SOLUȚII INFORMATICE PENTRU CERCETAREA VARIANTELOR ȘI TIPARELOR GENOMICE APLICÂND METODE DIN INGINERIA SISTEMELOR

Teză de doctorat – Rezumat

pentru obținerea titlului științific de doctor la
Universitatea Politehnica Timișoara
în domeniul de doctorat Ingineria Sistemelor

autor inf. Cristian-Grigore ZIMBRU

conducător științific Prof.univ.dr.ing. Ioan SILEA

luna 6 anul 2020

În lucrarea intitulată „Soluții informatice pentru cercetarea variantelor și tiparelor genomice aplicând metode din ingineria sistemelor” sunt prezentate o serie de metode pentru procesarea informațiilor genetice obținute după analiza secundară a datelor generate de dispozitivele/echipamentele de secvențiere[1].

Teza este structurată în trei părți având în total șase capitole (Fig. 1) repartizate astfel:

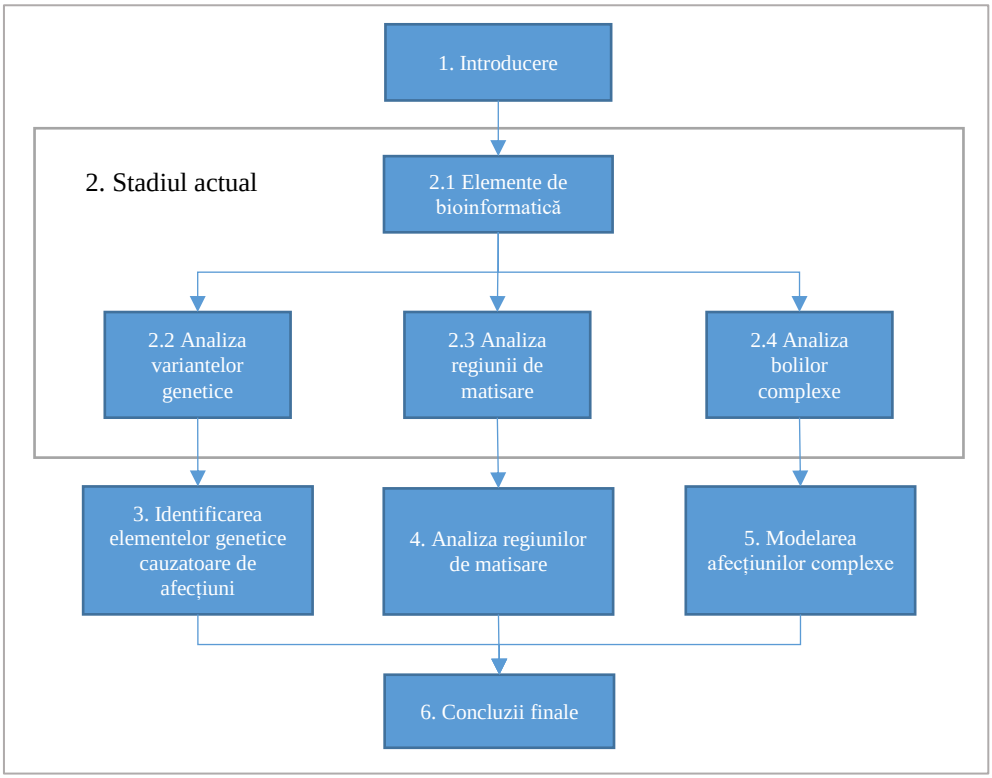

Fig. **Error! No text of specified style in document.** Arhitectura lucrării

1) prima parte conține capitolele 1 și 2 care au menirea de a prezenta tema și problematica tratată în lucrare;
2) partea principală formată din capitolele 3, 4 și 5 în care sunt prezentate soluțiile propuse

[1] HiSeq 2500, MiSeq, Sanger 3730xl.

pentru rezolvarea problemelor tratare în fiecare capitol;
3) ultima parte în care sunt prezentate concluziile finale și contribuțiile personale.

