Soluții informatice pentru cercetarea variantelor și tiparelor genomice aplicând metode din ingineria sistemelor

Teză destinată obținerii
titlului științific de doctor inginer
la
Universitatea Politehnica Timișoara
în domeniul Ingineria Sistemelor
de către

drd.inf. Cristian-Grigore ZIMBRU

Conducător științific:	Prof.univ.Dr.ing. Ioan SILEA
Referenți științifici:	Conf.univ.Dr.ing. Mihnea MOISESCU
	Prof.univ.Dr.med. Maria PUIU
	Prof.univ.Dr.ing. Radu-Emil PRECUP

Data susținerii tezei: 3 iulie 2020

Seriile Teze de doctorat ale UPT sunt:

1. Automatică
2. Chimie
3. Energetică
4. Ingineria Chimică
5. Inginerie Civilă
6. Inginerie Electrică
7. Inginerie Electronică și Telecomunicații
8. Inginerie Industrială
9. Inginerie Mecanică
10. Știința Calculatoarelor
11. Știința și Ingineria Materialelor
12. Ingineria sistemelor
13. Inginerie energetică
14. Calculatoare și tehnologia informației
15. Ingineria materialelor
16. Inginerie și Management

Universitatea Politehnica Timișoara a inițiat seriile de mai sus în scopul diseminării expertizei, cunoștințelor și rezultatelor cercetărilor întreprinse în cadrul Școlii doctorale a universității. Seriile conțin, potrivit H.B.Ex.S Nr. 14 / 14.07.2006, tezele de doctorat susținute în universitate începând cu 1 octombrie 2006.

Bd. Republicii nr. 9, RO-300159 Timișoara, TM
Tel./fax +40 (256) 403 823
e-mail: editura@edipol.upt.ro

Cuvânt înainte

Teza de doctorat a fost elaborată pe parcursul activității în cadrul Departamentului de Automatică și Informatică Aplicată al Universității Politehnica Timișoara și, totodată, activității în cadrul Centrului de Medicină Genomică (CMG) al Universității de Medicină și Farmacie „Victor Babeș” din Timișoara.

Prezenta lucrare este utilă cercetătorilor și medicilor care sunt interesați să aprofundeze metodologia identificării variantelor genetice cauzatoare de afecțiuni. Deși dezvoltarea unei metode universale pentru identificarea tuturor polimorfismelor genetice maligne reprezintă un scop dezirabil, momentan este utopic. În schimb, în lucrare sunt dezvoltate metode care permit (1) reducerea numărului de variante candidate, (2) identificarea variantelor care afectează procesul de matisare (variante menționate teoretic dar rar considerate în practică) și (3) identificarea relațiilor dintre variantele genetice și afecțiuni.

Doresc să adresez mulțumiri deosebite conducătorului de doctorat Prof.univ.Dr.ing. Ioan SILEA pentru sprijinul acordat pe întreaga perioadă a doctoratului, și mai ales pentru acceptarea îndrumării unei lucrări într-un domeniu multidisciplinar. Le mulțumesc colegilor de colectiv: Dr.ing. Antonius STANCIU, Dr.ing. Adriana ALBU și Dr.ing. Loredana STANCIU pentru îndrumarea, sprijinul moral și sfaturile pe care mi le-au oferit.

Îmi exprim profunda recunoștință față de doamna Prof.univ.Dr.med. Maria PUIU care mi-a prezentat oportunitatea realizării prezentei lucrări și mi-a oferit acces la platforma de cercetare pe care dânsa o coordonează. Le mulțumesc colegelor din cadrul Centrului de Medicină Genomică: Dr.med. Nicoleta ANDREESCU și Dr.med. Adela EMANDI-CHIRIȚĂ pentru sprijinul și îndrumarea acordate pe perioada dezvoltării tezei. Doresc să-i mulțumesc domnului Dr.med. Mihai D. NICULESCU (SUA) pentru observațiile critice și pentru timpul dedicat dezvoltării lucrărilor.

De asemenea, mulțumesc colegilor din cadrul Departamentului de Automatică și Informatică Aplicată (UPT) și din cadrul Centrului de Medicină Genomică (UMFT), a căror nume nu este menționat explicit, dar care m-au ajutat în această perioadă.

Mulțumesc soției mele, Cristina, care mi-a fost alături de la începutul doctoratului, pentru susținerea necondiționată și pentru observațiile critice cu privire la redactarea lucrării. Mulțumesc părinților și bunicilor pentru tot sprijinul acordat.

Timișoara, mai 2020 Drd.inf. Cristian-Grigore ZIMBRU

Făcând referire la cunoscuta frază a biologului *Theodosius Dobzhansky* despre evoluție, marchez acest moment evolutiv prin a-l dedica fiului meu Alexandru cu mesajul: *„când scriam această frază tu împlineai un anișor*".

ZIMBRU, Cristian-Grigore

Soluții informatice pentru cercetarea variantelor și tiparelor genomice aplicând metode din ingineria sistemelor

Teze de doctorat ale UPT, Seria X, Nr. YY, Editura Politehnica, 200Z, 146 pagini, 64 figuri, 15 tabele.

Cuvinte cheie: predicție genotip-fenotip, model de predicție matisare, modelare genetica steatoză, bioinformatică.

Rezumat,
Scopul general al lucrării constă în dezvoltarea și utilizarea unor metode informatice (*in silico*) care să permită identificarea elementelor genetice cauzatoare de afecțiuni. Contribuțiile țintesc trei categorii de probleme pentru determinarea afecțiunilor genetice la oameni, anume afecțiuni monogenice, afecțiuni cauzate de matisare și predicția afecțiunilor complexe. Prima parte a cercetării expune un set de metode care se concentrează asupra afecțiunilor cauzate de elemente genetice singulare. A doua parte prezintă un set de metode pentru identificarea regiunilor care favorizează matisarea și o metodă pentru calculare intensității semnalului pentru acest proces. Ultima parte a tezei se concentrează asupra dezvoltării unor modele, generate folosind metode din învățare automată, care identifică prezența steatozei.

Rezumat

Motivați de importanța informaticii aplicate în cercetarea sistemelor biologice, în particular cele pentru ființa umană și, în mare măsură, pentru depistarea și evoluția transmiterii unor caracteristici prin ADN/ARN, unele modalități noi de abordare și strategii aplicate sunt prezentate în această teză. Aflat la intersecția dintre multidisciplinar și multidomeniu, acest tip de cercetare are o dinamică complexă. Rezultatele obținute se bazează pe culegerea și prelucrarea sistematică a unor cantități mari de date și informații folosind modele și tehnologii informatice, dar și pe metode matematice de actualitate. Contribuțiile țintesc trei categorii de probleme pentru determinarea afecțiunilor genetice la oameni, anume afecțiuni monogenice, afecțiuni cauzate de matisare și predicția afecțiunilor complexe, având ca studiu de caz steatoza.

Prima parte a cercetării conține un set de metode care se concentrează asupra afecțiunilor cauzate de o singură genă sau de anumite variante genetice uninucleotidice. Prima metodă aplicată presupune identificarea caracteristicilor genotipului și asocierea acestora cu caracteristici ale fenotipului. Concret, se determină spațiul relațiilor dintre cele două tipuri de caracteristici, urmată de calcularea coeficienților de similaritate între acestea și caracteristicile fenotipului pentru fiecare afecțiune prezentă în baza de date. Rezultatul îl reprezintă lista posibilelor afecțiuni sau lista variantelor genetice care cauzează afecțiunea. Pentru îmbunătățirea rezultatelor, se aplică două metode de filtrare a variantelor genetice. Prima metodă presupune filtrarea pe baza predictorilor *in silico*, utilizând recomandările analizei de performanță a acestora. Cea de a două metodă presupune identificarea intervalelor de toleranță (folosind ecuațiile obținute în urma analizei statistice) și eliminarea fișierelor, respectiv a variantelor, care nu corespund acestor intervale. Folosirea metodelor permite identificarea afecțiunii unui pacient nediagnosticat sau identificarea cauzei genetice pentru un pacient diagnosticat.

A doua parte a cercetării are ca țintă analiza procesului de matisare din punct de vedere informatic, folosind baze de date genomice. În prima etapă s-au aplicat modele statistice pentru identificarea regiunilor de matisare, respectiv a regiunilor de prindere a spliceosomului, a sitului acceptor și a segmentului de pirimidine. Ulterior, s-a aplicat algoritmul *Needleman-Wunsch* pentru calcularea similarității secvenței țintă cu regiunile de matisare valide. Totodată, pentru identificarea unor situri reale, s-a dezvoltat o metodă capabilă să integreze toate semnalele de matisare. Această metodă indică diferențele de intensitate survenite în urma modificărilor genetice. Pentru analiză și validare s-a folosit setul de date *Homo Sapiens Splice Site Dataset*, aplicația *Human Splicing Finder* și alte resurse din literatură.

Ultima parte a lucrării este dedicată determinării unor modele matematice prin care se poate face predicția afecțiunilor complexe. Afecțiunea avută în vedere este steatoza hepatică. În primă fază, au fost generate modele folosind metodele consacrate în cadrul învățării automate, precum arborii decizionali sau *Random Forest*. Motivul pentru numărul limitat de metode se datorează nevoii de extragere a informațiilor relevante despre contribuția fiecărui SNP la afecțiunea finală. Prin urmare, metodele care generează modele de tipul *black-box* (rețele neuronale) au fost excluse din studiu. Pe lângă metodele „clasice”, s-a prezentat o metodă care se bazează pe diferențele dintre apariția SNP-urilor în populația afectată. De asemenea, este prezentată și metoda pentru obținerea parametrului dedicat filtrării zgomotului. În cele din urmă sunt prezentate hărțile de valori și modalitățile prin care se poate face predicția afecțiunii pe baza genotipului.

Mențiuni

O parte din cercetarea prezentată în lucrare a fost realizată în Centrul de Medicină Genomică al Universității de Medicină și Farmacie „Victor Babeș” din Timișoara, finanțat de POSCCE Project ID: 1854, cod SMIS: 48749, contract 677/09.04.2015;
și a fost susținută parțial de proiectul POC *Nutrigen*, SMIS: 104852, contract 91/09.09.2016, ID P_37-684.

Cuprins

Abrevieri

Termen	Descriere
ADN	Acid dezoxiribonucleic
AN	Adevărat negativ
AP	Adevărat pozitiv
ARN	Acid ribonucleic
ASSA	*Automated Splice Site Analyses* (aplicație)
BAM	Fișier rezultat în urma alinierii secvențelor de ADN (format fișier)
BED	Format specific stocării intervalelor unor secvențe (format fișier)
BRS	Regiunea de prindere a spliceosomului
BWA	Aplicație de aliniere care folosește algoritmul *Burrows-Wheeler*
CADD	*Combined Annotation Dependent Depletion* (predictor)
ClinVar	Bază de date care conține informații despre patogenitatea variantelor genetice
CSV	Fișier cu valori despărțite prin virgulă (C*omma Separated Values*)
DANN	Predicție pe baza CADD cu rețele neuronale (predictor)
dbNSFP	Bază de date care conține variante genetice adnotate cu o serie de predictori
dbSNP	Baza de date publică (arhivă) de SNP-uri
EBI	Institutului European de Bioinformatică
EMBL	*The European Molecular Biology Laboratory*
ESE	*Exonic splicing enhancers*
ESS	*Exonic splicing silencers*
ExAC	*The Exome Aggregation Consortium* (bază de date)
F1	Scorul pentru măsurarea acurateței F1
FASTQ	Fișier care conține secvențe de ADN rezultate în urma secvențierii
FATHMM	*Functional Analysis through Hidden Markov Models* (predictor)
FN	Fals negativ
FP	Fals pozitiv
GWAS	*Genome-Wide Association Studies*
HG37	GRC37, Genomul uman de referință versiunea 37
HG38	GRC38, Genomul uman de referință versiunea 38
HPO	Ontologia Fenotipului Uman (bază de date)
HS3D	*Homo Sapiens Splice Site Dataset* (bază de date)
HSF	*Human Splicing Finder* (aplicație)
IGSR	*The International Genome Samples Resource*
in silico	Indică faptul că analiza a fost realizată computațional
in vitro	Indică faptul că analiza a fost realizată în afara organismului viu
in vivo	Indică faptul că analiza a fost realizată într-un organism viu
ISE	*Intronic Splicing Enhancers*
ISS	*Intronic Splicing Silencers*
LSV	Regiune de matisare locală
M_ss	Media aritmetică între sensibilitate și specificitate
MAF	*Minor Allele Frequency*
mARN	Molecula matură de ARN mesager
M-CAP	*Mendelian Clinically Applicable Pathogenicity* (predictor)
MDD	*Maximal Dependence Decomposition*
MDL	*Minimum Description Length*
MED	*Maximum Entropy Distribution*

NAFLD	Boala ficatului gras non-alcoolic
NGS	Secvențiere de nouă generație (*Next Generation Sequencing*)
NIH	*National Institutes of Health*
NN	Rețele neuronale
OMIM	*Online Mendelian Inheritance in Man* (bază de date)
PCA	*Principal Component Analysis*
PGU	Proiectul Genomului Uman
Phevor	*Phenotype Driven Variant Ontological Re-ranking Tool*
PolyPhen	*Polymorphism Phenotyping* (predictor)
PROVEAN	*Protein Variation Effect Analyzer* (predictor)
PWM	*Position Weight Matrix*
REVEL	*Rare Exome Variant Ensemble Learner* (predictor)
ROC	*Receiver Operating Characteristic*
SBB	*Splicing Binary Base Model*
SBT	*Splicing Binary Tuple Model*
SGD	*Stochastic-Gradient Descent*
SIFT	*Sorting Intolerant from Tolerant* (predictor)
SNP	Varianta genetică uninucleotidică
snRNP	Riboproteine nucleare mici
splicing	Proces biologic prin care intronii sunt eliminați din structura pre-mARN-ului rezultând mARN-ul
SVM	*Support Vector Machine*
TSO	Kit de secvențiere *TruSight One*
UCSC	*University of California in Santa Cruz*
UK 100K	Proiectul 100.000 genomuri din Marea Britanie
VCF	Formatul fișierului care conține variante genetice
VEP	*Variant Effect Predictor*
WES	Secvențierea întregului exom
WGS	Secvențierea întregului genom

Convenții de scriere

Cuvintele scrise italic reprezintă denumiri ale unor aplicații (*Exomizer, Polymorphism Phenotyping*) sau reprezintă termeni care nu au o traducere standardizată în limba română.

Listă de figuri

Listă de tabele

Listă de anexe

1. INTRODUCERE

1.1. Domeniul tezei

De aproape șapte decenii cercetătorii încearcă să descopere totalitatea mecanismelor care stau la baza procesării informației din moleculele de acid dezoxiribonucleic (ADN) și cum influențează micile modificări ale acestei molecule comportamentul celulei umane. De la descoperirea structurii ADN-ului (în 1953, de către *Watson* și *Crick*, pe baza unei imagini realizate de către *Rosalind Franklin*) și până în prezent s-au modelat o serie de procese celulare, dar acest lucru nu s-a realizat pentru întreaga celulă umană (în totalitatea ei). Motivele neîndeplinirii acestui obiectiv se datorează volumului enorm de informații care sunt încapsulate în ADN, dar și din cauza cunoașterii limitate a proceselor biologice care decodifică aceste informații.

Paragraful publicat în lucrarea [1] din revista *Nature Genetics* (2019) surprinde foarte bine problematica din domeniul geneticii, respectiv domeniul bioinformaticii în acest moment. Prin urmare, paragraful este prezentat în întregime:

> *„The human genome comprises more than 3 billion base pairs. Recent technological advances have increased the mechanistic understanding of genome biology to an incredible degree. However, the complexity and sheer amount of information contained in DNA and chromatin remain roadblocks to complete understanding of all functions and interactions of the genome. Connecting genotype to phenotype, predicting regulatory function, and classifying mutation types are all areas in which harnessing the vast genomic information from a large number of individuals can lead to new insights. However, working in this large data space is challenging when conventional methods are used. Therefore, new and innovative approaches are needed in genome science to enrich understanding of basic biology and connections to disease."* [1]

Până nu demult, cercetătorii aveau la dispoziție o serie de echipamente limitate din punct de vedere al performanței pentru realizarea studiilor genomice de mare amploare. Totuși, avansul metodelor pentru secvențierea[1] ADN-ului, însoțită de scăderea substanțială a costurilor per analiză (Fig. 1.1), concomitent cu creșterea puterii de calcul a echipamentelor, a favorizat o integrare masivă a tehnologiei în domeniul biologiei moleculare. Acest fapt a condus la o explozie științifică și la crearea unor domenii adiacente, multidisciplinare, cum ar fi bioinformatica sau *system biology*, puse uneori sub umbrela domeniului biologiei computaționale.

Bioinformatica a apărut, în prima sa formă, la începutul anilor 1990 odată cu creșterea volumului de informații biologice. Aceasta, în cel mai simplu mod la acea vreme, a fost considerată ca fiind o denumire dată pentru simpla filtrare a datelor biologice. Desigur, după ce tipul informațiilor și metodele utilizate pentru extragerea datelor au început să fie tot mai diverse și mai complexe, rolul de filtrare a datelor a devenit cumva, implicit, un rol secundar. În prezent, bioinformatica îndeplinește o serie de roluri, de la colectarea și stocarea datelor până la dezvoltarea uneltelor

[1] În limba română acest proces poate fi întâlnit ca secvențializare sau secvențiere. Utilizarea termenului diferă în funcție de regiunea geografică sau în funcție de grupul de cercetare. Până în prezent acest termen nu a fost definit în DEX.

software pentru prelucrarea, interpretarea și vizualizarea acestora. Ca majoritatea disciplinelor, rolul bioinformaticii este acela de a utiliza informațiile avute la dispoziție pentru a contribui la cunoașterea diferitelor procese. Odată ce au fost extrase suficiente informații, pe baza lor se vor dezvolta diferite modele care prezic, cu un grad acceptabil, efectul acestora.

Prezenta lucrare se concentrează asupra unor probleme din ramura genomicii, care se preocupă de studierea ADN-ului și de asocierile dintre afecțiuni și modificările genetice. Acum câteva decenii, medicii luau în considerație o posibilă explicație genetică doar în cazul apariției unor modificări majore ale trăsăturilor unei persoane sau dacă afecțiunea putea fi urmărită în arborele genealogic. Toate acestea s-au schimbat după finalizarea Proiectului Genomului Uman (PGU). Proiectul, alături de alte proiecte care i-au urmat, a scos în evidență faptul că factorii genetici acționează în tandem și, susținuți de factorii de mediu, au un rol extrem de important în modul cum se manifestă afecțiunile complexe. Aceste afecțiuni precum: diabetul, cancerul, bolile cardiovasculare sau obezitatea, reprezintă grupul cel mai frecvent întâlnit pe întregul mapamond. Totuși, o parte dintre rezultatele obținute în domeniul genomicii le permite specialiștilor să aplice terapii din ce în ce mai țintite pentru tratarea acestor tipuri de afecțiuni. Desigur, scopul final este ca medicina să ofere un tratament personalizat fiecărui pacient, în funcție de semnătura sa genomică.

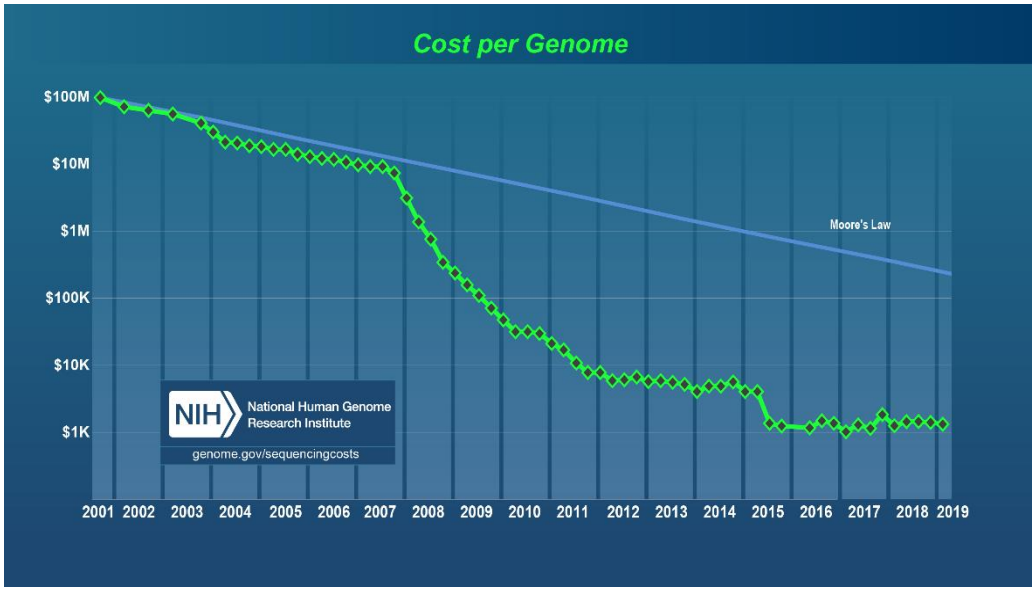


Fig. 1.1 Evoluția costului secvențierii unui genom în comparație cu legea lui *Moore* conform *National Institutes of Health*[2]

Este important, totuși, să menționăm că aceste terapii sunt rezultatele unor studii genomice și descoperirile din acest domeniu, care, în majoritatea cazurilor, presupun echipe mari de cercetători, perioade îndelungate de timp și o finanțare considerabilă. În general, medicamentele rezultate din aceste proiecte au nevoie de câtiva ani până când ajung să fie utilizate clinic.

[2] https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Este evident că nu doar elementele genetice sunt răspunzătoare pentru afecțiunile complexe, dar înțelegerea profundă a genomului ne permite să modelăm mecanismele celulare, care, la rândul lor, ne permit să explicăm cum comunică toate componentele organismului între ele și cum se poate deteriora starea de sănătate a unei persoane.

1.2. Scopul și obiectivele cercetării

În cadrul lucrării s-a urmărit atingerea unei serii de problematici importante care survin în procesul de diagnosticare a pacienților din punct de vedere genetic, și totodată, s-a dorit modelarea afecțiunilor pe baza markerilor genetici. Scopul general al lucrării constă în dezvoltarea unor metode informatice (*in silico*) care să identifice variantele genetice cauzatoare de afecțiuni. Desigur, identificarea tuturor variantelor este un obiectiv măreț, mai degrabă ideal spre utopic, în momentul redactării prezentei lucrări. Prin urmare, se tratează probleme specifice, precum variante aflate în zona de matisare sau variante asociate cu anumite caracteristici ale fenotipului.

Principalele teme tratate sunt:

- Determinarea afecțiunii utilizând informații despre fenotip și genotip;
- Identificarea afecțiunii sau a variantei genetice patogene 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.

1.2.1. Obiective principale

Pentru realizarea lucrării s-au propus următoarele obiective principale:

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;
2. Identificarea regiunilor de matisare folosind modele computaționale ale secvențelor de matisare și ale semnalelor activatoare și inhibitoare ale procesului aferent;
3. Generarea unui model computațional pentru predicția unei afecțiuni complexe (steatoză hepatică) folosind un set de markeri genetici.

1.2.2. Obiective specifice

Pentru Obiectivul 1 s-au propus următoarele obiective specifice:

1. Identificarea aplicațiilor și a lucrărilor care tratează problematica detecției afecțiunilor pe baza fenotipului sau pe baza genotipului;
2. Reducerea numărului de afecțiuni posibile ale unui pacient în funcție de panelul de gene țintit și în funcție de genotipul acestuia;
3. Identificarea strategiei optime pentru utilizarea predictorilor *in silico* pentru detecția variantelor genetice patogene;

4. Extragerea intervalelor de toleranță pentru variantele genetice, în funcție de panelul de gene țintit.

Pentru Obiectivul 2 s-au propus umătoarele obiective specifice:

1. Recenzia aplicațiilor care identifică regiunile de matisare și intensitatea semnalelor de matisare;
2. Analiza performanțelor metodelor pentru detectarea regiunilor de matisare și propunerea unor modele pentru identificare și extragerea acestora din regiunile intronice;
3. Dezvoltarea unei algoritm care permite calcularea intensității semnalului de matisare;

Pentru Obiectivul 3 s-au propus umătoarele obiective specifice:

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

1.3. Aplicabilitatea practică vizată de rezultatele cercetării

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 principal, seria de algoritmi (metode) va putea fi aplicată asupra genotipului unei persoane, mai precis asupra variantelor exonice, iar rezultatul acestora va consta în lista de afecțiuni posibile. Această listă va fi filtrată cu o metodă care ține cont de performanța predictorilor *in silico* și care va elimina variantele de artefact din fișier. Totodată, lista afecțiunilor va fi prioritizată în funcție de caracteristicile fenotipului.

Metoda vizată de al doilea obiectiv principal are în vedere un subset de variante genetice, mai precis modificările genetice care afectează procesul de matisare. Spre deosebire de variantele uninucleotidice clasice (exonice), variantele genetice care se află în regiunile de matisare pot afecta într-un mod mai semnificativ structura mARN-ului. Prin metodele dezvoltate în această secțiune se vizează identificarea și semnalarea acestor variante.

Prin ultimul obiectiv se dorește generarea unor modele care, pentru predicția cauzei, integrează mai mulți markeri genetici. Modelele obținute ar trebui să fie capabile să prezică, la o acuratețe de peste 80%, predispoziția genetică sau chiar prezența steatozei hepatice. Totodată, dacă markerii genetici furnizați sunt relevanți, metoda dezvoltată va permite modelarea și a altor afecțiuni complexe.

1.4. Conținutul lucrării

Lucrarea este structurată în trei părți având în total șase capitole împărțite astfel: (1) prima parte conține capitolele 1 și 2 care au menirea de a prezenta tema și problematica tratată în lucrare; (2) partea specială formată din capitolele 3, 4 și 5 în care sunt prezentate soluțiile propuse pentru rezolvarea problemelor tratate în

fiecare capitol; (3) ultima parte în care sunt prezentate concluziile finale și contribuțiile personale. În următoarele rânduri se va detalia conținutul fiecărui capitol pe scurt și se va prezenta modul în care se poate parcurge lucrarea (Fig. 1.2).

În primul capitol este prezentat domeniul tezei și oportunitatea redactării lucrării. De asemenea, se indică obiectivele și structura lucrării.

Capitolul al doilea este structurat pe patru subcapitole. Primul subcapitol (2.1) are menirea de a prezenta câteva noțiuni introductive pentru a-l familiariza pe cititor cu elemente din biologia moleculară și bioinformatică. Al doilea subcapitol (2.2) tratează problematica identificării genelor și a variantelor genetice cauzatoare de afecțiuni. De asemenea, tot aici sunt prezentate metode propuse în literatură și se va indica, pentru o parte dintre acestea, implementarea software. Al treilea subcapitol (2.3) conține informații despre matisarea care are loc la nivelului pre-mARN-ului. Sunt prezentate o serie de metode utilizate pentru identificarea regiunilor de matisare, dar și importanța acestor regiuni din punct de vedere clinic. În ultimul subcapitol (2.4) sunt prezentate metode din domeniul învățării automate utilizate în problematica afecțiunilor complexe. Tot aici se prezintă profilul genetic al steatozei.

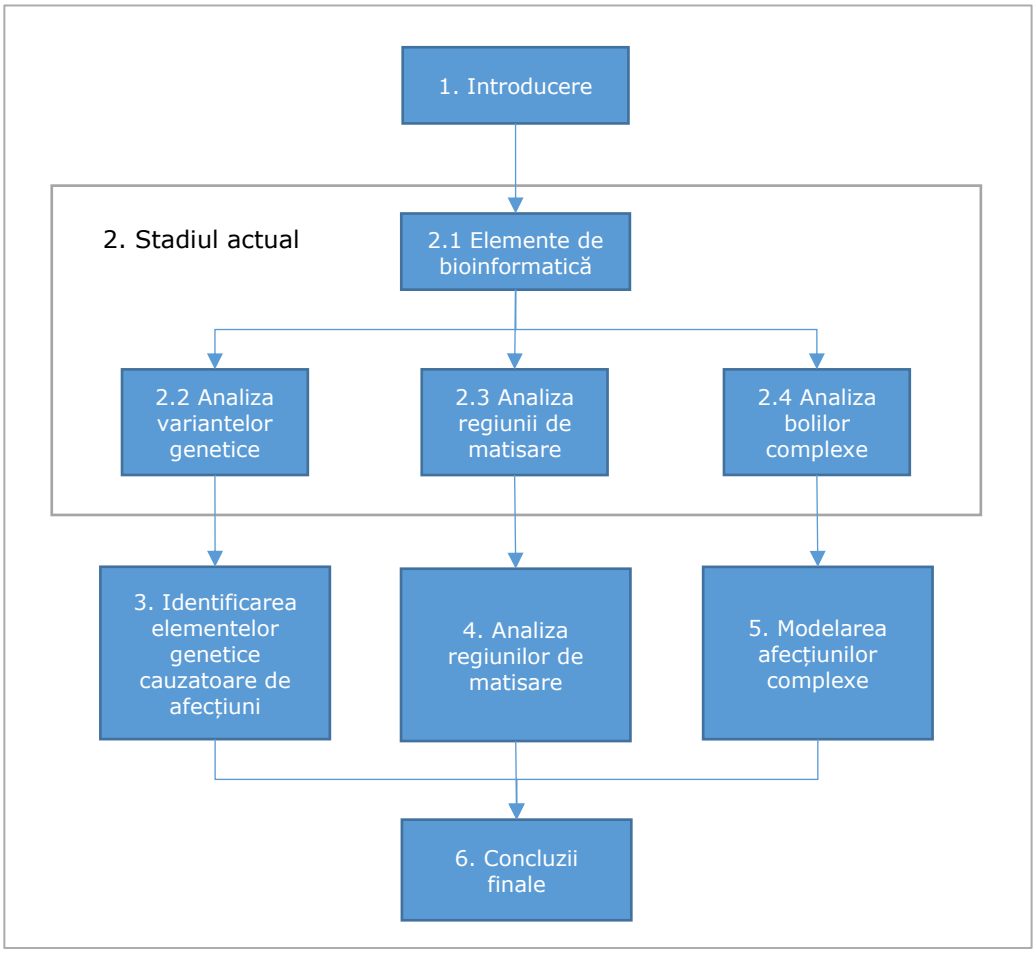


Fig. 1.2 Parcurgerea lucrării

În prima parte a capitolului al treilea sunt prezentate metode prin care se asociază caracteristicile fenotipului cu elementele din genotip rezultând astfel posibilele afecțiuni. De asemenea, este prezentat nivelul de asociere dintre gene și termenii standardizați corespunzători caracteristicilor fenotipului. Tot în cadrul acestui capitol este prezentată analiza performației predictorilor *in silico* și sunt indicate metode de utilizare a acestora pentru identificarea variantelor cauzatoare de afecțiuni. O altă metodă prezentată este identificarea afecțiunii pe baza semnăturii genotipului.

În capitolul patru este prezentată o analiza a regiunilor de matisare extrase din secvența genomului uman (hg19). Componența acestor regiuni este analizată statistic și sunt prezentate câteva modele pentru identificarea acestora. *Homo Sapiens Splice Site*, este o altă bază de date folosită pentru analiza regiunilor de matisare, pe care s-au efectuat testele diferitelor modele de predicție. În a doua parte a capitolului este prezentată o metodă pentru calcularea variației semnalului de matisare care apare în cazul unei variante genetice.

Capitolul al cincilea tratează bolile complexe și modelarea acestora, cu scopul predicției ulterioare, folosind variantele genetice. Pentru generarea modelelor de predicție s-a folosit steatoza ca exemplu. Metodele folosite pentru generarea modelelor au constat în clasificatoare multi-clasă sau în metode de tip *Ensemble*. În final este prezentată o metodă care are la bază frecvența stadiului afecțiunii sau simpla prezență a acesteia.

În ultimul capitol sunt prezentate concluziile, contribuțiile, publicațiile și posibile direcții de cercetare pentru viitor.

În funcție de subteme, pentru parcurgerea capitolului 3, Identificarea elementelor genetice cauzatoare de afecțiuni, este recomandată parcurgerea, dacă este cazul, a subcapitolelor 2.1 și 2.2. Pentru capitolul 4, Analiza regiunilor de matisare (*Splicing*), sunt sugerate subcapitolele 2.1 și 2.3. Iar pentru parcurgerea capitolului 5 se recomandă subcapitolele 2.1 și 2.4.

Teza are 146 de pagini, din care 116 pagini de conținut, 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 lucrările științifice la care autorul tezei este autor sau coautor.

2. STADIUL ACTUAL AL CERCETĂRII ÎN DOMENIU

2.1. Noțiuni preliminare

2.1.1. Genă

Cuvântul genă provine din limba greacă și este derivat din cuvintele *genesis* (care înseamnă naștere) sau *genos* (care denotă originea). În cadrul școlii de biologie clasică gena era considerată element de transmitere a unor caractere ereditare. Conform definiției biologiei moleculare actuale [2] respectiv NIH, gena este unitatea fundamentală și funcțională a eredității. Este important de reținut că, deși unele gene reprezintă tiparul pentru crearea unor molecule denumite proteine, o parte dintre acestea nu au acest rol. Cele din urmă vor rămâne la nivel de ARN și vor servi reglării anumitor procese celulare.

2.1.2. Polimorfism uninucleotidic (SNP)

Un polimorfism uninucleotidic (*Single Nucleotide Polymorphism*, SNP) reprezintă schimbarea unei baze azotate la o anumită poziție (*locus*) a unui genom, în contextul tezei, al genomului uman. Frecvența polimorfismelor diferă între grupurile de indivizi, fie că aceștia sunt grupați după etnie sau sunt grupați în funcție de regiunea geografică de unde provin. Aceste variații genetice sunt studiate pentru că unele dintre ele pot fi asociate cu modificări ale căilor metabolice, prin urmare pot fi cauza unor modificări ale fenotipului. Din totalul modificărilor genetice, SNP-urile sunt responsabile pentru 90% din variațiile genetice umane [3]. Desigur, nu toate SNP-urile sunt asociate cu afecțiuni sau cu risc pentru anumite afecțiuni. De exemplu, polimorfismele care codifică același aminoacid (sinonim) sunt considerate neutre. De asemenea, sunt și variante genetice care nu sunt sinonime, dar care par să nu aibă nici o manifestare fenotipică descoperită. Găsirea unei asocieri între un SNP sau un grup de SNP-uri și o afecțiune este esențială, deoarece permite medicilor să identifice pacienții cu un anumit risc pentru o anumită afecțiune de la începutul anchetei. Până în prezent, cercetătorii au efectuat numeroase studii de asociere pentru diferite afecțiuni, cum ar fi bolile cerebrale [4], cancerul [5] și diabetul [6], [7].

2.1.3. Tipuri de fișiere

Fișierul **FASTQ** conține înregistrări ale secvențelor cu nucleotide aferente fragmentelor de ADN. Fiecare înregistrare conține patru rânduri:

1. Locația unde se află pe lamă fragmentul de ADN;
2. Secvența de nucleotide (A, T, C, G);
3. Se repetă rândul 1;
4. Calitatea citirii fiecărei nucleotide de la poziția 2.

Acest fișier se generează imediat după ce procesul de secvențiere s-a încheiat. În el sunt prezente toate secvențele de ADN pe care secvențiatorul[3] le citește. În general demultiplexarea pacienților se face odată cu generarea fișierelor FASTQ.

Fișierul **BAM** este generat în urma alinierii (mapării) secvențelor din fișierul FASTQ la genomul de referință. Informațiile aferente unei înregistrări (secvențe) sunt multiple, precum: numele secvenței (inclusiv cromozomul, poziția de start, calitatea mapării și secvența CIGAR), calitatea secvenței, informații despre mapare și etichete specifice.

Fișierul **VCF** conține variantele genetice identificate în urma mapării secvențelor la genomul de referință. Fiecare variantă genetică conține coordonatele mapării (cromozom, poziție), ID-ul *rs*, nucleotida în genomul de referință, nucleotida identificată la pacient, calitatea, filtrul (PASS, Low_DP etc.) și câmpul INFO conține multiple informații precum: acoperirea, zigozitatea etc.

2.1.4. Genom de referință

După finalizarea procesului de secvențiere și extragerea fișierelor FASTQ, urmează procesul de aliniere a secvențelor la genomul de referință. Este important ca termenul *genom de referință* să nu fie înțeles greșit, anume ca fiind un *genomul normal* (inexistent) sau cel al unei persoane care nu prezintă nicio afecțiune. Acest genom de referință este format prin suprapunerea mai multor genomuri ale membrilor aceleași specii, iar dacă este cazul, la unele *locus*-uri, selectarea variantei considerate ancestrală. În prezent, pentru ființele umane, se folosește versiunea hg19 (GRCh37) făcându-se tranziția către hg38 (GRCh38). Prima versiune a genomului de referință a fost rezultatul Proiectului Genomului Uman (PGU) [8].

2.1.5. Tehnologii de secvențiere

În genomică, termenul de secvențiere face referire la un proces chimic care permite identificarea bazelor azotate din structura moleculei de ADN. Acest proces de secvențiere a fost realizat prima dată de către *Frederick Sanger*, în 1955, care a realizat secvențierea completă a aminoacizilor din insulină. Același cercetător, în 1977, propune secvențierea ADN-ului în [9], metodă care este folosită și în prezent ca standard în diagnosticarea clinică. Metoda este cunoscută sub denumirea de secvențiere *Sanger*. Platforma permite secvențierea fragmentelor de ADN cu o lungime de până la 800 de baze azotate. Avantajul secvențierii Sanger îl reprezintă acuratețea sa de 99.9% pentru identificarea corectă a bazelor azotate. Dezavantajul este consumul excesiv de resurse (timp și bani) pentru realizarea analizei. În cazul proiectelor de cercetare, unde este necesară secvențierea unui panel de gene sau a întregului genom, această metodă nu este ideală.

Dezavantajul menționat anterior a dus, în 2005, la apariția metodelor de secvențiere paralelă (secvențiere prin sinteză sau prin ligare) care au permis analizarea regiunilor de ADN de dimensiuni mari prin fragmentarea moleculei în bucăți mici. Secvențierea de nouă generație (*Next Generation Sequencing, NGS*) reprezintă tehnologia care a revoluționat cercetarea în domeniul genomicii. Pe piața internațională se găsesc mai multe platforme de secvențiere, dintre care Roche

[3] Reprezintă aparatul care efectuează citirea secvențelor de ADN. Exemplu: Illumina HiSeq 2500, Illumina MiSeq.

(2004), Illumina (2006), Ion Torrent (2010) [10]. Dintre acestea cea mai răspândită este platforma Illumina, compania fiind responsabilă pentru generarea a peste 80% din volumul de datele genomice [11]. Particularitățile fiecărei metode de secvențiere sunt explicate în detaliu în capitolul *DNA Sequencing Technologies* din [11].

În plus, tehnologia de secvențiere NGS a permis cercetătorilor să obțină cantități mari de date din genom, dar și din transcriptom. Datele obținute sunt, de obicei, aliniate la un genom de referință utilizând diferite unelte software [12], [13]. După aliniere, se extrag o serie de variante, inclusiv SNP-uri, care trebuie filtrate și clasificate de către specialiștii în biologie moleculară.

2.1.6. Fluxul de lucru pentru extragerea variantelor genetice

Analiza întregii secvențe de ADN facilitează descoperirea unui număr ridicat de variante genetice. Procesul de descoperire a variantelor presupune anumite etape precum filtrarea secvențelor de ADN și alinierea acestora la genomul de referință. Filtrarea inițială a secvențelor se poate realiza după diferite criterii, precum calitatea indicată de către aparatul de secvențiere. Prin urmare, secvențele care au calitatea medie sub Q20 sau secvențele care au un anumit număr de baze azotate sub Q20 se pot îndepărta. Desigur, acestea se pot refolosi ulterior pentru confirmarea unor variante genetice dacă au o acoperire[4] mică. De asemenea, secvențele se pot filtra și după lungimea fragmentelor de ADN. Dacă sunt secvențe care au lungimi prea mici (de exemplu, 20 bp) aceasta se poate exclude, deoarece o secvență cu un număr redus de baze azotate poate fi mapată la poziții multiple ale genomului de referință, ceea ce duce la creșterea artificială a adâncimii de secvențiere, iar în final pot apărea artefacte care mimează existența unei variante genetice.

Pentru alinierea secvențelor sunt disponibile o serie de unelte software dintre care cele mai cunoscute sunt BWA [14], Bowtie2 [15] și STAR [16]. Este important de precizat că BWA nu are funcții pentru alinierea secvențelor divizate (mARN) ceea ce înseamnă că nu poate fi folosit pentru secvențiere de tipul *RNA-seq*. Fapt valabil și pentru prima versiune a aplicației Bowtie. În cadrul lucrării se face referire doar la secvențele de ADN, ceea ce înseamnă că oricare dintre aceste unelte poate fi folosită.

După realizarea alinierii, Fig. 2.1, se obține fișierul BAM care conține toate secvențele de ADN mapate la genomul de referință. Următorul pas îl reprezintă marcarea secvențelor duplicate (pentru a evita interferențele). Din acest punct se pot extrage primele variante genetice cu aplicații dedicate precum GATK [17] sau *Freebayes* [18].

În prezent, pentru prioritizarea variantelor există mai mulți predictori disponibili, cum ar fi *PolyPhen* [19] și SIFT [20]. Un alt instrument mai complex este dezvoltat în cadrul proiectului *Ensemble Variant Effect Predictor* [21], care integrează o serie de instrumente, inclusiv cele menționate anterior. Instrumentele software sunt îmbunătățite continuu și extinse cu module noi de către cercetători [22].

[4] Adâncimea de secvențiere sau numărul de fragmente care confirmă varianta genetică

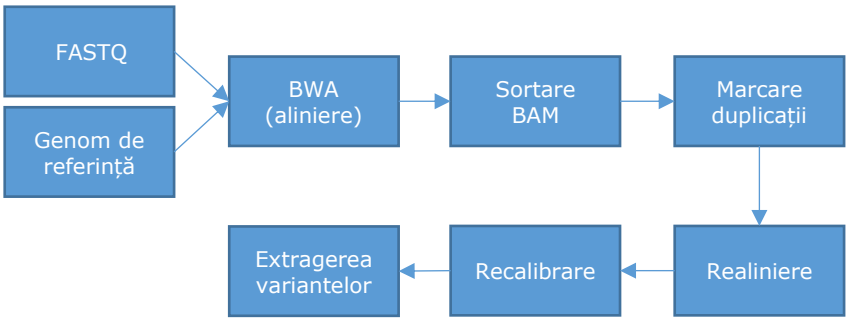


Fig. 2.1 Procesul de extragere a variantelor genetice

2.1.7. Învățarea automată

Învățarea automată este integrată tot mai mult în aplicațiile tehnice comerciale și în proiectele de cercetare. Acest domeniu se regăsește undeva la intersecția dintre statistică și tehnologia informației. Plaja de integrare este foarte diversificată, acoperind de la domeniile medicale (precum predicția diagnosticului bazat pe imagini) până la platformele electronice (precum cele ale rețelelor sociale). De obicei aceste metode sunt aplicate pentru prezicerea anumitor caracteristici care se bazează pe seturi mari de date.

