta. Un al doilea filtru de redundanță de 50% identitate a fost aplicat pentru obținerea datelor setului de antrenare, la nivelul segmentelor individuale LRR, obținându-se un set de ∼850 segmente LRR cu un grad sporit de divergență (PDB-LRR-50).

Suprapunerea 3D a diferitelor segmente LRR a arătat o asemănare topologică ridicată extinzând atât în amonte, cât și în aval, motivul minimalist $L_0XXL_3XL_5$ de 6aa cu cel puțin 5aa în ambele direcții (Fig. 3.1b). Prin urmare, un interval de 16aa de la poziția -5 la +10 în jurul poziției L_0 a fost denumit în continuare motivul extins. Un aspect important ce limitează analiza este distributia taxonomică foarte inegală a datelor 3D disponibile în comparație cu distribuția de bază echivalentă a bazelor de date de secvențe Uniref-50. Aproximativ jumătate din segmentele LRR din PDB-LRR-50 provin de la specii de mamifere, în timp ce în UniRef-50 ponderea proteinelor de mamifere este mai mică (3%) (Fig. 3.1d). Prin contrast, receptorii NLR de la plante sunt extrem de slab reprezentati in setul structural, cu o singură structură 3D (Wang et al., 2019b,a) raportată înainte de 2020, în timp ce majoritatea segmentelor LRR de la plante aparțin receptorilor RLP și RLK.