Scopul concret al lucrării constă în dezvoltarea și utilizarea unor metode informatice (*in silico*) care să fie capabile să identifice structurile genetice cauzatoare de afecțiuni. Contribuțiile țintesc evidențierea existenței variantelor și tiparelor genetice pentru: *(i)* afecțiuni monogenice, *(ii)* afecțiuni cauzate de matisare, dar și *(iii)* predicția afecțiunilor complexe (studiu de caz steatoza).

Ne vom referi în continuare, foarte sintetic, la aspectele esențiale care se regăsesc pe parcursul fiecărui capitol al tezei.

În primul capitol este prezentat domeniul tezei, oportunitatea elaborării lucrării, scopul și obiectivele urmărite. Orientarea tematică a lucrării vizează în principiu:

- Determinarea afecțiunii utilizând informații despre fenotip și genotip;
- Identificarea afecțiunii sau a variantei genetice cauzatoare pe baza semnăturii genetice;
- Reducerea numărului de variante genetice folosind anumite strategii de filtrare;
- Determinarea variației semnalului de matisare în urma modificărilor genetice;
- Identificarea regiunilor de matisare;
- Modelarea afecțiunilor complexe pe baza genotipului.

Pentru desfășurarea cercetărilor și realizarea lucrării s-au propus obiective majore cu subobiective aferente, prezentate mai jos.

1) Dezvoltarea unei metode sau a unui flux de lucru capabil să indice un număr redus de variante genetice care să explice caracteristicile fenotipului unui pacient, care include:
   a) Identificarea aplicațiilor și a lucrărilor care tratează problematica detecției afecțiunilor pe baza fenotipului sau pe baza genotipului;
   b) Reducerea numărului de afecțiuni posibile ale unui pacient în funcție de panelul de gene țintit și în funcție de genotipul acestuia;
   c) Identificarea strategiei optime pentru utilizarea predictorilor *in silico* pentru detecția variantelor genetice patogene;
   d) Extragerea intervalelor de toleranță pentru variantele genetice, în funcție de panelul de gene țintit.
2) Identificarea regiunilor de matisare folosind modele computaționale ale secvențelor de matisare și ale semnalelor activatoare și inhibitoare ale procesului aferent. Acesta presupune realizarea următoarelor sarcini:
   a) Recenzia aplicațiilor care identifică regiunile de matisare și intensitatea semnalelor de matisare;
   b) Analiza performanțelor metodelor pentru detectarea regiunilor de matisare și propunerea unor modele pentru identificare și extragerea acestora din regiunile intronice;
   c) Dezvoltarea unui algoritm care permite calcularea intensității semnalului de matisare;
3) Generarea unui model computațional pentru predicția unei afecțiuni complexe (steatoză) folosind un set de markeri genetici, este cel de-al 3-lea obiectiv și necesită trecerea prin:

a) Identificarea markerilor genetici relevanți pentru steatoză.
b) Studierea performanței modelelor generate de metodele utilizate în învățarea automată pentru predicția afecțiunilor complexe;
c) Propunerea unor modele de predicție a steatozei folosind metode care au la bază ansamblu de modele sau metode statistice.

Sub aspect practic, rezultatele obținute în urma cercetării ar trebui să faciliteze identificarea afecțiunii, în cazul pacienților nediagnosticați, sau identificarea variantelor genetice cauzatoare, în cazul pacienților a căror afecțiune este cunoscută dar nu este cunoscut motivul genetic. Concret, pentru primul obiectiv, seria de algoritmi (metode) va fi aplicată asupra genotipului unei persoane, mai precis asupra variantelor exonice, iar rezultatul acestora va consta în lista afecțiunilor posibile. Al doilea obiectiv are în vederea modificările genetice care afectează procesul de matisare. Metodele dezvoltate în această secțiune vizează identificarea modificărilor apărute în procesul de matisare. Ultimul obiectiv are ca finalitate generarea unor modele care ar trebui să fie capabile să prezică cu o acuratețe de peste 80% predispoziția genetică pentru steatoză sau chiar prezența acesteia.