Inițial, aplicațiile „inteligente” au fost realizate folosind reguli stricte. De exemplu, în cazul clasificărilor se bazau pe reguli *if-else* sau *switch-case*. Dezavantajul acestor sisteme este că funcționează corect doar dacă sistemul rămâne constant pe toată perioada de utilizare. Eventualele modificări ale sistemului atrag după sine și modificarea modelului care stă la baza lui. Avantajul modelelor obținute folosind metode de învățare automată este faptul că își pot schimba comportamentul în timpul funcționării pe baza datelor de intrare.

Învățarea automată are mai multe categorii de metode în funcție de caracteristica pe care o vrem detaliată. De exemplu, dacă ne referim la modul de învățare, algoritmii sunt de două feluri: învățarea supervizată și învățarea nesupervizată. Învățarea supervizată presupune ca algoritmul să învețe din setul de date care este format din caracteristicile de intrare și caracteristicile care urmează a fi prezise. Inițial algoritmul generează un model cu un set de antrenament. După generarea modelului, acestuia i se vor putea furniza caracteristicile unor înregistrări noi, care nu au fost folosite în procesul de obținere al modelului. Iar pe baza ieșirilor acestor date de intrare se va măsura performanța modelului. În general, metodele de învățare supervizată sunt bine înțelese. Partea consumatoare de timp este pregătirea setului de antrenament. În cazul metodelor de învățare nesupervizată, setul de date furnizate conține doar intrări, urmând ca modelul să extragă caracteristicile de interes pentru utilizator.

Spațiul problemelor pentru învățarea supervizată, în general, se poate împărți în două tipologii. Avem probleme de clasificare unde scopul este predicția clasei unei caracteristici, clasa fiind o valoare discretă. Al doilea tip de probleme este reprezentat de către regresii care presupun determinarea unei valori continue. În cazul învățării supervizate, generarea modelelor se realizează cu ajutorul unui set de date după care poate fi folosit pentru a prezice caracteristicile pentru un set de valori de intrare

nemaiîntâlnite până în acel punct. Este foarte important să avem în vedere ca modelul să nu se supra-potrivească (*overfitting*) pe setul de antrenament. De obicei, un model supra-potrivit va avea performanțe mult mai bune pe setul de antrenare decât pe setul de test. Setul de test reprezintă datele care sunt folosite strict pentru verificarea performanței modelului, nicidecum pentru generarea modelului. Când modelul are performanța mai slabă pentru setul de antrenament în comparație cu setul de test, înseamnă că modelul este sub-potrivit (*underfitting*). În această situație modelul a generalizat superficial informația. Situația ideală este când modelul are performanța asemănătoare pentru ambele seturi de date. Acest lucru se realizează prin ajustarea unor parametri specifici ai algoritmului de învățare. În general pentru determinarea performanței unui clasificator se apelează la o serie de metode, descrise în subcapitolul 2.1.8.

2.1.8. Evaluarea performanței unui model de clasificare

Evaluarea performanței unui model care clasifică înregistrările pe baza unor caracteristici este puțin mai dificilă decât evaluarea performanței unui model care are la bază o regresie. Desigur, înainte de evaluarea performanței și generarea modelului este important să împărțim datele în două subseturi. Unul dintre aceste subseturi va fi folosit pentru antrenarea (generarea) modelului, fiind cel mai consistent ca număr de înregistrări (de exemplu 80% din setul inițial), iar celălalt subset va fi folosit la final, când modelul este în formă finală, pentru evaluarea performanței.

Pentru evaluarea performanței unui model putem apela la validarea încrucișată. Aceasta ne va permite să măsurăm performanța chiar din stadiul de generare a modelului. În principiu, setul de antenare este împărțit într-un număr de subseturi. Aceste subseturi, cu excepția unuia, se vor folosi pentru crearea modelului. După ce modelul este creat, se verifică indicatorii de performanță cu subsetul exceptat. Procesul se repetă până când toate subseturile au fost folosite pentru verificarea performanței. La final, performanța fiecărui parametru este media valorilor obținute anterior. Este important ca fiecare parametru pentru măsurarea performanței să fie interpretat corect. Dacă măsurăm acuratețea, aceasta reprezintă numărul de clase identificate corect raportate la numărul total de valori din set. Calcularea acurateței se realizează folosind ecuația din (2.1-1), unde termenii ecuației reprezintă elementele matricei de contingență.

$$Acc = \frac{AP + AN}{AP + AN + FP + FN} \qquad (2.1\text{-}1)$$

Totuși, acuratețea unui model, singură, nu reprezintă o măsură de încredere deoarece aceasta poate crește artificial. Un exemplu în acest sens poate fi verificarea cu un set dezechilibrat de test (care conține prea multe înregistrări din aceeași clasă). Mai mult, dacă metodei i s-a furnizat un set dezechilibrat de date, favorizând o clasă, aceasta va tinde să se supra-potrivească modelului. Prin urmare, dacă avem un model care favorizează o anumită clasă, iar setul de test are un număr majoritar de înregistrări cu aceeași clasă, atunci valoarea acurateței crește artificial. Pentru evitarea acestui scenariu este recomandat să folosim un set care are un număr egal sau aproape egal de înregistrări din fiecare clasă. Desigur, sunt numeroase situații unde acest lucru nu este posibil. Ca să rezolvăm acest impas, se poate folosi matricea de contigentă, care conține informații despre valorile reale și cum au fost prezise de

către model. În cazul unui clasificator binar, matricea este formată din două rânduri și două coloane. Conținutul matricei este reprezentat de: (1) valorile clasei pozitive care au fost prezise corect (AP), (2) valorile clasei pozitive care au fost prezise greșit (FN), (3) valorile clasei negative care au fost prezise corect (AN) și (4) valorile clasei negative prezise greșit (FP). O altă pereche interesantă de indicatori pentru evaluarea performanței, este reprezentată de sensibilitate și specificitate, prezentate în ecuațiile de la (2.1-2).

$$Specificitatea = \frac{AN}{AN + FP}$$
$$Sensibilitatea = \frac{AP}{AP + FN} \qquad (2.1\text{-}2)$$

Pe lângă informațiile despre comportamentul clasei pozitive respectiv a celei negative, sensibilitatea și specificitatea, ne poate ajuta să selectăm un prag optim pentru parametrii modelului. O altă metodă prin care putem urmări performanța modelului, benefică și pentru ajustarea parametrilor, este curba ROC (*Receiver Operating Characteristic*). Curba se obține prin reprezentarea sensibilității în raport cu 1-specificitatea. Această metodă compară performanța modelului cu performanța unui model care ar face predicții aleatorii. Performanța este determinată prin valoarea ariei de sub curba generată, exemplu Fig. 2.2 Curba ROC pentru o metodă *Random Forest* în predicția steatozei (Knime Studio).

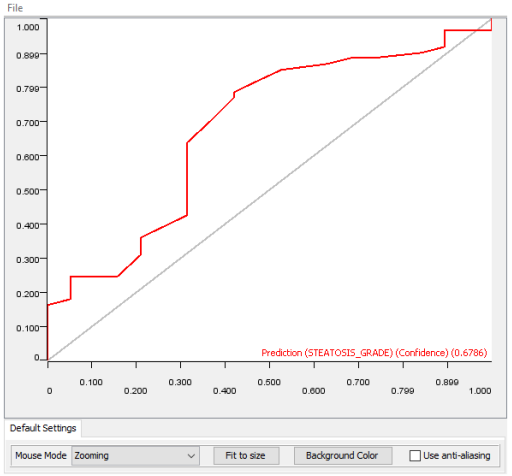


Fig. 2.2 Curba ROC pentru o metodă *Random Forest* în predicția steatozei (*Knime Studio*).

2.1.9. Arbori Decizionali

Modelele de tipul arborilor decizionali sunt utilizate atât pentru problemele de clasificare, cât și pentru cele de regresie. Desigur, pentru regresie arborii decizionali nu au o caracteristică lină precum regresia liniară sau regresia polinomială. Când vine vorba despre regresii, valoarea pe care o vor returna aceste modele va reprezenta media înregistrărilor din ultimul nod al căii prin arbore. Ca principiu fundamental, arborii decizionali funcționează ca o serie imbricată de instrucțiuni decizionale. O reprezentare grafică a unui arbore care identifică prezența steatozei se regăsește în

Fig. 2.3. Această metodă are o serie de avantaje. Unul dintre acestea este interpretabilitatea ușoară a modelelor, deoarece modelele sunt de tipul *white-box*. Un alt avantaj este faptul că modelele se pot genera chiar și cu date puține, în comparație cu alte metode folosite în învățarea automată. Desigur, această metodă are și dezavantaje. Modelele de tipul arborilor decizionali pot suferi modificări semnificative doar prin schimbarea câtorva înregistrări din setul de antrenare. Deși pot învăța din datele furnizate foarte ușor, arborii decizionali tind să se supra-potrivească (*overfit*). Prin urmare, performanța lor poate fi inferioară altor metode.

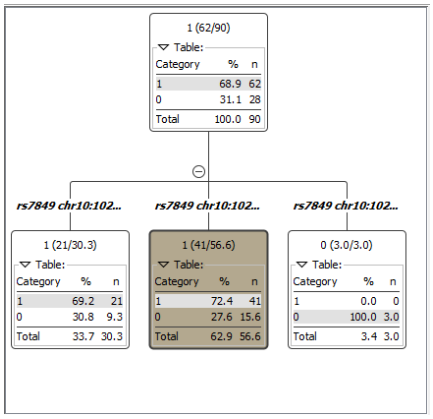


Fig. 2.3 Arbore decizional pentru predicția prezenței steatozei

Această metodă este utilizată într-o multitudine de domenii, în medicină, pentru diferite aplicații. Unele modele au fost generate pentru predicția pierderii în greutate [23] sau pentru calcularea costurilor de vaccinare [24] sau în genomică pentru detectarea secvențelor promotor [25]. Utilizarea acestor metode în genetică a început înaintea finalizării Proiectului Genomului Uman. *Salzberg* și colaboratorii săi au folosit arborii decizionali pentru predicția genelor din secvențele de ADN, metodă prezentată în [26]. O altă aplicație, denumită *Morgan*, a avut o acuratețe de 95%, cu o specificitate pentru bazele codante de 83% și o sensibilitate de 79%. Alt exemplu din domeniul genomicii (care folosește arborii decizionali) este prezentat în [27], unde metoda este folosită pentru a prezice funcția genică. Încă un exemplu, de data aceasta pentru predicția expresiei genice, este prezentat în [28]. Pe lângă predicția elementelor care țin de structura ADN-ului, arborii decizionali au mai fost folosiți pentru stabilirea studiului afecțiunilor. Un exemplu este cancerul, unde arborii decizionali s-au folosit pentru selectarea genelor care ajută la identificarea afecțiunii [29] sau pentru clasificarea acesteia [30].

În teză, arborii decizionali vor fi folosiți pentru extragerea informațiilor despre relația dintre o listă de SNP-uri și steatoză. De asemenea, au fost generate modele pentru testarea capacității de predicție a arborilor decizionali în identificarea stadiului steatozei. Toate acestea sunt tratate în capitolul 5, unde se mai analizează o serie de parametri care vor determina performanța acestei metode.

2.1.10. Metode de tip ansamblu (*Ensemble*)

Aurélain Géron, în cartea sa despre învățarea automată [31], afirmă că răspunsul agregat al 1000 de persoane, pentru o întrebare, este mai bun decât răspunsul unui expert. Desigur, această afirmație este *fuzzy*, dar în acest mod a dorit *Géron* să explice modelele ansamblu. De fapt, acest model reprezintă o mulțime de sub-modele generate de diverse metode de învățare automată care votează ieșirea finală a ansamblului. Nu este obligatoriu ca modelele din grup să fie generate de metode diferite. Este posibil să avem aceeași metodă cu ajutorul căreia să generăm un set de modele diferite și care vor acționa ca un grup. Un exemplu cunoscut este metoda *Random Forest*, care are ca structură primară arborele decizional. Practic, cu această metodă generăm o mulțime de arbori decizionali diferiți. Sunt diferiți deoarece fiecare arbore este antrenat cu un subset de date din setul de antrenare. Parametrii ajustabili ai arborilor decizionali se regăsesc și la metoda *Random Forest,* dar cea din urmă mai are câțiva parametri suplimentari (de exemplu, numărul de arbori decizionali din *Forest*).

2.1.11. Unelte software utilizate pentru învățarea automată

Odată cu dezvoltarea domeniului învățării automate, au fost dezvoltate o serie de unelte și biblioteci care să vină în sprijinul cercetătorilor și al companiilor pentru creșterea productivității. În cadrul capitolului 5 pentru analiza datelor și extragerea modelelor au fost folosite: platforma *KNIME Analytics* și *Anaconda* cu pachetele *scikit-learn*, *JupiterLab*, *NumPy*, *pandas*, *matplotlib*.

Knime Analytics este o platforma care are la bază programarea vizuală. Utilizatorul își creează metoda de analiză a datelor prin plasarea unor blocuri pe o planșă, metodă pe care o va executa ulterior. Cu ajutorul acesteia se pot procesa atât valori numerice, cât și imagini. Se pot realiza de la simple activități (precum procesarea de semnale) până la realizarea de rețele neuronale complexe.

Scikit-learn este un proiect *open source* care este în permanență dezvoltat și actualizat. Proiectul are o comunitate foarte activă, ceea ce ajută la rezolvarea rapidă a posibilelor probleme întâmpinate în dezvoltare. Biblioteca dispune de o serie de algoritmi folosiți în învățarea automată și este utilizată atât în mediul academic, cât și în industrie. Spre deosebire de *Knime*, în acest caz, utilizatorul este nevoit să aibă cunoștințe de programare într-un limbaj precum *Python*.

JupyterLab este un mediu interactiv care permite dezvoltarea de programe într-un navigator de Internet. *JupyterLab* este succesorul lui *Jupyter Notebook*, noutatea fiind integrarea unor facilități precum modulul de management al fișierelor.

NumPy este un pachet esențial pentru buna funcționare a bibliotecii *Scikit-Learn*. Acest pachet conține o mulțime de funcții matematice care permit lucrul cu structuri multidimensionale, operații de algebră liniară, generatoare de numere pseudoaleatoare etc.

Folosind limbajul de programare *Python*, biblioteca *Matplotlib* permite realizarea unor reprezentări grafice. Aceasta conține funcții de afișare de la simple grafice cu bare până la histograme și matrice de corelație. Este o unealtă utilă pentru vizualizarea rezultatelor. Mediul *JupyterLab* permite afișarea reprezentărilor, generate cu *Matplotlib,* chiar în navigatorul de Internet.

Pandas este o bibliotecă folosită când se lucrează cu tabele. Aceasta are ca structură de bază tipul *DataFrame* care se comportă asemănător unui tabel. Biblioteca vine cu o mulțime de funcții pentru prelucrarea datelor.

2.2. Elemente specifice analizei variantelor genetice

Diagnosticul și tratamentul bolilor este datorat, în egală măsură, ultimelor tehnologii folosite în laborator, cât și înțelegerii afecțiunilor de către specialiști. În particular, pentru înțelegerea cauzelor unei afecțiuni geneticienii caută variantele genetice care pot fi patogene. Pentru interpretarea și identificarea corectă a variantelor genetice sunt necesare o serie de progrese atât în diagnosticare, cât și în modelarea biologică a celulei umane. Problema aplicării etichetei de variantă patogenă este îngreunată de o serie de elemente, precum (1) factorii de mediu sau factorii epigenetici, (2) complexitatea rețelelor de interacțiuni moleculare ș.a.m.d.

Factorii de mediu reprezintă o cauză majoră pentru care o mare parte dintre afecțiuni nu au asociate elemente genetice cauzatoare. În timp, unii cercetători au catalogat anumite variante genetice ca fiind patogene doar ca, după doar câțiva ani, alți cercetători să infirme acest lucru. Evoluția încadrării unei variante genetice poate fi urmărită cel mai bine în baza de date *ClinVar* [32]. Un exemplu de polimorfism cu o astfel de traiectorie este reprezentat de varianta catalogată în dbSNP [33] cu identificatorul rs121912998, care a fost catalogată în 2015 ca fiind patogenă pentru cardiomiopatie, dar în 2016 și 2017 a fost considerată benignă.

Benign (Jul 4, 2017)	criteria provided, single submitter • Invitae Variant Classification Sherloc (09022015)	clinical testing	Cardiomyopathy dilated with woolly hair and keratoderma [MedGen | Orphanet | OMIM] Arrhythmogenic right ventricular cardiomyopathy, type 8 [MedGen | OMIM]	germline		Invitae	SCV000288553.3
Likely benign (Jun 14, 2016)	criteria provided, single submitter • ICSL Variant Classification 20161018	clinical testing	Cardiomyopathy, ARVC [MedGen]	germline	• PubMed (3) [See all records that cite these PMIDs]	Illumina Clinical Services Laboratory,Illumina	SCV000464829.2
Likely pathogenic (Dec 13, 2013)	criteria provided, single submitter • ACMG Guidelines, 2015 • ACMG Guidelines, 2015	clinical testing	Right ventricular cardiomyopathy [MedGen | Human Phenotype Ontology]	unknown		Centre for Mendelian Genomics,University Medical Centre Ljubljana Study description	SCV000492954.1
Pathogenic (Mar 27, 2015)	no assertion criteria provided	clinical testing	Cardiomyopathy [MedGen | Human Phenotype Ontology]	unknown	• PubMed (1) [See all records that cite this PMID]	Forensic Genetics Laboratory,Harris County Institute of Forensic Sciences - HCIFS-Postmortem genetic screening project Study description	SCV000263110.1

Fig. 2.4. Evoluția semnificației clinice a variantei rs121912998 în baza de date *ClinVar*

Determinarea cauzalității unor variante genetice pentru o anumită patologie se poate realiza prin studii de înlănțuire genetică sau studii de asociere, dar și în acest mod, având un lot de pacienți afectați și un lot de pacienți de control, studiile au anumite limitări. O limitare a acestor tipuri de studii este ținta regiunilor genomice care, de obicei, constă într-un set de gene sau de *loci* (poziții) pe ADN. Aceste regiuni sunt alese pe baza experienței echipei care scrie proiectul sau în funcție de rolul pe care îl îndeplinesc respectivele gene [34]. Pe de altă parte, un studiu care presupune secvențierea întregului genom, pentru toți pacienții, cu o acoperire decentă (50X), presupune costuri foarte mari. Mai mult de atât, secvențierea întregului genom ar avea ca efect obținerea unui set mare de variante, dintre care mare parte reprezintă zgomot. În ambele cazuri, după filtrarea variantelor și identificarea unor posibile variante patogene, sunt necesare studii de testare funcțională și de validare. Aceste investigații implică o serie de costuri financiare, proceduri care consumă timp și țin blocate resurse umane (personal de laborator). Prin urmare, această etapă, de cele mai multe ori, nu se realizează, rezultatele fiind publicate pe baza informațiilor existente în literatura de specialitate și pe baza experienței cercetătorilor [34].

Din fericire, nu toate afecțiunile necesită investigații care presupun secvențierea întregului genom sau căutarea unor variante patogene. *Pevsner* în [11] cataloghează afecțiunile, pe baza cauzei, ca fiind:

- Afecțiuni cauzate de o singură genă (sindromul *Marfan*);
- Afecțiuni complexe, care au implicate două sau mai multe gene (schizofrenia);
- Afecțiuni genomice, care sunt cauzate de anormalități genomice (sindromul *Down*);
- Afecțiuni cauzate de mediu (infecții).

Variantele genetice sunt de interes pentru afecțiunile cauzate de o singură genă și cele care implică mai multe gene. Majoritatea afecțiunilor care implică o singură genă fac parte din categoria bolilor rare. Odată ce afecțiunea este identificată clinic, procesul de validare genetică este relativ simplu. În schimb afecțiunile care implică gene multiple (precum bolile cardiovasculare, diabetul, steatoza) au o rată ridicată de manifestare în rândul populației. În acest caz, pentru identificarea variantelor cauzatoare, pe lângă cunoașterea variantelor existente și a funcției genelor, mai trebuie cunoscute și interacțiunile moleculare în care sunt implicate produsele genelor (proteinele) precum și căile metabolice în care activează.

2.2.1. Predictori

Pentru identificarea variantelor patogene se pot folosi unelte de predicție *in silico*, care au asociate scoruri de patogenitate pentru fiecare variantă genetică. Instrumentele inițiale, utilizate de cercetători, s-au bazat pe predicția efectului substituțiilor. Două instrumente care au fost dezvoltate timpuriu au fost *Sorting Intolerant from Tolerant* (SIFT) [20] și *Polymorphism Phenotyping* (PolyPhen)[19]. Acești predictori estimează efectul unei substituții la nivel de ADN (bază azotată) asupra stabilității și funcției proteinei. Un alt instrument care utilizează o metodă bazată pe aliniere pentru a prezice efectele dăunătoare ale unei variante este *Protein Variation Effect Analyzer* (PROVEAN) [35]. Există instrumente care fac predicția pe baza conservării secvenței ADN-ului între specii. Exemple precum *Mutation Taster* [36] și *Mutation Assessor* [37] prezic potențialul de risc al unei variante genetice pe baza modificării regiunii de matisare, conservarea între specii și pierderea funcției proteinei. Un alt instrument, *Functional Analysis through Hidden Markov Models* (FATHMM) [38], calculează toleranța proteinei prin combinarea conservării secvenței de ADN cu modelele *Markov* ascunse. Există și predictori care se bazează pe învățarea automată și calculează un scor pentru fiecare variantă pe baza altor caracteristici moleculare sau elemente de conservare. *Meta-Analytic Support Vector Machine* (MetaSVM) [39] și *MetaLR* sunt unelte care se bazează pe metode de tip *Ensemble*. Un alt predictor din această categorie este *Rare Exome Variant Ensemble Learner* (REVEL) [40] care este o aplicație creată pentru a fi folosită în detecția variantelor rare. *Mendelian Clinically Applicable Pathogenicity* (M-CAP) [41] și *Combined Annotation Dependent Depletion* (CADD) [42] sunt instrumente care au la bază unelte precum SIFT, *PolyPhen* și altele caracteristici pentru calcularea scorului de patogenitate. Un instrument similar cu CADD este DANN [43], diferența fiind că acesta din urmă folosește rețele neuronale pentru a determina scorul variantei genetice.

2.2.2. Proiecte genomice

Tehnologia NGS a revoluționat modul în care oamenii de știință și cadrele medicale cercetează și diagnostichează afecțiunile genetice. În ultimii ani această

tehnologie a fost integrată în tot mai multe centre de cercetare și spitale. Datorită scăderii costului de analiză, tot mai multe centre clinice introduc secvențierea ca metodă de rutină. Mai mult decât atât, dacă până acum se utiliza secvențierea unui panel de gene, încet, laboratoarele trec către secvențierea întregului exom sau uneori chiar a întregului genom. Desigur, secvențierea reprezintă doar primul pas, precum a fost descris în subcapitolul 2.1.6.

Ulterior secvențierii, variantele genetice trebuie verificate, filtrate și catalogate. În prezent, există o serie de instituții sau proiecte care facilitează aceste operațiuni. Exemple care se ocupă de catalogarea informației genetice, implementate la nivel național, sunt: *Health 2030 Genome Center* în Elveția, *Care4Rare* în Canada; sau la nivel internațional *The International Genome Samples Resource (IGSR)*. După ce ajung la maturitate majoritatea acestor inițiative oferă publicului acces la informații pentru a încuraja cercetarea. De exemplu, după finalizarea proiectului 1000 de genomuri, informații despre frecvența în populație a polimorfismelor a fost făcută publică. Ulterior a urmat 100.000 de genomuri [44] prin care cercetătorii speră să fi schimbat modul în care se va folosi secvențierea pentru diagnosticare. Afecțiunile urmărite în cardul UK 100K sunt bolile infecțioase, bolile rare și cancerul. Odată cu versiunea 7 a pachetului de date, UK 100K a reușit să secvențieze ținta de genomuri propuse. Desigur, acesta reprezintă doar un prim pas. Urmează faza de integrare a datelor fenotipice cu ajutorul cărora se speră identificarea variantele cauzatoare de afecțiuni și, bineînțeles, propunerea unor tratamente pentru aceste afecțiuni.

În ciuda tuturor eforturilor, găsirea variantelor patogene rămâne în continuare dificilă pentru multe afecțiuni. În literatură au apărut o serie de strategii pentru clasificarea variantelor și a genelor. Acestea ajută la identificarea celor mai bune opțiuni, proces cunoscut ca prioritizarea de variante. Dintre metodele actuale de prioritizare, o amintim pe următoarea, care se aplică pentru cazul general: (1) secvențierea simultană a unui lot de pacienți cu aceeași afecțiune și căutarea unor gene care sunt afectate la majoritatea indivizilor, (2) căutarea unor variante care se regăsesc la toți pacienții și prioritizarea pe baza frecvenței în populație, (3) analiza prin înlănțuirea genetică și analiza rețelelor de interacțiuni dintre gene. Primele trei strategii identifică, punctual, genele sau variantele presupuse ca fiind cauzatoare de afecțiunii. Analiza relațiilor dintre gene generează probabilitatea ca fiecare genă din genom să fie cauzală, subiect discutat în subcapitolul 3.2. Strategiile menționate anterior țin cont doar de caracteristicile moleculare ignorând intrările externe. O altă metodă care se dovedește a fi utilă este evaluarea *top-down*, prin care se evaluează fenotipul pacientului, iar pe baza evaluării se propune un set de gene posibil cauzatoare. Mai departe sunt prezentate o serie de instrumente utilizate în analiza computațională, bazate atât pe caracteristicile moleculare ale ADN-ului, cât și pe fenotipul pacienților. De asemenea, se vor explica punctele tari și punctele slabe ale acestor instrumente. Se va explica felul cum pot fi utilizate pentru a simplifica procesul de analiză, având la dispoziție fenotipul și fișierul cu variantele genetice ale pacientului. Aceste descrieri pot contribui la procesul de prioritizare a variantelor și de identificare a celor patogene.

2.2.3. Aplicații pentru adnotare și filtrare

În general, pentru un mic panel de gene, de exemplu *TruSightCardio* al celor de la *Illumina*, secvențiatorul generează variante de ordinul sutelor. În cazul întregului exom ne putem aștepta la sute de mii de variante, iar în cazul întregului genom fișierul poate ajunge la milioane. Nu toate aceste variante sunt patogene, multe dintre ele au fost conservate datorită selecției naturale, fiind de obicei specifice anumitor populații.

Prin urmare, 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. Spunem „ideal" pentru că nu sunt cunoscute toate variantele cu frecvență mare pentru fiecare populație din fiecare regiune geografică. România, de exemplu, face parte din această categorie. Pentru obținerea frecvenței ar fi nevoie de implementarea unui proiect asemănător *1000 Genomes* sau chiar mai bine *100K Genomes* care să cuprindă fiecare subregiune geografică sau fiecare subpopulație din acea regiune.

După filtrarea variantelor comune, variantele rămase sunt prioritizate pe baza patogenității lor. Realizarea prioritizării pe baza patogenității se poate realiza dacă variantele genetice au fost adnotate cu informații precum efectul lor asupra structurii proteinei, efectul asupra genelor și transcripțiilor rezultate din respectivele gene. Pentru a avea acces la aceste informații este necesar ca variantele să fie scrise în formatul *variant call* (VCF), care să reflecte coordonatele cromozomiale ale fiecărei variante (chr7: g.1234567T> G) și informația despre variantă la nivelul genei (c. 123G>C; p.Arg1190Cys în gena TTL). Informațiile sunt necesare pentru a pune în context modificările aduse genei și a produsului rezultat din aceasta, cu afecțiunea suspectată. Pe lângă informațiile de bază, unele instrumente, folosite pentru adnotare, oferă funcționalități suplimentare care permit filtrarea polimorfismelor în funcție de clasa variantelor genetice. Prin clasa variantelor înțelegem consecința funcțională a acesteia. De exemplu, mutații de aminoacizi (*missense*) sau codon stop (*nonsense*) pot fi rezultatul unui polimorfism uninucleotidic non-sinonim.

În plus, unelte precum ANNOVAR [45] permit adnotarea și filtrarea variantelor pentru a reduce numărul lor la un set ușor de gestionat. Pentru aceasta se pot aplica diverse criterii, eliminarea celor care au o frecvență mai mare de 1% în proiectul *1000 Genomes* sau în proiectul *ExAC* (6500 genomuri) sau mai nou *gnomAD* versiunea 3 (14156 de exoame și genoame) [46]. Se mai pot exclude și variantele care apar în dbSNP și despre care se știe că nu sunt patogene.

Jannovar este un alt instrument care permite adnotarea cu informații despre modelul de moștenire al variantei care poate fi folosit pentru o filtrare ulterioară. Asemănător cu *ANNOVAR*, *Variant Effect Predictor* [21], produsul celor de la Institutul European de Bioinformatică (EBI), poate fi folosit pentru adnotarea cu informații a fișierelor VCF. Acesta oferă mai multe moduri de utilizare. Se poate accesa printr-o interfață online sau poate fi descărcat un pachet *Pearl* care permite utilizarea în linie de comandă. Variantele acceptate pot fi într-o serie de formate. De asemenea, VEP oferă posibilitatea alegerii unui set de transcripții (*Ensemble*, *Gencode* sau *Refseq*) care va fi folosit pentru adnotare. Adnotarea variantelor se face cu informații despre frecvența în populație, clasa și predictori precum SIFT și *PolyPhen*. Avantajul acestei unelte, spre deosebire de ANNOVAR, este că permite filtrarea variantelor în interfața web a aplicației. Acest lucru este extrem de util specialiștilor, în special din domeniul medical, care nu sunt familiarizați cu sistemele de operare de tip UNIX.

2.2.4. Prioritizarea pe baza fenotipului

O altă strategie pentru situația în care specialistul (medic, cercetător) nu cunoaște diagnosticul unui pacient este să analizeze caracteristicile fenotipului. Există o serie de aplicații care pot filtra și prioritiza fiecare genă în funcție de ceea ce se observă la pacient. Totuși, pentru a defini fenotipul este necesar ca specialiștii să folosească același limbaj. În prezent există o serie de cataloage de termeni precum *PoSSuM, HPO, myPhenoBD, Face2Gene, MeDRA*. Cea mai cunoscută bază de date este „Ontologia Fenotipului Uman" (*Human Phenotype Ontology*, HPO). Aceasta are

ca scop catalogarea anomaliilor într-un mod care poate fi prelucrat cu ușurință în aplicații computaționale. Pe baza acestui catalog se pot determina asemănările și deosebirile dintre afecțiuni. Interogarea oricărei baze de date menționate anterior, folosind caracteristicile fenotipului, va furniza afecțiuniile care se potrivesc cel mai bine. Pentru realizarea operației se poate apela la o serie de aplicații, parte dintre acestea fiind descrise și analizate în [47] și prezentate în continuare.

Aplicația *Exomiser* [48] utilizează similaritatea descrierii fenotipului dintre pacient cu cel al fenotipului în șoareci, pentru toate genele din exom. Determinarea similarității se realizează cu algoritmul *PhenoDigm*. Acest algoritm notează genele cu un scor între 0 și 1, unde valoarea 1 reprezintă gene perfect similare. Pe lângă acest scor, se folosește și frecvența alelică obținută din proiectele *1000 Genomes* și *ExAC*. De asemenea, se folosesc predictorii de patogenitate SIFT, *PolyPhen* și *Mutation Taster*. Testul de performanță menționat anterior indică o acuratețe de 83% pentru transmiterea dominantă, iar pentru transmiterea recesivă o acuratețe de 66%.

Aplicația *PhenIX* [49] se bazează pe comparații între fenotipul pacientului și fenotipurile genelor patogene. Aplicația este destinată diagnosticării când sunt cunoscute genele care au asociate afecțiuni. Pentru determinarea similarității, aplicația utilizează algoritmul *Phenomizer*. Rezultatele prezentate în [47], pentru gene asociate cu afecțiuni, a arătat că 97% din genele cauzatoare au fost în topul genelor selectate.

O altă aplicație este *Phevor 2* [50] (*Phenotype Driven Variant Ontological Re-Ranking Tool*). Aplicația se folosește de rezultatele instrumentelor de predicție *in sillico*, adnotate cu ANNOVAR, pentru a prioritiza variantele genetice. Ulterior, genele pe care se află variantele se vor prioritiza în funcție de fenotip și de afecțiunile asociate cu acestea. Setul de date pentru adnotare folosit de această aplicație este format din HPO, ontologia bolilor (DO), ontologia genetică (GO) și ontologia fenotipului mamiferelor (MFO). Pentru a utiliza aplicația, utilizatorul trebuie să introducă termenii din ontologiile menționate anterior care se potrivesc fenotipului pacientului investigat. Ulterior introducerii datelor, *Phevor* generează o listă de gene suspecte. În cele din urmă, fiecare genă va primi un scor care se bazează pe modul de propagare a acesteia prin ontologii. În final, *Phevor* va combina scorul genetic cu informațiile adnotate de ANNOVAR pentru a realiza un top al genelor suspecte.

O altă aplicație online folosită pentru identificarea genelor cauzatoare de afecțiuni este *eXtasy* [51]. Aplicația folosește zece predictori, discutați anterior, și un scor de predicție a haploinsuficienței genelor pentru generarea unui scor general al patogenității. Metoda pe care se bazează aplicația constă în extragerea tuturor genelor asociate cu un termen HPO și generarea unui scor pentru acestea folosind algoritmul de similitudine *Endeavour*. Acest algoritm folosește o serie de metode pentru măsurarea similitudinii genetice (implicarea în aceleași căi metabolice, aceleași interacțiuni proteină-proteină sau similaritatea secvenței ADN). Scorul final atribuit de *eXtasy* este generat folosind metoda *Random Forest* care combină toate informațiile extrase pentru acea genă. Performanța aplicației *eXtasy* a fost evaluată folosind curba ROC pentru discriminarea între variantele cauzatoare de afecțiuni și variantele benigne. Această analiză, prezentată în lucrarea lui *Smedley* și *Robinson* [47], indică o îmbunătățire semnificativă comparativ cu metodele clasice de predicție precum *Mutation Taster*, *PolyPhen* sau SIFT. O limitare a acestei aplicații este faptul că prioritizează doar variante ne-sinonime. Există posibilități de extindere pe viitor, dacă vor fi disponibile baze de date, pentru variante necodante, sinonime și care se află în zona de matisare. O altă limitare este faptul că aplicația nu efectuează filtrarea pe baza frecvenței variantei genetice în populație. Este recomandat ca utilizatorul să efectueze această filtrare înainte de introducerea datelor în aplicație pentru

prioritizare. Operația nu este întotdeauna la îndemâna personalului medical când se efectuează o secvențiere de tipul WES sau WGS. Un avantaj major îl reprezintă faptul că *eXtasy* este o aplicație gratuită.

Phen-Gen este o altă aplicație, prezentată în [52], care permite prioritizarea variantelor genetice. Aceasta folosește o platformă prin care se compară variantele prezise ca fiind patogene și simptomele unui pacient cu datele deja stocate în baza internă de date. Aplicația mai permite și analiza variantelor non-codante, pentru analiza de tip WGS, prin determinarea proximității acestora față de secvențele codante. Spre deosebire de *eXtasy*, aplicația filtrează variantele care au o frecvență în populație mai mare de 1%. *Phen-Gen* mai are o funcționalitate interesantă prin care determină un model care ține cont de toleranța persoanelor sănătoase față de variantele patogene. Modelul a fost generat folosind baza de date a proiectului *1000 Genomes*. După efectuarea filtrării genelor și a variantelor genetice, acestea sunt analizate folosind *Phenomizer*. *Phenomizer* folosește termenii HPO pentru corelarea genelor și a simptomelor asociate pacientului. În cadrul analizei de performanță realizată în [47], prin care se simulează date asemănătoare cu *1000 Genomes*, Phen-Gen a reușit să identifice corect 88% dintre pacienți. Pentru descoperirea asocierilor, aplicația a identificat corect 56% dintre afecțiunile cu caracter dominant și 83% dintre cele cu caracter recesiv. Pentru 11 seturi de pacienți *trio* (pacient și părinți), care sufereau de dizabilități intelectuale încrucișate sau recesive, aplicația a reușit să identifice 81% dintre genele care au fost raportate ca fiind patogene în topul primelor zece gene prezise.

2.3. Elemente specifice analizei regiunilor de matisare (*splicing*)

2.3.1. Matisarea în genetică

Inițial se considera că, precum procariotele, proteinele din eucariote sunt formate după o secvență de baze azotate consecutive de pe structura ADN-ului. Acest lucru s-a dovedit a fi eronat, când în anul 1977 a fost comparată secvența unui mARN cu secvența ADN a unui adenovirus [53]. Secvența bazelor azotate din structura mARN-ului nu coincidea cu nicio secvență consecutivă de ADN, ceea ce a dus la descoperirea procesului de matisare. Prin acest proces pre-mARN-ul este tăiat în anumite puncte pentru a elimina unele segmente din secvență, iar restul segmentelor sunt concatenate pentru a forma ARN-ul matur. Segmentele de pre-mARN care au fost eliminate se numesc introni, iar secvențele care au fost concatenate pentru a forma mARN-ul se numesc exoni [54]. Determinarea granițelor dintre exoni și introni se bazează pe reacțiile biochimice dintre elementele *cis-acting*, care se regăsesc în secvența de nucleotide a ARN-ului, și elemente *trans-acting*, care reprezintă proteinele care se atașează de fragmentele de nucleotide.

Inițial, variantele care nu făceau parte din exonii unei gene erau ignorate; în ultimii ani, însă, specialiștii au început să acorde mai multă importanță variantelor care afectează procesul de matisare și să studieze rolul lor în manifestarea afecțiunilor. Analiza *RNA-seg* a unui pacient este metoda cea mai simplă și mai sigură de detectare a defectelor de *splicing*. Deși există diferite tehnici disponibile pentru detectarea erorilor de matisare, acestea nu sunt utilizate pe scară largă în clinici. Există o serie de motive, dintre care necesitatea unei aparaturi speciale (testul de *splicing in vitro*, testul de *splicing* a mini genelor etc.), personal specializat ș.a.m.d. Prin urmare, din punct de vedere clinic, cunoștințele despre acest proces abia acum încep să fie obținute și integrate în procesul de diagnosticare.

În prezent, metoda folosită pentru găsirea variantelor genetice care afectează regiunile de matisare este secvențierea ADN-ului. Avantajul acestui proces este faptul că permite specialistului să investigheze atât variantele genetice care se află în regiunile exonice, cât și variantele genetice care se află în regiunile limitrofe. Limitarea acestei metode este incertitudinea vizavi de manifestarea în ARN a modificării survenite în regiunea de matisare. Din literatură aflăm că la *homo sapiens* (oameni) regiunea de *splicing* nu este foarte bine conservată [54], în comparație cu bacteriile de exemplu. Un alt dezavantaj al secvențierii ADN-ului îl reprezintă multitudinea de variante care trebuie analizate de către specialiști. Alternativa ar fi testele specializate pentru detectarea regiunilor de matisare, dar acestea necesită personal specializat și timp îndelungat pentru obținerea rezultatelor, ceea ce implică o serie de costuri suplimentare, iar în majoritatea cazurilor depășesc limita superioară a bugetului destinat laboratoarelor. Prin urmare, testul ADN este, în prezent, cea mai rezonabilă soluție la îndemâna geneticienilor pentru numărul mare de pacienți investigați.

Pentru a îmbunătăți rezultatele în ceea ce privește regiunile de matisare, specialiștii apelează la instrumente software de predicție a regiunilor de *splicing*. Aceste instrumente permit restrângerea grupului de variante, eliminând variantele care au un potențial dăunător scăzut. Inițial, instrumentele software au fost folosite în proiecte de cercetare pentru găsirea genelor, dar ulterior, după finalizarea PGU, au fost respecializate pentru analiza regiunilor de matisare. Fiind soluții software, *in silico*, rezultatele obținute nu pot confirma cu certitudine că secvența de *splicing* este afectată. Prin urmare, medicii care folosesc aceste unelte, trebuie să fie atenți în

momentul în care aleg un astfel de instrument și să fie rezervați în punerea unui diagnostic doar bazându-se pe rezultatul acestora. Există numeroase astfel de aplicații, pentru aproape toate elementele de reglare a matisării. Un exemplu al reușitei acestor instrumente este *ESEfinder*, care a prezis pierderea funcției unei regiuni ESE aflate în gena SMN2 [55], [56].

2.3.2. Max Polimorfisme uninucleotidice care afectează matisarea

Regiunea de matisare conține trei componente importante: (1) punctul de excizie (*branch point, BRS*), tractul de pirimidine și situsul acceptor. Precum s-a demonstrat în multiple studii din literatura de specialitate, BRS-ul nu este foarte bine conservat [57]–[59]. Toate șabloanele (*URAY, UNA, UUNAN, CUNAN, YUNAY*) au în comun adenina, care reprezintă cea mai conservată bază azotată din BRS, și uracilul care precedă cu două poziții (-2) adenina. Pentru predicția BRS au fost dezvoltate aplicațiile *Human Splice Finder* și *SVM-BP finder*. Inițial aceste unelte au fost gândite doar pentru a identifica regiunile de matisare, respectiv pentru a detecta punctele de ramificație. Prin urmare, *SVM-BP* nu poate prezice efectul unui polimorfism uninucleotidic (*SNP*) în punctul de ramificație [60], desigur cu excepția variantelor care alertează nucleotidele bine conservate.

Aparenta problemă, lipsa conservării BRS-ului, se presupune că este rezolvată de către natură cu ajutorul tractului de baze azotate pirimidinice și situsul acceptor. Celula are doi factori de matisare care se vor prinde de tractul format din pirimidine. U2AF65 se va lega pe tract mai aproape de BRS pe când U2AF35 se va prinde în apropierea situsului acceptor (AG). Prin urmare, acești factori vor contribui la identificarea regiunii de matisare [61].