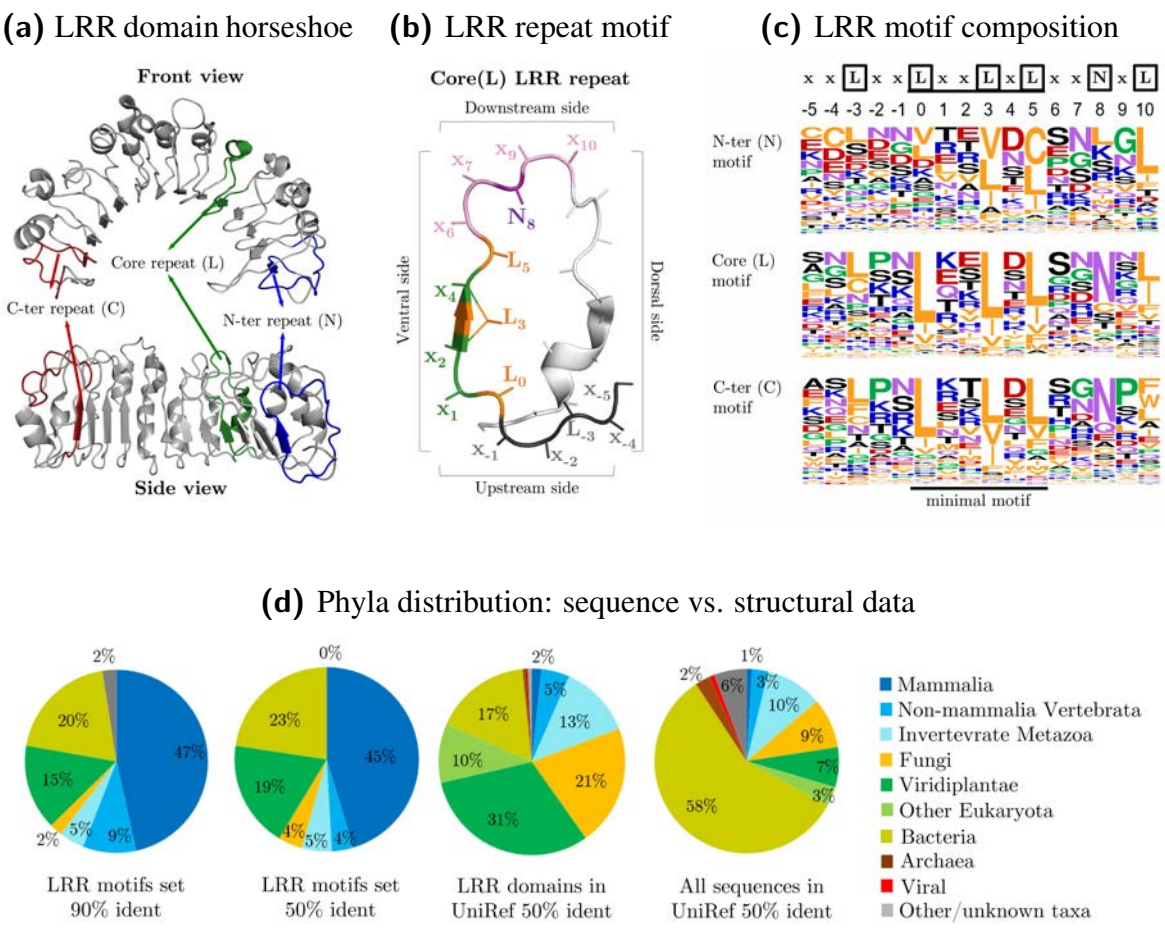


Fig. 3.1: **(a)** Arhitectura LRR exemplificată pe structura ZAR1 (PDB:6j5w). **(b)** Structura unui segment LRR. **(c)** Compoziția motivelor LRR N-ter, de mijloc si C-ter(setul PDB-LRR-50). **(d)** Distribuția taxonomica a seturilor de structuri 3D ale domeniilor LRR versus baza de date Uniref50. Figură derivată din (Martin et al., 2020a).

LRRpredictor a fost antrenat pe un set de proteine care cuprinde o colecție reprezentativă de 850 de motive LRR foarte diverse (la 50% identitate) și un set de proteine non-LRR din fiecare topologie CATH 3D. Pentru a suplimenta predictorul cu un context mai larg asupra secvenței, au fost utilizate profiluri de variabilitate (PSSM). Acestea sunt obținute din probabilitățile de tranziție a reziduurilor, condiționate de grupul de proteine din care aparțin și este de asteptat ca acestea să furnizeze informații de context și să sublinieze pozițiile cheie conservate, precum și relațiile dintre reziduuri. Pe lângă caracteristicile legate de secvență, au fost explorate și caracteristici structurale, precum structura secundară, accesibilitatea la solvenți și predicțiile de dezordine intrinsecă. Având în vedere dimensiunea redusă a setului de date structurale, au fost folosite diferite metode de samplare artificială.

Predictorul final optimizat - LRRpredictor - constă intr-un ansamblu de opt clasificatori ce

utilizează diferite tehnici de învățare supervizată și ce sunt agregate printr-o abordare de tip *soft voting*. Jumătate dintre estimatorii constituenți se bazează doar pe informații legate de secvență, în timp ce cealaltă jumătate utilizează atât secvențe, cât și caracteristici structurale. Etapa de antrenare a fost efectuată folosind o procedură de cross-validare pe 80% din setul de date, iar pentru evaluarea finală a performanței restul de 20% din date.

Atât în cazul etapei de cross-validare cat și de testare, scorurile de senzitivitate, precizie și F1 ale predictorului ansamblului variază în intervalul 85-97% pentru toate tipurile de motive LRR și în intervalul 89-98% atunci când numai motivele interne (L) sunt luate în considerare. De asemenea, LRRpredictor prezintă performanțe sporite comparativ cu alți predictori de motive LRR, cum ar fi LRRsearch (Bej et al., 2014) și LRRfinder (Offord and Werling, 2013). Fluxul LRRpredictor a fost testat pe patru clase de proteine solenoide - trimerice, pectat liază, ankyrin și armadillo - care prezină cea mai apropiată asemănare cu arhitectura LRR, utilizând seturi a câte 50 secvențe din fiecare clasă. LRRpredictor este capabil să facă distincția între motivele *reale* LRR și alte motive asemănătoare, nefiind obținute estimări fals pozitive (peste 50% probabilitatea) pentru niciunul din cele patru seturi.