Capitolul 2 este structurat în patru subcapitole. Primul subcapitol (2.1) are menirea de a prezenta câteva noțiuni introductive pentru a familiariza cititorul (inginer, informatician) cu elemente din biologia moleculară și bioinformatică. În cazul biologiei moleculare se explică câțiva termeni care vor fi folosiți destul de frecvent în conținutul tezei, precum; polimorfism uninucleotidic, locus, genă, genom de referință etc. După prezentarea termenilor genetici, se trece la prezentarea conceptelor de bioinformatică. Se descrie tehnologia utilizată pentru secvențierea ADN-ului, apoi se prezintă fluxul de lucru al analizei secundare. În ultima parte a subcapitolului se prezintă câteva informații din domeniul învățării automate și despre cum se evaluează performanța unui model generat prin astfel de metode.

Al doilea subcapitol (2.2) tratează problematica identificării genelor și a variantelor genetice cauzatoare de afecțiuni. Se cataloghează tipurile de afecțiuni conform [1] și este prezentată relativitatea informațiilor când vine vorba despre catalogarea unei variante genetice în baze de date precum ClinVar [2]. Pentru identificarea variantelor patogene se folosesc unelte de predicție *in silico*, care au asociate scoruri de patogenitate pentru fiecare variantă genetică. Se începe prin prezentarea predictorilor *in silico* precum: SIFT [3], CADD [4], PROVEAN [5] etc. De asemenea, sunt prezentate o serie de inițiative internaționale care se preocupă de procesarea și catalogarea informației genetice, precum proiectul *UK 100k Genomes* [6]. În continuare sunt prezentate metodele prin care se face filtrarea și selecția variantelor genetice. Un prim filtru, care se poate aplica acestor variante, este eliminarea celor care au o frecvență ridicată în populație, ideal cele din populația de unde provine pacientul. Pentru frecvența în populație a variantelor genetice se poate apela la baze de date precum 1000 Genomes sau Gnomad [7]. Tot în acest subcapitol sunt prezentate aplicații propuse în literatură pentru analiza variantelor genetice, precum Phevor2 [8], eXtasy [9] etc.

Subcapitolul (2.3) conține informații despre matisarea care are loc la nivelului pre-mARN-ului. La începutul subcapitolului se găsește o scurtă prezentare a procesului de matisare și a modului cum se analizează rezultatele în genetică. Deși, inițial variantele care nu făceau parte din exonii unei gene erau ignorate, în ultimii ani specialiștii au început să acorde mai

multă importanță variantelor care afectează procesul de matisare și să studieze rolul lor în manifestarea afecțiunilor. Pentru a îmbunătăți rezultatele în ceea ce privește regiunile de matisare, specialiștii apelează la instrumente software de predicție. Secvența de matisare conține trei componente importante: (1) punctul de excizie (branch point – BRS), tractul de pirimidine și situsul acceptor. Dintre acestea, BRS-ul și tractul de pirimidine nu sunt bine conservate [10]–[12], doar situl acceptor este conservat fiind format din bazele azotate AG. Pentru predicția regiunii de matisare sunt prezentate o serie de aplicații precum: *SplicePort*, *Automated Splice Site Analyses*, *MaxEnt* etc. De asemenea, mai sunt prezentate elementele reglatoare (*cis-acting*) ale matisării. Aceste elemente sunt reprezentate de secvențele dintre intron și exon, activatorii și inhibitorii intronici și activatorii și inhibitorii exonici.

În ultimul subcapitol (2.4) este prezentat un scurt studiu despre steatoză și markeri genetici care indică prezența acestei afecțiuni. Steatoza hepatică reprezintă acumularea de grăsime la nivelul țesutului hepatic și sunt cunoscute două forme ale acesteia (una datorată consumului de alcool, iar alta datorată altor factori – boala ficatului gras). În literatură sunt indicați o serie de factori genetici care pot predispune o persoană către această afecțiune. Principalul candidat pentru această afecțiune este polimorfismul rs738409 despre care se va discuta și în capitolul 5. Lucrarea [13] prezintă o analiză a descoperirilor genetice care sunt asociate cu boala ficatului gras, inclusiv rs738409. De asemenea, în acest subcapitol mai sunt prezentate metode din domeniul învățării automate utilizate în problematica afecțiunilor complexe. Inițial, studiile care tratau problema afecțiunilor complexe erau studii de asociere. O alternativă la această metodă o reprezintă metodele bayesiene sau metodele utilizate în învățarea automată. Arborii decizionali, metode de tip ansamblu, rețelele neuronale pot identifica ușor tiparele care apar în seturi de date multidimensionale.