Tractul de pirimidine este la rândul lui puțin conservat, singura pseudo-regulă fiind că secvența de ARN este formată dintr-un număr majoritar de piriminidine (uracil și citozină), U fiind preferat. Prin urmare, variantele genetice care apar în această regiune pot afecta în oarecare măsură procesul de matisare. Până în prezent nu s-au dezvoltat unelte speciale pentru analiza specifică a variantelor din tractul de pirimidine. Desigur există unelte care evaluează regiunile de *splicing* în ansamblu precum SPANR [62] și *IntSplice* [63], care teoretic ar putea sau ar trebui să detecteze variantele perturbatoare.

Dinucleotidele aferente sitului de acceptare, format din AG, sunt foarte bine conservate. De fapt, tiparul care corespunde acestei secvențe este nyag|G (| - reprezintă limita dintre intron și exon). Această regiune este esențială pentru prinderea factorului de matisare U2AF35, care va împiedica legarea aberantă a proteinei U2AF65. Oricare substituție a unei baze azotate din grupul de nucleotide AG va duce aproape sigur la anularea regiunii de matisare. Bineînțeles, acesta nu este singurul mod în care situl de acceptare poate să-și piardă funcția. Dacă în tractul de pirimidine avem o substituție sau o inserție a unei adenine (A) lângă o guanină (G), astfel încât să se formeze un nou grup AG, atunci grupul canonic își poate pierde funcția. Totuși, există unele condiții care au fost raportate în literatură, precum distanța sitului perturbator față de cel autentic de cel mult 21 de nucleotide în [64], sau 12 nucleotide [65] sau 14 nucleotide [59].

2.3.3. Elemente reglatoare în matisare

Elementele *cis*-activatoare le putem grupa în șase categorii [66]:

- *5′ splice site* reprezintă secvența de legătură dintre un exon și un intron;
- *3′ splice site* reprezintă secvența de legătură dintre un intron și un exon;
- *Exonic Splicing Enhancers* (ESE) reprezintă secvența care ghidează procesul de matisare;
- *Intronic Splicing Enhancers* (ISE) reprezintă secvența aflată pe intron care ghidează procesul de matisare;
- *Exonic Splicing Silencers* (ESS) reprezintă secvența din exon care împiedică procesul de matisare;
- *Intronic Splicing Silencers* (ISS) reprezintă secvența din intron care împiedică procesul de matisare;

Elementele *trans-acting* sunt reprezentate de spliceosom, care este format din cinci riboproteine nucleare mici (snRNP) și alte zeci de proteine care contribuie la procesele reglatoare din celulă [67].

Desigur, cel mai important moment în timpul procesului de matisare este identificarea și captarea semnalelor de *splicing* încorporate în secvența de pre-mARN de către spliceosom. Pentru identificarea regiunilor de matisare, 5′ și 3′, este nevoie să se identifice situsurile GT, respectiv AG. Aceste nucleotide sunt cele mai conservate, restul nucleotidelor din șablon sunt mai puțin conservate [68]. Deși s-au efectuat numeroase studii [57], [69] pentru identificarea unei regiuni unice, de consens, care să corespundă regiunii de matisare, acest lucru nu s-a reușit.

Deși procesul de matisare este descris ca fiind o funcție liniară, secvența de mARN care se obține după acest proces variază în majoritatea cazurilor. Acest lucru se datorează matisării diferențiale (alternative) care permite crearea unei secvențe diferite fără a afecta funcțiile proteinelor. Acest tip de matisare se poate realiza în mai multe moduri [70]:

- *Exon-skipping*, adică sărirea unui exon;
- *Mutually exclusive exons* este excluderea mutuală a exonilor, eliminarea unui exon din doi posibili, dar niciodată ambii exoni;
- *Intron retention* este reținerea unei zone intronice în secvența de mRNA;

Rolul matisării alternative este acela de a diversifica expresia genică în funcție de țesutul în care se găsește celula sau în funcție de stadiul de dezvoltare a organismului. Desigur, matisarea alternativă nu trebuie confundată cu posibilele variante genetice (din regiunea de matisare) care se dovedesc a fi dăunătoare. O variantă genetică (sau mai multe) pot crea o regiune mutantă de matisare sau pot altera o regiune autentică. Acest fapt va determina schimbarea locului de prindere a spliceosomului, ceea ce duce la o matisare aberantă. Un exemplu este substituția unei guanine (G) cu o timină (T) pe prima poziție a intronului 25 de pe gena DFNA1. Această modificare afectează regiunea conservată a secvenței de *splicing*, care determină o inserție de baze în mARN care, la rândul ei, duce la terminare prematură a mARN-ului al cărui rezultat este pierderea a 32 de aminoacizi [71]. Un alt exemplu bine documentat este varianta uninucleotidică prin care o citozină este înlocuită cu o timină pe exonul 7 al genei SMN2, pentru persoane care au deja deleții pe SMN1, ceea ce determină inactivarea unei regiuni ESE și crearea unei regiuni ESS, care în final duce la omiterea exonului 7, manifestându-se fenotipic ca atrofie musculară spinală [72].

Conform bazei de date privind mutațiile genetice umane, HGMD [73], care a început din anul 2013 să integreze mutații din regiunea de splicing, un procentaj de

9,2% din totalul mutațiilor, raportate ca fiind patogene, sunt în regiunea de matisare. Deși baza de date a raportat doar variantele patogene, în lucrări precum [74], [75] se afirmă că, dacă sunt luate în considerație și mutațiile care sunt raportate ca *missense,* se ajunge la procentaj de 22%. Acest fapt ne indică faptul că mutațiile în regiunile de matisare sunt mai numeroase decât se cunoaște în prezent.

2.3.4. Instrumente pentru predicția regiunilor de matisare

Înaintea Proiectului Genomului Uman, instrumentele de predicție pentru regiunile de matisare au fost utilizate pentru identificarea regiunilor limită dintre exon-intron. Ulterior, aceste instrumente au fost readaptate pentru predicția impactului transcripțional al variantelor genetice din aceste regiuni. Tranziția a fost motivată de necesitatea înțelegerii efectelor generate de variațiile genetice asupra proceselor celulare sau eventual descoperirea cauzelor anumitor afecțiuni. În prezent există o serie de unelte software care permit analizarea secvențelor genetice pentru detectarea regiunilor de *splicing*. Diferențele majore între aceste unelte sunt modul cum recunosc regiunile de matisare. Unele memorează o serie de secvențe, iar altele folosesc șabloane ale regiunii de *splicing* sau modelele statistice. Deși există un număr mare de unelte, ideile pe care se bazează aceste unelte nu sunt atât de diverse. O lucrare care tratează pe larg subiectul uneltelor care descoperă defectele de matisare a fost scrisă de *Jian* și colaboratorii săi [54]. În cele ce urmează se prezintă succint o serie de concepte și unelte folosite în acest scop.

Metoda *Position Weight Matrix* (PWM) propus de *Shapiro* și *Senapathy* în [69], presupune atribuirea unor scoruri pentru fiecare nucleotidă (A, T, C și G) și ordonarea secvenței folosind ponderi pentru fiecare poziție pe baza informațiilor din secvența de consens. Spre deosebire de următoarele modele, modelul PWM este simplu iar modelele generate sunt ușor de înțeles. Pozițiile fiind independente, schimbarea scorului la o anumită poziție nu are nici un impact asupra scorului pozițiilor vecine. Această metodă a fost folosită la baza dezvoltării aplicației *Splice-Site Analyzer Tool* [76]. O îmbunătățire a fost adusă modelului PWM în implementarea aplicației *SpliceView* [77] care examinează dependențele reciproce între nucleotide de pe diferite poziții. În prezent, PWM-ul este folosit pe scară largă pentru reprezentarea diferitelor tipare aferente secvențelor de ADN.

O altă metodă statistică pentru captarea tiparelor este *Maximal Dependence Decomposition* (MDD). Aceasta este folosită și la crearea arborilor decizionali care captează potențialele dependențe între poziții prin împărțirea setului de date în subseturi bazate pe dependență. Modelul MDD a fost incorporat în aplicația GENSCAN [78].

Pentru a rezolva problema modelelor anterioare, anume selectarea arbitrară a ponderilor pentru fiecare poziție, s-a apelat la învățarea automată. Prin instruirea cu seturi de date, rețele neuronale (NN) au reușit să identifice zonele de *splicing*. Exemple de aplicații care au folosit rețele neuronale sunt *NetGene2* [79] și *NNSplice* [80].

O altă tehnică de învățare automată care este folosită pentru detectarea regiunilor de matisare este *Suport Vector Machine* (SVM). Aplicația *SplicePort* folosește această metodă împreună cu un algoritm de generare a funcțiilor pentru a surprinde caracteristicile importante ale secvențelor [81]. Un dezavantaj major al învățării automate îl reprezintă supra-potrivirea pe setul de date de antrenare. O

metodă pentru a minimiza acest efect este utilizarea modelelor *Bayes*. Aplicația *SplicePredictor* a fost implementată folosind acest model.

Cele mai performante modele ale secvențelor de matisare au fost realizate prin folosirea *Maximum Entropy Distribution* (MED) [82]. Modelele create de această metodă iau în calcul dependențele între pozițiile adiacente, dar și între cele neadiacente. MED permite modificarea șablonului prin schimbarea setului de constrângeri, ceea ce evită problema supra-potrivirii. Aplicația *MaxEntScan* [83] folosește această metodă. Utilizatorii acestei aplicații pot folosi modelele implicite sau își pot crea propriile modele. Aplicația a fost folosită cu succes pentru predicția unei mutații pe gena ATM care este responsabilă pentru ataxia-telangiectazia [84]. *MaxEntScan* are opțiunea de folosire a altor modele PWM sau MDD.

Automated Splice Site Analyses (ASSA) [85] este un exemplu de aplicație care se bazează pe modele din teoria informațiilor. CRYP-SKIP [86] folosește un model de regresie logistică multiplă pentru a distinge exonii care sunt ignorați, astfel activând regiuni ascunse de matisare care sunt rezultatul unor mutații de *splicing*. *Spliceman* [87] prezice probabilitatea ca variantele din jurul regiunilor de matisare să afecteze *splicing*-ul prin analiza hexamerilor pe baza distribuțiilor poziționale.

Există si unelte care includ algoritmi multipli. Un exemplu este *Human Splicing Finder* [88], care afișează predicțiile despre matisare pe baza modelelor PWM și MED, precum și eventuale defecte ale regiunilor ESE și ESS. SROOGLE [89] este o platformă care integrează o serie de algoritmi de predicție pentru afișarea unor semnale de *splicing*.

Cu excepția ASSA, utilizatorul trebuie să furnizeze secvența pentru toate uneltele, făcând, astfel, mai puțin convenabilă aplicarea în clinică a acestor instrumente. ASSA poate localiza varianta bazată pe numele genei furnizate de utilizator, numărul de acces al mARN-ului sau numărul dbSNP.

2.3.5. Utilizarea aplicațiilor de predicție

O afișare clară a rezultatelor obținute de aplicațiile de predicție este importantă pentru aplicarea lor în clinică. În general instrumentele dau un scor care indică intensitatea semnalului de matisare. În majoritatea cazurilor un scor mare indică similaritate cu secvența de consens sau indică faptul că segmentul respectiv este unul de matisare. Scorul generat de aplicație are un caracter orientativ pentru că nu are o componentă cantitativă atașată, ci mai degrabă ajută la stabilirea unui prag pentru a elimina unele variante genetice. Desigur, există și posibilitatea eliminării unor variante patogene pentru că asupra procesului de *splicing* acționează și alți factori biologici [54].

Numărul publicațiilor care tratează subiectul predicției regiunilor de matisare, în experimente și cu rezultate *in vitro* sau *in vivo*, este scăzut. Prin urmare, utilizarea uneltelor de predicție a variației în regiunile de matisare a fost oarecum abandonată. *Houdayer* a făcut o evaluare sistematică a câtorva instrumente *in silico* în [90] pentru un număr de variante cu semnificație necunoscută. Cu un prag de 15% pentru *MaxEntScan* și cu un prag de 5% a reușit să obțină o sensibilitate de 96% și o specificitate de 83% [54]. Acest studiu nu este suficient pentru a generaliza. Este nevoie să se realizeze studii mai mari pe date genomice și transcriptomice. Din fericire, date pentru astfel de studii sunt disponibile (*GenBank, EMBL*). Costurile folosirii acestor unelte sunt substanțial mai mici decât verificarea variantelor genetice *in vitro* sau *in vivo.*

2.4. Elemente specifice analizei bolilor complexe

În acest subcapitol sunt rezumate o serie de informații care vor ajuta la parcurgerea capitolului 5, în care se tratează modelarea afecțiunilor complexe. Afecțiunea folosită ca exemplu pentru analiza de date este steatoza.

Afecțiunile complexe sunt cauzate de rezultatul interacțiunii genotipului cu factorii de mediu [91]. Printre aceste afecțiuni regăsim bolile cardiace, hipertensiunea, obezitatea, diabetul etc. Complexitatea este dată de faptul că un set de variante genetice, în combinație cu anumiți factori externi, predispun la manifestarea unei afecțiuni. Dar aceeași afecțiune, cu mici modificări ale fenotipului, poate fi manifestată pentru alt set de variante genetice și factori de mediu [92].

2.4.1. Steatoza

Steatoza hepatică reprezintă acumularea de grăsime la nivelul țesutului hepatic. Sunt cunoscute două forme de steatoză. Prima formă se datorează consumului de alcool, iar cea de a doua se datorează altor factori care nu au legătură cu consumul de alcool, precum cei genetici. Cea din urmă poartă numele steatoză hepatică non-alcoolică sau boala ficatului gras non-alcoolic, în literatura internațională *Nonalcoholic Fatty Liver Disease* (NAFLD). Această afecțiune se regăsește din ce în ce mai des în țările dezvoltate (Europa de Vest și Statele Unite). Dacă este însoțită și de o inflamație, atunci aceasta poate progresa către ciroză, care este caracteristică personalelor care consumă alcool în exces. Cauzele care determină apariția bolii ficatului gras non-alcoolic nu sunt foarte bine cunoscute, dar pare a fi asociată cu obezitatea, rezistența la insulină, hiperglicemia sau un nivel mai ridicat al grăsimilor în sânge [93].

În literatură sunt indicați o serie de factori genetici care pot predispune o persoană către această afecțiune. Un studiu prezentat de *Romeo* și colaboratorii săi în lucrarea [94] indică o posibilă conexiune cu variante genetice care se află în gena PNPLA3. Studiul a constat în analiza, folosind metoda *Genome-Wide Association Scan* (GWAS), a unui lot de 9229 de pacienți de diferite etnii. Studiul arată că SNP-ul rs738409, din gena PNPLA3, este puternic corelat cu un nivel ridicat de grăsimi hepatice și cu inflamație hepatică. Un alt studiu de către *Kotromen* și colaboratorii lui [95] prezintă un model care folosește o serie de parametri, precum diabetul de tip 2, sindromul metabolic (MetS), pentru a prezice prezența afecțiunii cu o specificitate de 71% și cu o sensibilitate de 86%. Autorii integrează și prezența variantei rs738409 pentru a crește performanța modelului, dar îmbunătățirea este mai mică de un procent (1%). Un alt grup care a realizat analiza variantei rs738409 în [96], de data aceasta pentru copii, confirmă că aceasta variantă este asociată cu severitatea steatozei non-alcoolice. Tot la copii, în [97] mai este confirmat faptul că varianta rs738409 este asociată cu predispoziția afecțiunilor de ficat. Pentru adulți au fost o serie de lucrări [98]–[101] care confirmă că această variantă este asociată cu steatoza. Recent a fost publicată o lucrare de către *Sookoiam* și *Pirola* [102] în care este prezentată o analiză a descoperirilor genetice care sunt asociate cu boala ficatului gras. Mai mult, autorii prezintă și o analiză a căilor metabolice (reglatoare) afectate.

Steatoza este considerată o afecțiune complexă. Aceasta are mai mulți factori care contribuie la manifestarea ei, precum a fost prezentat anterior. Acești factori pot fi genetici, exceptând varianta rs738409 de pe gena PNPLA3. Prin urmare printr-un studiu care este prezentat în capitolul 5, se va urmări identificarea altor variante genetice care pot fi asociate cu steatoza non-alcoolică. De asemenea, în lista variantelor genetice investigate este și rs738409.

2.4.2. Metode utilizate pentru detectarea afecțiunilor complexe

Odată cu evoluția procesului de secvențiere, metodele de analiză a datelor au fost mutate de la analiza clasică (cu creionul pe hârtie) către calculatoare. Inițial, o simplă stație de lucru era suficientă pentru analiza unui lot de pacienți, dar, odată cu accesibilitatea tehnologiei de secvențiere, a crescut și volumul de date pe care calculatoarele trebuiau să le analizeze. Acest lucru a condus către stații mult mai performante, dar și către dezvoltarea și aplicarea unor metode mai complexe și specializate.

Inițial, studiile care tratau problema afecțiunilor complexe erau studii de asociere. Practic, se studia frecvența apariției unor variante genetice între lotul de control, format din pacienți neafectați, și lotul de pacienți afectați. O frecvență mai ridicată în lotul celor afectați indica faptul că acea variantă poate fi asociată cu afecțiunea sau poate este asociată cu o altă variantă apropiată care este cauzatoare. Desigur, prezența variantei nu semnifică direct prezența respectivei afecțiuni, ci reprezintă doar un risc pentru afecțiune. Mai recent, aceste studii s-au transformat în studii de asociere pe întregul genom (*Genome-Wide Association Studies*, GWAS).

Cea mai cunoscută metodă pentru validarea unei variante într-un studiu de asociere este reprezentat de calcularea valorii *p (p-value)* din testul ipotezei nule (*null hypothesis,* H_0). Această valoare se calculează folosind o serie de metode precum regresia liniară, regresia logistică etc. Desigur, valoarea *p* va fi diferită pentru fiecare SNP dintr-un studiu. Există și o serie de limitări. Valoarea *p*, pentru același SNP, diferă de la studiu la studiu. Testul mai poate fi afectat de numărul și genotipul pacienților înscriși în studiu [103].

O alternativă la această metodă o reprezintă metodele bayesiene sau metodele utilizate în învățarea automată. Arborii decizionali, metode de tip *Ensemble*, rețelele neuronale (NN) pot identifica mai ușor tiparele care apar în seturi de date multidimensionale. Un studiu realizat de *Garcia* [104] demonstrează că analiza cu ajutorul arborilor decizionali poate produce modele predictive bune, cu o precizie de peste 85%, pentru determinarea unor caracteristici ale fenotipului. Într-o altă lucrare, redactată de *Uppu* și colaboratorii lui [105], autorii discută diferite metode ale învățării automate utilizate pentru studierea asocierilor la nivelul genomului. De asemenea, în prezenta lucrare se vor atinge diferite aspecte ale învățării automate utilizate atât pentru filtrarea variantelor genetice, cât și pentru identificarea afecțiunilor complexe. Astfel de implementări găsim pentru o serie de afecțiuni precum scleroză multiplă (NN) [106], *Alzheimer* (NN) [107], cancer ovarian (SVM) [108].

3. IDENTIFICAREA ELEMENTELOR GENETICE CAUZATOARE DE AFECȚIUNI

3.1. Identificarea variantelor genetice

Pentru extragerea variantelor genetice ale unui pacient trebuie să apelăm la fluxul prezentat în subcapitolul 2.1.6. Acest proces va genera un fișier care va conține toate variantele genetice. În acest caz, variante genetice reprezintă atât SNP-uri cât și *indel*-i. Cele din urmă se regăsesc sub formă de inserții sau ștergeri („deleții") ale unor baze azotate din molecula de ADN.

3.2. Identificarea genelor cauzatoare pe baza simptomelor

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, aceștia sunt nevoiți să identifice care sunt genele responsabile pentru afecțiunile de care suferă pacientul. În general, problema nu este lipsa datelor sau lipsa informațiilor, ci procesarea, filtrarea și prioritizarea acestora. Un mod în care medicii pot aduna informații suplimentare este să apeleze la o serie de instrumente software regăsite în diverse forme de la aplicații apelabile din linia de comandă (ANNOVAR) până la baze de date online care au o interfață foarte ușor de utilizat (*Ensemble, Variant Effect Predictor*).

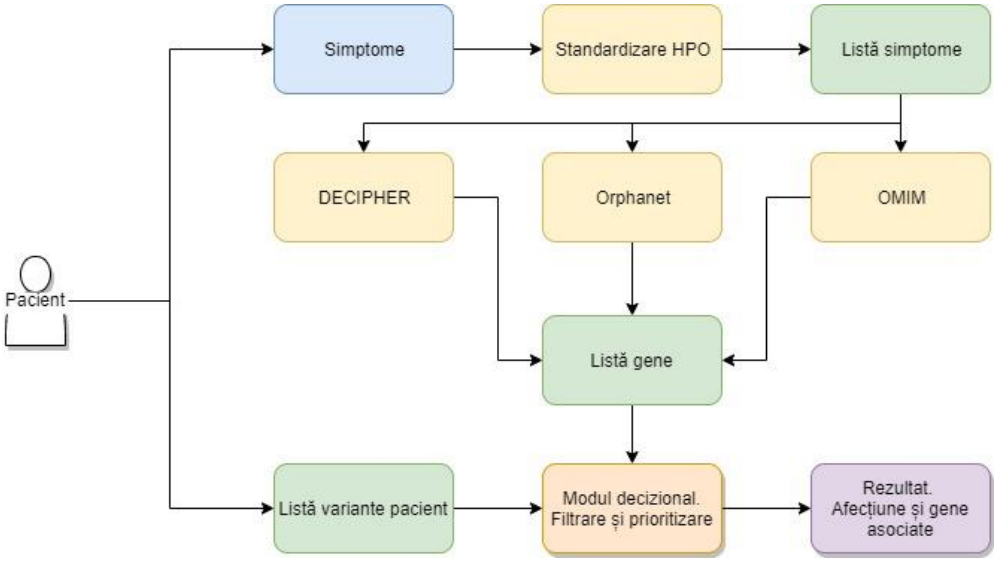


Fig. 3.1 Diagrama analizei afecțiunii pe baza simptomelor și variantelor genetice

În cele ce urmează vom propune un mod automat de extragere a afecțiunilor pe baza simptomelor și identificarea genelor sau un mod de extragere a posibilelor

gene care cauzează aceste afecțiuni. După identificarea unor posibile gene sau afecțiuni, acestea vor fi corelate cu gene obținute din fișierul de secvențiere al pacientului pentru identificarea genelor cauzatoare.

Pentru realizarea unui sistem automat de identificare a posibilelor gene cauzatoare de afecțiuni, se vor folosi mai multe baze de date disponibile online. Primele informații disponibile imediat clinicienilor la întâlnirea cu pacientul sunt simptomele acestora. Pentru a obține o listă de simptome vom folosi baza de date *Human Phenotype Ontology* (HPO). Această bază de date reprezintă un dicționar al termenilor care descriu anumite caracteristici ale fenotipului. Lista afecțiunilor o vom obține din bazele de date: *Online Mendelian Inheritance in Man* (OMIM), *Orphanet* și DECIPHER.

În Fig. 3.1 este prezentată schema generală a sistemului pentru analiza simptomelor și a variantelor genetice. În prima fază, medicul identifică simptomele și le introduce conform standardizării HPO. După acest pas, sistemul va avea la dispoziție o listă de simptome standardizate. Această listă va fi folosită mai departe pentru extragerea listei de afecțiuni care sunt asociate cu aceste simptome. Pe baza listei de afecțiuni se va întocmi o listă de gene. Fiecărei gene i se va asocia o pondere internă care va indica nivelul asocierilor multiple. În primă fază această pondere, P_{gena}, se va calcula prin împărțirea întregului la numărul de afecțiuni care sunt asociate cu acesta, precum este redat în ecuația (3.2-1). Aceeași metodă se poate aplica și pentru simptome, astfel având o pondere, $P_{simptom}$ (3.2-2), asociată fiecărui simptom.

$$P_{gena} = \frac{1}{numarul\ afectiunilor\ asociate} \tag{3.2-1}$$

$$P_{simptom} = \frac{1}{numarul\ afectiunilor\ asociate} \tag{3.2-2}$$

3.2.1. Distribuția afecțiunilor, simptomelor și a genelor

Pentru început s-a făcut o evaluare a numărului de caracteristici ale fenotipului care sunt asociate cu fiecare genă. Gena care are cele mai multe caracteristici asociate este LMNA, aceasta regăsindu-se în 415 fenotipuri. Următoarea în listă este KRAS, care are 310 fenotipuri asociate. La polul opus avem 42 de gene care sunt asociate cu un singur fenotip. În Tabelul 3.1 avem prezentate în partea stângă primele zece cele mai asociate gene cu caracteristicile ale fenotipului, iar în partea dreaptă a tabelului ultimele zece gene. De asemenea, pentru fiecare genă s-a calculat și indicele P_{gena}.

Tabelul 3.1 Primele zece gene asociate cu mai multe caracteristici ale fenotipului (stânga) și zece gene asociate cu o singură caracteristică (dreapta).

Nr	Gena	Caracteristici (fenotipuri)	P_{gena}	Nr	Gena	Caracteristici (fenotipuri)	P_{gena}
1	LMNA	415	0.00241	1	ND3	1	1
2	KRAS	310	0.00322	2	XIST	1	1
3	FGFR2	298	0.00335	3	MPZL2	1	1
4	FGFR1	289	0.00346	4	ARHGEF6	1	1

5	FLNA	277	0.00361	5	IFNAR2	1	1
6	FGFR3	269	0.00371	6	AGTR2	1	1
7	ELN	259	0.00386	7	AK7	1	1
8	COL2A1	243	0.00411	8	RELB	1	1
9	BRAF	242	0.00413	9	CYP2C19	1	1
10	FBN1	237	0.00421	10	AMTN	1	1

Aceeași metodă a fost aplicată și pentru calcularea ponderii pentru simptome, $P_{simptom}$. În cazul simptomelor, anomalia sistemului nervos ocupă topul având aproape 3000 de gene asociate cu acesta. De altfel, nouă dintre cele zece caracteristici sunt anomalii. Acest lucru este oarecum de înțeles pentru că anomaliile tind să caracterizeze ceva general. La polul opus s-au găsit 1980 de caracteristici ale fenotipului care sunt asociate cu o singură genă. În Tabelul 3.2 avem prezentate primele zece caracteristici ale fenotipului care au asociate cele mai multe gene (de la 1a până la 10a). De asemenea, mai avem prezentate o parte din caracteristicile fenotipului care au asociate o singură genă (1b până la 10b).

Tabelul 3.2 Caracteristicile fenotipului care au asociate cele mai multe gene (1a-10a) și caracteristicile care au asociate cele mai puține gene (1b-10b)

Nr.crt	Simptome (fenotip)	Gene	$P_{simptom}$
1a	Anomalia sistemului nervos	2955	0.000338
2a	Anormalitatea fiziologiei sistemului nervos	2749	0.000364
3a	Anomalii ale capului sau gâtului	2377	0.000421
4a	Anomalii ale capului	2354	0.000425
5a	Anomalia sistemului osos	2312	0.000433
6a	Anomalii ale ochilor	2294	0.000436
7a	Anomalii ale morfologiei osoase	2228	0.000449
8a	Moștenire recesivă autosomală	2218	0.000451
9a	Anomalii ale morfologiei sistemului nervos	2212	0.000452
10a	Anomalii ale feței	2159	0.000463
1b	Anormalitatea celui de-al doisprezecelea nerv cranian	1	1
2b	Nivelul crescut de D-tritol în plasmă	1	1
3b	A 5-a vertebră lombară hipoplazică	1	1
4b	Morfologie areolară anormală	1	1
5b	Număr crescut de celule B	1	1
6b	Rigiditate generalizată de dimineață	1	1
7b	Aphakia congenitală	1	1
8b	Vertebre sub formă de pere	1	1
9b	Spasticitatea mușchilor faringieni	1	1
10b	Microptalm unilateral	1	1

Pentru calcularea ponderii fiecărei afecțiuni s-au folosit termenii HPO, astfel încât să se calculeze nivelul de complexitate necesar pentru ca afecțiunea să fie luată în considerație. În cazul acesta sindromul *Williams* este determinat de cei mai mulți factori HPO, având asociate 172 de caracteristici ale fenotipului. Numărul afecțiunilor mono-caracteristice a fost de 199. În Tabelul 3.3 regăsim o listă cu afecțiunile care au asociate cei mai mulți termeni HPO și o listă cu afecțiunile care au asociate câte un singur termen HPO. O parte dintre afecțiunile care au asociate o singură caracteristică HPO au asociată o singură genă (1b–4b). Dacă ne uităm în lista

afecțiunilor care au multipli termeni HPO, vom observa că majoritatea fac parte din lista bolilor rare. De exemplu, sindromul *Lowe* are o frecvență de 1 la 500.000 de oameni; cu sindromul *Simpson-Golabi-Behmel* au fost diagnosticați peste 250 de oameni pe întregul glob; sindromul *Williams* apare la 1 din 7.500 de oameni, iar sindromul *Schwartz-Jampel* a fost raportat doar în 150 de cazuri în literatura de specialitate. Pentru a afla informații suplimentare despre sindroamele menționate se poate folosi baza de date disponibilă pe situl *National Institutes of Health*, la rubrica *Genetics Home Reference*.

Tabelul 3.3 Afecțiuni care au mai mulți termeni HPO (1a-10a) și afecțiuni care au câte un singur termen HPO (1b-10b)

Nr.crt	Afecțiune	Caracteristici	$P_{afectiune}$
1a	Sindromul *Williams*	172	0.0058
2a	Sindromul *Simpson-Golabi-Behmel*	127	0.0078
3a	Sindromul *Rubinstein-Taybi*	122	0.0081
4a	Sindromul *Wiedemann-Rautenstrauch*	121	0.0082
5a	Sindromul deleției 22q11.2	119	0.0084
6a	Sindromul *Lowe*	118	0.0084
7a	Sindromul *Myhre*	118	0.0084
8a	Sindromul *Schwartz-Jampel*	108	0.0092
9a	Sindromul *Rothmund-Thomson*	107	0.0093
10a	MELAS	105	0.0095
1b	Afecțiunea *Hirschsprung*, HSCR3	1	1
2b	Eșecul Spermatogenic 20; SPGF20	1	1
3b	*Retinitis Pigmentosa* 81; RP81	1	1
4b	Surditate; Dominant Autosomal 74; DFNA74	1	1
5b	Dominanta Oculară	1	1
6b	Megalodactilia	1	1
7b	Anodonția	1	1
8b	Osteoscleroză	1	1
9b	Hipospadias	1	1
10b	Diabetul zaharat, dependent de insulină, 10	1	1

3.2.2. Filtrarea afecțiunilor pe baza simptomelor

Odată ce sistemul are toate caracteristicile fenotipului introduse, acesta poate întocmi o listă de gene care sunt de interes. Pentru generarea listei de gene, sistemul trebuie să aibă acces la lista completă a asocierilor dintre caracteristicile fenotipului și gene. De asemenea, este necesară lista completă a asocierilor dintre caracteristicile fenotipului și afecțiuni.

Inițial se întocmește mulțimea caracteristicilor fenotipului, $L_{simptome}$, care va fi folosită pentru obținerea unei liste provizorii de afecțiuni. Pentru ca o afecțiune A să se regăsească în mulțimea presupuselor afecțiuni, aceasta ar trebui să aibă toate caracteristicile asociate, $L_{caracteristici}$, incluse în $L_{simptome}$. Cu alte cuvine, mulțimea caracteristicilor afecțiunii, $L_{caracteristici}$, trebuie să fie aceeași cu mulțimea creată de către specialist, $L_{simptome}$ (cazul ideal). În mod uzual, însă, specialistul nu are acces la toate simptomele pacientului. Este posibil ca unele simptome să se obțină după o serie de investigații care necesită o perioadă îndelungată de analiză (IRM, CT etc.). Pe de altă parte, niciodată pacientului nu i se vor solicita toate testele posibile pentru

a avea o descriere completă a fenotipului. Prin urmare, $L_{simptome}$ va fi incompletă în majoritatea cazurilor. Rezultă că verificarea unu la unu nu este eficientă în situația practică. Pentru a rezolva acest handicap se va calcula un coeficient de similaritate, CS_{AS}, al $L_{caracteristici}$ cu $L_{simptome}$, precum este prezentată în (3.2-3).

$$CS_{AS} = 1 - \frac{card\ (L_{caracteristici} - L_{simptome})}{card\ (L_{caracteristici})} \quad (3.2\text{-}3)$$

Acest coeficient de similaritate ne va prezenta doar câte dintre caracteristicile unei afecțiuni se regăsesc între cele introduse de către cadrul medical. Prin urmare, dacă lista caracteristicilor unor afecțiuni se află într-o ierarhie de includere, acest coeficient va indica potrivire perfectă, chiar dacă lista simptomelor conține mai multe caracteristici. Pentru a verifica relația inversă, dacă lista simptomelor este acoperită în întregime, s-a definit coeficientul de similaritate între $L_{simptome}$ cu $L_{caracteristici}$, denumit CS_{SA}. Ecuația pentru calcularea acestui coeficient este prezentat în (3.2-4).

$$CS_{SA} = 1 - \frac{card\ (L_{simptome} - L_{caracteristici})}{card\ (L_{simptome})} \quad (3.2\text{-}4)$$

3.2.3. Filtrarea simptomelor pe baza genelor

După întocmirea unei liste de posibile afecțiuni este important să evaluăm dacă simptomele se regăsesc în lista genelor. Pentru efectuarea acestei verificări vom apela la metoda descrisă în Fig. 3.2. Pentru fiecare afecțiune din lista afecțiunilor se va întocmi o listă a caracteristicilor care o definesc. Pe baza acestei liste se va genera o mulțime de gene care reprezintă reuniunea tuturor genelor asociate cu toate caracteristicile afecțiunii. De asemenea, se va înregistra și numărul apariției fiecărei gene în această reuniune. Identificarea genei sau genelor cauzatoare este limitată în primă fază de numărul genelor care au fost țintite în procesul de secvențiere. În unele cazuri se realizează secvențierea unui număr restrâns de gene, un panel. Acest panel poate conține de la câteva zeci până la câteva mii de gene. Ideal, se dorește secvențierea întregului exom (toate genele). Dacă situația nu este ideală, vom aplica o mască folosind panelul țintit de gene. Această mască are menirea de a elimina genele care nu se regăsesc în panel. Lista genelor rezultate din acest proces va fi comparată cu genele identificate în genotipul pacientului ca având variante genetice posibil dăunătoare.

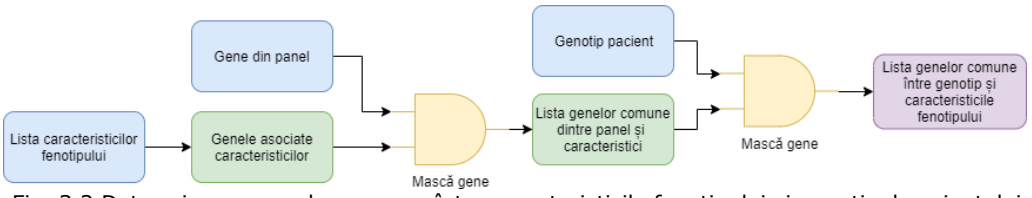


Fig. 3.2 Determinarea genelor comune între caracteristicile fenotipului și genotipul pacientului

Pentru a exemplifica acest proces vom analiza un pacient care suferă de insuficiență hepatică. În cazul acestui pacient, afecțiunea nu este cauzată de consumul excesiv de alcool sau de consumul de droguri. Prin urmare, se caută o explicație genetică. Dacă interogăm baza de date care conține lista caracteristicilor fenotipice pentru insuficiență hepatică vom obține o listă de 16 caracteristici. Pe baza caracteristicilor vom obține o listă de gene implicate direct sau indirect cu cele 16 caracteristici, anume de 2730 gene. Peste această listă vom aplica o mască de 4800 gene, specifică unui panel comercial. După aplicarea măștii, lista genelor de interes conține 1956 de gene. Următorul pas este extragerea genelor care au fost găsite în urma secvențierii. Desigur, înainte de acest pas se va face o filtrare primară prin care se exclud: (1) variantele genetice sinonime, (2) variantele care nu se află în zona exonică sau în zona de matisare. Dacă aplicăm și cea de-a doua mască, lista variantelor obținute de la pacient, vom obține 706 gene finale care sunt asociate cu insuficiența hepatică.

După efectuarea procedurii prezentate în Fig. 3.2, se caută care dintre simptomele asociate patologiei sunt confirmate de către lista genelor aferente pacientului. Primul pas îl reprezintă identificarea insuficienței hepatice în lista de termeni HPO. În această listă regăsim insuficiența hepatică sub denumirea de *Hepatic Failure*. Dacă se face o căutare vom identifica următorii termeni: *HP:0001399 Hepatic Failure*, *HP:0006554 Acute Hepatic Failure, HP:0100626 Chronic Hepatic Failure, HP:0004448 Fulminant Hepatic Failure*. Din lista prezentată, ne dăm seama că insuficiența hepatică (*HP:0001399*) este o categorie care include celelalte trei subcategorii. Insuficiența hepatică acută (*HP:*0006554) reprezintă o afecțiune care se manifestă într-un ritm alert, efectele fiind vizibile în intervale de timp de ordinul zilelor sau săptămânilor. Insuficiența hepatică cronică (*HP:0100626)* are un ritm de manifestare mai lent, efectele acesteia fiind vizibile în timp, după luni sau chiar ani. Termenul de fulminantă se folosește pentru a indica apariția encefalitei hepatice, în cazul *HP:0004448*, aceasta apare în 8 săptămâni de la primele simptome. În acest punct este relevant să știm faptul că pacientul investigat nu suferă de la insuficiență hepatică cronică, în cele mai multe cazuri fiind rezultatul consumului excesiv de alcool. Caracterul fulminant al afecțiunii nu este cunoscut. Prin urmare vom filtra rezultatele pentru insuficientă hepatică acută (*HP:0006554)*.

Dacă interogăm baza de date pentru insuficiență hepatică acută se obține următoarea listă de gene: LARS, CYC1, SCYL1, JAK2, MST1, MPV17, TRMU, FAH, PORCN, NBAS, XIAP, CACNA1S, TCF4, IKZF1, HADH, ATP7B, HLA-B, SH2D1A, GFM1, F5, EIF2AK3, GPR35, ACAD9, MEFV, RYR1, VPS13A. Acum vom verifica aceste gene în lista genelor găsite la pacient. Genele identificate sunt următoarele: MST1, NBAS, TCF4, HADH, ATP7B, HLA-B, GFM1, F5, EIF2AK3, MEFV, RYR1. Mai departe pentru evaluarea genelor vom folosi indicele P_{gena}, valorile pentru acest indice fiind prezentate în Tabelul 3.4. Dacă facem o ierarhie observăm că variantele găsite în genele GFM1, F5, NBAS ar fi prioritare în comparație cu variantele din RYR1, TCF4, HLA-B.

Tabelul 3.4 Lista genelor asociate cu insuficiența hepatică acută și ponderea acestora.

Genă	Caracteristici asociate	P_{gena}
MST1	49	0,0204
NBAS	42	0,0238
TCF4	129	0,0077
HADH	43	0,0232

ATP7B	62	0,0161
HLA-B	108	0,0092
GFM1	23	0,0434
F5	33	0,0303
EIF2AK3	82	0,0121
MEFV	93	0,0107
RYR1	176	0,0056

În Tabelul 3.5 avem prezentată o listă de posibile variante genetice care ar putea cauza insuficiența hepatică. Desigur aceste variante genetice nu contribuie toate la afecțiune. Folosind acest tabel, mai este necesar să se facă un studiu al populației prin care să se elimine variantele care sunt comune în rândul acesteia. În Tabelul 3.5 avem reprezentată, pe coloana AF, frecvența în populația globală.

Tabelul 3.5 Variantele genetice posibil candidate pentru insuficiență hepatică

Chr	Start	End	Ref	Alt	Genă	AF
chr1	169510475	169510475	G	T	F5	0.1628
chr1	169519049	169519049	T	C	F5	0.9827
chr2	15519924	15519924	C	T	NBAS	0.0228
chr2	15607842	15607842	T	C	NBAS	0.4949
chr2	15674686	15674686	T	C	NBAS	0.5155
chr2	15676686	15676686	G	A	NBAS	0.6191
chr2	88874891	88874891	C	A	EIF2AK3	0.7678
chr2	88882942	88882942	T	A	EIF2AK3	0.2177
chr2	88895123	88895123	T	C	EIF2AK3	0.7036
chr3	49723735	49723735	C	G	MST1	0.0086
chr3	49724183	49724183	C	G	MST1	0.1744
chr3	49726028	49726028	T	C	MST1	0.2809
chr3	49726070	49726070	G	A	MST1	0.142
chr3	49935526	49935526	T	G	MST1R	0.0055
chr3	158366900	158366900	G	A	GFM1	0.5752
chr3	158409262	158409262	C	T	GFM1	0.4034
chr4	108931039	108931039	T	C	HADH	0.8666
chr13	52511606	52511606	G	A	ATP7B	0.5449
chr13	52524488	52524488	T	C	ATP7B	0.5467
chr13	52544805	52544805	C	G	ATP7B	0.4194
chr13	52548140	52548140	A	C	ATP7B	0.4155
chr16	3293888	3293888	C	T	MEFV	0.4128
chr16	3293922	3293922	A	T	MEFV	0.394
chr18	53303101	53303101	C	G	TCF4	0.9988
chr19	38983180	38983180	G	T	RYR1	0.0458
chr19	38991640	38991640	C	G	RYR1	0.2345
chr19	38993372	38993372	A	G	RYR1	0.3475
chr19	38997459	38997459	G	C	RYR1	0.3159

3.3. Analiza performanței aplicațiilor de predicție *in silico*

O parte dintre informațiile prezentate în acest studiu au fost publicate în lucrarea intitulată *Performance Evaluation of in Silico Predictors for the Classification of ClinVar Variants* [109].

3.3.1. Pregătirea datelor

Baza de date *ClinVar* [32] a fost folosită pentru evaluarea performanței aplicațiilor de predicție în determinarea patogenității variantelor genetice. Setul de date, format din variantele genetice și clasa de patogenitate a acestora, a fost organizat în două fișiere. Primul fișier a conținut variantele genetice clasificate ca fiind patogene, iar cel de-al doilea fișier a conținut variante care erau clasificate benigne. Variantele care erau clasificate în două sau mai multe categorii, de exemplu patogen sau probabil patogen, au fost eliminate din setul de date. În plus, au fost selectate doar polimorfismele mononucleotidice (SNP) care au fost în regiunea exonică. De asemenea, variantele aflate pe cromozomul Y și pe ADN-ul mitocondrial au fost eliminate deoarece unele instrumente testate nu ofereau suport pentru aceste regiuni. După procesul de selecție, setul de date a fost format din 30.678 variante patogene și 15.397 variante benigne. Adnotarea a fost realizată cu ANNOVAR [45], folosind baza de date dbNSFP 35a [110], [111].