De asemenea, a fost evaluată capacitatea de extrapolare a predictorului pe diferite clase de proteine ce conțin domenii LRR specifice sistemului imun: 4 seturi conținând receptori citosolici (CNL, TNL) si extracelulari (RLK, RLP) de la plante și 2 seturi de la vertebrate - NLR citosolice și TLR extracelulare. Motivele LRR identificate folosind LRRpredictor prezintă o acoperire bună a domeniilor LRR adnotate în baza de date Interpro (Mitchell et al., 2019) în toate cele șase seturi de date. Aproximativ 75% din CNL-uri și 50% din TNL-uri nu prezinta nici o adnotare de motive LRR in Interpro, în timp ce adnotările receptorilor extracelulari acoperă între 30-80% din domeniul LRR. Acoperirea obținută de LRRpredictor este semnificativ mai mare, cu valori de peste 90% per domeniu LRR în două treimi din fiecare set. Analiza motivelor LRR identificate sublineaza particularități specifice fiecarei clase de receptori imuni (Fig. 3.2). În timp ce intervalul minim al motivului - $L_0XXL_3XL_5$ - este invariabil între seturi, o variabilitate crescută se poate observa in afara acestei regiuni, in special in grupurile CNL și TNL, în timp ce receptorii extracelulari prezintă un motiv extins.

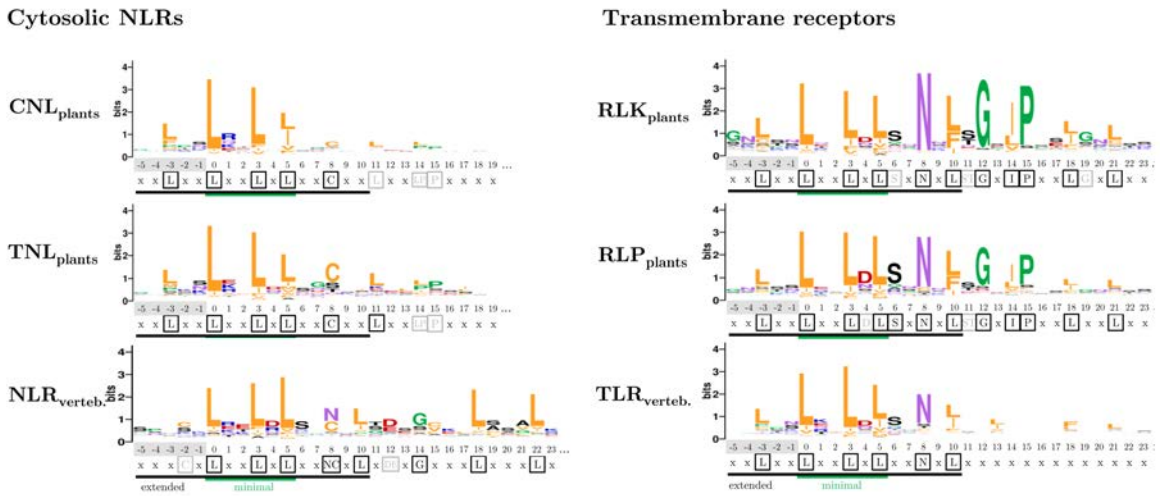


Fig. 3.2: Consensus al motivelor LRR în diferite clase de proteine. Variabilitatea este exprimată ca entropie relativă, înălțimea literelor fiind proporțională cu gradul de conservare. Figură derivată din (Martin et al., 2020a).

3.3 Conclusion

Rezultatele prezentate in acest capitol indică faptul că pachetul software LRRpredictor prezintă o performanță bună pe structurile 3D disponibile si o bună capacitate de extrapolare pe domenii LRR ale diferitelor clase de proteine cu rol in sistemul imun, în special in cazul receptorilor CNL si TNL de la plante, care sunt caracterizati de o iregularitate crescută a motivelor LRR și sunt slab reprezentati în setul de date structurale existente. Predictorul prezintă o bună rată de acoperire a domenilor LRR adnotate in baza de date Interpro, semnificativ mai mari în comparație cu alte metode de adnotăre a motivelor LRR. Mai mult, motivele LRR identificate urmează proliful de secvență raportat anterior al fiecărei clase de receptori imuni investigate.