Subiectele **capitolului 3** pun accent pe câteva metode care contribuie la îmbunătățirea detecției genelor, respectiv a variantelor genetice cauzatoare de afecțiuni. În majoritatea cazurilor, geneticienii au la dispoziție câteva tipuri de informații despre pacienți, precum fișierul cu variantele genetice, simptomele pacientului și eventual istoricul familiei. Pe baza acestor informații, medicii sunt nevoiți să identifice care sunt genele responsabile pentru afecțiunile de care suferă pacientul. Prima metodă prezentată presupune o asociere a caracteristicilor fenotipului cu elementele genotipului rezultând astfel o listă a posibilelor afecțiuni. Pentru realizarea unui sistem automat de identificare a genelor cauzatoare de afecțiuni, s-au folosit mai multe baze de date disponibile online. Pentru obținea listei de simptome s-a folosit baza de date Human Phenotype Ontology (HPO), iar pentru obținerea listei afecțiunilor s-au folosit bazele de date: Online Mendelian Inheritance in Man (OMIM), Orphanet și DECIPHER. Pentru ierarhizarea genelor, respectiv a variantelor, se propune utilizarea unor coeficienți de similaritate. După calcularea ponderilor și a coeficienților de similaritate, se aplică genotipul pacientului împreună cu caracteristicile fenotipului pentru a determina lista de posibile afecțiuni. Această metodă are rezultate satisfăcătoare dar nu complet.

Pentru a îmbunătății performanța metodei prezentate anterior, se poate apela la predictori *in silico*. Deși în literatura de specialitate se găsesc diverse teste de performanță a acestor predictori, din păcate nu sunt realizate pentru aceiași bază de date. Prin urmare, s-a efectuat un test de performanță care să aibă o bază de date comună. Baza de date utilizată a fost ClinVar, iar metricile folosite au fost precizia, scorul F_1 și media dintre specificitate și

sensibilitate. În ceea ce privește rezultatele, predictorii CADD și DANN au identificat cele mai multe variante patogene și au avut scoruri asociate pentru mai mult de 95% din setul de date. Dezavantajul acestor predictori a fost faptul că aveau o specificitate relativ scăzută. REVEL, MetaSVM și PolyPhen-2 HVAR au avut cele mai bune performanțe generale, calculate cu media aritmetică dintre specificitate și sensibilitate. Ca rutină pentru clasificarea corectă a SNP-urilor, variantele genetice patogene ar putea fi determinate cu instrumentele cu sensibilitate ridicată (CADD și DANN) și apoi se pot folosi predictorii echilibrați (REVEL, MetaSVM, PolyPhen) pentru a le prioritiza [14].

În subcapitolul 3.4 a fost tratată detecția erorilor de calitate și cantitate a variantelor genetice identificate în urma procesului de secvențiere. Se aplică o metodă pentru identificarea erorilor folosind intervale de toleranță. Intervalele prezentate au caracter orientativ, dar în același timp acestea pot fi folosite ca referință pentru detecția erorilor. Este indicat ca fiecare laborator să folosească astfel de metode pentru determinarea calității secvențierii. Pe lângă posibilele erori, aceste intervale pot semnala anumite cauze ale afecțiunii precum consangvinitate.

În **capitolul 4** au fost analizate elementele secvenței ARN care fac parte din procesul de matisare, anume secvența și semnalele de matisare. Aceste elemente au fost prelucrate din perspectiva ADN-ului.