În fișierul cu variante patogene, după adnotare, majoritatea aplicațiilor aveau un număr de 17.000 de înregistrări. Excepțiile au fost MutationTaster, LRT, CADD și DANN, care au avut în jur de 30000 de înregistrări. Pentru fișierul benign, numărul de variante adnotate a fost cam în același interval de valori, în jur de 6.000 de variante pentru fiecare aplicație. Excepția a fost M-CAP, care a avut doar 1.237 de adnotări din totalul de 15.397. Pentru fiecare predictor, numărul de înregistrări era format dintr-un amestec de variante genetice din diferite regiuni ale ADN-ului (gene diferite, cromozomi diferiți) astfel evitându-se o repetiție a condițiilor asemănătoare.

Pentru analiza fișierului cu SNP-urile patogene, adnotat, doar variantele genetice clasificate ca fiind dăunătoare au fost luate în considerație. Semnalizarea patogenității variantei genetice se face diferit pentru fiecare aplicație. Unele aplicații folosesc litera *D* pentru a indica starea dăunătoare (SIFT, *PolyPhen-2*, LRT, FATHMM, PROVEAN, M-CAP, MetaSVM, MetaLR), în timp ce altele folosesc litera *H* (*Mutation Assessor, Mutation Taster*). Câteva dintre acestea (CADD, DANN, REVEL) aveau asociate praguri de patogenitate, astfel că dacă aceste valori depășeau acel prag, varianta era considerată patogenă. Pentru CADD, care are un interval de valori cuprins între 0-99, pragul este 20. Astfel, dacă scorul indicat de aplicație era mai mare decât 20 înseamnă că varianta genetică este patogenă. Intervalul de scor al predictorului DANN este cuprins între 0 și 1, deci o variantă este considerată patogenă dacă are scorul peste valoarea 0,9. Pragul pentru REVEL, pentru care o variantă era considerată cauzatoare de afecțiuni, este 0,7. SNP-urile clasificate în clasele de mijloc (de exemplu, probabil patogenă) au fost considerate inexacte și au fost ignorate.

Când fișierul cu variantele benigne, adnotat, a fost procesat pentru evaluarea performanței, numai clasificarea benignă a fost considerată valabilă. Clasificarea ambiguă a fost respinsă. Adnotarea pentru variantele benigne a fost litera *T* pentru SIFT, FATHMM, MetaSVM, MetaLR, M-CAP; litera *B* pentru PolyPhen-2; litera *N* pentru *Mutation Taster*, LRT, PROVEAN; litera *L* pentru *Mutation Assessor* sau, pentru valori, pragul de 20 pentru CADD, 0,9 pentru DANN și 0,5 pentru REVEL.

Pentru o evaluare cât mai riguroasă a performanței, trebuie abordată problema variantelor care au valori lipsă. Prin valori lipsă se înțeleg variantele genetice din setul de date care aveau un punct (valoare lipsă) după procesul de adnotare. Pentru calcularea caracteristicilor statistice, valorile lipsă au fost ignorate. Prin urmare, dimensiunea eșantionului a fost diferită pentru fiecare aplicație. Totuși, acest lucru nu afectează scopul experimentului, acesta fiind determinarea abilității unei aplicații să clasifice o variantă ca fiind patogenă sau benignă. Scopul nu a fost categorisirea unui număr cât mai mare de variante. Cu alte cuvine, scopul a fost unul calitativ și nu cantitativ.

3.3.2. Evaluarea performanței

Pentru fiecare aplicație au fost calculați o serie de indicatori de performanță. Sensibilitatea, ecuația (2.1-2), s-a folosit pentru determinarea proporției de variante patogene reale din numărul variantelor patogene prezise. Specificitatea, ecuația (2.1-2), s-a folosit pentru determinarea proporției de variante benigne reale din totalul variantelor benigne prezise. Precizia a fost calculată pentru a evalua raportul relevant și cel irelevant al rezultatelor. Media armonică (scorul F_1), prezentată în ecuația (3.3-1), este utilizată pentru a compara clasificatorii.

$$\mathrm{F}_1 = \frac{2 \times \text{precizie} \times \text{sensibilitate}}{\text{precizie} + \text{sensibilitate}} \tag{3.3-1}$$

Deoarece media armonică se concentrează pe determinarea clasei pozitive, am utilizat suplimentar media aritmetică între sensibilitate și specificitate, prezentată în

(3.3-2). Parametrii folosiți în ecuația $Mean_{ss}$ (AP, AN, FP, FN) reprezintă valorile matricei de contingență (de exemplu, AP, adevărat pozitiv).

$$Mean_{ss} = \frac{\frac{AP}{AP + FN} + \frac{AN}{AN + FP}}{2} \tag{3.3-2}$$

3.3.3. Rezultatele evaluării performanței

Rezultatele analizei statistice pentru setul de date cu variante patogene, clasa pozitivă, sunt prezentate în Tabelul 3.6. Valorile indicatorilor de performanță au fost calculate în raport cu numărul total de variante adnotate prezente în fișier după etapa de preprocesare. Numărul total de variante adnotate este prezentat în coloana *variante adnotate* a tabelului. Coloana adevărat pozitiv (AP) conține numărul de variante patogene care au fost clasificate corect. Coloana fals negativ (FN) reprezintă numărul de variante patogene care au fost clasificate ca benigne. În unele cazuri (de exemplu, LRT) suma valorilor AP și FN nu este aceeași cu numărul total de variante adnotate. Diferența este reprezentată de clasificarea ca variante incerte, iar acestea au fost ignorate pentru calcularea indicatorilor de performanță.

Tabelul 3.6 Rezultatele analizei variantelor patogene din *ClinVar*

Predictor	Variante adnotate	Adevărat pozitiv	Fals negativ	Sensibilitate [%]
SIFT	17263	15328	1935	88.8
PolyPhen-2 HDIV	17631	14162	1568	90.0
PolyPhen-2 HVAR	17631	12763	2321	84.6
Mutation Taster	30319	12418	605	95.4
LRT	28988	21210	5886	78.3
Mutation Assessor	17088	6269	2025	75.6
FATHMM	17346	12531	4815	72.2
PROVEAN	17468	14641	2827	83.8
MetaSVM	17653	14075	3578	79.7
MetaLR	17653	13992	3661	79.3
M-CAP	17214	16661	553	96.8
CADD	30443	29274	1169	96.2
DANN	30443	29879	564	98.1
REVEL	17653	12939	2143	85.8

Același proces a fost aplicat și pentru fișierul cu variantele benigne; rezultatele sunt prezentate în Tabelul 3.7. În acest caz, coloana *variante adnotate* conține numărul de variante benigne, clasa negativă, care au avut o adnotare validă. Coloana *adevărat negativ* reprezintă variantele benigne identificate corect, iar *fals pozitiv* reprezintă variantele benigne identificate ca fiind patogene.

Tabelul 3.7 Rezultatele analizei variantelor benigne din *ClinVar*

Predictor	Variante adnotate	Adevărat negativ	Fals pozitiv	Specificitate [%]
SIFT	6006	4117	1889	68.6
PolyPhen-2 HDIV	6062	3955	1243	76.1
PolyPhen-2 HVAR	6062	4635	795	85.4
Mutation Taster	6223	2012	2032	49.8
LRT	5333	3650	1383	72.5
Mutation Assessor	5753	1861	109	94.5
FATHMM	5985	4436	1548	74.1
PROVEAN	6026	4820	1205	80.0
MetaSVM	6158	5619	538	91.3
MetaLR	6158	5510	647	89.5
M-CAP	1237	411	825	33.3
CADD	6295	4043	2251	64.2
DANN	6295	2569	3725	40.8
REVEL	6158	5624	159	97.3

Cele mai bune aplicații pentru clasa pozitivă, măsurate prin sensibilitate, sunt DANN (98,1%), M-CAP (96,8%) și CADD (96,2%). Este important de menționat că M-CAP are doar 56% din setul de date adnotat, în timp ce CADD și DANN au aproximativ 90%. Deși toți trei sunt meta-predictori și au integrat în modelele lor predictori precum SIFT și *PolyPhen-2*, unele dintre variantele clasificate corect de către SIFT și *PolyPhen-2* nu au fost clasificate corect de către acestea, așa cum este prezentat în Fig. 3.3. Predictorii cu cele mai mici scoruri, pentru clasa pozitivă, sunt FTHMM (72,2%), *Mutation Assessor* (75,6%) și LRT (78,3%).

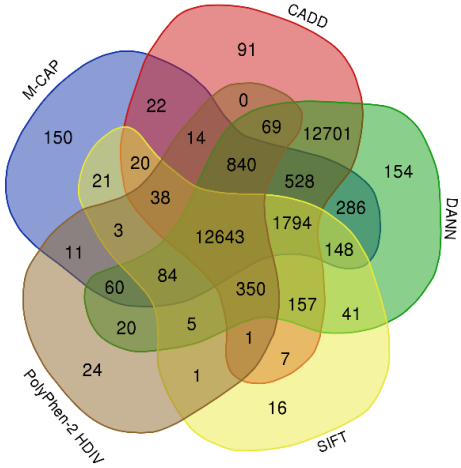


Fig. 3.3 Numărul de variante patogene suprapuse, clasificate în mod corect de către CADD, DANN, M-CAP, SIFT și *PolyPhen-2*

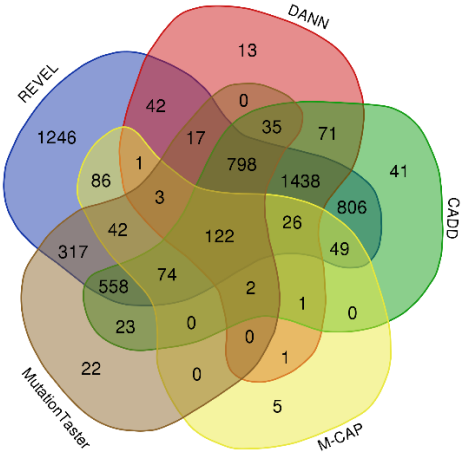


Fig. 3.4 Numărul de variante benigne suprapuse, clasificate corect de către CADD, DANN, M-CAP, REVEL și *Mutation Taster*

Instrumentele cu cele mai mici scoruri pentru predicția clasei negative, măsurate prin specificitate, sunt *Mutation Taster* (49,8%), M-CAP (33,3%) și DANN (40,8%). Este destul de evident că în cazul instrumentelor DANN și *Mutation Taster* lipsa adnotării nu este cauza. Aceste aplicații au adnotate peste 85% dintre variantele genetice din fișier. Cu toate acestea, analiza intersecției variantelor identificate corect, realizată în Fig. 3.4, relevă faptul că unele dintre variantele benigne detectate de REVEL nu au fost detectate de cele mai bune instrumente de la clasa pozitivă (CADD, DANN și M-CAP). Dacă ne uităm la rezultatele pentru DANN și M-CAP, putem presupune că acestea sunt supra-potrivite pentru clasa pozitivă. Cei mai buni predictori pentru detectarea variantelor benigne sunt *Mutation Assessor* (94,5%), REVEL (97,3%), MetaSVM (91,3%).

Versiunile *PolyPhen-2* au obținut un scor bun pentru sensibilitate și specificitate. *PolyPhen-2* HVAR se remarcă printr-un scor echilibrat 84,6%, respectiv 85,4%. Un alt instrument echilibrat este PROVEAN cu sensibilitate de 83,8% și specificitate de 80,0%. În ceea ce privește ceilalți predictori din listă, aceștia au obținut, de asemenea, o performanță decentă în raport cu dimensiunea setului specific de date.

Dacă scorul F_1 este utilizat pentru a evalua performanța, predictorii de top sunt M-CAP (96%), CADD (94,5%) și DANN (93,3%), așa cum sunt prezentate în Fig. 3.5. LRT, FATHMM și *Mutation Assessor* au avut cele mai mici scoruri F_1. Principalul

contribuitor al acestui rezultat este faptul că toți cei trei predictori au scoruri mici pentru sensibilitate.

Motivul pentru care $Mean_{ss}$ a fost propus pentru măsurarea performanței este datorat faptului că ia în considerație clasa negativă, mai specific valoarea pentru adevărat negativ. Cu această măsură, REVEL (91,5%), MetaSVM (85,5%) și *PolyPhen-2* HVAR (85%), au cele mai bune performanțe, ceea ce înseamnă că sunt mai echilibrate. Predictorii cu cea mai mică performanță sunt M-CAP (65%), DANN (69,5%) și *Mutation Taster* (77%).

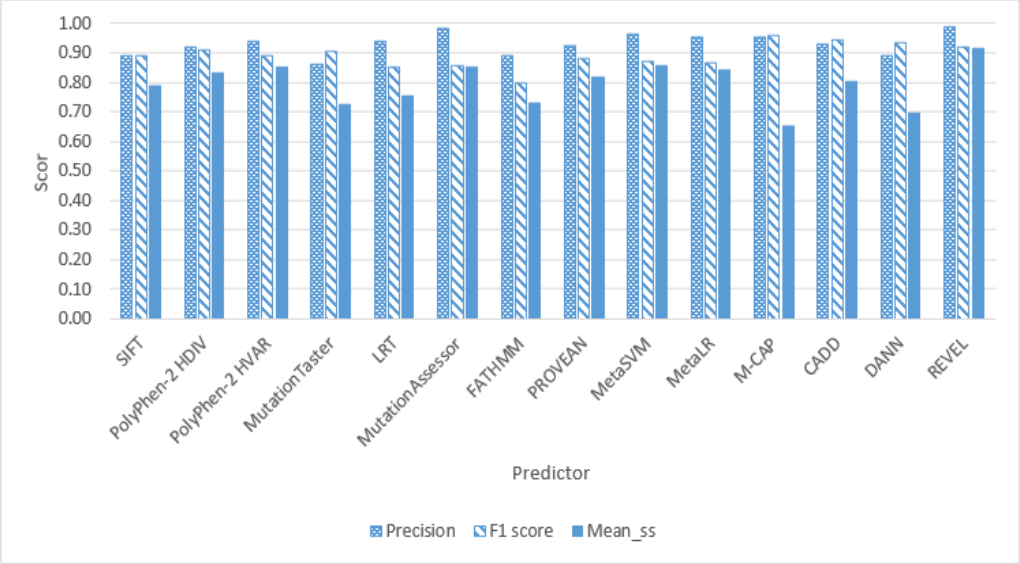


Fig. 3.5 Valorile obținute de predictori pentru fiecare indicator de performanță (precizie, F1, M_ss)

Există unele limitări la care este supusă această analiză a performanței. Una dintre aceste limitări este reprezentată de baza de date *ClinVar* care suferă actualizări frecvente. Unele variante genetice considerate benigne, în versiunea folosită pentru evaluarea performanței, de-a lungul timpului, ar putea fi evaluate și incluse în lista variantelor patogene, dacă cercetările ulterioare demonstrează patogenitatea lor. Pentru a diminua acest risc, au fost luate în considerație doar variante care au fost raportate fie patogene, fie benigne. Variantele care au avut o clasificare mixtă au fost ignorate. O altă limitare este adnotarea efectuată cu ANNOVAR folosind setul de date dbNSFP. Unii predictori aveau valori lipsă pentru un număr semnificativ de variante genetice. Valorile lipsă au fost excluse și nu au fost incluse pentru măsurări. S-ar putea ca unele dintre aceste variante să aibă un scor asociat în setul de date specific predictorului, dar care nu a fost integrat în dbNSFP.

3.4. Detecția erorilor de secvențiere din analiza fișierelor VCF

3.4.1. Erori în procesul de secvențiere

Cea mai folosită metodă de secvențiere a ADN-ului este secvențierea prin sinteză. Această metodă are o rată de erori de aproximativ 0.1% pentru fiecare nucleotidă. Aceste erori constituie în general SNP-uri. Sursele generării erorilor pot fi multiple, precum interferențe de culori, interferențe între clustere apropiate [112], desincronizarea bazelor azotate sau *dimming* [113]. Erorile de interferențe apar deoarece spectrul emis de floufori diferitelor baze azotate, din doua clustere învecinate, se suprapun. Desincronizarea bazelor azotate apare când într-un ciclu de secvențiere, la o secvența de ADN monocatenar, se adaugă două sau mai multe baze azotate în loc de una. Este valabil și dacă nu se adaugă nicio bază azotată. Efectul de *dimming* (sau întunecare) se datorează grupării unui număr mic de secvențe de ADN într-un cluster sau datorită deteriorării induse de laser asupra ADN-ului [114].

Pentru a preîntâmpina posibilele erori de citire a variantelor genetice, în acest studiu se vor urmări o serie de indicatori specifici în fișierele VCF. Indicatorii urmăriți au fost interschimbările dintre bazele azotate, mai precis tranzițiile și transversiile. Alt element urmărit a fost zigozitatea variantelor genetice și calitatea variantelor care au fost peste pragul minim. De asemenea, au mai fost analizate și caracteristicile *indel*-ilor și a variantelor complexe.

3.4.2. Materiale și metode

Pentru determinarea indicatorilor de interes au fost folosite trei grupuri de fișiere. Aceste fișiere au provenit din două surse diferite. Două dintre grupuri au aceeași sursă, iar pe baza acestora se va determina intervalul valorilor pentru indicatori. Cel de-al treilea grup va fi folosit pentru validarea rezultatelor. Primul grup de fișiere conține 106 probe biologice pentru care s-a efectuat analiza unui panel de gene asociate cu bolile cardiace (Cardio), conținând 175 de gene. Al doilea grup de fișiere (TSO) conține 74 probe biologice pentru care s-a efectuat secvențierea unui panel de gene cu caracter mai general, având 4.813 gene. Cel de-al treilea grup (WES) conține 16 probe biologice pe care s-a efectuat secvențiere *full exome*, care a avut ca țintă totalitatea genelor. Generarea fișierelor VCF a fost realizată cu aplicația recomandată de producător, iar pragul minim de calitate a fost setat la 20 (GQ).

Pentru a observa dacă fișierele care conțin erori au caracteristici diferite, comparativ cu cele valide, au fost adăugate câteva fișiere invalide, care conțin diferite erori precum: (1) adâncimea de secvențiere prea mică, (2) calitatea citirii sub pragul indicat sau (3) un număr redus de variante genetice. În cadrul primului grup au fost inserate trei probe, în cadrul celui de-al doilea grup fost inserate o serie de fișiere care nu aveau filtru de calitate, iar în al treilea grup au fost inserate două fișiere.

Fișierele au fost adnotate cu *ANNOVAR*, astfel conținând o serie de informații extrase din bazele de date, precum gnomAD, ExAC și predicatorii de patogenitate precum SIFT, CADD etc.

Extragerea datelor din fișiere CSV a fost realizată cu un script dezvoltat în cadrul laboratorului, folosind limbajul de programare *Python*. Pentru analiză, fișierele au fost stocate pe rând, în funcție de grupuri, într-un director de unde erau prelucrate de către acest script. Cu excepția antetului, toate liniile dintr-un fișier au fost preluate una câte una. Procesarea unei linii constă în identificarea câmpurilor unde se găsește

baza azotată specifică genomului de referință și varianta genetică asociată probei biologice. În cazul în care câmpurile conțineau caractere specifice nucleotidelor (A, T, C ,G) atunci acestea erau considerate schimbări singulare și erau contorizate la tipul tranziției, de exemplu A>G. În schimb, dacă în locul unei nucleotide s-a identificat caracterul „-" acesta a fost interpretat ca *indel*. În funcție de poziția pe care se afla caracterul „-", varianta genetică putea fi considerată inserție, dacă se afla pe poziția nucleotidei de referință, sau putea fi deleție, dacă se afla pe poziția nucleotidei aferentă probei biologice. Dacă înregistrarea nu se regăsea în niciuna dintre situațiile menționate anterior și unul dintre cele două câmpuri conțineau multiple nucleotide, atunci era considerată o variantă complexă. Pentru calcularea zigozității, se căutau pe fiecare linie caracterele specifice variantelor homozigote sau variantelor heterozigote, acestea neavând o poziție fixă în fișierul CSV.

3.4.3. Analiza grupurilor de fișiere

Pentru calcularea câmpurilor de interes, grupurile de fișiere au fost analizate separat. Din Fig. 3.6 observăm că valorile pentru interschimbările dintre nucleotide rămân într-un interval restrâns, nefiind diferențe majore. Excepție de la regulă fac fișierele invalide adăugate pentru testare. Desigur, valorile absolute pot să difere datorită procesului de pregătire a probelor biologice sau din motive care depind de aparatul de secvențiere. Pentru a scăpa de aceste interferențe, numărul variantelor se poate raporta la numărul total de variante din fișier.

În Fig 3.7 sunt reprezentate rezultatele pentru interschimbările nucelotidelor aferente grupului Cardio. Se poate observa că tranzițiile nucleotidelor sunt mai frecvente decât transversiile, ceea ce este firesc. O tranziție între nucleotide reprezintă o interschimbare între o purină cu o altă purină sau o pirimidină cu o altă pirimidină, adică (A>G, G>A, C>T și T>C). Transversiile reprezintă interschimbarea unei baze azotate purinice cu una pirimidinică. În Fig. 3.7 sunt prezentate rezultatele interschimbărilor raportate la numărul total de variante din fișier. Valorile încercuite reprezintă fișierele care au ieși din zona de cuartilelor, iar acestea reprezintă fișierele invalide.

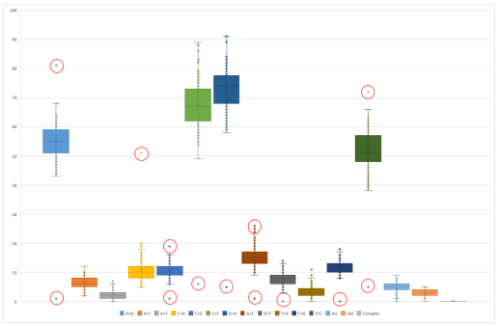


Fig. 3.6 Valorile absolute ale interschimbărilor dintre nucleotide obținute din fișierele aferente grupului *Cardio*.

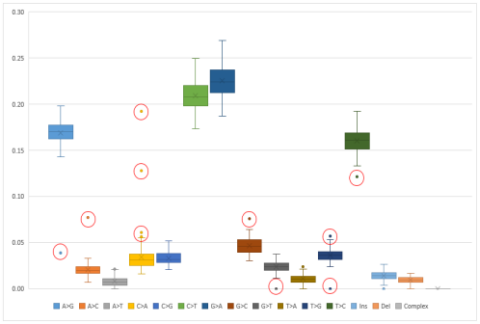
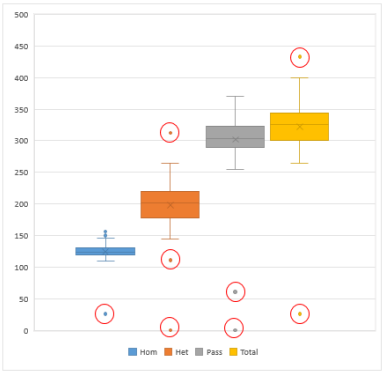
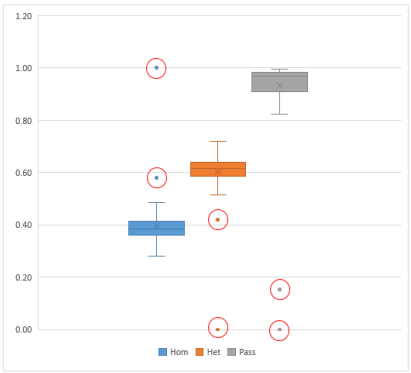


Fig. 3.7 Valorile raportate ale interschimbărilor dintre nucleotide obținute din fișierele aferente grupului *Cardio*.

În Fig. 3.8 regăsim rezultatele zigozității variantelor. Din aceste rezultate reiese faptul că variantele genetice heterozigote sunt mai numeroase decât variantele homozigote.



a) Valori absolute



b) Valori raportate

Fig. 3.8 a) Valorile absolute pentru variantele: homozigote, heterozigote; și calitate. b) Valorile raportate la numărul total de variante din grupul *Cardio*.

Raportul ideal dintre variantele homozigote este de 40%, iar pentru cele heterozigote este de 60%. Asta înseamnă că în raportul de secvențiere pentru *Cardio* ar trebui să regăsim indicatorul *hom/het* în jurul valorii 1,5. Dacă analizăm intervalul observăm că, în cazul inferior, indicatorul pentru panelul *Cardio* ar putea ajunge în apropierea valorii 1. Acest interval mare se poate explica prin numărul de gene care sunt secvențiate pentru fiecare grup. Prin urmare, datorită numărului mic de gene țintite vom avea un număr mai mic de variante genetice, ceea ce înseamnă că fiecare SNP are o pondere statistică mai mare. Raportul dintre variantele homozigote și cele heterozigote ajută la identificarea cazurilor de consangvinitate. Dacă privim Fig. 3.8 b), observăm o înregistrare care are 58% variante homozigote și 42% variante heterozigote, adică un raport 0,72. În acest caz, pacientul ar trebui să intre sub suspiciunea de consangvinitate.

În Fig. 3.9 și Fig. 3.10, regăsim intervalele pentru grupul TSO. În cazul acesta, intervalele sunt mai restrânse în jurul anumitor valori. Acest lucru este datorat numărului ridicat de variante care se regăsesc în cele 4.800 de gene. În cazul reprezentării valorilor absolute se pot vedea fișierele care nu au avut impus un prag de calitate, astfel generându-se un număr mult mai mare de variante genetice. Dar acest lucru nu se regăsește la reprezentarea raportului.

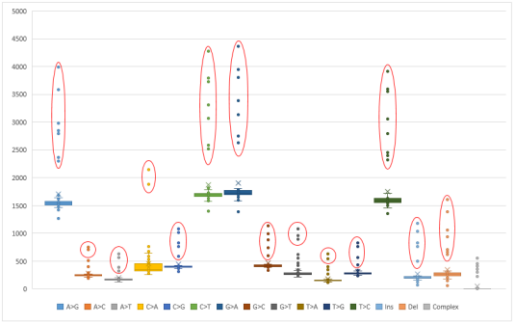


Fig. 3.9. Valorile absolute ale interschimbărilor dintre nucleotide, obținute din fișierele aferente grupului TSO.

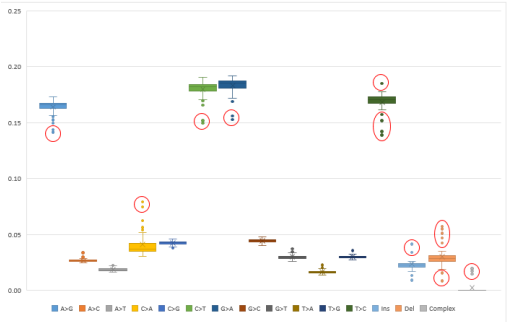
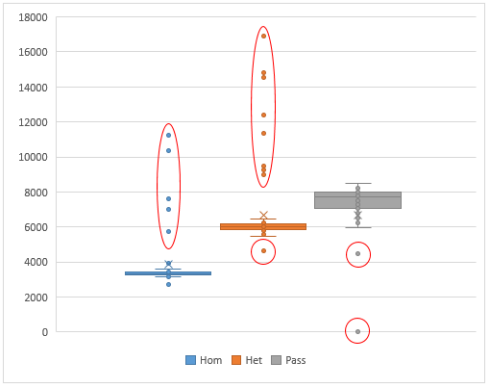


Fig. 3.10 Valorile raportate ale interschimbărilor dintre nucleotide, obținute din fișierele aferente grupului TSO.



a) Valori absolute

b) Valori raportate

Fig. 3.11 a) Valorile absolute pentru variantele: homozigote, heterozigote. b) Valorile raportate la totalul variantelor din grupul TSO.

În Fig. 3.11 avem reprezentată zigozitatea variantelor din grupul TSO. Spre deosebire de *Cardio*, intervalele sunt mai depărtate unele de altele. În cazul raportului, heterozigotele sunt peste 60%, iar homozigortele sunt sub 50%. În acest caz, raportul *het/hom* este în jurul valorii 1,7. Asemănător ca în cazul interschimbărilor, valorile sunt grupate în jurul unor intervale, excepție făcând fișierele invalide. Desigur, acestea sunt vizibile doar în Fig. 3.11 a). Din aceste date putem trage concluzia că erorile cauzate de calitate pot fi identificate cel mai ușor prin compararea numărului de variante.

Valorile pentru grupul WES, sunt prezentate în Fig. 3.12 și Fig. 3.13. Asemănător cu grupul TSO, intervalele sunt restrânse în jurul unor valori excepție făcând doar fișierele invalide care nu aveau filtrul de calitate aplicat. Dacă analizăm toate fișierele observăm că unele tranziții sunt mai frecvente decât altele. Între perechile A>G și G>A, ultima este cea mai frecvent întâlnită, iar in cazul citozinei și timinei tranziția C>T este cea mai întâlnită. Dintre *indel*-i, deleții se pare că sunt mai frecvente decât inserțiile.

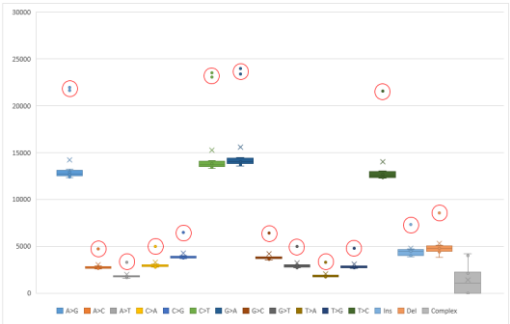


Fig. 3.12 Valorile absolute ale interschimbărilor dintre nucleotide obținute din fișierele aferente grupului WES.

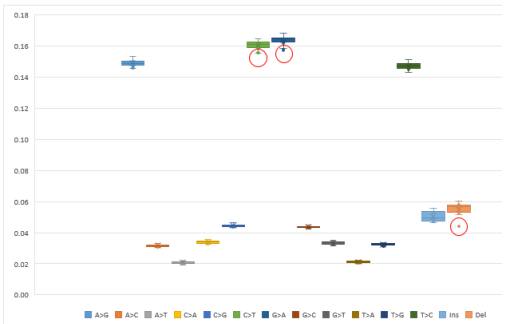


Fig. 3.13 Valorile raportate ale interschimbărilor dintre nucleotide obținute din fișierele aferente grupului WES.

a) Valori absolute b) Valori raportate

Fig. 3.14 a) Valorile absolute pentru variantele: homozigote, heterozigote; și calitate. b) Valorilor raportate la totalul de variante din grupul WES.

În cazul raportului dintre variantele homozigote și cele heterozigote, acesta se menține în jurul valorii de 1,7, asemănător cu cel al grupului TSO. În Fig. 3.14 avem reprezentate valorile pentru zigozitatea variantelor în grupul WES.

Odată cu transformarea valorilor absolute în valori procentuale se induce și un posibil caz de omitere. Dacă numărul de variante detectate este scăzut, ceea ce poate semnala o problemă de amplificare, dar procentul variantelor este în intervalul calculat, atunci acest caz de eroare nu va fi semnalat. Prin urmare, este necesară verificarea atât a intervalului valorilor absolute, cât și a intervalului procentual.

Intervalul calculat în funcție de numărul total de variante, este benefic pentru că indică o eroare sistematică a echipamentului. De asemenea, intervalul ne ajută să urmărim ponderea fiecărui tip de interschimbare din totalul variantelor genetice, ceea ce poate fi benefic pentru clinicienii care caută variante genetice rare. De asemenea, intervalul procentual al zigozității le poate indica geneticienilor anumite lucruri despre proba biologică. De exemplu, dacă variantele homozigote sunt mai numeroase decât cele heterozigote atunci poate fi vorba despre un caz de consangvinitate sau de rudenie apropiată.

Acest test de diagnoză a fișierelor VCF are o serie de beneficii, dar și o serie de limitări. Valorile obținute în acest studiu sunt informale și nu sunt universale. Fiecare laborator trebuie să facă propriile studii pentru a stabili care sunt intervalele valorilor absolute, dar și intervalele procentuale. În al doilea rând, dacă într-un fișier se găsește o abatere de la intervalul calculat, acest lucru nu implică obligatoriu existența unei erori de secvențiere. În schimb, această abatere poate indica o cale mai ușoară pentru rezolvarea cazului clinic.

3.4.4. Determinarea intervalelor de toleranță

Pentru fiecare interschimbare genetică, calcularea intervalelor de referință se realizează cu ajutorul cuartilelor mulțimii valorilor ordonate crescător. Cu alte cuvinte, se calculează valoarea mediană (Q2). Din grupul valorilor superioare acesteia, se calculează mediana superioară (Q3). Pentru obținerea medianei inferioare se va folosi

grupul de valori inferior medianei Q2, altfel obținându-se Q1. Pentru calcularea limitei inferioare (Li) a intervalului folosim formula (3.4-1), iar pentru calcularea limitei superioare (Ls) a intervalului folosim formula (3.4-2). Dacă în urma calculării limitei inferioare aceasta este negativă, atunci ea va deveni 0. Pentru grupurile de fișiere analizate, aceste valori au fost determinate și au fost reprezentate în Fig. 3.15.

$$Li = q1 - 1.5 * (q3 - q1) \tag{3.4-1}$$

$$Ls = q3 + 1.5 * (q3 - q1) \tag{3.4-2}$$

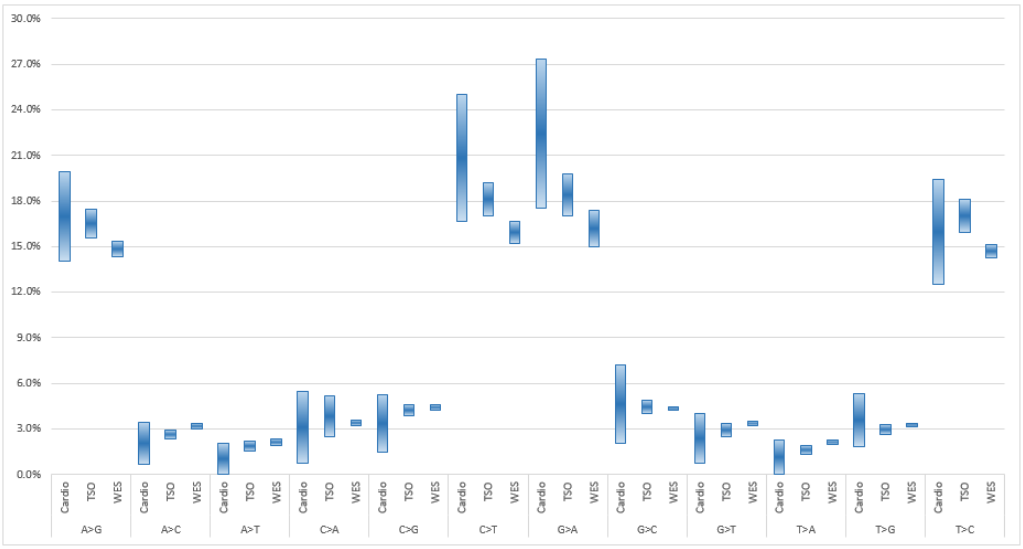


Fig. 3.15 Limitele intervalelor de acceptare aferente interschimbărilor bazelor azotate pentru fiecare grup (*Cardio*, TSO, WES).

În concluzie, pentru detecția erorilor de cantitate este recomandat să se folosească valorile absolute ale interschimbărilor și ale zigozității. Pentru determinarea calității procesului de secvențiere este recomandat să se folosească valorile raportate la numărul total de variante. De asemenea, pentru identificarea cazurilor de consangvinitate este recomandat să se folosească tot valorile procentuale. Desigur, în cazul panelurilor de gene mici intervalele pentru valorile raportate vor fi mai mari, iar pentru WES sau WGS intervalele vor fi mai restrânse.

Intervalele prezentate în Fig. 3.15 sunt orientative, dar în același timp acestea pot fi folosite ca referință pentru detecția erorilor. Este indicat ca fiecare laborator să folosească astfel de metode pentru determinare calității secvențierii. Desigur, fiecare laborator va avea setul de intervale proprii, dar în teorie acestea ar trebui să fie asemănătoare cu intervalele prezentare anterior. Diferențele pot apărea datorită aparatului de secvențiere sau datorită fluxului de lucru utilizat pentru extragerea variantelor genetice.

3.5. Concluzii de capitol și contribuții proprii

În cadrul subcapitolului 3.2 a fost studiată strategia de predicție a afecțiunii în funcție de termenii HPO, furnizați de către cadrul medical, și variantele genetice identificare la pacient. Această metodă se bazează pe trei elemente: afecțiunea pacientului, variantele cauzatoare și termenii HPO. Oricare element poate fi prezis dacă celelalte două sunt complete sau cvasi-complete.

În ceea ce privește studiul din subcapitolul 3.3, predictorii CADD și DANN au identificat cele mai multe variante patogene (96,2%, respectiv 98,1%) 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 (64,2% și 40,8%). REVEL (91,5%), MetaSVM (85,5%) și *PolyPhen-2* HVAR (85%) au avut cele mai bune performanțe generale în funcție de 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 folosiți predictorii echilibrați (REVEL, *MetaSVM*, *PolyPhen*) pentru a le prioritiza.

În subcapitolul 3.4 a fost analizată detecția erorilor de calitate și cantitate ale variantelor genetice identificate în urma procesului de secvențiere. Se propune o metodă pentru identificarea erorilor folosind intervale de toleranță. Pe lângă posibilele erori, aceste intervale pot semnala anumite cauze ale afecțiunii precum consangvinitate.

Contribuții personale:

1. Dezvoltarea unei metodologii 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.
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 unor cauze, precum consangvinitatea.

4. ANALIZA REGIUNILOR DE MATISARE (*SPLICING*)

Precum a fost prezentat în subcapitolul 2.3, matisarea este procesul prin care se îndepărtează regiunile intronice pentru obținerea mARN. Acest proces a fost conservat de către celulă de-a lungul timpului și este esențial pentru evoluția acesteia [115]. Matisarea se poate realiza în două moduri. Matisarea obișnuită, care îndepărtează intronii și concatenează exonii în ordinea consecutivă găsită pe ADN, și pe de altă parte, *splicing*-ul alternativ care concatenează exonii din secvență în ordinea apariției, dar în acest caz putându-se elimina exoni sau chiar reține introni [116]. Se estimează că peste 90% din genele umane sunt matisate alternativ [116], [117], ceea ce duce la o crește a diversității expresiei genelor și a proteinelor codificate de acestea [70].

Ca majoritatea proceselor biologice, matisarea nu este perfectă. Deși spliceosomul (mecanismul care execută matisarea) realizează acest lucru printr-o serie de interacțiuni între ARN și proteine [118], uneori apar erori, care pot duce la manifestarea unei afecțiuni sau la modificarea unor trăsături fenotipice. Înțelegerea modului de funcționare al matisării, împreună cu factorii care determină acest proces, poate oferi o imagine asupra modului de funcționare a sistemului de reglare al celulelor și, desigur, poate oferi soluții pentru corectarea anumitor afecțiuni.

În mod normal, în procesul de *splicing*, regiunile intronice sunt îndepărtate. În schimb o serie de variante găsite în zonele intronice sunt raportate ca fiind cauzatoare de afecțiuni. De exemplu, o mutație intronică, descrisă în [119], cauzează funcționarea defectuoasă a proteinei ABCA3. Această malfuncție se datorează unei tranziții C>T heterozigote care creează o regiune nouă de *splicing*. În [120] a fost raportată o mutație T>G în intron care duce la „exonizarea" zonei intronice. O variantă genetică găsită în regiunea intronică a genei FBN2 este considerată responsabilă pentru o boală rară numită arahnodactilie congenitală contractuală [121]. Date fiind rezultatele publicate în aceste lucrări, putem considera că există polimorfisme (SNP), în regiuni intronice îndepărtate de situl de matisare, care pot perturba activitatea spliceosomului și pot altera procesul de *splicing*.

Studiul descris în continuare a urmărit analizarea trăsăturilor componentelor siturilor de *splicing* aferente cromozomului 21. Cu informațiile obținute s-a urmărit identificarea secvențelor similare cu cele de matisare în regiunile intronice. Deși în prezent există o serie de aplicații pentru determinarea regiunile de *splicing*, nu există nicio metodă sau un protocol standardizat pentru interpretarea și validarea acestora [122]. Datorită complexității datelor generate de secvențierea unui genom, identificarea *in silico* a joncțiunilor reprezintă un instrument pentru eficientizarea identificării *in vitro* a unor astfel de situri de matisare.

O parte dintre informațiile prezentate în acest studiu au fost publicate în lucrarea intitulată *Splice Site Pattern Analysis and Identification of Similar Sequences in the Deep Intron Areas of Human Chromosome 21* [123].

4.1. Analiza regiunilor de matisare

4.1.1. Unelte software și hardware

Procesarea datelor a fost realizată pe un sistem cu procesor AMD A10-6800B, 8GB de RAM cu un sistem de operare Windows 10, pe 64 de biți. Limbajul de programare utilizat a fost *Python* 3.5.1 cu biblioteca suplimentară *BioPython* 1.66. Datele folosite au fost descărcate de pe site-ul UCSC *Genome Browser* [124]. Informațiile preluate au reprezentat trei fișiere: (1) secvența de ADN a cromozomului 21 pentru *Homo Sapiens*, versiunea HG38; (2) un fișier care conține toate pozițiile exonilor din cromozomul 21 și (3) un fișier care conține toate pozițiile intronice din cromozomul 21.

4.1.2. Convenții de termeni folosiți pentru studiu

În contextul experimentului, termenul exonul 5' (cinci prim) se referă la exonul care se găsește în amontele intronului curent (Fig. 4.1.). Termenul exonul 3' (trei prim) se referă la exonul care este în avalul intronului curent. Pentru realizarea acestui experiment parcurgerea secvențelor de ADN s-a făcut în două direcții. Prima direcție de căutare este dinspre nucleotida 35 către exonul 3', reprezentată în Fig. 4.1 ca direcția A. Cea de-a doua direcție, direcția B, parcurge secvența dinspre exonul 3' spre nucleotida 35. Termenul pirimidină se referă la acizii nucleici citozină (C), timină (T) și uracil (U). Un tract de pirimidine se referă la o secvență de ADN compusă în principal din pirimidine.