În concluzie, LRRpredictor este un instrument ce își propune să asiste cercetarea structurală în înțelegerea interacțiunii secvență-structură-funcție a receptorilor imuni - esențială în domeniul ingineriei receptorilor imuni si al detectării patogenilor.

Capitolul 4

NLRexpress - o colecție de predictori ML pentru detectarea motivelor NLR

4.1 Introduction & context

Capitolul anterior descrie dezvoltarea LRRpredictor (Martin et al., 2020a) - un predictor de motive LRR conceput pentru a adresa motivele LRR caracterizate de o iregularitate crescută, cum ar fi în cazul receptorilor NLR de la plante. Acesta constă într-o colecție de 8 clasificatori individuali ce utilizează diferite strategii de eșantionare artificială și de învățare automată. Clasificatorii LRRpredictor utilizează variabilitatea profilurilor PSSM, construite folosind bazele de date globale de proteine Uniprot-20 și predicțiile proprietăților structurale - ambele necesitând resurse computaționale ridicate, ceea ce face ca LRRpredictor să fie mai puțin fezabil pentru scanarea seturilor mari de date. Un instrument rapid capabil să scaneze întregul proteom sau transcriptom al unei specii și să adnoteze motivele funcționale cheie este util în analiza comparativă de secvență, în procesul de discriminare între secvențele NLR complete si cele ce prezintă unul sau mai multe motive absente, în generarea de modele 3D mai acurate și în analiza modificărilor suprafețelor de interacție proteină-proteină.

Pentru a reduce aceste limitări, au fost investigate modele de rețele neuronale capabile să reducă timpul de execuție și resursele de calcul necesare cu un impact minim asupra performanței. Suplimentar motivelor *LxxLxL* ce descriu fiecare segment repetitiv LRR, analiza a fost extinsă pentru a include și motivele de secvență prezente in alte domenii specificife receptorilor NLR, precum domeniile NBS și CC - deoarece aceste regiuni

conservate prezintă un rol vital, cum ar fi legarea ADP/ATP, interacția inter-domenii sau în stabilitatea arhitecturii 3D (Wang et al., 2019a; Ma et al., 2020).

Acest capitol prezintă pachetul software NLRexpress - o colecție de predictori ce utilizează tehnici de învățare automată (ML), conceput pentru a identifica motivele de secvență specifice proteinelor de rezistență în domeniile CC, NBS și LRR și pentru a se putea scala scanării de seturi mari de date. Fluxul de lucru a fost folosit pentru a scana un set de ∼34,300 receptori NLR de la plante , iar motivele de secvență detectate au fost clusterizate și analizate pentru a identifica corelațiile între motive folosind tehnici de învățare nesupravegheată.

4.2 Rezultate și discuții

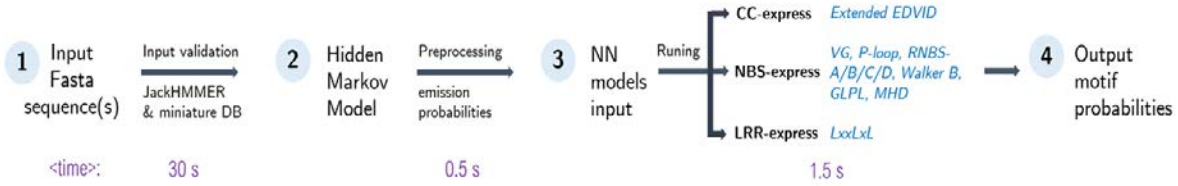
NLRexpress cuprinde o colecție de 11 clasificatori ce utilizează rețele neuronale antrenate pentru a detecta motivele individuale de secvență specifice receptorilor NLR de la plante. NLRexpress este organizat în 3 module de predicție, după cum urmează: (i) CCexpress - pentru motivul EDVID extinse; (ii) NBSexpress - pentru motivele *VG/hhGRE*, *P-loop*, *Walker-B*, *A/B/C/D-RNBS*, *motive GLP*L și *MHD*; (iii) LRRexpress - motive LxxLxL.