Prin studiul realizat în subcapitolul 4.1 s-au identificat unele regiuni de matisare care au două sau mai multe secvențe care corespund regiunii de prindere (BRS) a spliceosomului. Experimentul a fost structurat pe două etape. Prima etapă a constat în definirea unui model pentru secvențele de matisare, bazându-ne pe lucrări din literatură [10] precum și pe rezultatele analizei a peste 11000 de secvențe de matisare din cromozom 21. A doua etapă a constat în folosirea definiției modelului generat la pasul 1 pentru a identifica pseudo-regiuni de matisare în regiunile intronice. Rezultatele obținute în prima etapă au indicat un grad de redundanță a regiunilor de prindere pentru anumiți exoni, fapt care se poate datora regiunii de pirimidine sau care poate avea o semnificație biologică - acesta necesitând o investigație mai amănunțită. Regiunile de prindere au fost adesea localizate în apropierea pozițiilor 16 și 28, în amontele exonului 3 prim. Modelul unei regiuni de matisare, în urma rezultatelor obținute în prima etapă, este format din: (1) o regiune de prindere care are modelul *YTnAy,* (2) o regiune acceptor AG și (3) o regiune de pirimidine care este formată din 75% baze azotate pirimidinice cu o lungime de cel puțin 17 baze azotate [15]. Folosind acest model, au mai fost identificate, *in silico,* secvențe intronice care sunt similare ca structură cu regiunile de matisare. Rolul biologic al acestor secvențe nu poate fi validat *in silico,* dar pot fi testate în experimente *in vitro* sau *in vivo* [16].

Studiul de la punctul 4.2 a constat în analiza secvențelor de matisare din baza de date *Homo Sapiens Splice Site Dataset* folosind diverse metode. În urma analizei s-au prezentat o serie de informații statistice despre structura secvențelor de matisare și s-au generat o serie de modele care au avut menirea să valideze aceste regiuni. Modelele prezentate inițial au avut la bază ecuații generate din structura regiunilor de matisare (ordinea nucleotidelor, tupli etc). Acuratețea predicției acestor modele a fost cuprinsă între 70% și 80%. Deși este un nivel decent, acuratețea țintită a fost în jurul valorii de 90%, performanța metodei *MaxEnt*. În ultima parte a capitolului a fost propusă o metodă pentru detecția regiunilor de matisare care are la bază

distanța față de secvențele vecine. Pentru calcularea distanței s-au analizat o serie de metode, iar cea aleasă în final a fost Needleman-Wunsch. Folosind această metodă s-a realizat o analiză computațională pentru determinarea unui optim al lungimii secvenței și un optim al numărului de vecini. Rezultatele au indicat că secvența ar trebui să fie de 20 de nucleotide iar numărul de vecini este 9. Folosind aceste valori s-a reușit obținerea unei acurateț̦e de 85.61%.

Scopul studiul din subcapitolul 4.3 a fost dezvoltarea unei metode care să permită adnotarea fișierelor VCF cu informații despre variațiile semnalului de matisare. În prima fază s-au adunat bazele de date cu secvențe de nucleotide considerate ca fiind semnale pentru procesul de matisare. Aceste secvențe, în forma inițială, au avut diverse criterii pentru calcularea intensității semnalului, ceea ce limita utilizarea lor simultană. Prin urmare, s-au dezvoltat o serie de ecuații care au permis determinarea intensității semnalului unei secvențe. În continuare, pentru calcularea diferenței de amplitudine între două secvențe, cea inițială și cea care conține modificarea genetică, se calculează media intensității vectorului de poziții nenule aferent secvenței. Calcularea amplitudinii se realizează atât pentru semnalul de amplificare a matisării, cât și pentru semnalul de inhibare a matisării. Direcția generală este dată de analiza acestor două componente. Validarea metodei s-a realizat pe o bază de date care conține secvențele genetice (normale și modificate) și care conține indicația comportamentului procesului de matisare. În plus, rezultatele corespund cu informațiile detaliate indicate de *Human Splicing Finder*. Metoda poate fi utilizată pentru filtrarea și prioritizarea variantelor genetice [17].