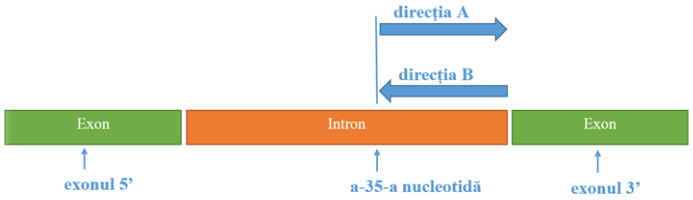


Fig. 4.1. Poziționarea exonilor în funcție de intron

4.1.3. Analiza regiunilor intronice

Studiul a fost realizat în două etape. În prima etapă s-a efectuat o analiză a secvențelor de ADN care preced exonul 3' pentru determinarea definiției unei regiuni de matisare. În cea de-a doua parte, s-a folosit definiția regiunii de *splicing*, determinată la faza precedentă, pentru a găsi secvențe similare în regiunile intronice.

În prima etapă, toate secvențele intronice au fost extrase din fișierul în care se afla secvența ADN a cromozomului 21. Pozițiile de început și de sfârșit ale intronului au fost obținute din fișierul BED descărcat de pe portalul UCSC [124]. Regiunile intergenice au fost ignorate prin potrivirea poziției finale a exonului cu poziția de start a intronului sau prin potrivirea poziției finale a intronului cu poziția inițială a exonului. Dacă gena a fost adnotată în browser-ul UCSC pe catena anti-sens, informația genei

fiind în direcția opusă, secvența a fost inversată folosind metoda *reverse complement* prezentă în biblioteca *BioPython*.

După obținerea secvenței intronice, s-a efectuat o căutare a regiunilor valide de matisare. Un sit valid de *splicing*, în acest studiu, are în componență: (1) o regiune de prindere a spliceosomului (*Branch Point*, BRS), care este format după modelul *yTnAy*, echivalent cu *yUnAy* în ARN, unde: *y* reprezintă o pirimidină, *U* reprezintă uracil, *n* reprezintă oricare nucleotidă, *A* reprezintă adenină și *T* reprezintă timină; (2) un sit acceptor compus din nucleotidele AG, care este situat lângă startul exonului 3' și (3) o secvență bogată în pirimidine care are în componență minim 4 pirimidine (conform lui *Gao* [57] pirimidinele ar trebui să fie predominante). Structura acestor componente se bazează parțial pe rezultatele prezentate în [57] unde s-au examinat, *in vitro*, secvențele de prindere (BRS) în 20 de gene umane și s-a concluzionat că modelul pentru BRS este *YUnAy*.

Dacă în urma căutării s-a găsit o potrivire, pentru secvența respectivă s-au calculat și s-au stocat un set de valori, mai precis distanța de la secvența punctului de prindere la exonul 3' și numărul de apariții pentru fiecare nucleotidă pentru tractul de pirimidine. Aceste valori au fost extrase pentru determinarea unor măsurători aferente secvenței regiunii de matisare.

Căutarea regiunii de matisare a fost realizată utilizând trei scenarii, așa cum este ilustrat în Fig. 4.2. În scenariul 1, obiectivul a fost găsirea pozițiilor pe care se află BRS-ul, ignorând restricțiile regiunii de pirimidine. Direcția căutării a fost realizată de la exonul 3’ către nucleotida 35, ilustrată în Fig. 4.1 ca direcția B. În scenariul 2, scopul a fost găsirea unui tipar cu toate limitele exprimate în paragraful anterior. Parcurgerea s-a realizat ca direcția B. În scenariul 3, căutarea a pornit de la nucleotida 35 spre exonul 3’, reprezentat în Fig. 4.1 ca direcția A. La fel ca în scenariul precedent, s-a urmărit identificarea tuturor componentelor dintr-o regiune de *splicing*.

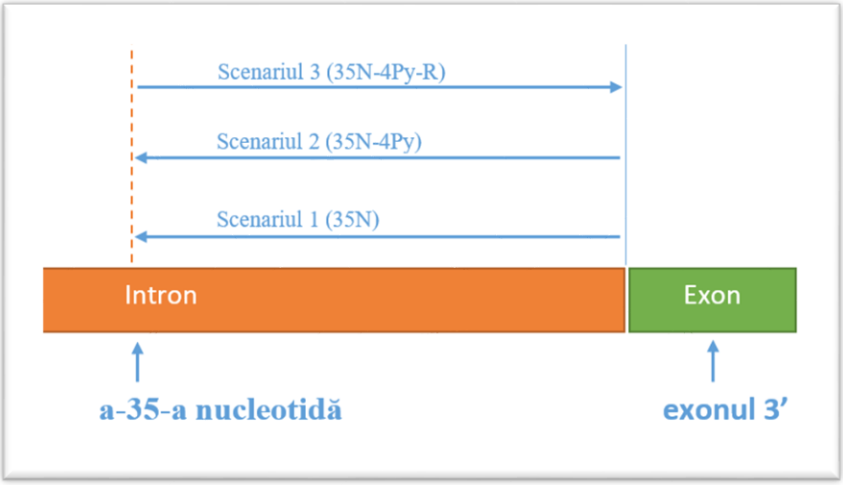


Fig. 4.2. Scenariile de căutare a regiunilor de prindere (*Branch Point*, BRS)

Folosind configurația scenariului 1, s-au identificat 6966 de regiuni de matisare în cei 11.143 de introni (aproximativ 62,5% dintre introni au avut siturile de *splicing* valide). Cea mai întâlnită secvență BRS a fost TTTAT, precum este prezentat în Tabelul 4.1. La poziția 8, în amonte de exonul 3', s-au găsit 882 secvențe BRS așa cum este ilustrat în Fig. 4.3. Cu alte cuvinte, din 6.966 de situri de *splicing* identificate

corect, 882 au avut modelul *yTnAynAG*, care este compus din regiunea de prindere *yTnAy*, o nucleotidă și acceptorul AG.

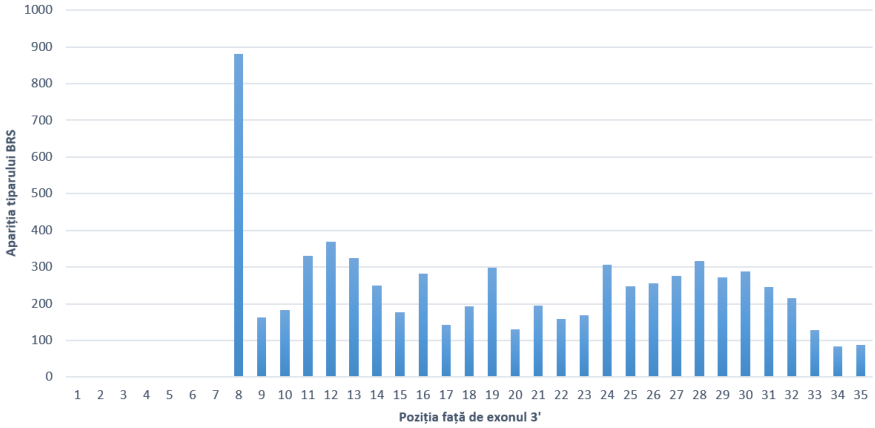


Fig. 4.3. Distribuția regiunilor *BRS* în primele 35 de poziții pentru scenariul 1

Concentrația de regiuni BRS din scenariul 1, de pe poziția 8 predecesoare exonului 3', poate indica faptul că această zonă a acționat ca o posibilă regiune de prindere a spliceosomului în cadrul evoluției speciei sau poate chiar are un rol în procesul de matisare. În plus, în Fig. 4.3 se observă că de la poziția 9 până la poziția 15 există o regiune asemănătoare unei distribuții normale, cu punctul maxim la poziția 12. O altă regiune asemănătoare poate fi observată de la poziția 22 până la poziția 35, cu maximul la poziția 28. Poziția 24 este o excepție de la această distribuție. Aceste locații ar putea fi poziții ancestrale ale punctelor de prindere ale spliceosomului, care au migrat datorită procesului evolutiv.

Tabelul 4.1. Regiunile de prindere (BRS) detectate în cele trei scenarii

Index	Secvență	Număr de apariții		
		scenariul 1	scenariul 2	scenariul 3
1	TTTAT	1069	887	730
2	CTGAC	616	598	644
3	TTCAT	577	483	370
4	CTCAC	562	526	578
5	TTTAC	528	342	373
6	TTCAC	446	315	264
7	TTAAT	424	456	544
8	CTCAT	398	384	365
9	TTAAC	392	332	329
10	TTGAT	372	308	318
11	CTGAT	310	339	386
12	TTGAC	299	254	291
13	CTAAT	281	330	363
14	CTTAT	265	287	256
15	CTTAC	255	176	155
16	CTAAC	172	207	258
Total		6966	6224	6224
% din toți intronii		62.5	55.8	55.8

În scenariul 2, concentrațiile de BRS (indicate la scenariul 1) nu mai sunt prezente. Totuși, o anumită concertare are loc în jurul poziției 16 și a poziției 28. În scenariul 3, o zonă de concentrare poate fi observată, Fig. 4.4, între poziția 25 și poziția 33, cu maximul la poziția 31.

Rata de detecție a regiunilor de matisare pentru scenariul 2 și scenariul 3 a fost aceeași, de 55,8%. În schimb, dacă se analizează Fig. 4.4 se observă că, deși rata de detecție este aceeași, valoarea pentru fiecare poziție este diferită. Acest lucru se datorează secvențelor BRS care au fost găsite pe poziții multiple în secvența de 35 nucleotide din amontele exonului 3'. Dintre cele 6.224 de situri detectate, 1.209 au un BRS de rezervă, adică pe aceeași secvență de ADN există două sau mai multe regiuni de prindere a spliceosomului. În plus, același lucru, funcția redundantă a BRS, poate fi observată când se calculează diferența dintre scenariul 1 și scenariile 2 și 3. Numărul de BRS identificate corect, în scenariul 1, a fost 6.966. De la poziția 8 până la poziția 11, numărul de situri detectate a fost de 1.556. Prin urmare, atunci când se scad regiunile de prindere din totalul numărului de BRS, rezultatul este de 5.410 regiuni. Acesta ar trebui să fie numărul punctelor de prindere detectate în scenariile 2 și 3. Diferența de 814 poate fi atribuit redundanței regiunilor de prindere a spliceosomului.

Cea mai comună secvență pentru BRS, în cazul scenariului 2 și 3, este TTTAT, Tabelul 4.1. Din primele șase regiuni de prindere din Tabelul 4.1, numai o regiune are două purine prezente, CTGAC. O explicație poate fi faptul că secvența din apropierea exonului este bogată în baze azotate pirimidinice, astfel că o regiune de matisare este valabilă dacă apare o singură bază azotată, adenina. Așa pot fi explicate copiile de rezervă a BRS-ului.

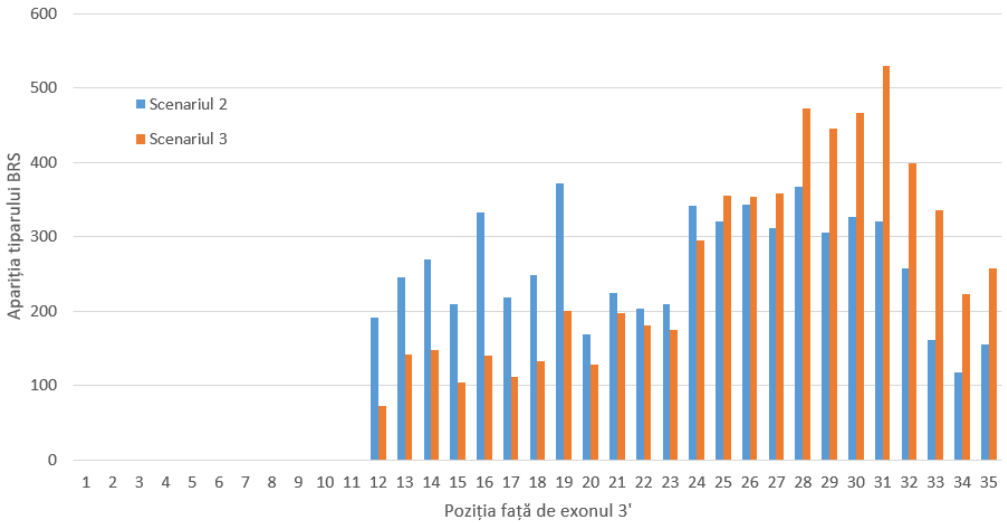


Fig. 4.4. Distribuția regiunilor BRS în primele 35 de poziții pentru scenariul 2 și scenariul 3

În faza a doua a studiului, utilizând parametrii stocați într-un fișier CSV, obținuți în urma scenariilor 2 și 3, au fost determinate mai multe caracteristici pentru structura și lungimea sitului de matisare. Folosind pozițiile locului de prindere și noile informații despre structura regiunii de matisare, s-a efectuat o căutare pentru secvențe similare cu definiția actualizată a sitului de matisare, în toate regiunile intronice ale cromozomului 21. Ultimele 35 de nucleotide din fiecare intron au fost

ignorate deoarece aceste regiuni au fost utilizate în prima fază pentru calcularea valorilor statistice care definesc situl de *splicing*. În plus, în conformitate cu fișierele UCSC, în aceste regiuni se găsesc regiuni valide de matisare.

Analiza regiunii de pirimidine, care a fost efectuată pentru secvențele care corespundeau definiției regiunii de matisare, indică lungimea medie a acestora ca fiind de 17 nucleotide, ca în Tabelul 4.2. Raportul bazelor azotate purinice, pentru cele două scenarii, este de 25% și 23%. Raportul bazelor pirimidinice este de 75% și 77%.

Tabelul 4.2. Compararea regiunilor de pirimidine din scenariul 2 și scenariul 3.

Direcție	Scenariul 2		Scenariul 3	
Numărul de introni	6224		6224	
	Raport AG	Raport CT	Raport AG	Raport CT
Media	0.25	0.75	0.2343	0.7657
Deviația standard	0.1162	0.1162	0.1291	0.1291
Variația	0.0135	0.0135	0.0167	0.0167
Lungimea medie	17		17	

Pe baza rezultatelor prezentate în paragrafele anterioare, a fost stabilită o definiție actualizată a modelului pentru regiunea de matisare. Noua definiție a acestei regiuni este formată din (1) o regiune de prindere care are modelul *YTnAy*, (2) o regiune acceptor AG și (3) regiunea de pirimidine care este formată din 75% baze pirimidinice și are o lungime de cel puțin 17 baze azotate. Modelul nu poate depăși 35 de baze azotate. Pentru această definiție au fost luate în considerație valorile medii din Tabelul 4.2.

Folosind definiția actualizată pentru regiunea de matisare, au fost testate trei configurații pe secvențele intronice. În primul caz, definiția regiunii de *splicing* au fost utilizate precum este în definiție. Cu această configurație s-au găsit 33.703 de secvențe, prezentate în Tabelul 4.3, care erau similare cu o regiune de matisare. Acest rezultat reprezintă o medie de trei (3) secvențe similare în fiecare intron al cromozomului 21.

În cel de-al doilea caz, definiția regiunii de pirimidine a fost modificată astfel încât lungimea minimă să fie de 11 baze azotate iar raportul minim al pirimidinelor de 63%. Cu această configurație s-au găsit 306.500 secvențe similare, care reprezintă, în medie, 30 de secvențe per intron. În cel de-al treilea caz, regiunea de pirimidine este caracterizată cu o lungime minimă de 23 baze și un raport minim de pirimidine de 82%. Cu această configurație s-au găsit 3.001 secvențe similare, care reprezintă 0,3 secvențe per intron. Valorile 11 și 23 pentru lungimea minimă a pirimidinei au fost calculate prin adăugarea sau scăderea deviației standard (st = 6) din lungimea medie (17).

Tabelul 4.3. Numărul de secvențe similare regiunii de *splicing* găsite în cele 3 cazuri.

Cazul	Raportul pirimidinelor	Lungimea regiunilor de pirimidine	Numărul de secvențe găsite
1	0.75	17	33703
2	0.63	11	306500
3	0.82	23	3001

4.1.4. Interpretarea rezultatelor

Acest studiu a fost realizat exclusiv *in silico*, care adaugă anumite limitări, o serie dintre acestea fiind prezentate în literatură [125]–[127]. Secvențele regiunilor de matisare care au fost găsite în acest studiu nu au fost analizate utilizând baze de date cum ar fi [128], [129]. Prin urmare, nu se poate confirma că secvențele care au fost găsite în a doua parte au rol biologic.

Pentru a doua fază a studiului, lungimea medie a regiunii de pirimidine și raportul pirimidinelor au fost utilizate pentru a defini secvența de matisare. Motivul utilizării acestor valori a fost acela de a diminua generarea rezultatelor fals pozitive. Desigur, nu înseamnă că nu apar regiuni de matisare valide având valori mai scăzute pentru tractul de pirimidine de exemplu. Dacă ar fi necesar să calculăm probabilitatea identificării unei regiuni de matisate, cu definiția actualizată, probabilitate de 0,25 pentru o nucleotidă și 0,5 pentru pirimidină, detecția ar fi inferioară valorii 10^{-6}.

Totuși, chiar cu aceste limitări, au fost identificate, în medie, trei secvențe similare regiuni de matisare per intron. Desigur, aceste rezultate nu indică dacă regiunile din introni chiar sunt valide din punct de vedere biologic. Unele dintre aceste secvențe pot fi regiuni de *splicing* local (LSV) așa cum sunt descrise în [130], sau pot fi regiuni folosite în *splicing*-ul alternativ care nu au fost investigate sau ar putea fi responsabile de *splicing* aberant [131].

Acesta este primul studiu, *in silico*, din ceea ce știm deocamdată, care încearcă să identifice în mod sistematic regiuni de matisare în regiunile intronice ale unui întreg cromozom. Pentru a dovedi definitiv faptul că aceste secvențe sau o parte dintre acestea sunt relevante biologic, sunt necesare studii care implică experimente *in vivo* sau *in vitro*.

4.2. Analizarea secvențelor de matisare din baza de date *Homo Sapiens Splice Site Dataset*

4.2.1. Validarea regiunilor de matisare folosind modele bazate pe poziția nucleotidelor

Baza da date HS3D a fost creată de către *Pollastro* și *Ramspone* [132] pentru a servi ca o colecție standard pentru determinarea acurateței modelelor de predicție. Această colecție conține secvențe cu regiuni de matisare validate și o serie de secvențe invalide. În Tabelul A1.1 din Anexa 1 sunt prezentate caracteristicile generale ale apariției bazelor azotate pe fiecare poziție și procentul care îl ocupă acestea din totalul numărului de secvențe valide de matisare. În Tabelul A1.2 al aceleiași anexe regăsim caracteristici pentru secvențele care sunt considerate false. În ambele cazuri au fost ignorate bazele azotate care nu corespundeau notațiilor de bază (A, T, C , G). Pentru calcularea procentajului pozițional pentru fiecare bază azotată s-a folosit ecuația prezentată în (4.2-1), în care $P_i(x)$ reprezintă procentajul pozițional pentru baza azotată *x* de pe poziția *i*, $N_i(x)$ reprezintă numărul total de baze azotate *x* aflate pe poziția *i*, iar N_i^* reprezintă numărul total de baze azotate de pe poziția *i*.

$$P_i(x) = \frac{N_i(x)}{N_i^*} \tag{4.2-1}$$

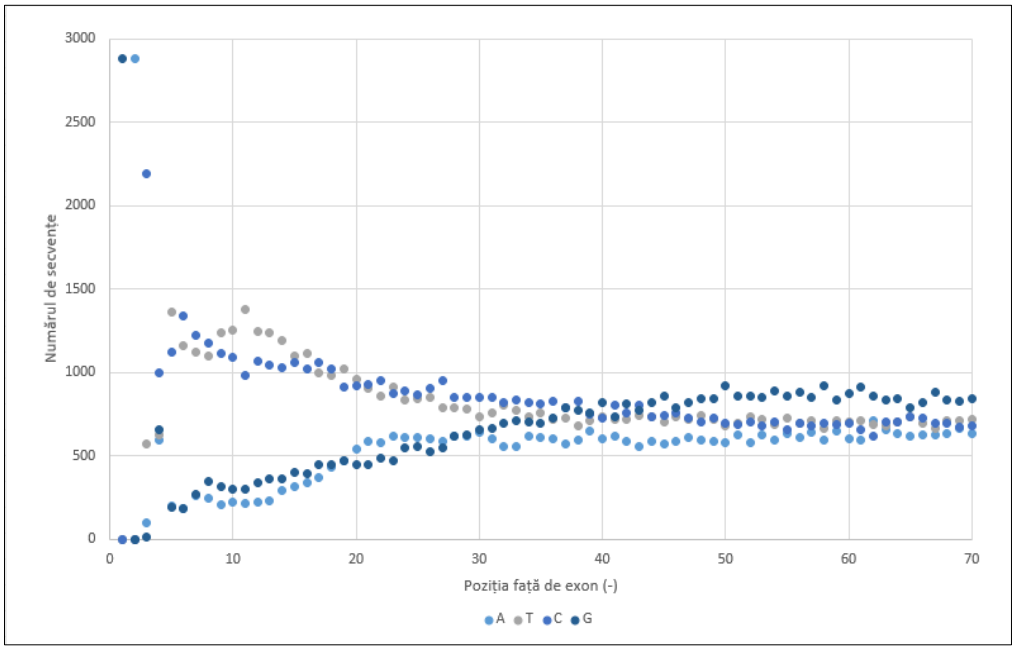


Fig. 4.5. Reprezentarea bazelor pentru fiecare poziție a secvențelor valide de *splicing*

Aceste tabele ne ajută să extragem caracteristicile care diferențiază cele două grupuri. Concret, în Fig. 4.5 se observă că guanina (G) tinde să fie identificată într-un

număr mai mare între pozițiile 40 și 70. După aceste poziții numărul guaninei scade odată cu apropierea de situl acceptor (AG, pozițiile 1 și 2). Timina și citozina, după poziția 40, trec prin procesul invers față de guanină, numărul acestora crescând odată cu apropierea de situl acceptor. În ceea ce privește distribuția bazelor azotate pentru secvențele invalide, acestea au o distribuție uniformă fiind prezente cam în același număr la toate pozițiile, ca în Fig. 4.6.

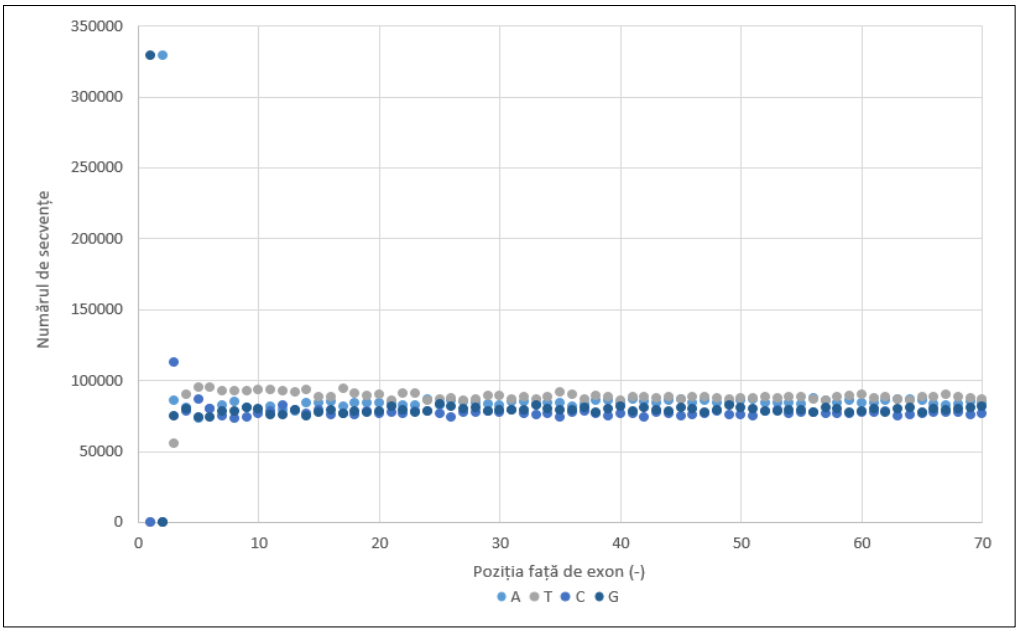


Fig. 4.6. Reprezentarea bazelor azotate pentru fiecare poziție a secvențelor invalide de *splicing*

Din aceste informații putem extrage câteva elemente prin care se poate îmbunătăți modelul de predicție a regiunii de matisare, respectiv al pierderii funcției de *splicing*. Prima ipoteză care se poate enunța, înaintea sitului acceptor este necesar ca bazele azotate citozina și timina să aibă o proporție mai mare decât adenina și guanina. O altă ipoteză este ca guanina să aibă o proporție ușor mai ridicată între pozițiile 40 și 70 în raport cu situl acceptor. De asemenea, în completarea primei ipoteze se poate postula că, dacă nivelul guaninei este asemănător cu cel al citozinei și timinei, atunci cel mai probabil nu avem de a face cu o regiune de matisare. Prin urmare, putem dezvolta un model de predicție binar care să aibă valoarea 1 când analizăm o secvență validă de matisare altfel să aibă valoarea 0. Bazându-ne pe cele enunțate anterior descriem modelul SBB *(Splicing Binary Base Model)* în ecuația (4.2-2).

$$SBB(seq) = \begin{cases} 1 \text{ dacă } \sum_{i=3}^{30}\{1 \mid x_i \in \{C,T\}\} > \sum_{i=3}^{30}\{1 \mid x_i \in \{A,G\}\} \bigwedge \sum_{i=40}^{70}\{1 \mid x_i = G\} > \dfrac{\sum_{i=40}^{70}\{1 \mid x_i \in \{A,C,T\}\}}{3} \\ 0 \text{ altfel} \end{cases}$$

unde $x_i \in seq$ (4.2-2)

Dacă testăm acest model pe datele care conțin regiuni autentice de *splicing*, dintre cele 2.880 de secvențe doar 1.586 au fost confirmate ca fiind valide. Din baza de date cu secvențe invalide, având un total de 329.374 de secvențe, un număr de 73.669 au fost identificate ca fiind regiuni valide. Modelul în forma actuală are o sensibilitate de 55% și specificitate de 77% pentru secvențele negative.

O posibilă explicație pentru performanța slabă a modelului poate fi a doua condiție care presupune ca guanina să fie predominantă între pozițiile 70 și 40. Această condiție nu are un rol semnificativ. Prin urmare, dacă eliminăm acest element din model, precum este prezentat în (4.2-3), atunci pentru secvențele valide vom avea un grad de predicție de 96%. Din cele 2.880 de secvențe, 2.774 au fost detectate ca fiind valide. Pentru predicția secvențelor invalide, nivelul de predicție scade la 46%, de unde concluzionăm că modelul este prea permisiv.

$$SBB(seq) = \begin{cases} 1 \text{ dacă } \sum_{i=3}^{30} \{1 \mid x_i \in \{C,T\}\} > \sum_{i=3}^{30} \{1 \mid x_i \in \{A,G\}\} \\ 0 \text{ altfel} \end{cases}$$

unde x_i reprezintă o bază azotată din secvența seq. (4.2-3)

Un alt model care se poate dezvolta este bazat pe secvențe consecutive de baze azotate. În cazul de față, vom avea în vedere secvențele a câte două baze azotate, în total 16 combinații posibile. În Anexa 1, în Tabelul A1.3 și Tabelul A1.4 sunt prezentate rezultatele apariției tuplilor de baze azotate pentru secvențele valide, respectiv invalide, de *splicing*. Această reprezentare ne va da un nivel mai ridicat de granularitate asupra secvențelor de ADN. În Fig. A1.1 și Fig. A1.2 din Anexa 1, sunt prezentate graficele cu reprezentarea tuturor tuplilor pentru secvențele valide, respectiv invalide, pentru fiecare poziție în raport cu situl acceptor. Având în vedere că cele două fișiere conțin un număr diferit de secvențe, este necesară scalarea după numărul total de secvențe. Această scalare nu va afecta caracteristicile prezentate în Fig. A1.1 și Fig. A1.2, dar ne va permite să observăm diferența dintre secvențele valide și cele invalide, pentru același tuplu. În Tabelul A1.5 sunt prezentate rezultatele diferențelor dintre cele două fișiere. Dacă valoarea este pozitivă înseamnă că pentru secvențele valide apariția tuplului respectiv este mai mare. Altfel, dacă valoarea este negativă, înseamnă că tuplul respectiv apare mai frecvent în secvențele invalide. În tabel au fost evidențiați doar tuplii care au o diferență mai mare de 6%. Această valoare a fost aleasă prin divizarea totalului disponibil (100) la numărul de variabile (16) care este de aproximativ 6%. Deci, dacă diferența este de 6% înseamnă că tuplul respectiv apare de cel puțin două ori mai frecvent.

În urma analizării rezultatelor obținute în Tabelul A1.5 putem observa că diferențele majore sunt prezente în ultimele 20 de poziții. Tuplii formați din bazele azotate purinice, adenina și guanina, scad în frecvență, ca în Fig. 4.7. Dintre cei trei tupli (AG, AA și GG), ultimul are caracteristica mai interesantă. În secvențele valide, perechea GG, între pozițiile 40 și 70, are o frecvență ridicată, după care între pozițiile 20 și 3 are o frecvență scăzută. Între pozițiile 40 și 20 fiind o descreștere a frecvenței. În schimb, pentru secvențele invalide de matisare se observă că perechea GG se regăsește constant pe fiecare poziție. La perechile primidinice, citozină și timină, se poate observa un comportament opus. Acești tupli se întâlnesc mai frecvent pe ultimele 20 de poziții ale secvențelor valide în comparație cu secvențele invalide, precum este redat în Fig. 4.8.

Întorcându-ne la modelul SBB (4.2-2), acesta îndeplinește, oarecum, cele observate în paragraful anterior, doar că nivelul de detaliere este mai general. Prin urmare, se poate propune un alt model, mai specific, care să ia în considerație tuplii în locul bazelor azotate.

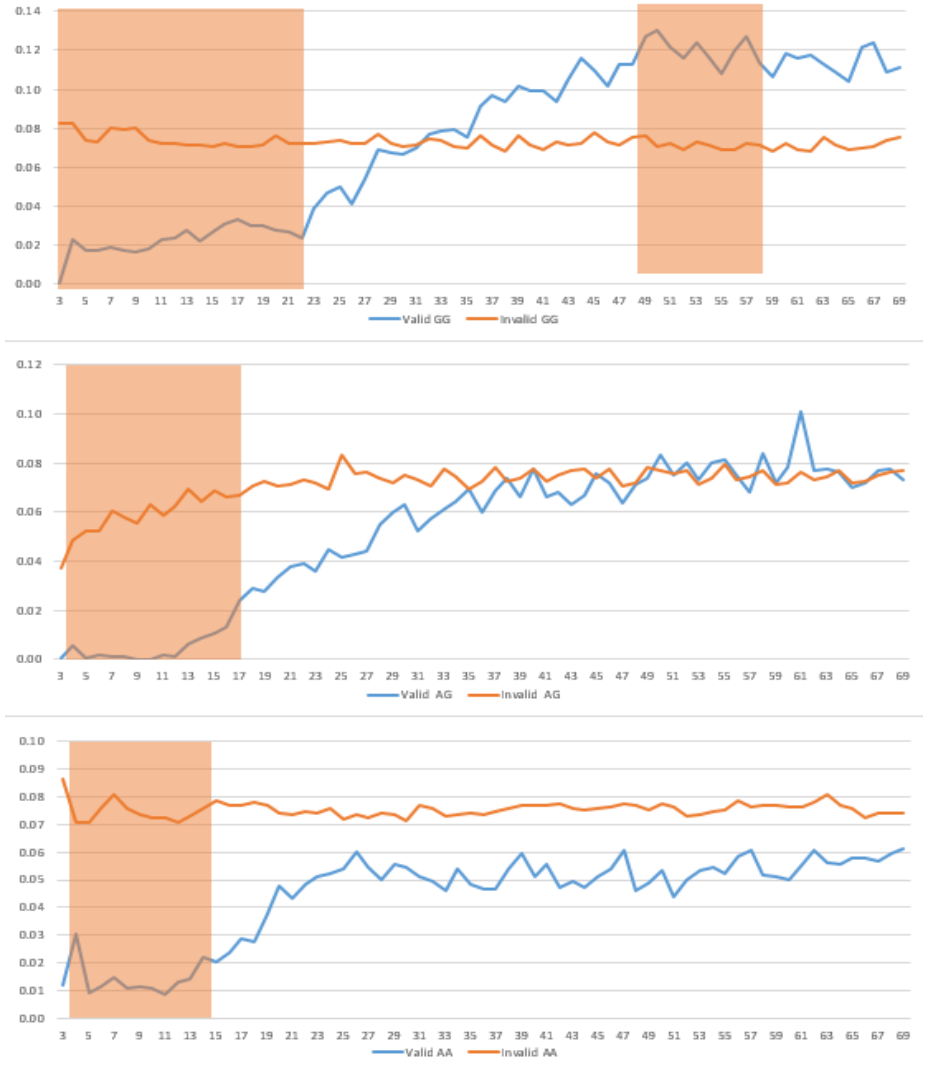


Fig. 4.7. Diferențe între apariția tuplilor, formați din baze azotate purinice (A,G), din secvențele de *splicing* valide și cele invalide.

În (4.2-4) avem prezentat modelul de predicție SBT (*Splicing Binary Tuple Model*), care va indica o secvență validă de *splicing* doar dacă tuplii formați din baze azotate pirimidinice (TT, TC, CT, CC) se regăsesc în număr mai mare decât tuplii formați din baze azotate purinice (AG, GG, AA, GA).

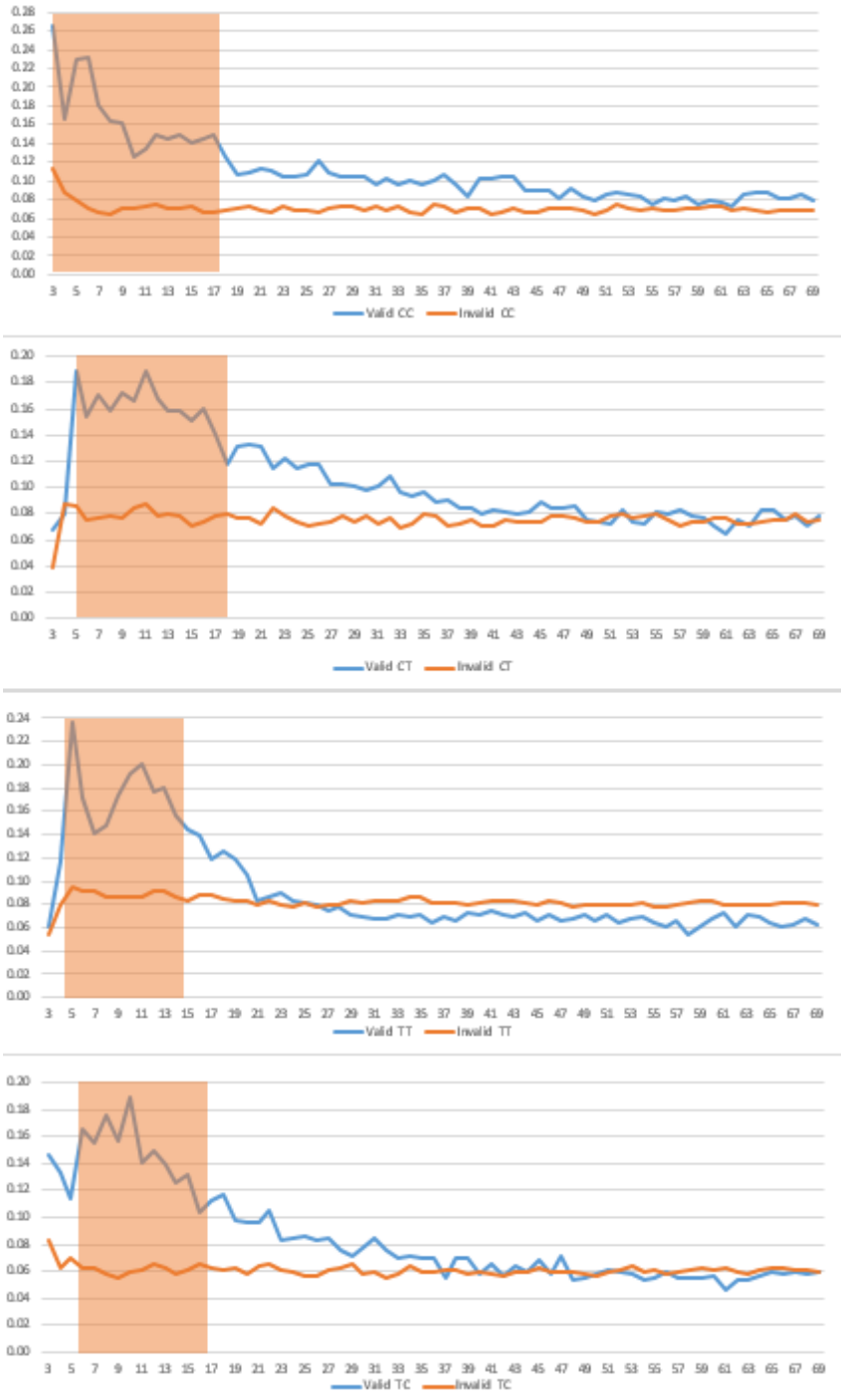


Fig. 4.8. Diferențe între apariția tuplilor, formați din baze azotate primidinice (C,T), din secvențele de *splicing* valide și cele invalide.

$$SBT(seq) = \begin{cases} 1 \text{ dacă } \sum_{i=3}^{20}\{1 \mid x_i \in \{TT, TC, CT, CC\}\} > \sum_{i=3}^{20}\{1 \mid x_i \in \{GG, GA, AG, AA\}\} \\ 0 \text{ altfel} \end{cases} \quad (4.2\text{-}4)$$

unde x_i reprezintă un tuplu din secvența seq.

Din cele 2.880 de secvențe valide avute la dispoziție, modelul SBT a reușit să identifice corect 2.811, acesta având o rată de detecție de 97,6%. Când acest model a fost pus să indice secvențele care sunt valide din fișierul cu secvențe invalide, a indicat 47% ca fiind regiuni autentice de *splicing*. Comparativ cu modelul SBB care a avut 55% pentru pozitiv și 77% pentru negativ, iar în formă simplificată 96%, respectiv 47%, modelul SBT, care are o rată de detecție 97,6% pentru pozitive și 53% pentru negative, este mai performant. Dar, ca și în cazul modului SBB, este prea permisiv având o un număr ridicat de rezultate fals pozitive.

Desigur, pe baza datelor reprezentate în Fig. 4.7, tuplul GG, observăm un dezechilibru care se poate fi exploatat pentru obținerea unor predicții mai bune. Se poate observa că pentru secvențele valide de *splicing*, tuplul GG apare mai frecvent între pozițiile 70 și 40 și scade în frecvență între pozițiile 20 și 3. Dacă adăugăm acest element modelului, se obține o rată de detecție a secvențelor invalide de 64%, dar rata de detecție pentru secvențele valide scade la 77%. Desigur, au fost testate și alte ipoteze, precum nivelul constant al numărului bazelor azotate, când secvențele sunt invalide, dar nivelul valorii de detecție se modifică marginal.

În continuare, numărul de apariții al fiecărui tuplu, pe o anumită poziție, a fost transformat în procentaj raportat la valoarea maximă înregistrată pe poziția respectivă. Astfel, tuplul cu cel mai mare număr de apariții va avea valoarea 1, iar restul tuplilor se vor raporta la acesta, cum se poate vedea în Tabelul A1.6. Pentru procesul de transformare s-a folosit ecuația prezentată în (4.2-5), unde $P(tuplu)_i$ reprezintă raportul tuplului la poziția *i*, $N(tuplu)_i$ reprezintă numărul de apariții al tuplului la poziția *i*, iar $NMax(i)$ reprezintă valoarea maximă înregistrată la poziția *i*.

$$P(tuplu)_i = \frac{N(tuplu)_i}{NMax(i)} \quad (4.2\text{-}5)$$

În urma acestor modificări din 2.880 de secvențe valide s-au identificat corect 2.721, adică o rată de detecție de 94,4%. Când modelul a fost pus să indice secvențele valide din fișierul cu secvențe invalide, a indicat 32% ca fiind regiuni autentice de matisare. În continuare, modelul este ușor permisiv, dar este cel mai bun dintre cele prezentate.

4.2.2. Validarea regiunilor de matisare folosind metoda *MaxEnt*

Precum s-a discutat în subcapitolul 2.3.4, metoda *MaxEnt* este cea mai eficientă pentru predicția regiunilor de matisare. În implementarea realizată de *Burge* în [82], pentru detecția regiunilor de matisare 3 prim (3'), metoda are nevoie de o secvență de ADN care conține 20 de nucleotide din intron (inclusiv situl acceptor AG) și trei nucleotide din exon. Pentru regiunile de matisare 5 prim (5') sunt necesare trei

nucleotide din exon și șase din intron. În continuare vom analiza performanța acestei metode pentru regiunile de matisare 3' din baza de date HS3D. Dacă aplicăm metoda pe întreaga bază de date, pentru regiunile valide MaxEnt a identificat corect 2.841, iar incorect doar 39. Din baza de date cu regiuni de matisare invalide, metoda a identificat corect 210.734, iar incorect 118.640. Cu alte cuvinte, *MaxEnt* are o sensibilitate de 98.6%, specificitate de 63.9% și o acuratețe de 64.2%. În această situație acuratețea este distorsionată datorită numărului mare de înregistrări invalide. Dacă s-ar folosi seturi egale de înregistrări, probabil acuratețea ar fi în apropierea valorii de 80%. În Fig. 4.9 sunt reprezentate valorile pentru cele 2.880 de secvențe valide și 2.880 de secvențe invalide.

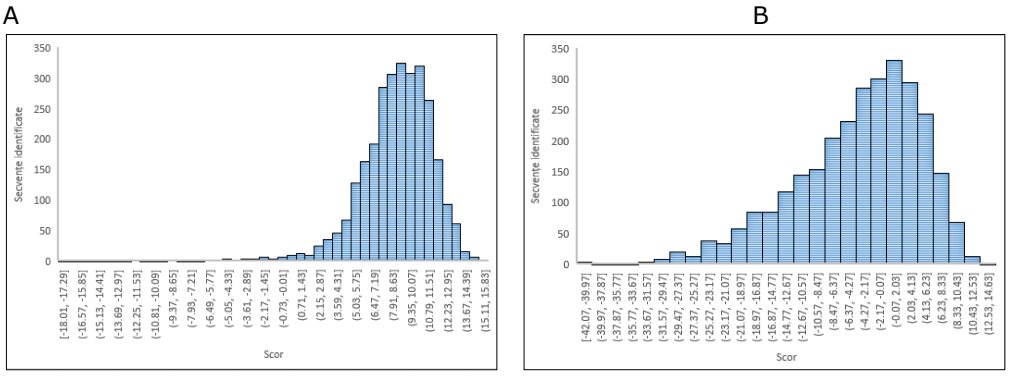


Fig. 4.9 Distribuția valorilor generate de *MaxEnt* pentru secvențele de matisare valide (A) și invalide (B)

Desigur, dacă ajustăm pragul peste care secvențele sunt considerate valide atunci rezultatele se schimbă. De exemplu, dacă se ia valoarea 3 ca prag, atunci pentru secvențele valide au fost identificate corect 2.781 iar incorect 99. Pentru cele invalide, au fost identificate corect 263.293 iar incorect 66.081. Cu aceste valori avem o sensibilitate de 96.5% și o specificitate de 79.9%.