Fluxul de lucru NLR-express este prezentat în Fig. 4.1 și pornește de la secvențele de proteine introduse de utilizator în format FASTA. Primul pas constă în generarea datelor de input necesare modelelor de predicție, care includ profilele de variabilitate HMM. Predictorii sunt rulati in mod individual, fiecare dintre ei returnând ca output valoarea estimată a probabilității de a începe motivul dat pentru fiecare poziție a secvenței furnizate.

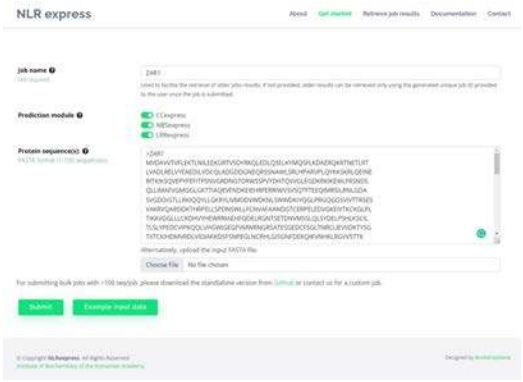
Multe dintre metodele de predicție dezvoltate în ultimii ani ce utilizează drept date secvențe de proteine, se bazează pe utilizarea caracteristicilor deduse din modele HMM, generarea cărora prezintă costuri computaționale ridicate din cauza dimensiunilor mari a bazelor de date de proteine utilizate precum Uniprot-20 sau Uniclust-30. Pentru a reduce drastic timpul de calcul al aceastei etape, a fost investigată utilizarea unei baze de date de căutare miniaturizate, orientată spre receptorii NLR pentru a obține cel mai bun compromis între performanță și timpul de execuție. Antrenarea celor 11 modele NN corespunzătoare fiecărui motiv a fost efectuată pornind de la un set de motive CC, NBS și LRR din receptori NLR de la plante, folosind o procedură de cross-validare pentru optimizarea parametrilor.

În cazul modulului NBSexpress, cele 9 motive prezintă un nivel de conservare semnificativ

(a) Fluxul de lucru NLRexpress



(b) Input utilizator (captură de ecran)



(c) Pagina de rezultate (captură de ecran)

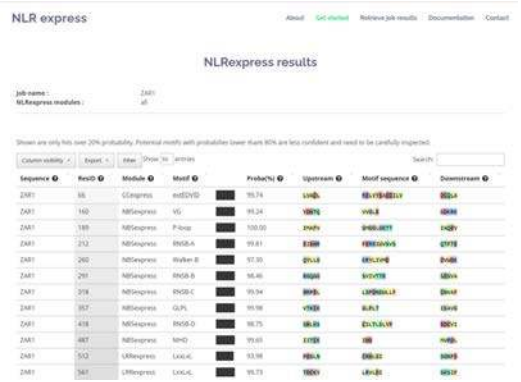


Fig. 4.1: **(a)** Principalele etape ale fluxului de lucru NLRexpress si timpul mediu de execuție al fiecărei etape. **(b)** NLRexpress - captură de ecran (https://nlrexpress.biochim.ro).

mai mare, fapt ce se reflectă în performanța predictorilor individuali cu scoruri de precizie și senzitivitate de peste 96%. Modulul LRRexpress prezintă un grad de precizie și senzitivitate echilibrate, cu scoruri F și G de 92% pe setul de testare. În continuare a fost investigată capacitatea de extrapolare pe alte clase de proteine care conțin domenii LRR. Pe setul PDB-LRR contînând ∼2000 motive LRR cu structura 3D cunoscută, descris anterior și în (Martin et al., 2020a), LRRexpress prezintă un scor general F1 de aprox. 92% atunci când se iau în considerare numai motivele LRR de mijloc, în timp ce doar ∼88% când sunt incluse segmentele marginale cu un grad de iregularitate mai ridicat.