O serie de modele pentru predicție a steatozei pe baza unei liste de marker-i genetici, sunt prezentate în **capitolul 5**. Primul subcapitol prezintă materialele utilizare și lista de variante genetice folosite pentru realizarea modelelor. În al doilea subcapitol este realizată o analiză descriptivă a înregistrărilor din baza de date în funcție de variantele genetice țintite. Tot în acest subcapitol este realizată analiza corelațiilor dintre variantele genetice și gradul steatozei.

Al treilea subcapitol prezintă o serie de modele de predicție pentru gradele diferite ale acestei afecțiuni. Primul model investigat este cel generat cu ajutorul metodei *Stohastic-Gradient Descent* (SGD). Pentru predicția individuală a stadiului, modelul SGD are o acurateț̦e medie de aproximativ 70%, dar pentru predicția celor cinci stări ale steatozei, simultan, această valoare scade semnificativ, fiind de 25%. În continuare pentru predicția multi-clasă s-au folosit arborii decizionali. Pentru determinarea configurației optime, o serie de parametri au fost analizați. Scorul mediu al acurateței, pentru steatoză cu stadii multiple, a fost de 30%. Tot în acest subcapitol, s-a încercat modelarea stadiului steatozei folosind un ansamblu de arbori decizionali (*Random Forest*). Valoarea acurateței în acest caz a fost de 36%, marginal mai ridicată decât în cazul unui singur arbore decizional.

În subcapitolul patru gradul complexității este redus prin eliminarea stadiilor steatozei și înlocuirea acestora cu simpla prezență a afecțiunii. În această situație, folosind arborii decizionali s-a reușit obținerea unui model care a reușit o acurateț̦e de 91%, scorul mediu fiind de 81%. Aceeași performantă a fost obținută și în cazul ansamblului de arbori decizionali [18]. Precum a fost prezentat în literatură [19], [20], ambele metode au generat modele care indicau faptul că SNP-ul rs738409 este asociat cu steatoza.

Arborii decizionali au avut rezultate mai bune în cea de-a doua fază, când stadiul steatozei a fost redus. Acest lucru poate indica faptul că mărimea eșantionului folosit a fost prea

mic pentru a modela ieșirea, dar a fost optimă pentru a determina dacă patologia este prezentă. În plus, nu s-a ținut cont de diferența dintre subiecții de sex masculin și cei de sex feminin. Este posibil ca anumite SNP-uri să fie relevante, mai mult sau mai puțin, pe baza sexului. O configurație, cu cea mai bună precizie pentru arborii decizionali, a fost folosirea funcției Gini Index cu fasonare MDL și cu metodă de eșantionare aleatorie.

În cadrul subcapitolului cinci s-a dezvoltat o metodă care generează modele de predicție în funcție de frecvența apariției și „expertiza" fiecărui SNP. Folosind această metodă, starea steatozei fiind binară, s-a obținut o acuratețe de 82%. Acest model are o performanță mai scăzută decât cea a arborilor decizionali și cea a modelelor de predicție de tip ansamblu. Totuși, metoda prezintă câteva avantaje precum generarea hărților de vot care permit identificarea mult mai ușor a relațiilor dintre SNP-uri și afecțiune. De altfel, o serie de relații au fost evidențiate pentru steatoză. De exemplu, SNP-urile rs2167444 și rs7848 aflate pe gena SCD, in stare heterozigotă, par să aibă o afinitate pentru stadiul zero al steatozei; ori SNP-urile de pe gena ABCB4 par să indice o afinitate pentru starea 2 a steatozei.