4.2.3. Validarea regiunilor de matisare folosind distanța secvențelor vecine

În continuare se va încerca gruparea secvențelor de ADN pe baza distanței dintre acestea. Pentru început se va face o evaluare vizuală a 200 de secvențe (100 valide și 100 invalide) pentru a identifica grupurile de secvențe și cum sunt acestea intercalate în dendogramă. *Knime* a fost folosit pentru a genera reprezentările grafice. Schema pentru realizarea dendogramelor, Fig. 4.10, conține patru blocuri. *File Reader* preia secvențele de ADN din fișier, *String Distance* indică metoda pentru calcularea distanței dintre două șiruri de caractere, blocul *Hierarchical Clustering* va calcula matricea distanțelor dintre secvențele de ADN iar blocul *Hierarchical Cluster View* va genera dendograma aferentă matricei.

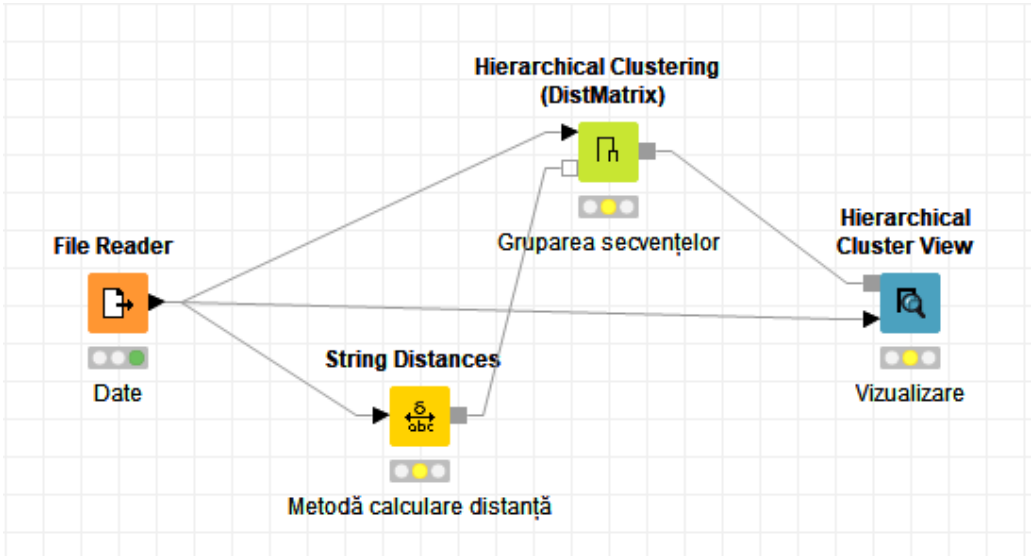


Fig. 4.10 Schema *Knime* pentru realizarea dendogramelor

Dacă analizăm rezultatele pentru distanțele dintre ultimele 20 de nucleotide ale intronului, Fig. 4.11, și ultimele 20 de nucleotide ale intronului și 3 nucleotide din exon, Fig. 4.12, observăm că în ultimul caz avem mai puține secvențe fără vecini din aceeași categorie. Adică în ultimul caz secvențele sunt mai bine grupate. Desigur, acesta este un mod empiric pentru găsirea unei soluții optime. Dar în acest mod am putut observa că deși este o diferență de 3 nucleotide diferențele dintre dendograme sunt relevante.

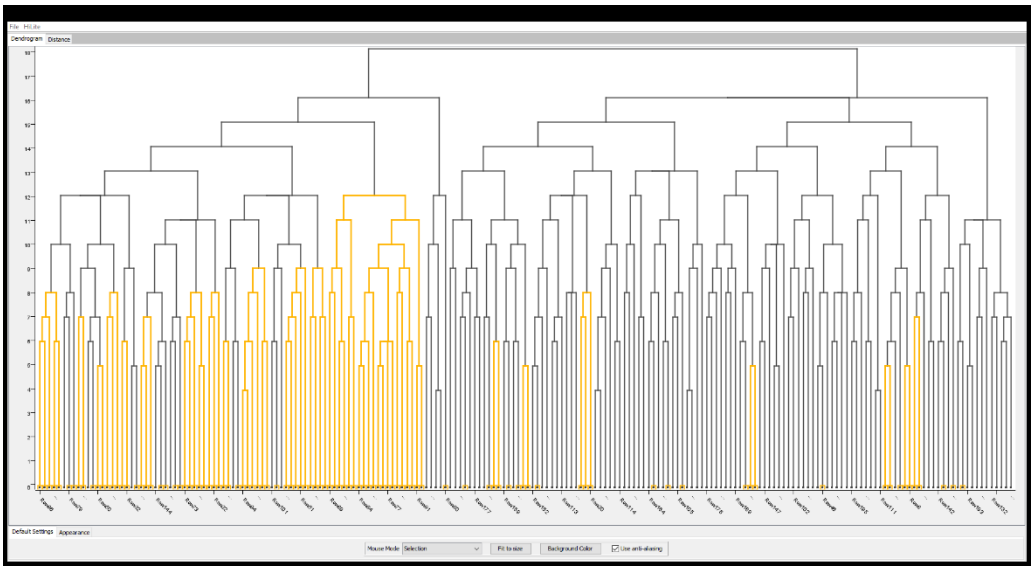


Fig. 4.11 Dendograma cu ultimele 20 de nucleotide din intron

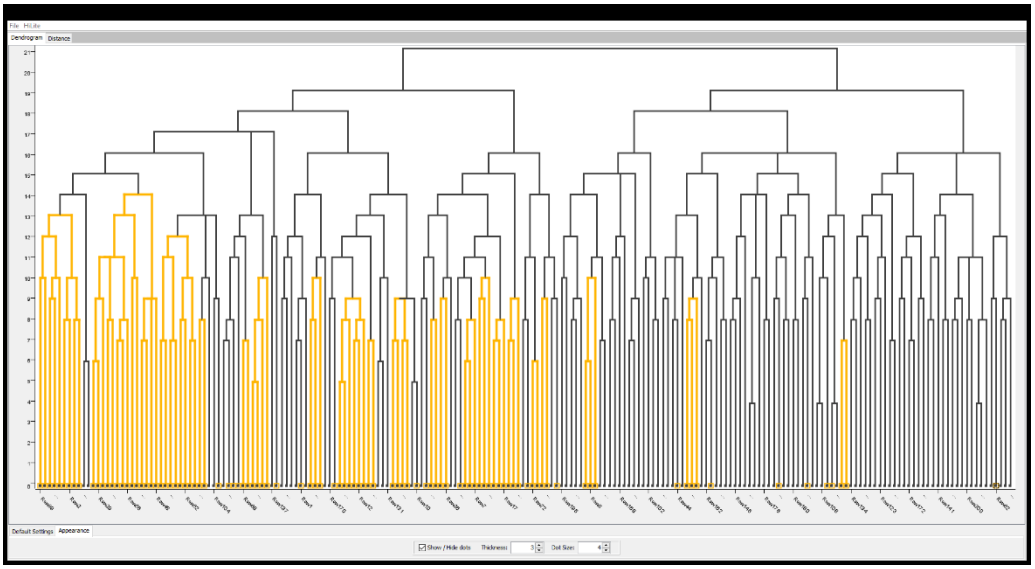
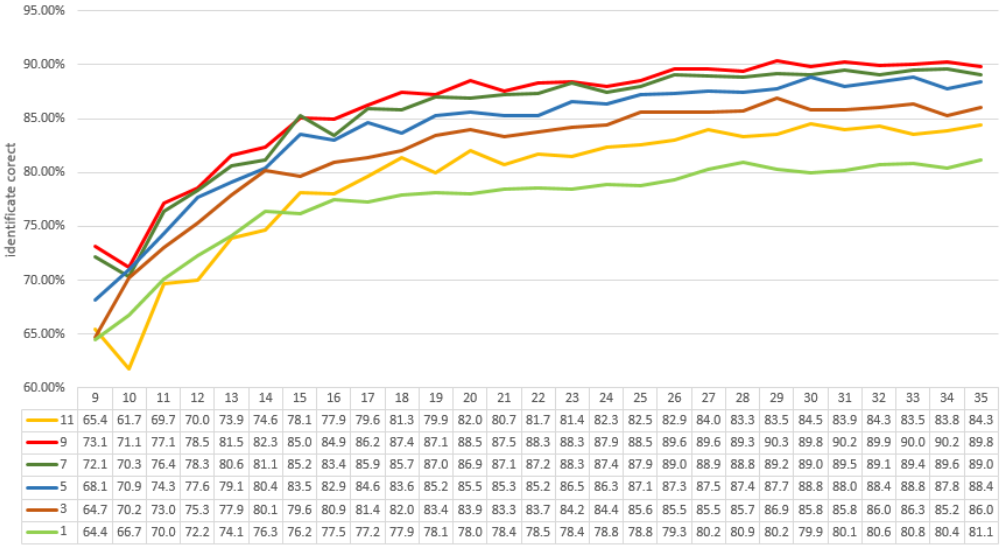


Fig. 4.12 Dendograma cu ultimele 20 de nucleotide intronice și 3 nucleotide din exon

Pentru obținerea celei mai bune configurații este nevoie să parcurgem întreg spațiu valorilor pentru fiecare variabilă de intrare. Lista de variabile este formată din: numărul de nucleotide incluse în analiză, poziția de final, numărul de vecini și metoda folosită pentru a calcula distanța.

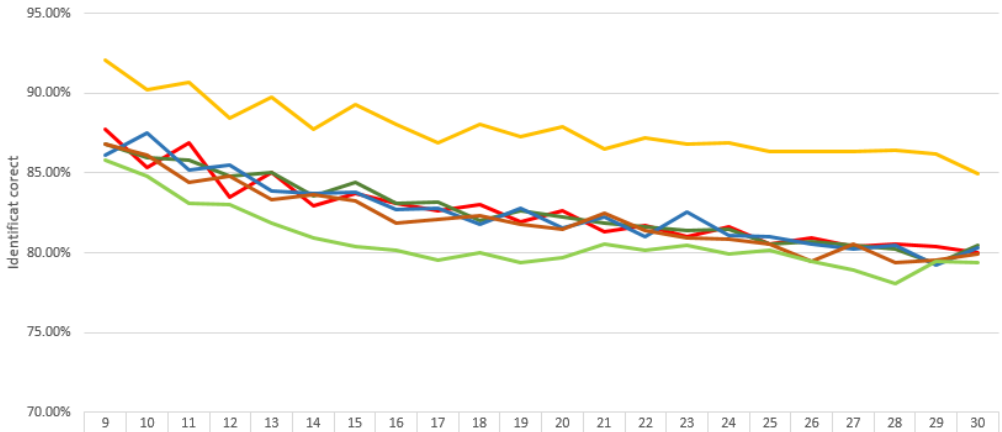
În cele ce urmează sunt prezentate rezultatele analizei secvențelor de matisare folosind o serie de intervale pentru fiecare parametru. Pentru numărul nucleotidelor s-a folosit un interval cuprins între 9 și 35 de nucleotide intronice. S-a început de la 9, deoarece s-a ținut cont de dimensiunea regiunii de prindere a spliceosomului și situl acceptor, elemente despre care s-a discutat în subcapitolul 4.1. S-a ales valoarea 35, deoarece rezultatele din subcapitolul 4.1 indică o scădere a regiunilor de prindere după această poziție. De asemenea, în literatură nu s-au găsit referințe care să indice o BRS care se află mai departe de nucleotida 35 înaintea exonului. Poziția de final a fost selectată în așa fel încât să includă primele 3 nucleotide din exon. Analiza s-a realizat pentru configurația cu următorul număr de vecini: 1, 3, 5, 7, 9 și, respectiv, 11. S-au luat în considerație doar numerele impare de vecini pentru a evita nevoia luării unei decizii suplimentare în cazul situației 50/50. Pentru calcularea distanței dintre secvențele de matisare s-a folosit algoritmul *Needleman-Wunsch* [133]. S-au luat în considerație și alte metode precum *Hamming, Levenshtein* etc., dar această metodă a indicat rezultate superioare în testele preliminare.

În Fig. 4.13 avem prezentate rezultatele pentru analiza secvențelor valide de matisare. Dacă se analizează din perspectiva numărului de vecini, un număr de 9 vecini a reușit de fiecare dată să ofere o rată de predicție superioară. Acesta a fost urmat de 7 respectiv 5 vecini. Pentru secvențele invalide de matisare, prezentate în Fig. 4.14, net superior a fost un număr de 11 vecini. A fost urmat de 7, 9 și 5 vecini. Dacă analizăm numărul de nucleotide aflate în introni un optim se găsește la valoare 20. Caracteristica rezultatelor este asimptotică, valorile medii superioare tind să se afle spre finalul abscisei. Configurația cu valoare medie superioară celorlalte se află la poziția 20 pentru 9 vecini. Acuratețea detecției este de 85.61%, din care detecția secvențelor invalide este de 82.76% iar detecția secvențelor valide este de 88.54%.



	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
11	65.4	61.7	69.7	70.0	73.9	74.6	78.1	77.9	79.6	81.3	79.9	82.0	80.7	81.7	81.4	82.3	82.5	82.9	84.0	83.3	83.5	84.5	83.9	84.3	83.5	83.8	84.3
9	73.1	71.1	77.1	78.5	81.5	82.3	85.0	84.9	86.2	87.4	87.1	88.5	87.5	88.3	88.3	87.9	88.5	89.6	89.6	89.3	90.3	89.8	90.2	89.9	90.0	90.2	89.8
7	72.1	70.3	76.4	78.3	80.6	81.1	85.2	83.4	85.9	85.7	87.0	86.9	87.1	87.2	88.3	87.4	87.9	89.0	88.9	88.8	89.2	89.0	89.5	89.1	89.4	89.6	89.0
5	68.1	70.9	74.3	77.6	79.1	80.4	83.5	82.9	84.6	83.6	85.2	85.5	85.3	85.2	86.5	86.3	87.1	87.3	87.5	87.4	87.7	88.8	88.0	88.4	88.8	87.8	88.4
3	64.7	70.2	73.0	75.3	77.9	80.1	79.6	80.9	81.4	82.0	83.4	83.9	83.3	83.7	84.2	84.4	85.6	85.5	85.5	85.7	86.9	85.8	85.8	86.0	86.3	85.2	86.0
1	64.4	66.7	70.0	72.2	74.1	76.3	76.2	77.5	77.2	77.9	78.1	78.0	78.4	78.5	78.4	78.8	78.8	79.3	80.2	80.9	80.2	79.9	80.1	80.6	80.8	80.4	81.1

Fig. 4.13 Analiza rezultatelor pentru secvențele de matisare valide



	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
11	92.08	90.24	90.66	88.40	89.76	87.74	89.31	88.09	86.91	88.09	87.29	87.92	86.53	87.19	86.81	86.88	86.39	86.32	86.35	86.42	86.18	84.93
9	87.71	85.38	86.91	83.47	85.00	82.92	83.75	83.13	82.64	83.06	81.98	82.67	81.32	81.70	81.04	81.63	80.52	80.97	80.42	80.59	80.42	80.00
7	86.81	85.94	85.83	84.83	85.03	83.54	84.44	83.13	83.16	82.01	82.67	82.26	81.84	81.63	81.42	81.49	80.59	80.73	80.45	80.28	79.34	80.45
5	86.11	87.53	85.21	85.52	83.89	83.75	83.82	82.71	82.81	81.77	82.81	81.53	82.22	81.04	82.57	81.11	81.04	80.59	80.21	80.45	79.27	80.31
3	86.84	86.15	84.41	84.83	83.33	83.61	83.26	81.88	82.12	82.33	81.77	81.49	82.50	81.39	80.94	80.87	80.56	79.48	80.52	79.41	79.58	79.97
1	85.80	84.79	83.09	83.06	81.88	80.90	80.38	80.17	79.58	80.00	79.38	79.72	80.52	80.17	80.49	79.97	80.14	79.48	78.96	78.06	79.44	79.41

Fig. 4.14 Analiza rezultatelor pentru secvențele invalide de matisare

4.3. Determinarea semnalelor de matisare pe baza unor laitmotive din secvența de nucleotide

Matisarea este un proces important prin care intronii sunt îndepărtați pentru crearea ARN-ului matur. Acest proces este rezultatul unor interacțiunii între proteine mici. Unele dintre aceste proteine se leagă la anumite zone ale pre-mARN-ului care au o secvență de ARN specifică. Mici alterări ale acestor regiuni pot duce la diferite tipuri de boli [117]. Matisarea alternativă este procesul prin care exonii pot fi incluși sau excluși din mARN-ul final. Acest proces se desfășoară în peste 90% din pre-mARN-ul genelor umane [116] și este controlat de semnalele de matisare. Schimbările acestor semnale de reglare a matisării pot modifica *splicing*-ul alternativ, astfel încât să conducă la o expresie anormală a genei care va genera proteine defectuoase. Înțelegerea procesului de reglare a matisării ajută la identificarea și vindecarea diferitelor afecțiuni [134].

În general, majoritatea cadrelor medicale consideră că toate informațiile relevante despre procesul de *splicing* sunt regăsite în regiunea de matisare. Însă cercetările au arătat că acest lucru nu este adevărat [135]. Procesul de matisare este controlat într-o anumită măsură prin îmbinarea elementelor de reglare. Aceste elemente pot fi găsite fie în zona intronică, fie în zona exonică [136]. Având în vedere că elementele de reglare sunt prezente și în exoni, secvențele codificatoare au funcții care depășesc pe cel evident, anume tiparul pentru proteină. Deși s-au efectuat o serie de studii științifice asupra semnalelor de matisare, proporțiile în care secvențele codificatoare au un rol în *splicing* nu sunt cunoscute [137]. În principiu, o secvență codificatoare conține semnale exonice de *splicing* (ESR) [137] sub formă de amplificatoare exonice de matisare (ESE) și inhibitoare exotice de matisare (ESS) [138]. Orice modificări în regiunile ESR sunt cunoscute ca având efecte asupra fenotipului uman, după cum sugerează mai mulți autori [56], [75], [139].

De-a lungul timpului au fost propuse diferite instrumente pentru analiza efectelor variantelor genetice umane asupra ESR. *Human Splicing Finder* (HSF) [88] este un instrument care permite utilizatorilor să investigheze variații la nivel de secvență, oferind informații detaliate despre schimbările care au avut loc. Există, de asemenea, ESEFinder [55], RESCUE-ESE [140], EX-SKIP[141], toate fiind instrumente online și pot fi accesate cu ușurință. Unele se concentrează pe secvențe specifice (ESEFinder, RESCUE-ESE), în timp ce HSF integrează mai multe secvențe simultan și oferă informații detaliate despre schimbări. Principalul dezavantaj este că trebuie introdusă fiecare variantă în platformă pentru a obține rezultate. Acest lucru nu este întotdeauna foarte convenabil, de exemplu atunci când se efectuează studii care presupun secvențierea întregului exon.

Ținta acestui studiu a fost de a genera o metodă de analiză bazată pe datele publicate în literatură, care ar oferi o indicație generală despre schimbările semnalelor ESR când apare o variație genetică. Rezultatul metodei ar trebui să fie ușor de interpretat și de integrat în fișierele care conțin mai multe înregistrări cu variații genetice. Metoda propusă va ajuta procesul de filtrare a variantelor și prioritizarea acestora, atât în cercetare, cât și în diagnosticul clinic.

O parte dintre informațiile prezentate în acest studiu au fost publicate în lucrarea intitulată *Determining Splicing Signal Variation in Humans by Analyzing the Regulatory Splicing Motifs* [142].

4.3.1. Determinarea intensității semnalului

Secvențele de reglare a matisării (ESE și ESS) au fost obținute dintr-o serie de surse prezente în literatură. Aceste grupuri de secvențe au fost obținute folosind metode computaționale diferite, astfel încât în forma finală publicată, secvențele și scorurile asociate au avut reprezentări diferite. Laitmotivele pentru inhibare publicate de *Sironi* [143] (Si-ESS) și laitmotivele de activare publicate de *Cartegni* [55] (Ca-ESE) au fost prezentate utilizând matrice cu pondere poziționlă (PWM). Tiparele (PESE și PESS) identificate de *Zhang* și *Chasin* [144] și tiparele (EIE și IIE) identificate de *Zhang* în [145], au fost determinate prin utilizarea metodelor de comparație, iar acestea sunt reprezentate ca o succesiune de tupli formați din secvență și scor. Lista activatorilor RESCUE-ESE publicat de *Burge* [140] și lista Fas-ESS publicate de *Wang* [138] au fost reprezentate sub formă de secvențe de ADN simple. În cadrul acestui studiu s-a folosit reprezentarea ADN-ului care a fost utilizată pentru toate secvențele.

Semnalul global a fost împărțit în două componente, indicatorul de activare (semnalul amplificator) și indicatorul de inhibare (semnalul inhibitor). Ambii indicatori pot avea valori pozitive sau negative. O valoare pozitivă pentru indicatorul de amplificare înseamnă că semnalul de activare pentru secvența țintă (secvența cu variație genetică) a crescut în comparație cu secvența de referință. Dacă indicatorul de amplificare este negativ înseamnă că semnalul de activare a scăzut, iar acest fapt va conduce la o diminuare a șansei de prindere a factorului de amplificare a matisării pentru acea secvență. Dacă indicatorul inhibitor a avut o valoare pozitivă, înseamnă că s-a mărit șansa de legare a factorului de inhibare, iar dacă indicatorul are o valoare negativă, atunci aceasta a scăzut.

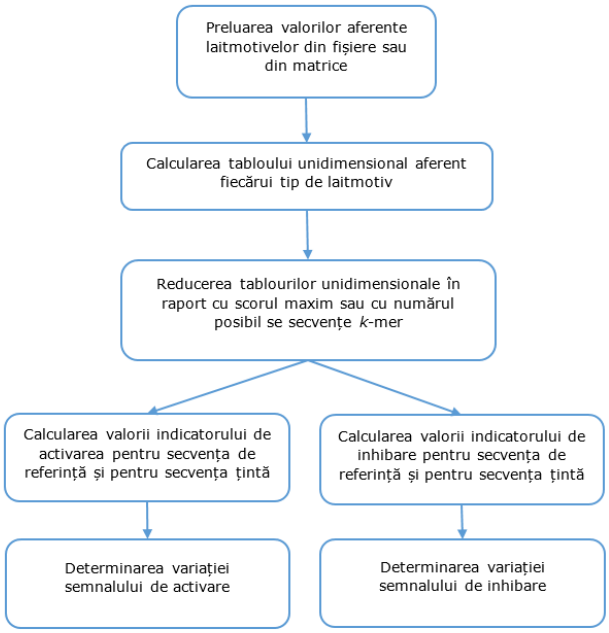


Fig. 4.15. Etapele pentru determinarea variației semnalelor de activare și inhibare

Pentru determinarea semnalului final de amplificare și a semnalului final de inhibare, a fost alocată o submulțime de semnale pentru fiecare secvență ADN. Calcularea intensității fiecărui semnal a fost realizată cu o serie de operații precum este prezentat în Fig. 4.15.

Pentru reprezentările PWM (Si-ESS și Ca-ESE), intensitatea semnalului a fost calculată pe baza matricei și a pragului furnizat de autori conform ecuației (4.3-1), unde S_{sec} reprezintă scorul secvenței, a reprezintă matricea cu pondere pozițională iar sec_i reprezintă baza azotată de la poziția *i*.

$$S_{sec} = \sum_{i=0}^{n} a_{sec_i,\ i} \tag{4.3-1}$$

Dacă scorul pentru secvența de referință și scorul pentru secvența țintă au fost sub pragul minim, atunci cele două valori ale semnalelor vor fi ignorate. Altfel, dacă unul dintre semnale este peste pragul minim, valoarea semnalului este adunată la elementele vectorului corespunzător fiecărui indicator conform ecuației (4.3-2). Chiar dacă doar una dintre secvențe are scorul peste pragul minim, scorul se va lua în considerație pentru ambele secvențe.

$$v_i = v_i + S_{sec} \tag{4.3-2}$$

unde v_i reprezintă al *i*-lea element din vectorul indicatorului, iar S_{sec} reprezintă scorul secvenței. Indicatorul ia valori de la poziția primei baze azotate din secvență până la ultima poziție din secvență.

Pentru laitmotive care au fost determinate prin utilizarea metodelor comparative și au avut un scor asociat (v_{sec}) pentru fiecare secvență (PESE, PESS, EIE, IIE), valoarea semnalului S_{sec} a fost calculată împărțind scorul secvenței (v_{sec}) cu cel mai mare scor din acel grup (v_{max}) conform ecuației (4.3-3).

$$S_{sec} = \frac{v_{sec}}{v_{max}} \tag{4.3-3}$$

Pentru tiparele care au doar secvențe (Fas-ESS și Rescue-ESE), semnalul a fost calculat prin împărțirea numărului de secvențe valide la numărul de secvențe posibile în raport cu varianta genetică. Desigur, numărul de secvențe posibile reprezintă chiar numărul de baze azotate aflate într-o secvența validă. În cazul Fas-ESS și Rescue-ESE secvențele sunt *6-mer*, adică șase baze azotate. Prin urmare, scorul unei secvențe este raportul dintre numărul de secvențe valide și numărul de k-mer al secvențelor precum este redat în ecuația
(4.3-4).

$$S_{sec} = \frac{nr_{valid}}{k-mer_{metoda}} \tag{4.3-4}$$

De exemplu, pentru un singur polimorfism (SNP), calculul ar presupune numărul de secvențe detectate împărțit la șase. Dacă s-au găsit mai multe variații genetice consecutive, atunci căutarea secvențelor valide trebuie realizată de la prima secvență *k*-mer care conține prima variantă pe ultima poziție până la ultima secvență *k*-mer care conține ultima variantă pe prima poziție.

Pentru calcularea completă a variației cauzate de un SNP, este necesar ca lungimea secvenței de referință și a secvenței țintă să fie de 15 nucleotide. Această

dimensiune se datorează lungimii celor mai mari laitmotive, adică opt nucleotide (poziții). Ca varianta genetică să fie pe fiecare poziție a secvenței, sunt necesare șapte nucleotide în fiecare direcție (aval și amonte). Secvențele de ADN au fost analizate în direcția pozitivă a catenei, de la 5 prim la 3 prim.

După calcularea celor două seturi aferente fiecărui indicator, s-a trecut la determinarea variației. Variația semnalului (vf_i) a fost calculată ca diferența dintre valorile semnalului amplificatorului din secvența țintă (vt_i) și valorile semnalului amplificatorului din secvența de referință (vr_i) precum în (4.3-5). Aceeași metodă a fost aplicată și pentru semnalul de inhibare. Valoarea finală a variației indicatorului de amplificare (EI) și valoarea finală a indicatorului de inhibare (SI) au fost determinate ca medie a valorilor diferite de zero din vectorul de variație a fiecărui indicator. Pe baza acestor două valori se va determina dacă fenotipul de matisare este amplificat sau este inhibat.

$$ {}_1^n vf_i = vt_i - vr_i \tag{4.3-5}$$

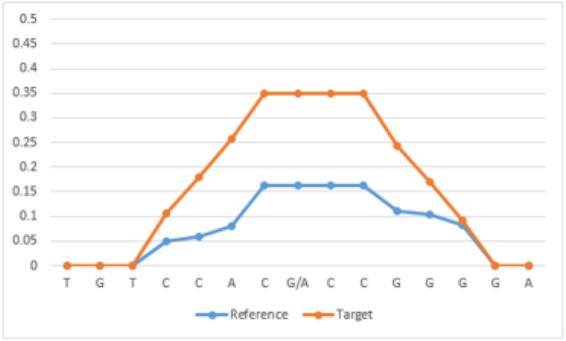
4.3.2. Validarea metodei

Testarea modelului pentru determinarea variației semnalului de amplificare respectiv de inhibare s-a realizat folosind setul de date publicat ca material suplimentar în lucrarea lui *Zhang* și *Chasin* [144]. Setul de date este format din două grupuri de variații genetice. Primul grup avea variante genetice care se găsesc doar în gena HPRT, iar cel de-al doilea avea variante care se găsesc în diferite gene. Toate rezultatele pentru fenotipurile variațiilor genetice ale genei HPRT au corespuns cu fenotipurile de matisare găsite în datele de testare. O parte dintre rezultate sunt prezentate în Tabelul 4.4, rândurile 1-5. Setul de date conține, în cea mai mare parte, fenotipuri de inhibare; singurele excepții au fost variațiile genei MAPT, Tabelul 4.4, rândurile 7-9.

Tabelul 4.4. Rezultatele modelului de predicție pentru semnalele de amplificare respectiv inhibare a matisării

Id	Genă	Secvență	Fenotip	Semnalul de amplificare	Semnalul de inhibare
1	HPRT	TGAAATT[C/-]CAGACAA	▼	-6.3	0.0
2	HPRT	TGAAATT[CC/TT]AGACAAGT	▼	-11.9	10.1
3	HPRT	AGTTGTT[G/A]GATTTGA	▼	-1.5	1.0
4	HPRT	GTTGTTG[G/A]ATTTGAA	▼	-1.5	7.0
5	HPRT	AATACTT[C/T]AGGGATT	▼	-8.0	16.0
6	HPRT	TGCCCTT[G/T]ACTATAA	▼	-2.8	2.4
7	MAPT	TAATTAA[T/G]AAGAAGC	▲	36.0	4.5
8	MAPT	AGGATAA[T/C]ATCAAAC	▲	2.7	0.0
9	MAPT	CTGGATCT[T/C]AGCAAC	▲	4.0	-0.5
10	F8	AATTTGG[C/G]GGGTGGA	▼	0.0	26.2
11	BRCA1	AGATGCT[G/T]AGTATGT	▼	-16.9	6.7
12	PMM2	CAGCCAA[G/A]AAGAACG	▼	-30.2	-4.5
13	ADA	TGTCCAC[G/A]CCGGGGA	▼	13.1	0.0
14	ATM	TGACCTC[G/A]AAACAGC	▼	3.6	0.0
15	CFTR	AATGGGA[T/G]AGAGAGC	▼	4.0	3.7

A



B

tif (value 0-100)	Linked SR protein	Mutant Motif (value 0-100)	Variation
	SRp40	ccacacc (90.90)	New site
	SF2/ASF (IgM-BRCA1)	cacaccg (81.77)	New site
	SF2/ASF	cacaccg (80.78)	New site
	SRp40	acacccgg (83.29)	New site
g (91.92)	SF2/ASF (IgM-BRCA1)	cacccggg (89.69)	-2.43 %
g (91.92)	SF2/ASF	cacccggg (86.43)	-5.97 %
g (83.81)	SF2/ASF (IgM-BRCA1)	cacccggg (89.69)	+7.02 %
g (83.81)	SF2/ASF	cacccggg (86.43)	+3.13 %

Motif value (0-100) reference sequence	Mutant motif	Motif value (0-100) mutant sequence	Variation
12.88			Site broken -100

Enhancer motif reference sequence	Enhancer motif mutant sequence	Variation
	accggg	New Site

C

and mutant values. Wild Type value is taken as reference.

eference	Reference silencer (value 0-100)	Sironi Mutant motif	Mutant silencer (value 0-100)	Variation
		Motif 3 - TCTCCCAA	tgtcccca (62.63)	New site
T/A] GGGG	cgcccggg (62.60)			Site broken -0.05

sition	Silencer motif reference sequence	Silencer motif mutant sequence	Variation
	gcccggg		Site broken

Fig. 4.16 Comparare rezultatele metodei cu HSF pentru gena ADA

Dintre cele 25 de variante genetice prezente în cel de-al doilea grup de gene, pentru 22 efectul asupra matisării a fost determinat corect. Trei dintre acestea având fenotipul de matisare greșit (Tabelul 4.4, rândurile 13-15). Cea mai evidentă variantă genetică a fost cea din gena ADA (înregistrarea 13), care are un efect de inhibare al matisării, dar modelul a indicat o activitate de amplificare cu 13 puncte (Fig. 4.16 A). HSF a fost utilizat pentru a obține a treia estimare. HSF a prezis că au fost create trei

secvențe noi Ca-ESE (Fig. 4.16 B), iar o secvență a avut o creștere a intensității semnalului. Un laitmotiv EIE a fost de asemenea creat, iar un laitmotiv PESE a fost distrus. În categoria inhibitoarelor, o secvență Si-ESS a avut o scădere a intensității și s-a creat un nou Si-ESS (Fig. 4.16 C). În plus, un motiv IIE a fost deteriorat.

4.3.3. Analiza rezultatelor

Rezultatele pentru toate înregistrările din setul de date ale genei HPRT au fost în concordanță cu cele indicate de autori. Prezentarea unei analize statistice cu rezultatele obținute nu este relevantă în acest caz deoarece valoarea de referință pentru fenotipul de matisare este doar o indicație direcțională. În cazul genei HPRT, toate variantele genetice au indus o scădere a semnalului de îmbinare. În cazul metodei prezentate, o scădere a fenotipului de îmbinare este semnalată într-o varietate de moduri. De exemplu, secvența care conținea o deleție (înregistrarea 1 din Tabelul 4.4) a primit o valoare negativă pentru indicatorul de activare de -6,36 și nu a indicat nicio modificare pentru indicatorul de inhibare. Pentru secvența care are două substituții consecutive (Tabelul 4.4, înregistrarea 2), valorile calculate au arătat un indicator de activare scăzut (-11.97) și un indicator de inhibare ridicat (10.13). O altă situație în gena F8 (Tabelul 4.4, înregistrarea 10), în care indicatorul activator nu a fost afectat, dar valoarea indicatorului inhibitor a crescut cu 26,2 puncte. Toate aceste comportamente indică o scădere a fenotipului de matisare. Opusul este valabil pentru creșterea fenotipului de matisare. De exemplu, în gena MAPT (Tabelul 4.4, înregistrarea 7), deși ambii indicatori sunt pozitivi, semnalul activator a crescut cu o valoare semnificativ mai mare decât semnalul de inhibitor.

După cum s-a menționat anterior, metoda a indicat o creștere a semnalelor de matisare pentru trei variante genetice, anume în genele ADA, ATM și CFTR. În cazul ADA, metoda a indicat o creștere semnificativă a semnalului amplificatorului. Pe baza calculelor, algoritmul nu a considerat că variația secvenței inhibitor a fost suficient de relevantă pentru a o raporta. Aceeași variație a secvenței a fost simulată în HSF pentru compararea rezultatelor. Rezultatul HSF a prezis că s-au creat noi secvențe de activare cu un scor mare și că a apărut o creștere pentru o secvență existentă (Fig. 4.16 B). Algoritmul prezice același efect într-un mod comprimat, cu o creștere de 13,1 puncte. În secțiunea de inhibitoare, HSF prezice două secvențe deteriorate și crearea unei noi secvențe de inhibare a matisării. În secțiunea *Sironi* a HSF (Fig. 4.16 C), secvența creată și cea deteriorată au aproape același scor. Înseamnă că aceasta este o schimbare nulă. Secvența deteriorată de la IIE, *GCCGGG*, nu are un scor în HSF. Cu toate acestea, după o verificare în setul de date, scorul acestui motiv este 6,82, ceea ce reprezintă un scor redus comparativ cu maximul de 54. Având în vedere cele prezentate, am considerat că algoritmul a calculat o valoare corectă cu datele disponibile. Aceeași procedură a fost făcută și pentru variațiile genetice ale ATM și CFTR. În ATM, s-au creat șase noi secvențe de activare și au fost distruse trei. La secvențele de inhibare, nu au fost indicate modificări. În gena CFTR, o secvență de amplificare a fost distrusă și au fost create opt noi. În secțiunea inhibitoare, au fost create trei noi secvențe și a fost distrusă una. În ansamblu, algoritmul a funcționat bine în determinarea direcției generale a indicatorilor și corespunde cu predicțiile HSF.

Ar trebui luat în considerație faptul că algoritmul calculează o valoare medie a pozițiilor din secvență care sunt nenule și faptul că valoarea fiecărei poziții este determinată pe baza rezultatelor din diferite metode. Datorită acestora, importanța poziționlă a unei secvențe a fost diluată într-o anumită măsură. Pentru utilizatorul final, informațiile detaliate, precum cele oferite de HSF (Fig. 4.16 B și C), sunt

pierdute. Cu toate acestea, în faza inițială a unui proiect, cum ar fi prioritizarea variantelor genetice, detaliile privind secvențele sunt irelevante și sunt greu de urmărit. Soluția de a afișa informațiile generale este mai utilă pentru personalul medical. După ce se selectează un mic grup de variante genetice, se pot utiliza instrumente precum HSF pentru a afișa informații detaliate pentru fiecare variantă genetică.

În ceea ce privește valoarea indicatorilor, în majoritatea cazurilor, o variație de unul sau două puncte nu este o indicație a schimbării semnificative a fenotipului de matisare. Cu toate acestea, atunci când o variantă genetică creează noi secvențe de activare a matisării și distruge alte secvențe de activare, valoarea afișată indică numai modificarea semnalului. Această schimbare a semnalului reprezintă diferența dintre scorurile secvențelor activatoare create și ale celor distruse în raport cu secvența de referință. O variație de peste douăzeci de puncte reprezintă, de obicei, o schimbare semnificativă la nivelul secvenței, care ar putea reprezenta o schimbare în fenotipul de *splicing* și care poate produce sau nu o schimbare a fenotipului general al pacientului. Modelul nu poate decide în acest moment dacă o variație a semnalului de îmbinare va determina modificări fenotipice.

Există câteva limitări privind această metodă pentru determinarea variației semnalului de matisare. Scorurile secvențelor, folosite pentru determinarea intensității, obținute din literatură, au fost calculate prin metode diferite. Calcularea indicatorilor finali dintr-un amestec de valori determinate diferit poate induce erori prin calibrarea greșită a acestora în ponderea indicatorului final. În plus, unele secvențe nu au valori pentru pragul de relevanță. În aceste cazuri, relevanța secvenței a fost determinată în raport cu scorul cel mai mare. În plus, Fas-ESS și Rescue-ESE nu au avut scoruri asociate și au fost folosite ca raport de secvențe. Ponderea pe care fiecare tip de secvență o are în valoarea finală a indicatorului este diferită. De exemplu, Fas-ESS și Rescure-ESE au o valoare maximă de 14,2 puncte în fiecare indicator (semnal de inhibare pentru Fas-ESS și semnal activare pentru Rescue-ESE). Valoarea indicatorului nu se limitează la o sută de puncte. Teoretic, acesta poate depăși acest prag. Totuși, acest lucru este dificil de realizat printr-o singură variație genetică între două secvențe de ADN. Pentru a obține o valoare mai mare ca o sute de puncte, secvența țintă trebuie să aibă mai multe substituții poziționale față de secvența de referință, într-un interval de ± 7 nucleotide în raport cu o nucleotidă.

O altă limitare este lipsa informațiilor detaliate pe care această metodă nu le oferă utilizatorului final. Așa cum am menționat, utilizatorul final nu are acces la metoda internă de calculare a indicatorilor. În acest caz, informațiile sensibile pot trece neobservate. Dat fiind faptul că scorurile multiple se combină pentru a forma un indicator, într-o oarecare măsură, contextul biologic și specificitatea secvențelor sunt diluate. Această metodă este utilă pentru o abordare generală, dar este recomandată să fie asistată cu alte instrumente (HSF) care sunt mai specifice și oferă mai multe detalii.

Beneficiul utilizării instrumentelor și metodelor care oferă informații despre ESR este că variantele genetice sinonime pot fi luate în considerație. Faptul că *Savisaar* și *Hurst* au afirmat în raportul lor [137], că variantele sinonime sunt nefuncționale poate fi uitat. În plus, *Caceres* și *Hurst* au estimat în lucrarea lor [136] că cel puțin 4% dintre mutațiile sinonime sunt dăunătoare. Integrarea acestor variabile (indicatorii ESR) în procesul de prioritizare a variantelor genetice poate avea implicații enorme. Până în prezent au fost luate în considerație numai variante genetice care nu sunt sinonime pentru modelarea bolilor complexe, dar cu noile perspective asupra rolului variantelor sinonime lucrurile s-ar putea schimba și strategiile de prioritizare ar trebui să fie actualizate.

4.4. Concluzii de capitol și contribuții proprii

În studiul din capitolul 4.1 s-au identificat unele regiuni de matisare care au două sau mai multe secvențe care corespund regiunii de prindere (BRS) a spliceosomului. Acest lucru se poate datora regiunii de pirimidine care facilitează procesul de formare a BRS. Secvențele regiunilor de prindere sunt adesea localizate în apropierea poziției 16 și 28 în amonte de exonului 3'. Au mai fost identificate, *in silico*, secvențe intronice care sunt similare ca structură cu regiunile de matisare. Rolul biologic al acestor secvențe poate fi testat în continuare utilizând experimente *in vitro* sau *in vivo*.

Studiul din capitolul 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 baza ecuații generate din structura regiunilor de matisare. Acuratețea predicției acestor modele a fost cuprinsă între 70% și 80%. În ultima parte a capitolului a fost propusă o metodă pentru detecția regiunilor de matisare care are la bază o distanță față de secvențele vecine. Pentru calcularea distanței s-au analizat o serie de metode, iar cea aleasă a fost *Needleman-Wunsch*. Folosind această metodă am realizat o analiză computațională pentru determinarea unui optim al lungimii secvenței și un optim al numărului de vecini. Rezultatele au indicat un optim pentru 20 de nucleotide și 9 vecini. Folosind aceste valori s-a reușit o acuratețe de 85.61%.

Scopul studiul din capitolul 4.3 a fost dezvoltarea unei metode care să permite adnotarea fișierelor VCF cu informații despre variațiile semnalului de matisare. În prima fază s-au adunat bazele de date cu secvențe considerate semnale pentru procesul de matisare. Aceste secvențe, în forma inițială, nu puteau fi combinate simultan. Prin urmare, s-au dezvoltat o serie de ecuații care au permis utilizarea lor simultană pentru determinarea intensității semnalului unei secvențe. În continuare, pentru calcularea diferențelor 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 amplificarea a matisării, cât și pentru semnalul de inhibarea a matisării. Direcția generală este dată de analiza acestor două componente. Validarea metodei s-a realizat pe o bază de date care conținea secvențele genetice și care conținea 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.