Un test auxiliar a fost evaluarea comportamentului LRRexpress pe alte arhitecturi solenoide non-LRR care conțin motive de secventa similare cu *LxxLxL* dar care prezintă o organizare 3D diferită, acestea putând constitui o sursă de confuzie/predicții fals pozitive. Pentru aceasta, au fost utilizate cele 4 seturi de referință descrise anterior (ankyrin, pectate lyazes, trimeric și armadillo), ce contin câte 50 de secvențe din fiecare clasă. Pe aceste seturi, LRRexpress este capabil să clasifice corect motivele *L**L*L* atunci când apar în afara contextului arhitecturii LRR, cu aproape nicio estimare fals pozitivă, desi seturile conțin fiecare între 1000-2700 de motive LxxLxL asemănătoare celor din domeniile LRR.

Fluxul de lucru NLRexpress a fost utilizat pentru a scana un set de 34314 receptori NLR de la plante selectate la un nivel de redundanță de 90% identitate. Motivele individuale identificate au fost supuse unei proceduri de clusterizare fie pe baza metricilor de similitudine a aminoaciziilor, fie pe baza proprietăților fizico-chimice generale (hidrofobicitate, volum și sarcină electrostatică).

Având în vedere proximitatea celor 9 motive NBS în spațiul 3D - șapte dintre acestea participând activ la formarea situsului de legare ADP/ATP - analiza regiunilor individuale luate separat ar ascunde relațiile relevante care se formează la distanțe mari în secvență. Prin urmare, motivele NBS identificate au fost extrase, concatenate și clusterizate colectiv pe baza unei măsuri de similitudine a aminoacizilor la diferite praguri de identitate, așa cum este descris în secțiunea metode. La o limită de identitate de 55%, aproximativ 85% din secvențe sunt grupate în top zece cele mai mari grupuri. Cele mai invariante motive sunt, așa cum era de așteptat, cele implicate direct în legarea ADP/ATP, în special în regiunea P-loop, Walker-B, B-RNBS în subdomeniul NBD și motivele GLPL și MHD în subdomeniile ARC1 și ARC2, în timp ce caracteristici specifice fiecarui cluster se conturează în cazul motivelor: hhGRE și A/C/D-RNBS.

Deoarece motivele LRR prezintă cea mai mare variabilitate dintre motivele specifice receptorilor NLR, a fost analizata in continuare maniera in care variabilitatea motivului se distribuie în funcție de poziția în domeniul LRR. Un set de motive LxxLxL excizate (aproximativ 65000) au fost supuse clusterizarii folosind tehnci de învățare nesupervizată utilizând un embedding ce descrie proprietățile lor fizico-chimice (hidrofobicitate, sarcină și dimensiune). O predilecție puternică pentru motivele cu sarcină pozitivă este observată în toate clasele NLR pe regiunea primelor 4 segmente, motivul LRR cel mai frecvent fiind de tipul *LRxLxL*. In cazul receptorilor TNL, primul segment LRR prezintă o preferință pentru un mediu acid în poziția 1 a motivului, pe când în cazul receptorilor RNL, o preferință ridicată pentru tipul *LRxLxL* este observată în cazul primului și celui de-al treilea segment LRR.

4.3 Concluzii

Rezultatele prezentate în acest capitol sublinează faptul că NLRexpress - o colecție de estimatori conceputi pentru a detecta motivele CC, NBS și LRR specifice receptorilor NLR

de la plante - prezintă o performanță bună in cadrul testelor realizate și ar putea fi de folos în asistența diferitor tipuri de investigații experimentale ale receptorilor NLR. Pe lângă aplicațiile în realizarea de modele structurale 3D, datorită îmbunătățiri vitezei de calcul, acesta este fezabil pentru analiza secvenței la scară largă, cum ar fi scanarea unui întreg proteom al unei specii sau în analizele comparative de secvență pe seturi mari de ortologi.

Contribuții personale

Publicații

Publicații în calitate de autor principal

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

Publicații in calitate de co-autor

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

Pachete Software

- **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/

Participări conferințe

Prezentări orale

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

Prezentări tip poster

- 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)

Rezultatele prezentate în aceasta teză au beneficiat de suportul financiar din partea 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 și PN-III-P4-IDPCE-2020-2444 și din partea Academiei Române, Programul IBAR: ”Structural and systemic research in immunobiology and gerontomics”.

Mulțumiri

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.

Bibliografie

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.