În **capitolul 6** sunt prezentate concluziile, contribuțiile personale și direcții viitoare de dezvoltare. Principalele contribuții personale:

1. Dezvoltarea unei metode pentru determinarea variantelor genetice patogene în funcție de caracteristicile fenotipului și a variantelor detectate la pacienți;
2. Realizarea unui studiu pentru identificarea celei mai bune metode de folosire a predictorilor *in silico* în filtrarea variantelor genetice. Prezentarea rezultatelor și sugerarea unor strategii de prioritizare;
3. Propunerea unei metode pentru determinarea intervalelor de toleranță utilizată în detecția erorilor de secvențiere. Această metodă poate fi folosită și pentru identificarea rapidă a cauzelor unor afecțiuni, precum consangvinitatea;
4. Realizarea unui studiu asupra tuturor intronilor din cromozomul 21 pentru generarea unui model statistic al secvenței de matisare;
5. Implementarea unei metode informatice pentru detecția secvențelor de matisare parazite aflate în regiunile intronice;
6. Dezvoltarea unei metode pentru calcularea variației semnalului de matisare în cazul modificării secvenței ADN;
7. Identificarea secvențelor redundante pentru prinderea spliceosomului în regiunea de matisare;
8. Prezentarea unei metode pentru detecția regiunilor de matisare în funcție de distanța dintre secvența țintă și secvențele vecine folosind algoritmul Needleman-Wunsch;
9. Realizarea unui studiu statistic care prezintă structura regiunii de matisare;
10. Determinarea unor modele *in silico* folosind arbori decizionali pentru predicția prezenței steatozei, respectiv determinarea stadiului acesteia pe baza genotipului;
11. Dezvoltarea unei metode pentru predicția prezenței steatozei și a stadiului acesteia pe baza frecvenței variantelor genetice;
12. Implementarea și validarea metodelor anterior menționate folosind date din cadrul Centrului de Medicina Genomică (UMFT).

Totodată, în teză sunt referite, din cele **14** lucrări publicate de către autorul tezei ca prim-autor sau coautor, cele **5** care au validat rezultatele cercetărilor. Sintetic, după nivelul de impact la care s-au comunicat diversele rezultate științifice, gruparea lucrărilor publicate este următoarea:

- 2 lucrări în reviste indexate Web of Science (ISI) – cumulat **FI 7.2**
- 3 lucrări în volume ale unor manifestări științifice (proceedings) indexate Web of Science (ISI)
- 2 lucrări în reviste de specialitate indexate BDI (IEEE Xplore) – supuse pentru integrare în Web of Science (ISI)
- 7 lucrări în volumele unor manifestări științifice.

Consistența rezultatelor cercetărilor din lucrarea „Splice Site Pattern Analysis and Identification of Similar Sequences in the Deep Intron Areas of Human Chromosome 21" a condus la obținerea premiului 3 în cadrul conferinței EHB 2017, iar urmare a lucrării „Detection of high-risk intron areas that can cause splicing errors" a fost obținut un grant de tip *Young Scientist Fellowship*.

Teza are 146 de pagini, din care: 116 pagini structurate în 6 capitole, 10 pagini de bibliografie și 20 de pagini dedicate anexelor. Lucrarea conține 64 figuri și 151 de titluri bibliografice. O parte dintre contribuțiile prezentate au fost publicate în lucrări științifice la care autorul tezei este autor sau coautor, iar altele vor face obiectul unor viitoare lucrări și colaborări.

[1] Jonathan Pevsner, *Bioinformatics And Functional Genomics, 3rd edition*. 2015.

[2] M. J. Landrum *et al.*, "ClinVar: public archive of interpretations of clinically relevant variants," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D862-868, Jan. 2016, doi: 10.1093/nar/gkv1222.

[3] P. C. Ng and S. Henikoff, "SIFT: predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.

[4] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, Jan. 2019, doi: 10.1093/nar/gky1016.

[5] Y. Choi and A. P. Chan, "PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels," *Bioinformatics*, vol. 31, no. 16, pp. 2745–2747, Aug. 2015, doi: 10.1093/bioinformatics/btv195.

[6] M. Caulfield *et al.*, "The National Genomics Research and Healthcare Knowledgebase." Aug. 21, 2019, doi: 10.6084/m9.figshare.4530893.v5.

[7] K. J. Karczewski *et al.*, "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes," *bioRxiv*, p. 531210, Aug. 2019, doi: 10.1101/531210.