Contribuții personale:

1. Realizarea unui studiu asupra tuturor intronilor din cromozomul 21 cu intenția de a identifica regiunile de matisare parazite din regiunile intronice;
2. Dezvoltarea unei metode pentru calcularea variației semnalului de matisare în cazul modificării secvenței ADN respectiv ARN;
3. Identificarea unor secvențe redundante pentru prinderea spliceosomului în regiunea de matisare;
4. Prezentarea unei metode pentru detecția regiunilor de matisare în funcție de distanța dintre secvența țintă și secvențele vecine;
5. Studiu statistic în care se prezintă structura regiunii de matisare;
6. Dezvoltarea unor modele de predicție pentru regiunile de matisare folosind algoritmul *Needleman-Wunsch*.

5. MODELAREA GRADULUI STEATOZEI FOLOSIND MARKERI GENETICI

O scurtă introducere a modelării afecțiunilor complexe și a unor algoritmi utilizați în învățarea automată, a fost realizată în subcapitolul 2.4 și subcapitolul 2.1.

5.1. Materiale utilizate

Pentru efectuarea calculelor necesare determinării modelului matematic aferent steatozei hepatice, s-a folosit un sistem cu procesor AMD A10-6800B, 8 GB de RAM, cu un sistem de operare Windows 10, pe 64 de biți. Platformele software folosite au fost *Knime* 3.4.2 și *Anaconda* împreuna cu bibliotecile disponibile. Datele folosite au fost obținute de la Centrul de Medicină Genomică din Timișoara. Aceste date reprezintă înregistrările unor pacienți care suferă de steatoză, în diferite etape. De asemenea, fiecare înregistrare are asociat un număr de polimorfisme (SNPs) suspectate că influențează afecțiunea. În cazul studiului, obiectivul principal a fost generarea unui model care, pe baza prezenței variantelor genetice, să catalogheze stadiul afecțiunii.

Baza de date folosită dispune de 400 de înregistrări, fiecare având asociate 56 de coloane. Prima coloană din fișier conține stadiul afecțiunii, iar în restul coloanelor reprezintă genotipul pacientului. Fiecare poziție din genotip are trei stări posibile prin care se indică dacă alela respectivă este homozigotă, heterozigotă sau referință. Starea afecțiunii a fost reprezentată prin cinci stări începând de la -1 până la 3. Stările de la zero la trei (0-3) reprezintă rezultatul histologic (stadiul afecțiunii, clasificat de către specialist), iar lipsa afecțiunii a fost reprezentată cu valoarea minus unu (-1).

În Tabelul 5.1 sunt prezentate informații suplimentare despre variantele genetice folosite pentru realizarea modelului. Tabelul va fi folosit în continuare pentru analiza diferențelor între frecvența variantelor în grupul investigat și frecvența variantelor în populația lumii. *RS* reprezintă ID-ul variantei asociat în dbSNP, *Chr* reprezintă cromozomul, *Locus*-ul reprezintă poziția pe cromozom (versiunea GRCh37), *Ancestral* reprezintă nucleotida ancestrală, MAF reprezintă frecvența alelei minore iar *High* MAF reprezintă cea mai ridicată frecvență întâlnită la o anumită populație.

Tabelul 5.1 Poziția pe cromozom și frecvența în populația lumii a fiecărei variante genetice

RS	Chr	Locus	Alelă	Ancestral	MAF [%]	High MAF [%]
rs1149222	7	87073775	G/T	G	0.38	0.41
rs2071645	7	87105276	G/C	C	0.34	0.46
rs31672	7	87059699	C/T	C	0.36	0.46
rs4148811	7	87101486	T/G	G	0.31	0.49
rs9655950	7	87033561	C/T	C	0.34	0.48
rs1202283	7	87082292	G/A	G	0.35	0.5
rs2854117	11	116700142	T/C	T	0.5	0.5
rs12676	3	53857803	A/C	C	0.13	0.33
rs2289209	3	53852835	C/T	T	0.12	0.37
rs4563403	3	53850814	C/T	T	0.25	0.49

rs4687591	3	53864407	A/G	G	0.46	0.49
rs6807783	3	53859662	G/C	C	0.39	0.39
rs7634578	3	53876728	C/T	T	0.17	0.5
rs881883	3	53847805	A/G/T	A	0.42	0.49
rs1557502	22	51013998	C/T	C	0.45	0.5
rs1557503	22	51013072	G/A	G	0.19	0.43
rs470117	22	51009953	C/T	C	0.33	0.49
rs7238	22	51007488	A/G	A	0.28	0.5
rs2526678	11	61623793	G/A	G	0.13	0.43
rs526126	11	61624885	G/C/T	G	0.32	0.34
rs10135928	14	64866439	T/C	C	0.06	0.22
rs1801133	1	11856378	G/A	G	0.25	0.5
rs2066471	1	11860458	C/G/T	C	0.11	0.28
rs4846048	1	11846252	G/A	G	0.29	0.49
rs4846052	1	11857951	T/C	T	0.49	0.5
rs7525338	1	11862332	C/T	C	0.02	0.11
rs868014	1	11849447	A/G	G	0.07	0.28
rs11557927	10	102121816	T/G	T	0.09	0.18
rs11599710	10	102105788	G/A	G	0.1	0.19
rs12247426	10	102115327	C/G	G	0.12	0.47
rs2167444	10	102124744	T/A	T	0.13	0.27
rs7849	10	102122603	T/C	C	0.27	0.48
rs10120572	9	108077756	T/G	G	0.12	0.5
rs10820799	9	108092216	A/C	C	0.14	0.49
rs193008	9	108042806	T/C	C	0.27	0.46
rs328006	9	108039808	G/C	C	0.24	0.5
rs440290	9	107987290	T/C	C	0.28	0.43
rs443094	9	108016685	G/C	C	0.22	0.5
rs7018875	9	108077434	C/A	A	0.14	0.49
rs9891119	17	40507980	A/C	A	0.39	0.5
rs1580820	3	195966258	G/A	A	0.07	0.26
rs4898190	X	24607933	A/C	C	0.01	0.08
rs1109859	17	17424333	G/A	A	0.25	0.5
rs12103822	17	17418432	C/G/T	G	0.04	0.23
rs16961845	17	17432456	C/T	C	0.13	0.46
rs4244593	17	17420218	T/A/G	G	0.43	0.5
rs4479310	17	17405504	C/T	C	0.4	0.48
rs7214988	17	17491836	C/G	C	0.15	0.49
rs7946	17	17409560	C/T	C	0.42	0.48
rs8068641	17	17480187	A/G	G	0.24	0.46
rs936108	17	17439793	C/T	C	0.37	0.5
rs13342397	17	17460926	T/C	T	0.12	0.45
rs6502603	17	17445680	G/T	G	0.42	0.5
rs2281135	22	44332570	G/A	A	0.26	0.49
rs738409	22	44324727	C/G	C	0.26	0.49

5.2. Analiza descriptivă a caracteristicilor genotipului

Înaintea realizării unui model de predicție, a fost necesară analiza datelor care urmează să stea la baza modelului. Cea mai la îndemână unealtă pentru această operațiune a fost reprezentarea grafică a celor 55 de caracteristici asociată cu fiecare înregistrare. Pentru asta s-a folosit biblioteca *matplotlib* cu ajutorul căreia s-au generat histogramele pentru fiecare caracteristică, redată în Fig. 5.1.

Fig. 5.1. Reprezentarea stărilor în care se regăsesc variantele genetice

Din reprezentările realizate în Fig. 5.1, se poate observa că o serie de caracteristici au o varietate scăzută a stărilor. Cu alte cuvinte, o stare a variantei genetice este majoritară, exemplu PEMT_rs12103822. Această situație poate fi favorabilă dacă starea unei caracteristici este puternic corelată cu starea (grad) afecțiunii. Altfel, dacă nu există o corelație puternică atunci caracteristica poate fi

eliminată din procesul de modelare, prin urmare reducând dimensiunea modelului final.

Pentru a determina caracteristicile care au o varietate redusă, s-a calculat frecvența stărilor pentru fiecare caracteristică. Variantele genetice cu valoare mai mare de 0.8 sunt: SLC44A1_rs10120572, MTHFD1_rs10135928, SLC44A1_rs10820799, SCD_rs11557927, SCD_rs11599710, PEMT_rs12103822, SCD_rs12247426, PEMT_rs13342397, CHKB_rs1557503, PEMT_rs16961845, CHDH_rs2289209, FADS2_rs2526678, PCYT1B_rs4898190, SLC44A1_rs7018875, PEMT_rs7214988, MTHFR_rs7525338, CHDH_rs7634578, MTHFR_rs868014. Prin urmare, s-a studiat corelația dintre aceste variante și stadiul afecțiunii. Operația s-a realizat folosind funcția *corr* din biblioteca *Pandas.* Rezultatele, folosind metodele *Pearson*, *Kendall* și *Spearman*, sunt prezentate în Fig. 5.2. Se observă că niciuna dintre variantele genetice nu prezintă o corelație puternică cu stadiul afecțiunii. Prin urmare, caracteristicile pot fi excluse din procesul de generare a modelului. Menținerea lor în procesul de modelare poate induce mai mult zgomot decât utilitate. De asemenea, reducând numărul de variabile scad și șansele de *overfitting* al modelului.

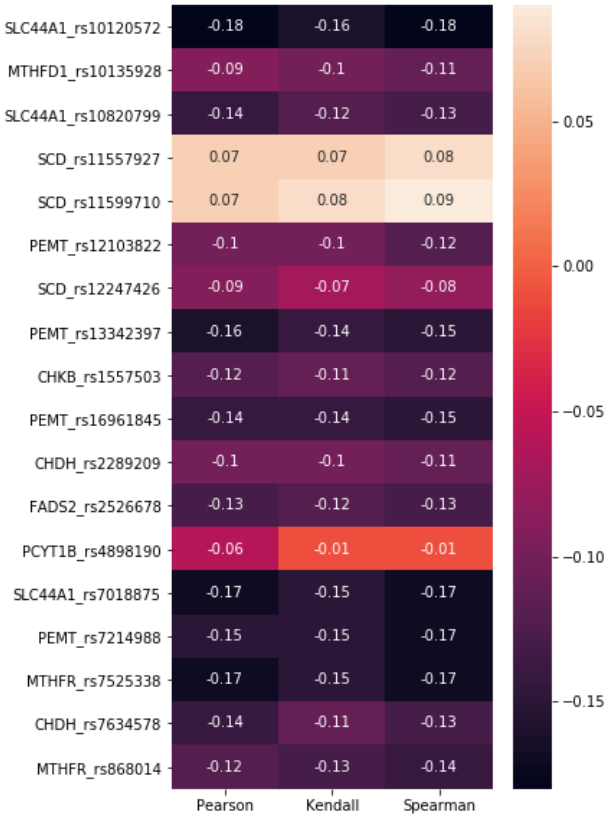


Fig. 5.2 Corelația dintre starea SNP-urilor cu varietate redusă și stadiul steatozei

Pentru a scoate în evidență și alte posibile corelații, analiza a fost extinsă pe întregul eșantion de variante genetice. În figura Fig. 5.3 s-a realizat matricea de corelații a variantelor genetice, inclusiv a stadiului steatozei. Pentru generarea

matricei s-a folosit metoda *Pearson*. Diferența dintre metodele *Pearson* și *Spearman*, respectiv *Kendall*, precum s-a văzut anterior, este marginală.

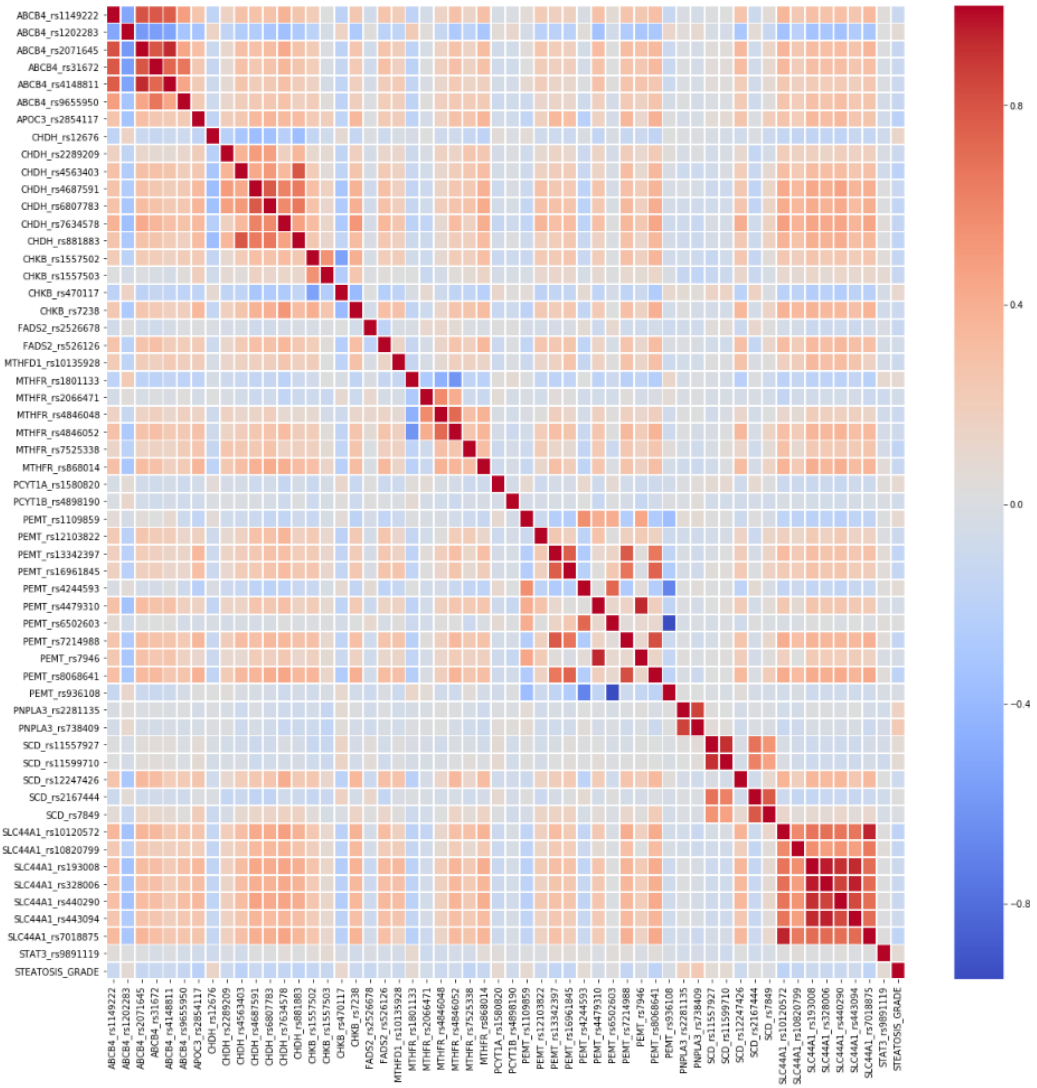


Fig. 5.3 Corelația dintre variantele genetice și stadiul steatozei

În cazul variantelor genetice care se află pe aceeași genă se observă un grad de corelație. De exemplu, variantele de pe SLC44A1, PNPLA3, CHDH, ABCB4 au un grad ridicat de corelație pozitivă. Pentru gene ABCB4, varianta rs1202283 are un grad ridicat de corelație negativă cu restul variantelor din această genă. Același lucru îl întâlnim și la gena PEMT, unde rs936108 este invers corelat cu rs6502603. Corelațiile pot fi explicate într-o oarecare măsură de procesul biologic denumit încrucișare cromozomială (*crossing-over*).

Dacă ne uităm la relațiile dintre gene, putem observa că genele SLC44A1 și CHDH au gradul cel mai ridicat de corelație. Tot de SLC44A1 și CHDH sunt ușor corelate ABCB4, PEMT, MTHFR. Gena SCD și PNPLA3 sunt cel mai puțin corelate gene.

Dacă ne întoarcem la Tabelul 5.1, observăm că aceste gene se află pe cromozomi diferiți, deci în cazul acesta nu mai poate fi vorba despre încrucișare cromozomială. Relația cu stadiul steatozei nu arată foarte promițător, neavând corelații puternice cu niciuna dintre variantele genetice respectiv cu nicio genă.

5.3. Metode utilizate pentru generarea modelului

Datele avute la dispoziție reprezintă valori discrete asociate cu starea de zigozitate a variantelor genetice și al stadiului steatozei. Valorile nu au o caracteristică continuă. Prin urmare, este evident că avem de a face cu o problemă de clasificare. De asemenea, putem genera modele folosind metode deja consacrate în învățarea automată precum arbori decizionali, K-*nearest neighbors*, regresie liniară sau logistică etc.

În acest studiu, modelul final va fi un model generat *offline*, prin urmare se vor folosi doar datele avute la dispoziție. Modelul, cel puțin în contextul actualei lucrări, este antrenat local și este menit să permită evaluarea stării unui pacient, fără să sufere modificări în timpul procesului de predicție.

Pentru a obține un model cât mai general, care să fie capabil să genereze rezultate corecte pentru înregistrări noi, datele folosite pentru realizarea modelului trebuie să fie reprezentative. Dacă înregistrările nu sunt stratificate din punct de vedere al numărului de înregistrări per stare atunci modelul generat va fi dezechilibrat, astfel favorizând o anumită stare. În cazul de față, în baza de date numărul înregistrărilor cu stări diferite nu este egal. Prin urmare, pentru a putea genera un model echilibrat, toate stările afecțiunii trebuie să aibă un număr reprezentativ de înregistrări.

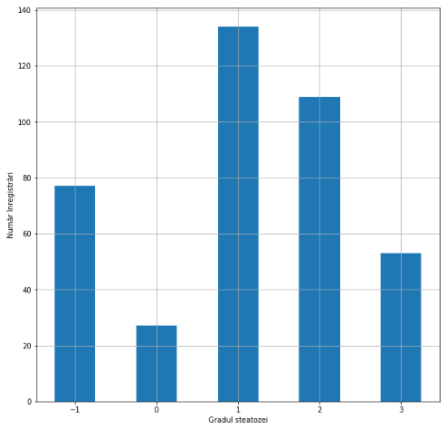


Fig. 5.4 Numărul de înregistrări în funcție de gradul steatozei

Pentru a genera setul de test și setul de antrenare s-a folosit clasa *StratifiedShuffleSplit* din pachetul *Scikit-Learn*. Înregistrările au fost împărțite stratificat, 20% dintre înregistrări au fost mutate în setul de test, iar 80% în setul pentru generarea modelului.

5.3.1. Extragerea unui model folosind clasificatori binari

Pentru început vom analiza performanța unui model generat de un clasificator *Stochastic-Gradient Descent* (SGD) pentru fiecare stare a afecțiunii. În acest caz trebuie să readaptăm cele două seturi de înregistrări în așa fel încât să obținem date binare. Înregistrările cu stadiul afecțiunii țintit de predictor vor avea valoarea *adevărat,* iar în caz contrar valoarea *fals.* Performanța modelului de clasificare a fost evaluată folosind trei metode: validare încrucișată, curba dintre sensibilitate și specificitate și curba ROC.

Pentru realizarea validării încrucișate setul de date rezervat realizării modelului a fost împărțit în trei subseturi. Fiecare două subseturi au fost folosite pentru a genera un model. Ulterior, folosind al treilea subset, s-a verificat care este acuratețea acestuia. Prin acuratețe, înțelegem numărul total de valori identificate corect raportat la numărul total de cazuri. Rezultatele prezentate în Fig. 5.5 par promițătoare. Prin metoda SGD, utilizată precum un clasificator binar, s-a reușit o acuratețe medie de 69,7%. Pentru detectarea unui stări, aceste modele au o performanță decentă, dar nu se pot folosi simultan. Dacă sunt folosite în tandem, în unele cazuri, s-ar contrazice deoarece sunt independente. De asemenea, dacă se urmăresc datele expuse în Fig. 5.4 și Fig. 5.5 se poate observa cum cantitatea de date și performanța sunt oarecum invers proporționale. De exemplu, gradul 0, care are cele mai puține înregistrări (27 în total, 22 în setul de antrenare, 5 în subset), are scorul acurateței cel mai mare (83%), iar gradul 1 care are 107 înregistrări are o acuratețe de 55%. Prin urmare, se presupune ca fiind un caz de *overfitting*.

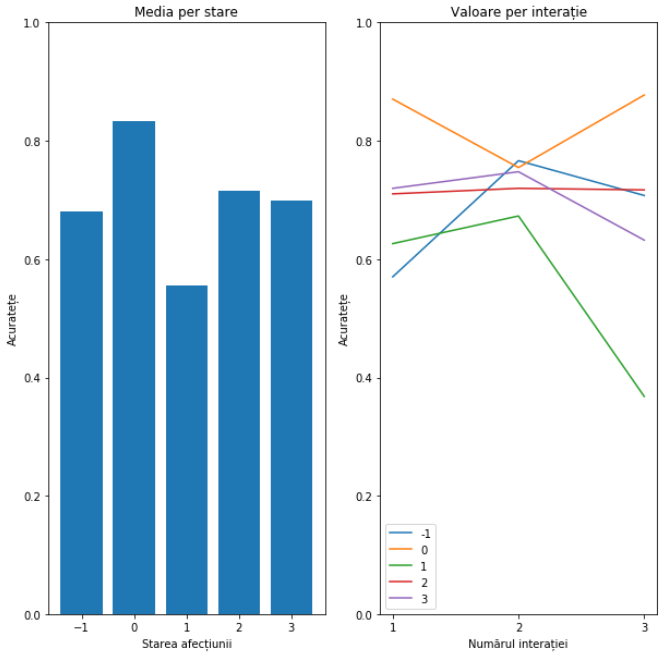


Fig. 5.5 Rezultatele validării încrucișate cu trei subseturi de date pentru metoda SGD

Pentru a realiza o predicție, metoda SGD, în etapa de generare a modelului, va stabili un prag în funcție de cele 55 de variabile de intrare. Acest prag va determina

în cele din urmă dacă o înregistrare este catalogată pozitivă sau negativă. Dacă valoarea generată de model este peste acest prag, atunci înregistrarea este considerată aparținând clasei pozitive, altfel clasei negative. Prin modificarea pragului putem îmbunătăți anumiți parametri de performanță în defavoarea altora, adică putem muta înregistrările în diferite categorii ale tabelei de contingență. Doi indicatori ai performanței sunt precizia și sensibilitatea.

Metoda de clasificare SGD are la bază un algoritm de optimizare iterativ, prin urmare pragul ales reprezintă o valoare optimă. Dacă ar trebui să analizăm un model ideal, am constata că acesta are valori în tabela de contingență doar în categoriile adevărat pozitiv (AP) și adevărat negativ (AN). Categoriile fals negativ (FN) și fals pozitiv (FP) ar fi nule. Desigur, această situație se regăsește în realitate foarte rar. Prin urmare, este important să controlăm înregistrările care fac parte din categoriile FN și FP.

Pentru a scădea numărul de înregistrări din categoria FN puteam ajusta performanța sensibilității modelului, iar pentru a controla categoria FP putem ajusta precizia. În reprezentările grafice din Fig. 5.6 este redată evoluția celor doi parametri pentru fiecare stare a patologiei, la diferite valori ale pragului.

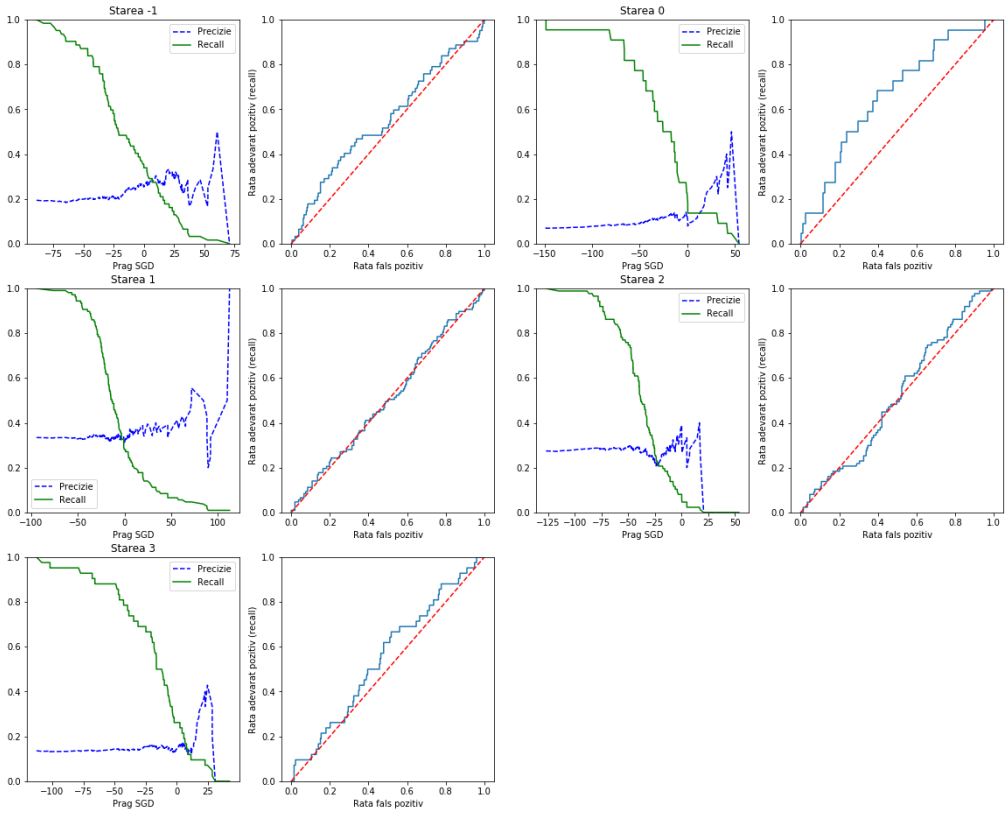


Fig. 5.6 Precizia și sensibilitate în raport cu pragul SGD și curba ROC.

Pentru a genera un model cât mai echilibrat este necesar să identificăm punctul optim dintre cei doi parametri. În funcție de model, prin echilibru înțelegem

direcția în care dorim să îndreptăm modelul. Adică, trebuie să alegem între un model care să aibă o rată FP scăzută sau o rată FN scăzută.

În cazul de față, pentru toate stările, sensibilitatea are o caracteristică bizară. De asemenea, acest indicator de performanță nu atinge valoarea maximă, excepție făcând starea 1. În schimb, precizia scade odată cu creșterea pragului SGD, acesta fiind un comportament normal, deoarece scade numărul înregistrărilor adevărat pozitive. Pentru a explicarea comportamentului sensibilității este nevoie să evaluăm numărul de înregistrări din categoriile adevărat pozitiv și fals pozitiv.

Ideal, prin creșterea pragului SGD ne așteptăm să scadă numărul înregistrărilor din categoria fals negativ și să crească numărul celor adevărat pozitive. În acest caz, observăm cum caracteristica crește până la o anumită valoare, după care scade abrupt. Pentru a înțelege fenomenul, s-a reprezentat grafic, în Fig. 5.7, valorile prezise de model în raport cu valorile autentice. Desigur, s-a ținut cont că modelul bazat pe SGD va calcula o valoare pentru fiecare înregistrare. Valoarea, în funcție de pragul ales, poate reprezenta clasa pozitivă sau clasa negativă. Prin urmare, dacă analizăm reprezentările pentru stările -1, 0, 2, și 3 observăm cum valorile adevărate sunt „umbrite" de valorile false. Dacă deplasăm pragul SGD, paralel cu abscisa, am obține caracteristica sensibilității modelului. În aceste condiții puteam explica faptul că sensibilitatea nu atinge valoarea maximă pentru stările menționate anterior. Starea 1 reprezintă excepția, aceasta având o înregistrare în clasa pozitivă pentru care modelul indică o valoarea mai mare decât maximumul clasei negative.

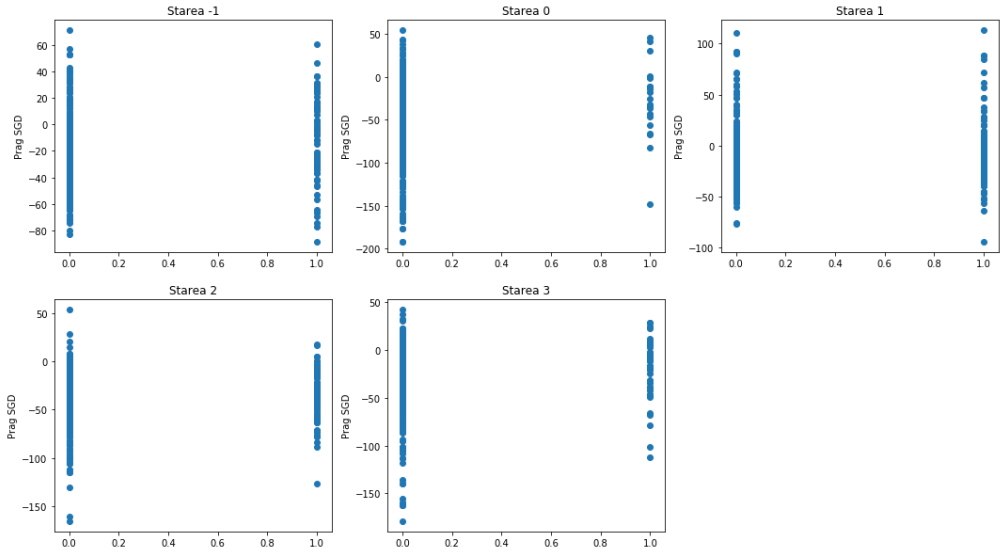


Fig. 5.7 Valorile prezise de model în raport cu valorile actuale

Dacă analizăm curbele ROC asociate celor vinvi modele de clasificare, observăm oarecum același lucru, anume că înregistrările din categoria AP sunt direct proporționale cu cele din categoria FP. De altfel, deoarece această curbă evaluează relația dintre sensibilitate și specificitate, putem spune că clasificatoarele, cu excepția celui pentru starea 0, au o performanța asemănătoare unui clasificator aleatoriu, reprezentat prin linia roșie întreruptă în Fig. 5.6. Clasificatorul pentru starea 0 are o performanță mai bună decât unul aleatoriu, dar este departe de performanța unui clasificator ideal.

5.3.2. Generarea unor modele folosind clasificarea multiclasă

În subcapitolul anterior s-a realizat clasificarea fiecărei stări, individual. În continuare, folosind SGD s-a încercat generarea modelului capabil să clasifice toate stările. De fapt, SGD va genera câte un model binar pentru fiecare stare și va aplica un sistem de votare unul contra toți (*one-vs-all*). Practic, toate modele de clasificare binară vor genera un scor, iar starea selectată va fi cea cu cel mai mare scor. Folosind această metodă, performanța modelului este redusă, acuratețea fiind de 25%. Pentru a înțelege motivul performanței scăzute, a fost generată matricea de contigență din Fig. 5.8. În cazul primei reprezentări, figura A, avem matricea de contigență cu valorile absolute. Rândurile reprezintă valorile autentice (reale) iar coloanele reprezintă valorile prezise. În rezultatele din Fig. 5.8 B, avem matricea de contigentă cu valorile normalizate în funcție de numărul total de înregistrări ale stării reale.

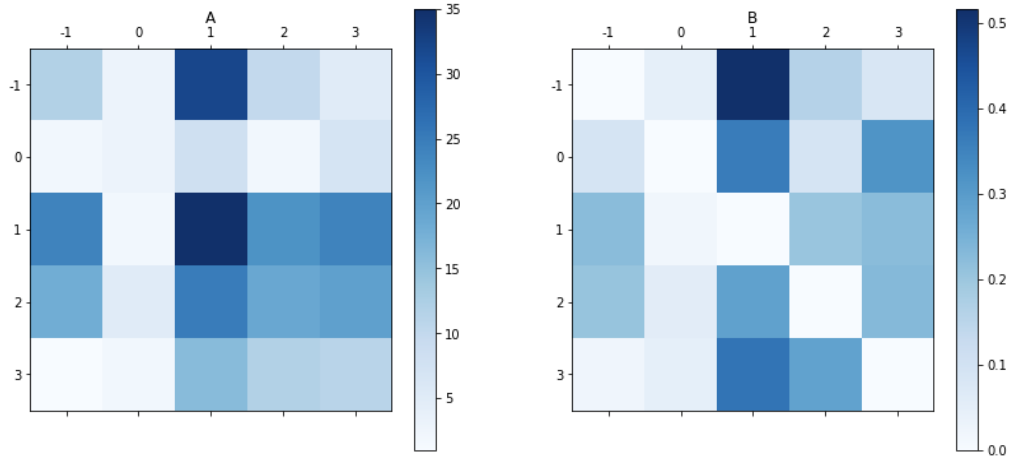


Fig. 5.8 Matricea de contingență pentru SGD: A e nenormalizată, B e normalizată

Din reprezentarea matricelor de contigentă ne dăm seama că starea 1 a afecțiunii este suprareprezentată, prin urmare, modelul tinde să facă *overfitting* către această stare. De asemenea, pentru că starea 0 este subreprezentată, modelul nu reușește să o identifice foarte bine. Date fiind rezultatele acestor modele, este evident că prin această metodă nu se poate crește foarte mult performanța de predicție. Desigur, acest demers a facilitat obținerea de informații suplimentare despre setul de date.

O altă metodă cunoscută pentru abilitățile de clasificare sunt arborii decizionali. Această metodă este dinamică și poate fi folosită pentru o serie de sarcini complexe de modelare. Având în vedere că vectorul cu valorile de intrare conține 55 de variabile, fiecare având 3 clase posibile, iar ieșirea sistemului are 5 clase, se așteaptă ca performanța arborilor decizionali să fie superioară SGD. Pentru această metodă, analiza datelor a fost realizată atât folosind *Anaconda*, cât și *Knime*. În continuare este prezentată, în paralel, metodologia de lucru și rezultatele obținute folosind ambele platforme. O parte din rezultatele acestei etape au fost prezentate de către autor în lucrarea [146].

Există o serie de beneficii pentru utilizarea arborilor decizionali în modelarea afecțiunilor complexe. Pentru început, prin această metodă se vor observa mult mai

bine relațiile dintre SNP-uri, spre deosebire de SGD, unde variabilele sunt considerate independente. În cazul arborilor decizionali, variabilele sunt poziționate ierarhic astfel creându-se relații de precondiție. Alt motiv pentru utilizarea arborilor decizionali este reprezentarea relativ simplă, intuitivă, a acestora în ceea ce privește prelucrarea datelor. Mai mult decât atât, modelul generat permite utilizatorului să identifice vizual relațiile, fiind un model de tipul *white-box*.

Pentru platforma *Knime*, fișierul de intrare, în format CSV, a fost integrat în fluxul de analiză, prezentat în Fig. 5.9, prin blocul *File Reader*. Deoarece valorile numerice din fișier erau discrete, fiecare reprezentând o stare a afecțiunii sau a SNP-ului, fără nicio semnificație continuă, au fost convertite în caractere prin blocul *Number to String*. Cele 55 de variabile reprezintă genotipul pacientului. Desigur, fiecare valoare din acest vector reprezintă o anumită clasă. După conversie, datele sunt transmise unei bucle care efectuează operația de antrenare, respectiv de măsurare a performanței printr-o repetare de 50 de ori a acestor operații. Următorul bloc din fluxul de analiză este partiționarea datelor. Acest bloc extrage două subseturi de date, prima reprezentând 80% din setul de date, iar al doilea set reprezentând restul de 20%. Primul subset, de 80%, este folosit pentru generarea arborelui decizional. După construirea arborelui (modelului), acesta este testat în blocul *Decision Tree Predictor* cu al doilea subset de date. În cele din urmă rezultatele sunt stocate într-un fișier CSV. Metodologia de implementare a algoritmului pentru generarea arborilor decizionali, în *Knime*, este descrisă în [147].

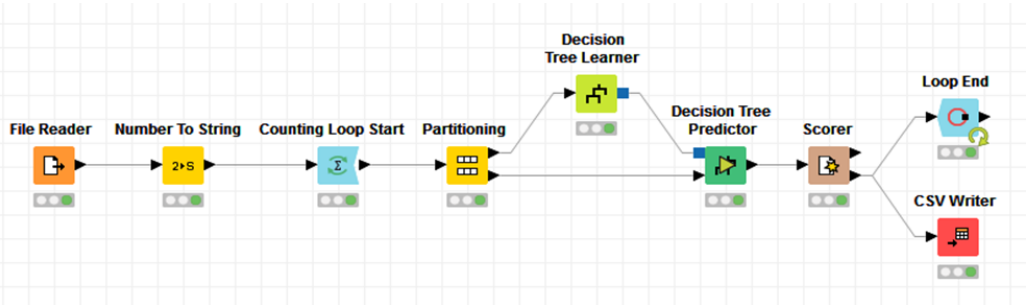


Fig. 5.9 Schema de conectare a blocurilor din *Knime* pentru utilizarea arborilor decizionali

Din literatură este cunoscut faptul că dacă se realizează mici schimbări în datele de antrenament, acestea pot duce la diferențe semnificative în structura arborelui, astfel încât performanța modelului rezultat se poate schimba semnificativ. Pentru a evita un subset de antrenare care generează un arbore la limitele maxime ale performanței, s-a adăugat bucla prin se asigură obținerea unei valori medii. De asemenea, la fiecare iterație se vor stoca valorile minime, maxime și medii pentru parametrii de interes. Acești parametri sunt acuratețea predicției, adică numărul valorilor care au fost corect clasificate din subsetul de testare precum și coeficientul Cohen (k), care măsoară acordul *inter-rate* pentru elementele arborelui.

Studiul s-a realizat în două etape. În prima etapă, fișierul de intrare a fost utilizat cu modificări minore, care au constat în umplerea spațiilor libere și transformarea în format CSV. În cea de-a doua etapă, stadiul final al afecțiunii a fost redus, având o stare binară, prezentă sau nu. În fiecare fază, au fost opt configurații posibile datorită următorilor trei parametri. Primul parametru al configurației a fost atributul de ramificare. Acest parametru are două opțiuni: câștigul informațional (*Gain Ratio*) și indexul *Gini*. Al doilea parametru al experimentului a fost metoda de reducere (*pruning*) cu două opțiuni disponibile: una fără *pruning* și cealaltă de tip *Minimum*

Description Length (MDL) descrisă de *Mehta* și colaboratorii în [78]. Al treilea parametru a fost eșantionarea, care avea de asemenea două opțiuni. Prima opțiune a fost prelevarea aleatorie, care extrage înregistrări aleatorii din setul de date, iar a doua opțiune a fost eșantionarea stratificată, care ar extrage înregistrări aleatorii bazate pe stadiul patologiei. Acuratețea a fost calculată ca procentaj al cazurilor clasificate corect (etape) din totalul cazurilor testate.

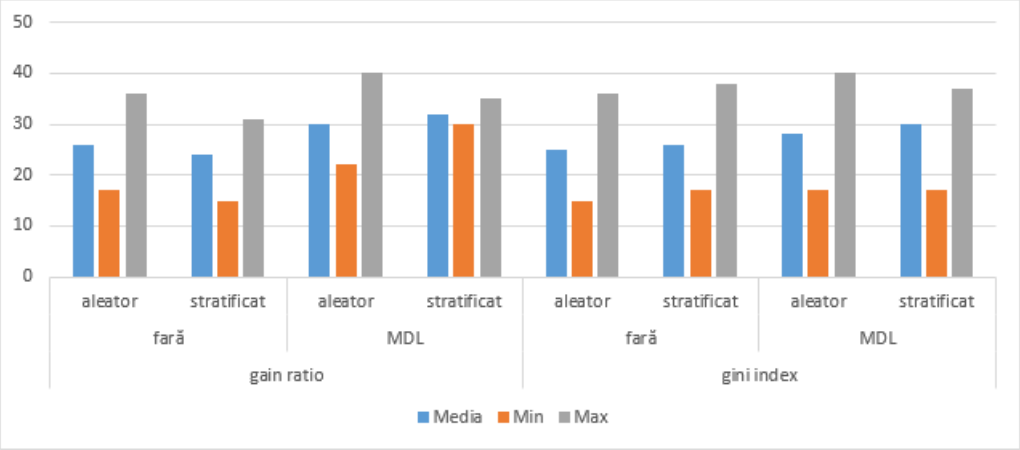


Fig. 5.10 Rezultatele predicției arborilor decizionali folosind *Knime* (Etapa I)

În prima fază, când stadiul afecțiunii avea cinci stări, rezultatele medii de predicție sunt sub 32%, prezentate în Fig. 5.10. Rata maximă de predicție, 40%, a fost realizată utilizând metoda de reducere MDL și prelevarea aleatorie, care ne poate duce cu gândul la un posibil *overfitting*. Rata minimă de predicție, 15%, a fost obținută neutilizând trunchierea arborelui. În această fază, coeficientul mediu al lui Cohen (k) a fost în jurul valorii zero (0) în toate cele opt configurații, prezentate în Tabelul 3.1. Valorile minime și maxime pentru k au fost -0,18, respectiv 0,18.

Tabelul 5.2. Rezultatele pentru prima etapă

Etapă	Atributul de ramificare	Metoda de reducere	Prelevarea	Acuratețea [%] (medie/min/max)	Coeficientul Cohen (k) (mediu/min/max)
Fișier fără modificări	Gain ratio	Fără	aleator	26 / 17 / 36	0.03 / (-0.12) / 0.18
	Gain ratio	Fără	stratificat	24 / 15 / 31	0 / (-0.11) / 0.10
	Gain ratio	MDL	aleator	30 / 22 / 40	0 / (-0.10) / 0.02
	Gain ratio	MDL	stratificat	32 / 30 / 35	0 / (-0.05) / 0.03
	Gini index	Fără	aleator	25 / 15 / 36	0.02 / (-0.10) / 0.14
	Gini index	Fără	stratificat	26 / 17 / 38	0.02 / (-0.10) / 0.18
	Gini index	MDL	aleator	28 / 17 / 40	0 / (-0.12) / 0.18

Folosind platforma *Anaconda*, am realizat același experiment. La verificarea încrucișată pentru trei subseturi am obținut performanțe de 35%, 29% respectiv 32%. Rezultatele se află în același interval precum cele obținute pe platforma *Knime*.

Desigur, dacă nu precizăm nicio restricție, atunci probabil metoda va genera un model care va memora toate înregistrările.

În cazul platformei *Knime*, parametrii puși la dispoziție pentru arborii decizionali, au fost baleiați pentru a crește performanța modelului. În cazul *Anaconda*, mai specific pachetul *scikit-learn*, acesta ne permite să mai ajustăm un set de variabile care nu erau disponibile în *Knime*. Spre exemplu: numărul minim de înregistrări care sunt necesare unei nod, numărul minim de înregistrări necesare pentru ca un nod să se dividă.