[8] M. V. Singleton *et al.*, "Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families," *Am. J. Hum. Genet.*, vol. 94, no. 4, pp. 599–610, Apr. 2014, doi: 10.1016/j.ajhg.2014.03.010.

[9] A. Sifrim *et al.*, "eXtasy: variant prioritization by genomic data fusion," *Nat. Methods*, vol. 10, no. 11, pp. 1083–1084, Nov. 2013, doi: 10.1038/nmeth.2656.

[10] K. Gao, A. Masuda, T. Matsuura, and K. Ohno, "Human branch point consensus sequence is yUnAy," *Nucleic Acids Res.*, vol. 36, no. 7, pp. 2257–2267, Apr. 2008, doi: 10.1093/nar/gkn073.

[11] D. A. Bitton *et al.*, “LaSSO, a strategy for genome-wide mapping of intronic lariats and branch-points using RNA-seq,” *Genome Res.*, p. gr.166819.113, Apr. 2014, doi: 10.1101/gr.166819.113.
[12] A. J. Taggart, A. M. DeSimone, J. S. Shih, M. E. Filloux, and W. G. Fairbrother, “Large-scale mapping of branchpoints in human pre-mRNA transcripts *in vivo*,” *Nat. Struct. Mol. Biol.*, vol. 19, no. 7, pp. 719–721, Jul. 2012, doi: 10.1038/nsmb.2327.
[13] S. Sookoian and C. J. Pirola, “Genetics of Nonalcoholic Fatty Liver Disease: From Pathogenesis to Therapeutics,” *Semin. Liver Dis.*, vol. 39, no. 2, pp. 124–140, May 2019, doi: 10.1055/s-0039-1679920.
[14] C. G. Zimbru, N. Andreescu, A. Albu, A. Chirita-Emandi, A. Stanciu, and M. Puiu, “Performance Evaluation of in Silico Predictors for the Classification of ClinVar Variants,” in *2019 E-Health and Bioengineering Conference (EHB)*, Nov. 2019, pp. 1–4, doi: 10.1109/EHB47216.2019.8969963.
[15] C. G. Zimbru *et al.*, “Splice site pattern analysis and identification of similar sequences in the deep intron areas of human chromosome 21,” in *2017 E-Health and Bioengineering Conference (EHB)*, Jun. 2017, pp. 145–148, doi: 10.1109/EHB.2017.7995382.
[16] Cristian Zimbru, Nicoleta Andreescu, Adela Chirita-Emandi, Antonius Stanciu, Ioan Silea, Maria Puiu, “Detection of high-risk intron areas that can cause splicing errors,” *Adv. Lect. Course Syst. Biol.*, p. p 74, Mar. 2016.
[17] C. G. Zimbru, A. Albu, N. Andreescu, A. Chirita-Emandi, and M. Puiu, “Determining Splicing Signal Variation in Humans by Analyzing the Regulatory Splicing Motifs,” in *2019 E-Health and Bioengineering Conference (EHB)*, Nov. 2019, pp. 1–4, doi: 10.1109/EHB47216.2019.8969983.
[18] C. G. Zimbru, N. Andreescu, A. Chirita-Emandi, I. Silea, M. Puiu, and M. D. Niculescu, “Analysis of decision tree performance in predicting the relationship between a scored outcome and multiple single nucleotide polymorphisms,” in *2017 E-Health and Bioengineering Conference (EHB)*, Jun. 2017, pp. 57–60, doi: 10.1109/EHB.2017.7995360.
[19] P. J. Meffert *et al.*, “The PNPLA3 SNP rs738409:G allele is associated with increased liver disease-associated mortality but reduced overall mortality in a population-based cohort,” *J. Hepatol.*, vol. 68, no. 4, pp. 858–860, Apr. 2018, doi: 10.1016/j.jhep.2017.11.038.
[20] S. Grimaudo *et al.*, “PNPLA3 rs738409 C>G variant predicts occurrence of liver-related events and death in non-alcoholic fatty liver,” *Dig. Liver Dis.*, vol. 51, pp. e5–e6, Feb. 2019, doi: 10.1016/j.dld.2018.11.045.