Este important să optimizăm acești parametri, altfel modelul este predispus spre *overfitting*. Arborii decizionali, spre deosebire de alte metode, nu presupun nimic despre datele care vor sta la baza modelului. Dacă un model este realizat folosind arborii decizionali, și este lăsat să genereze un model fără constrângeri, acesta va adapta structura arborelui (modelului) la datele care îi sunt puse la dispoziție. Lipsa ajustării duce la un model care este supra-potrivit.

În continuare, s-a efectuat evaluarea performanței modelelor generate în funcție de evoluția a trei parametri: adâncimea arborelui, numărul minim de înregistrări asociat unei frunze, numărul minim necesar pentru diviziunea unui nod. În cazul adâncimii arborelui, pentru setul de antrenare acuratețea modelului crește odată cu numărul de niveluri, dar scade pentru subsetul de testare, grafic reprezentat în Fig. 5.11 (*max_depth*). Acest lucru este explicabil, deoarece modelul memorează înregistrările. Pentru numărul minim de înregistrări pentru divizarea unui nod și numărul de înregistrări asociat unui nod, acuratețea pentru setul de antrenare scade odată cu creșterea valorii absolute. Deși efectul de memorare este redus prin această metodă, acuratețea pentru subsetul de testare are o creștere marginală.

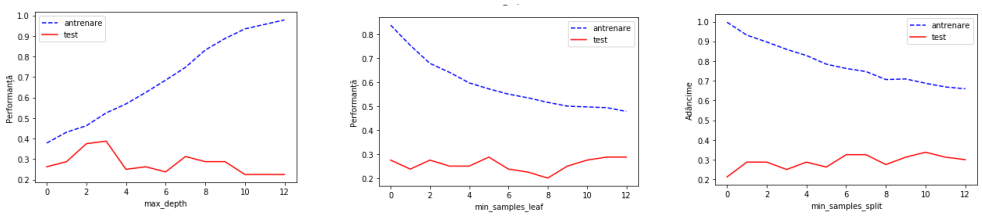


Fig. 5.11 Evoluția performanței modelului datorată modificării parametrilor specifici arborelui decizional

În schimbul prezentării unui arbore decizional (modelul), care ar avea dimensiuni mari, apelăm la extragerea parametrilor importanți din acesta. Folosind această metodă vom scoate în evidență setul de SNP-uri importante care stau la baza modelului. Importanța este reprezentată subunitar, valori între 0 și 1, unde 0 reprezintă faptul că SNP-ul nu contribuie la generarea modelului iar 1 reprezintă faptul că SNP-ul face distincția perfectă între stările afecțiunii. Desigur, valoarea 0 nu este absolut lipsită de semnificație pentru model. Această lipsă se poate datora și faptului ca două SNP-uri sunt puternic corelate, iar doar unul dintre acestea a fost selectat pentru generarea arborelui decizional, informația utilă fiind furnizată de primul SNP.

Pentru generarea reprezentării importanței SNP-urilor s-a folosit funcția dedicată din pachetul *SKlearn*. Reprezentarea din Fig. 5.12 indică faptul că informația utilă pentru generarea unui model este comprimată în doar 11 SNP-uri, din cele 55 disponibile. Cel mai relevant SNP este cel aflat pe gena SLC44A1 precedat de PEMT. Desigur, precum am menționat anterior, acest lucru nu înseamnă că cele 44 de

SNP-uri nu pot contribui pentru obținerea unui model cu performantă satisfăcătoare. În plus, acest model a fost generat folosind o serie de constrângeri, precum: adâncime este 4, numărul minim de înregistrări per frunză este 4 și numărul minim pentru diviziunea nodului este 4. Este evident că aceste valori nu reprezintă rețeta pentru un model care are un maxim de acuratețe.

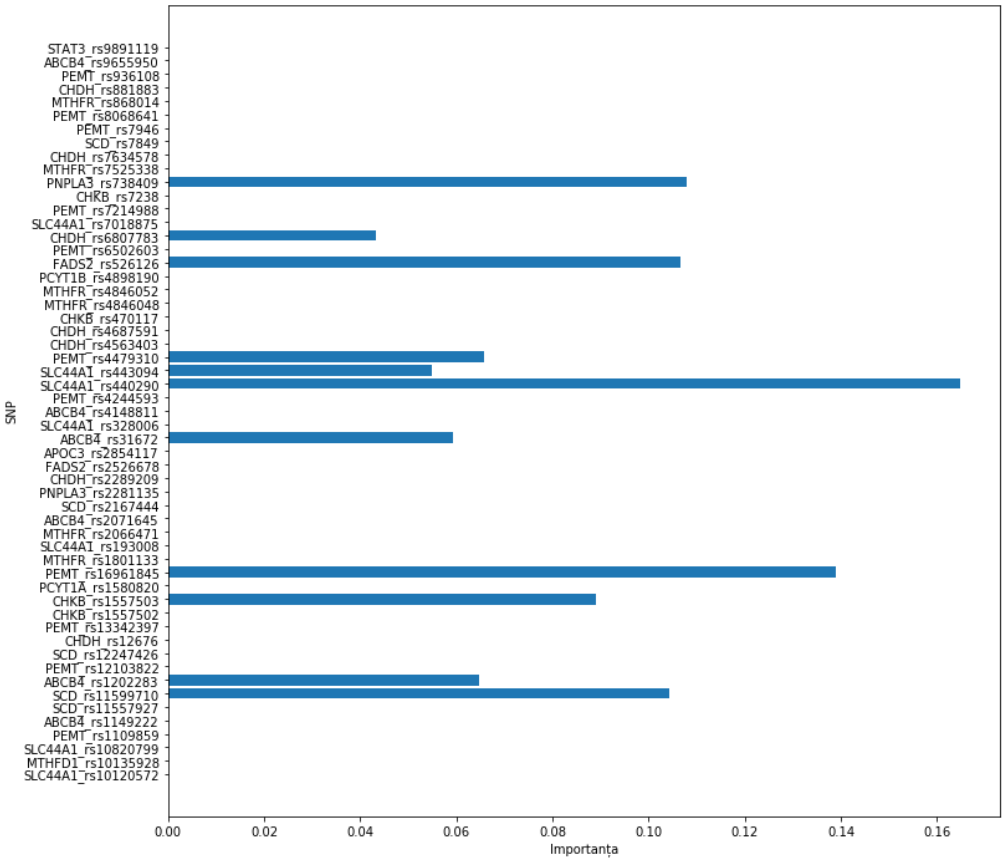


Fig. 5.12 Importanța *marker*-ilor genetici pentru generarea arborelui decizional

Pentru a vedea toate SNP-urile reprezentative va trebui să generăm o serie de modele și, desigur, să măsurăm nivelul mediu de importanță al fiecărui SNP. Astfel, s-au generat modele având adâncimea cuprinsă în interval 2–5, numărul minim de înregistrări per frunză cuprins în intervalul 2–6 și numărul minim pentru diviziunea nodului în intervalul 4–8. Au fost alese aceste intervale pentru că în interiorul acestora se găsesc zonele de maximum, dacă parametrii sunt analizați independent (Fig. 5.11). Procesul de generare a modelului și de stocare a vectorului de relevanță este iterat de 100 de ori. Valorile parametrilor au fost generate aleatoriu în respectivele intervale. Rezultatele prezentate în Fig. 5.13, relevă faptul că SNP-ul de pe gena PEMT are o relevanță mai ridicată față de SLC44A1. Desigur, noile rezultate arată că avem 19 SNP-uri care contribuie la generarea modelelor.

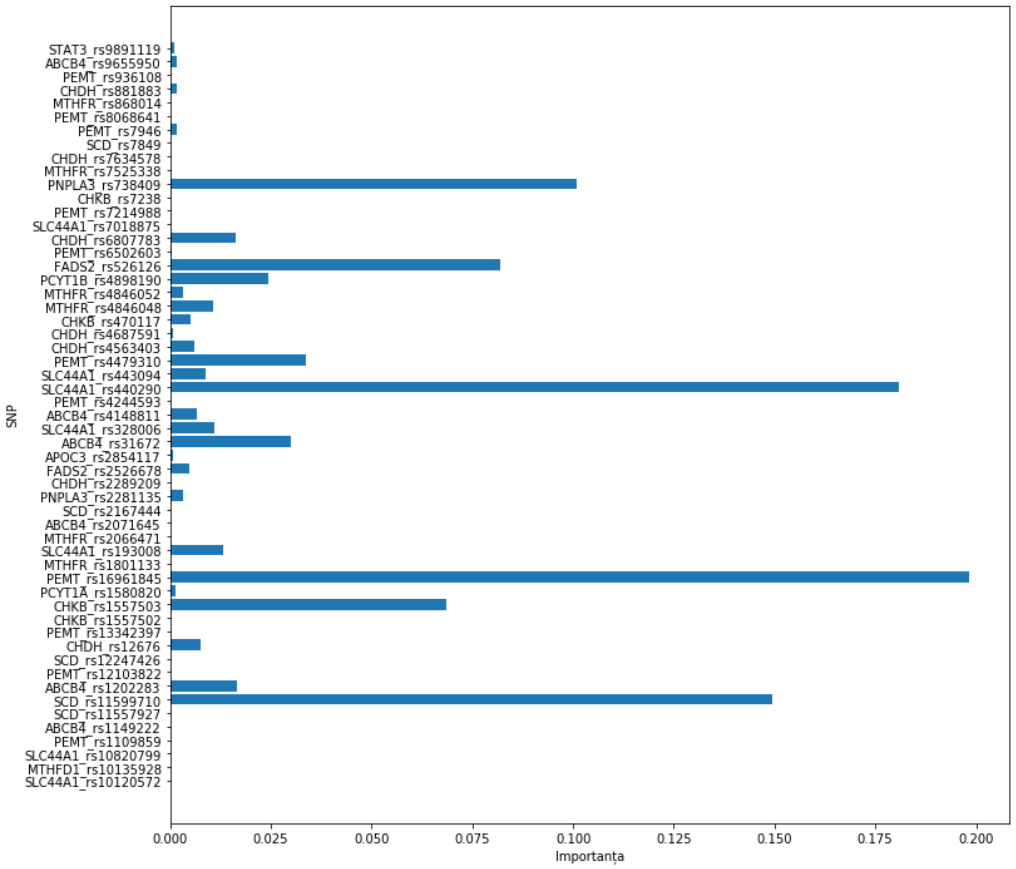


Fig. 5.13 Importanța medie a *marker*-ilor pentru arborele decizional

În subcapitolul 5.2, s-a analizat corelația dintre variantele genetice și stadiul steatozei. Ca urmare, s-a generat o listă de variante, care aveau o varietate scăzută, și s-a indicat că acestea reprezintă zgomot pentru realizarea unui model. Dacă revenim asupra acestora, observăm că următoarele SNP-uri sunt ignorate și în realizarea arborilor decizionali: SLC44A1_rs10120572, MTHFD1_rs10135928, SLC44A1_rs10820799, SCD_rs11557927, PEMT_rs12103822, SCD_rs12247426, PEMT_rs13342397, CHDH_rs2289209, SLC44A1_rs7018875, PEMT_rs7214988, MTHFR_rs7525338, CHDH_rs7634578, MTHFR_rs868014. Excepție de la această listă sunt: SCD_rs11599710, CHKB_rs1557503, PEMT_rs16961845, FADS2_rs2526678, PCYT1B_rs4898190. Dacă luăm în considerație și rezultatele din Fig. 5.13, lista de variante care reprezintă zgomot ar fi mult mai mare. Pe de altă parte, trebuie să avem în vedere și faptul că dacă două sau mai multe SNP-uri sunt puternic corelate, atunci doar unul dintre acestea va fi considerat pentru procesul de determinare al modelului.

Dacă reunim toate informațiile pe care le avem despre setul de date ne dăm seama că unele dintre aceste SNP-uri nu contribuie la obținerea unui model performant, acestea reprezentând doar zgomot. O metodă pentru scăderea nivelului de zgomot furnizat de înregistrări este reducerea dimensiunii datelor, detaliată în subcapitolul 5.3.3.

5.3.3. Reducerea dimensiunii parametrilor din înregistrările aferente steatozei

În general, pentru modelele care se bazează pe metode de învățare automată și care au înregistrări de intrare cu mii de caracteristici, este recomandat să se elimine unele dintre acestea sau să fie grupate pentru a fi mai bine corelate cu rezultatul vizat. Procesul de reducere nu implică eliminarea caracteristicii respective, din contră, aceasta este reunită cu altă caracteristică sau cu un grup și formează o caracteristică nouă. În cazul de față, SNP-urile pot fi grupate pentru a forma caracteristici noi. Evident că această abordare nu presupune automat că noile modele vor avea o performanță mai bună.

Pentru reducerea dimensiunii folosim metoda Analizei Componentelor Principale (*Principal Component Analysis*, PCA). În Fig. 5.14 este redată evoluția calității informației în funcție de numărul dimensiunilor asociate fiecărei înregistrări. Din reprezentare se observă că pentru a menține o varietate a informației de peste 0,9, numărul minim de dimensiuni trebuie să fie 24.

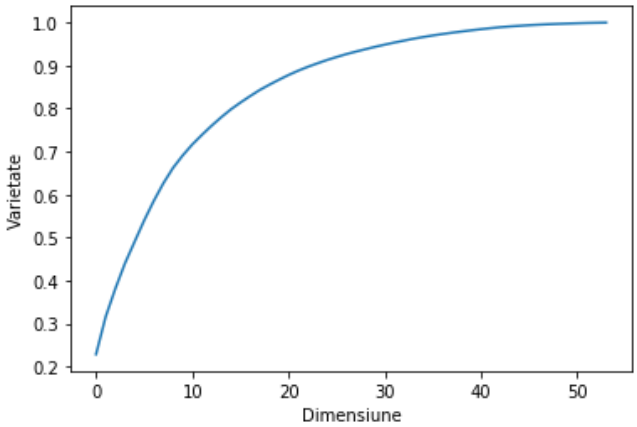


Fig. 5.14 Varietatea informației în raport cu dimensiunile înregistrărilor folosind metoda PCA

Având în vedere faptul că, folosind 24 de caracteristici compuse, am stocat peste 90% din informație, se pot genera noi modele de arbori decizionali și ulterior se poate evalua performanța acestora. După generarea noilor modele, odată cu scăderea numărului de parametri, a scăzut marginal și nivelul de acuratețe al modelelor. Prin urmare, în acest caz, reducerea numărului de caracteristici nu aduce îmbunătățiri, cu excepția faptului că generarea modelelor este mai rapidă.

5.3.4. Predicția gradului steatozei folosind metode *Ensemble*

Metodele de tip *Ensemble* permit combinarea mai multor modele pentru a crea o metodă mai eficientă de predicție. Cele mai cunoscute metode *Ensemble* sunt *Random Forest* și *Gradient Boosted Decisional Trees* (GBDT). Precum am văzut în subcapitolele anterioare, arborii decizionali tind să memoreze înregistrările. Pentru a remedia acest defect, se pot genera o serie de arbori decizionali care vor lua decizii

împreună. Bineînțeles, toți arborii vor fi antrenați diferit astfel încât să nu existe arbori identici. Crearea unor arbori diferiți nu este o sarcină foarte grea, deoarece prin mici modificări ale datelor de antrenare, metoda va schimba structura suficient de mult.

Random Forest are o serie de parametri care pot fi ajustați în așa fel încât să ofere performanța cea mai ridicată. Pe lângă parametrii specifici unui arbore decizional (adâncime, numărul minim de înregistrări per frunză, numărul minim pentru diviziunea nodului, ș.a.m.d.) *Random Forest* are și câțiva parametri specifici precum: numărul de arbori, *bootstrap* și scorul *out-of-bag*. Pentru a genera configurația ideală, probabil ar trebui să se genereze toate combinațiile posibile pentru toți parametrii.

Totuși, înainte rulării unei asemenea analize, care ar presupune o mulțime de resurse, vom face o verificare liniară (valori consecutive) pentru o serie de parametri. La fel ca în cazul arborilor decizionali vom alege parametrii *max_depth, min_samples_leaf*, *min_samples_split,* la care se mai adăugă numărul de arbori din metodă, numărul maxim de caracteristici considerate pentru obținerea arborilor și aplicarea metodei de extragere a înregistrărilor, *bootstrap*.

Configurația parametrilor pentru căutarea valorii optime a unuia dintre parametri este următoarea: adâncime maximă 6, numărul minim de înregistrări per frunză 4, numărul minim pentru diviziunea unui nod 4, numărul maxim de caracteristici 20, *bootstrap* activ, numărul de arbori 30. Pentru a verifica performanța modelelor s-a folosit validarea încrucișată (*3-fold*) și scorul de predicție pentru setul de test.

Pentru numărul de arbori s-a făcut o căutare între 1 si 200, reprezentat în Fig. 5.15 A. În acest caz se poate observa că acuratețea predicției validării încrucișate, media a trei subseturi din setul de antrenare, este aproape de rezultatele predicției pentru setul de test. În intervalul 0–50, setul de testare are o valoare ușor mai scăzută față de validarea încrucișată, ceea ce indică o supra-potrivire a modelelor. În schimb, în intervalul 125–200 observăm că rezultatele pentru setul de test sunt deasupra celor aferente validării încrucișate, iar acest lucru indică o sub-potrivire (*underfitting*). Intervalul ideal, unde atât rezultatele pentru setul de test cât și pentru cel de validare încrucișată sunt suprapuse, este 50–125. Desigur, este vorba despre diferențe minore. Pentru adâncimea arborilor, s-a ales intervalul de căutare între 2 și 50. Intervalul potrivit pentru acest parametru pare a fi între 2 și 10, Fig. 5.15 B. În restul intervalului modelul tinde să se supra-potrivească. Pentru limitarea numărului de caracteristici pe care un arbore le poate lua în considerație, s-a ales intervalul 0–50. În acest caz, rezultatele celor două subseturi sunt intercalate, indicând o generalizare bună a modelului. Pentru testarea numărului minim de înregistrări per frunză, Fig. 5.15 D, numărul minim pentru divizarea unui nod, Fig. 5.15 E, și pentru a verifica efectul de *bootstrapping*, Fig. 5.15 F, s-a selectat un interval între 2 și 10. Numărul minim per frunză este influențat și de setul de antrenare. În acest caz, dacă se alege un interval mai mare de 10, apare efectul de supra-potrivire. Odată cu creșterea numărului minim necesar pentru divizarea nodului, performanța scade.

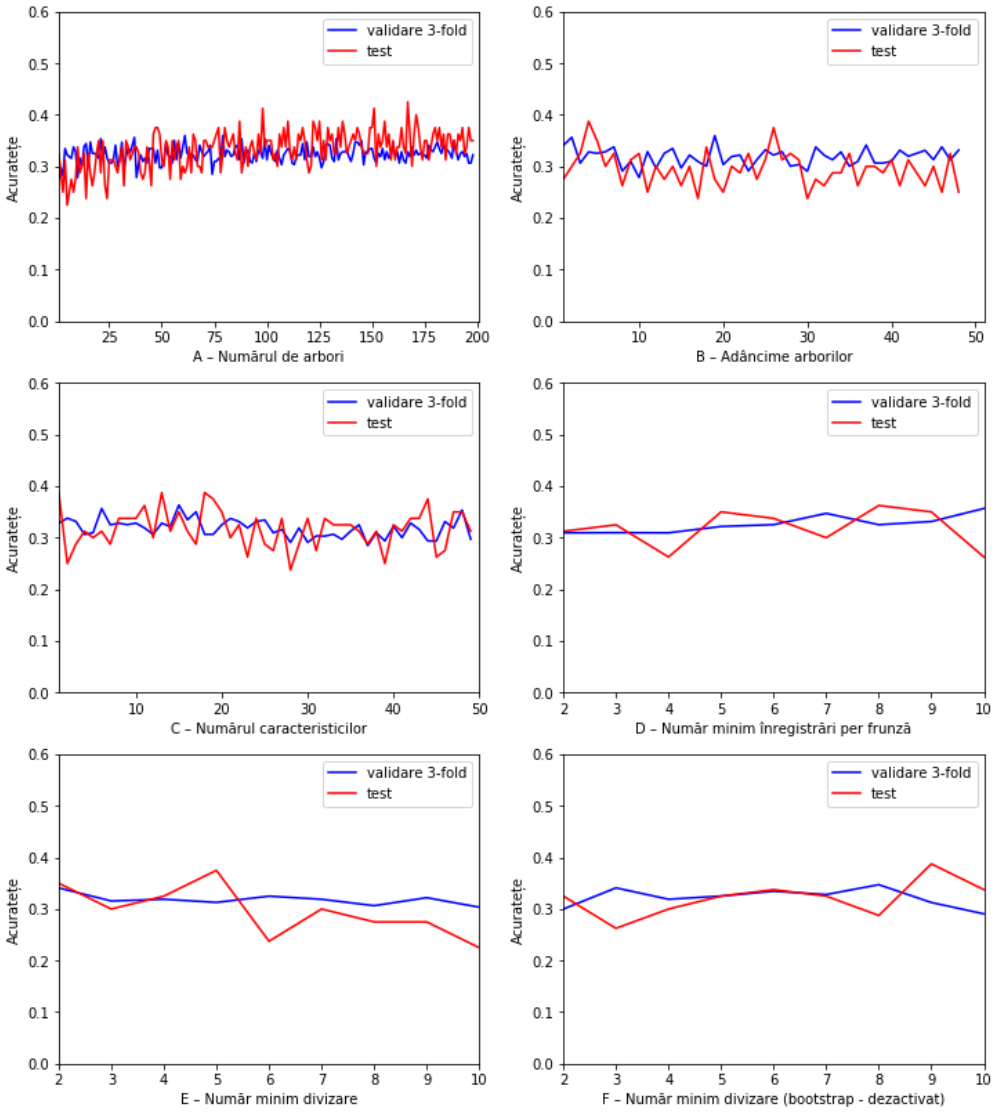


Fig. 5.15 Acuratețea modelelor *Random Forest* în funcție de evoluția parametrilor interni

Dacă dorim o abordare mai complexă ar trebui să folosim o metodă de optimizare a parametrilor pentru căutarea unui maxim, în caz ideal, un maxim global. O metodă pentru realizarea acestui deziderat este căutarea aleatorie în spațiul parametrilor sau căutarea integrală pentru intervalele folosite. Aceste metode necesită mai multe resurse. Dacă luăm intervalele sugerate anterior, avem aproximativ 100.000 de cazuri de evaluat. Dacă folosim metoda *GridSearchCV,* obținem un model cu performanța medie de 33% pentru validarea încrucișată, iar pentru setul de test, 36%. Configurația obținută are adâncimea maximă 6, numărul minim de înregistrări pe frunză 3, divizarea nodului 2 și are 108 arbori decizionali.

Rezultatul nu reprezintă un maximum global, ci unul local. Probabil, maximumul global este cu câteva procente peste această performanță. Dacă se modifică datele de antrenare sau subseturile din validarea încrucișată, rezultatele, cel mai probabil, se vor schimba. Prin urmare și performanța se va schimba.

Dacă analizăm SNP-urile considerate relevante, Fig. 5.16, observăm că importanța SNP-urilor este mult mai dispersată comparativ cu situația când se folosea un singur arbore decizional. SNP-ul PEMT_rs16961854 este al doilea ca importanță în timp ce PNPLA3_rs738409 este cel mai relevant, invers comparativ cu Fig. 5.13. Acest rezultat nu este absolut, el se poate schimba odată cu datele de antrenare a modelului.

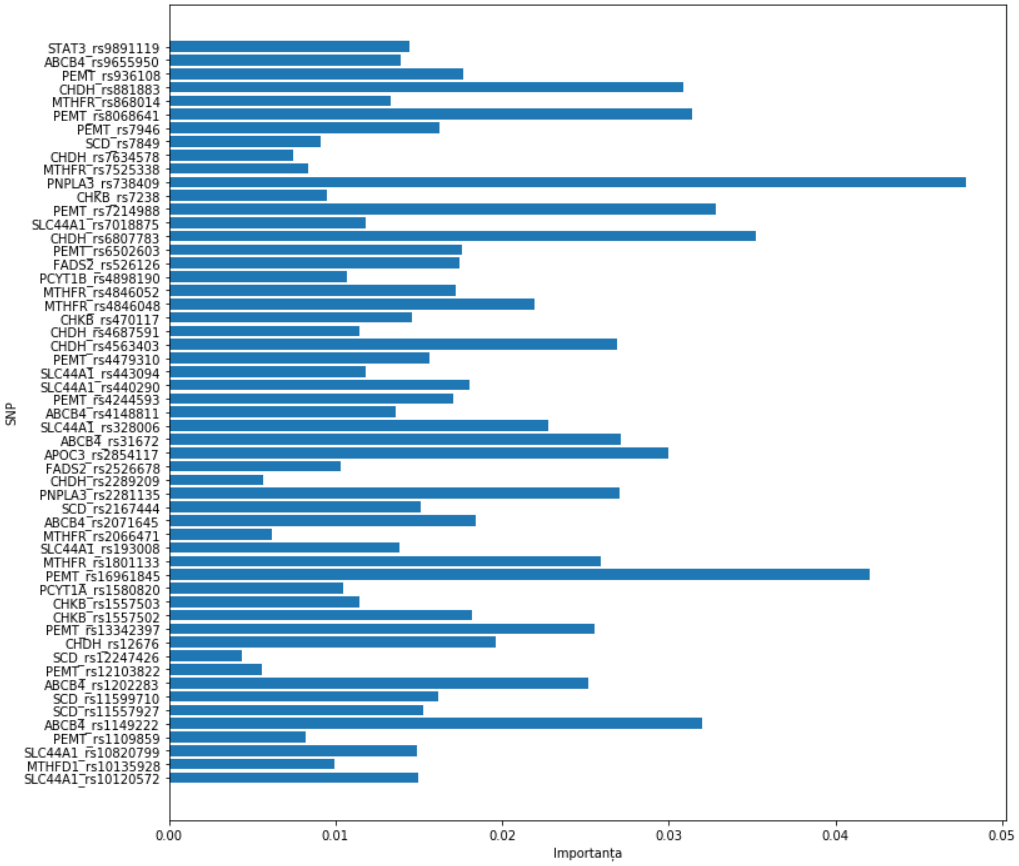


Fig. 5.16 Importanța medie a SNP-urilor, folosind metoda *Random Forest*

5.4. Determinarea prezenței steatozei prin reducerea dimensiunii stadiului afecțiunii

Performanța obținută până în acest punct nu este extraordinară. Desigur, dacă luăm în considerație faptul că afecțiunea are cinci stări posibile, o funcție aleatorie ar avea, statistic și cu o bază de date suficient de mare, o acuratețe de aproximativ 20%. Din această perspectivă, modelele obținute, cu performanțe mai mari de 35%, pot fi considerate o ușoară îmbunătățire, dar departe de a fi utilizabilă în practica clinică.

În acest subcapitol, se va reduce problema de determinare a stadiului steatozei prin reducerea stadiului la o formă binară. Stadiul 0, 1, 2, 3 au devenit 1 sau *adevărat*, iar -1, adică lipsa steatozei, a devenit 0 sau *fals*. Variantele genetice care definesc genotipul au rămas în continuare cu trei stări posibile:

- Starea 0, variantă comună în populație, în formă homozigotă, *wild type*;
- Starea 1, una dintre alelele e *wild type*, iar cealaltă este diferită, heterozigot;
- Starea 2, ambele alele au nucleotida diferită de cea *wild type*.

5.4.1. Rezultatele determinării unui model de predicție folosind arborii decizionali

După modificările prezentate anterior, am aplicat aceleași opt configurații posibile care au fost testate și în subcapitolul 5.3.2, secțiunea arbori binari. Acuratețea maximă în acest caz este semnificativ mai bună, 91% (Fig. 5.17). Această performanță a fost obținută utilizând metoda de reducere MDL, selecție aleatorie și coeficientul *Gini*. Putem presupune că scorul se datorează și prelevării aleatorii care a permis o oarecare supra-potrivire. Performanța minimă de 53% a fost obținută utilizând indexul *Gini*, fără fasonare și prelevare aleatorii, cel mai probabil sub-potrivire. Coeficientul Cohen, mediu, a fost de 0,07. Valorile minime și maxime pentru *k* au fost -0,23, respectiv 0,37. Rezultatele complete sunt prezentate în Tabelul 5.3.

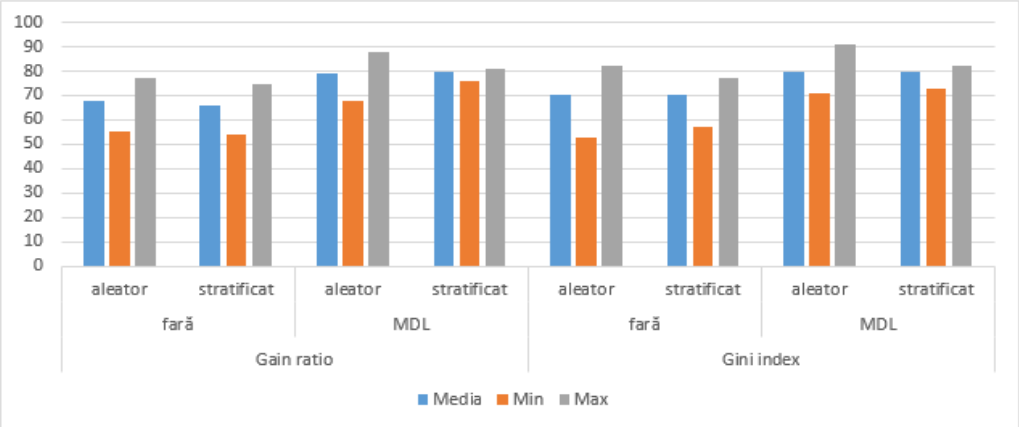


Fig. 5.17 Rezultatele predicției arborilor decizionali folosind *Knime* (etapa a doua)

Tabelul 5.3. Rezultatele performanței arborilor binari folosind platforma *Knime* (etapa a doua)

Etapa	Atributul de ramificare	Metoda de reducere	Prelevarea	Acuratețea [%] (medie/min/max)	Coeficientul Cohen (*k*) (mediu/min/max)
Fișier cu starea binară a afecțiunii	*Gain ratio*	fără	Aleatoriu	68 / 55 / 77	0 / (-0,23) / 0,21
	Gain ratio	fără	Stratificat	66 / 54 / 75	(-0,01) / (-0,22) / 0,27
	Gain ratio	MDL	Aleatoriu	79 / 68 / 88	(-0,01) / (-0,09) / 0,12
	Gain ratio	MDL	Stratificat	80 / 76 / 81	(-0,01) / (-0,08) / 0,02
	Gini index	fără	Aleatoriu	70 / 53 / 82	0,07 / (-0,17) / 0,37
	Gini index	fără	Stratificat	70 / 57 / 77	0,06 / (-0,20) / 0,29
	Gini index	MDL	Aleatoriu	80 / 71 / 91	0,04 / (-0,10) / 0,32
	Gini index	MDL	Stratificat	80 / 73 / 82	0,04 / (-0,12) / 0,20

Dacă analizăm importanța SNP-urilor pentru determinarea stadiului afecțiunii prin compararea rezultatelor din Fig. 5.12 și Fig. 5.18, observăm că există o serie de asemănări. Varianta PNPLA3_rs738409 se află în lista *marker*-ilor genetici importanți în ambele cazuri. SLC44A1_rs328006 și SCD_rs2167444, în cazul stadiului binar, capătă un rol mult mai important în timp ce în cazul stărilor multiple acestea sunt lipsite de importanță. SLC44A1_rs440290, PEMT_rs13342397 și SCD_rs11599710, care în cazul stărilor multiple reprezentau cei mai importanți *marker*-i genetici, în această configurație își pierd importanța. Desigur, dacă se ia în considerație și matricea de corelații din Fig. 5.3, observăm că variantele SLC44A1_rs328006 și SLC44A1_rs440290 sunt puternic corelate, având o valoare de corelație de aproximativ 0,8. Același lucru se poate spune și despre variantele SCD_rs2167444 și SCD_rs11599710.

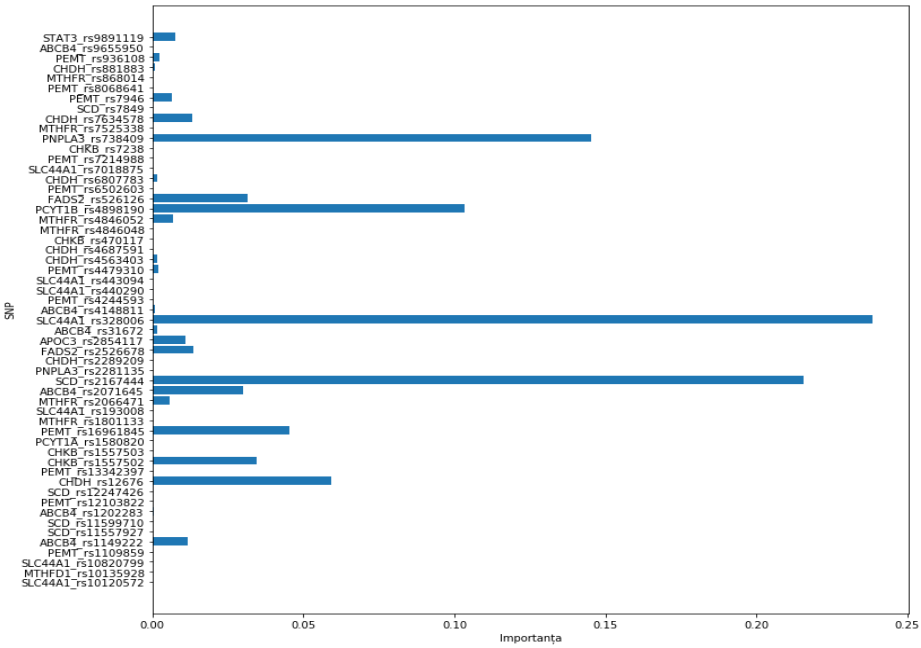


Fig. 5.18 Importanța medie a SNP-urilor pentru generarea arborilor binari (stadiu binar)

5.4.2. Rezultatele determinării unui model de predicție folosind metoda *Random Forest*

În cazul al doilea s-a folosit *Knime Studio* pentru evaluarea performanței metodei *Random Forest*. Schema de conectare a blocurilor este prezentată în Fig.5.19. Datele sunt preluate prin blocul *File Reader* după care sunt transformate în clase (pentru a evita interpretarea ca valori continue) și vor fi trimise către bucla de prelucrare. În buclă, setul de date este împărțit în două subseturi, primul conținând 80% dintre înregistrări, iar cel de-al doilea conținând 20%. Bucla este iterată de 100 de ori pentru a afla media, minimumul și maximumul acurateței.

Dacă analizăm rezultatele din Tabelul 5.4, se observă că media acurateței este undeva în jurul valorii de 80%. Dacă comparăm performanța unui arbore decizional, folosind metoda MDL, cu un set de arbori, generați cu *Random Forest*, observăm că aceasta crește marginal. În ceea ce privește modelele generate nu pare a fi o diferență semnificativă în ceea ce privește atributul de ramificare. O ușoară îmbunătățire apare dacă folosim coeficientul *Gini*. Dacă ne uităm la numărul de niveluri permise, din nou nu apar diferențe semnificative. Diferențele apar la tipul de eșantionare folosit. Dacă se folosește eșantionarea aleatorie reușim să obținem modele cu o acuratețe mai bună. Se poate presupune că este vorba despre *overfitting*, dar în acest caz eșantionarea nu joacă un rol atât de semnificativ, deoarece sunt o mulțime de serii de arbori. În schimb când ne uităm la valorile medii observăm că acestea sunt asemănătoare cu eșantionarea stratificată. O altă diferență apare la valoarea minimă, care este mai mică în cazul eșantionării aleatorii.

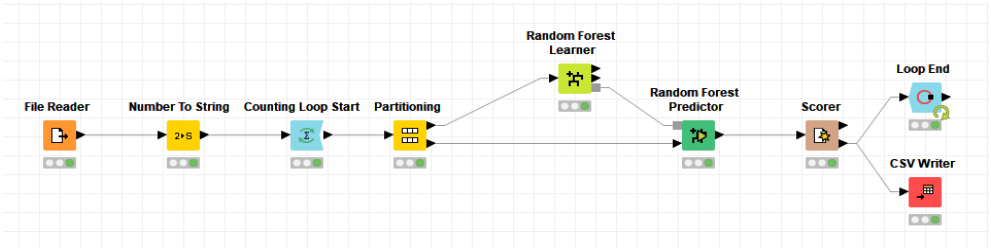


Fig.5.19. Schema bloc a sistemului de predicție folosind metoda *Random Forest* din *Knime Analytics Platforms*

Tabelul 5.4. Rezultatele metodei *Random Forest* pentru platforma *Knime*

Etapa	Atributul de ramificare	Numărul maxim de niveluri	Eșantionare	Acuratețea [%] (medie/min/ max/stdev)	Coeficientul Cohen (k) / 10^2 (mediu/min/max/ stddev)
Steatoză valori binare	*Information gain*	Automat	Stratificat	81 / 75 / 85 / 2	6 / (-10) / 28 / 8
	Information gain ratio	Automat	Stratificat	81 / 76 / 85 / 2	8 / (-8) / 33 / 10
	Gini index	Automat	Stratificat	81 / 76 / 83 / 2	7 / (-8) / 25 / 8
	Information gain	Automat	Aleatoriu	81 / 72 / 91 / 4	6 / (-4) / 27 / 8
	Information gain ratio	Automat	Aleatoriu	80 / 70 / 87 / 4	7 / (-6) / 30 / 8

	Gini index	Automat	Aleatoriu	81 / 72 / 88 / 4	6 / (-6) / 31 / 8
	Gini index	3	Aleatoriu	81 / 72 / 91 / 4	1 / (-4) / 17 / 4
	Gini index	5	Aleatoriu	81 / 68 / 91 / 4	4 / (-6) / 27 / 6
	Gini index	50	Aleatoriu	81 / 66 / 90 / 4	6 / (-8) / 30 / 9

Dacă repetăm analiza folosind *sklearn* obținem același rezultat, acuratețe 81%. La fel ca în capitolul 5.3.4, s-a făcut analiza parametrilor aferenți metodei *Random Forest*, prezentată în Fig. 5.20. În cazul numărului de arbori din model, diferența apare doar când numărul acestora este mai mic decât cinci, acuratețea scăzând marginal la 75%. În rest performanța este constantă. De asemenea, acuratețea modelelor este constantă chiar dacă schimbăm adâncimea arborilor, numărul minim de înregistrări pentru divizarea unui nod, numărul de caracteristici, ș.a.m.d. Acest lucru se datorează multiplilor arbori din model care reușesc să ofere un echilibru al performanței.

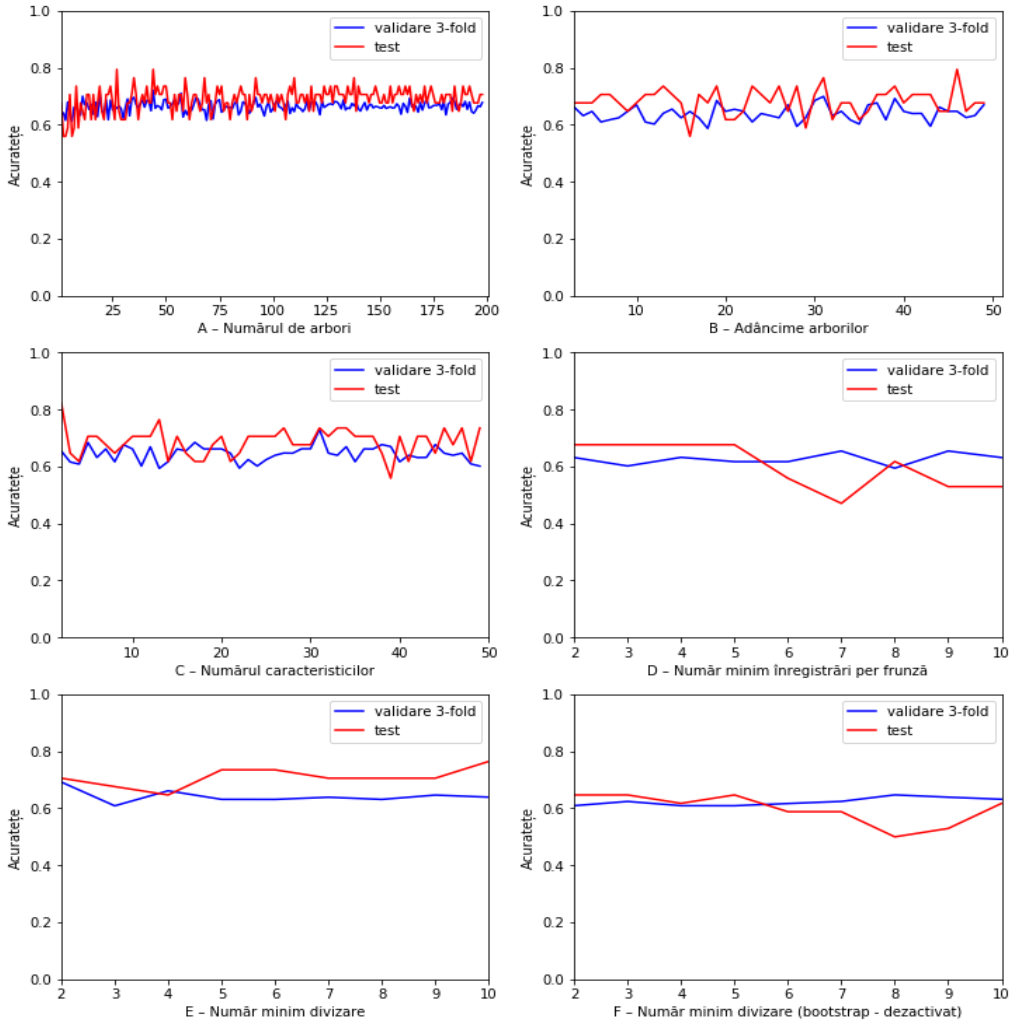


Fig. 5.20. Analiza acurateței metodei *Random Forest* prin modificarea parametrilor interni

Dacă facem o comparație între importanța SNP-urilor din Fig. 5.16 (când afecțiunea are stări multiple) și din Fig. 5.18 (când afecțiunea este binară) observăm că există diferențe în ceea ce privește lista celor mai importante variante. În cazul afecțiunii multiple cel mai relevant SNP a fost PNPLA3_rs738409, în timp ce în starea binară acesta este al doilea ca importanță. Cea mai relevantă variantă genetică în cazul stării binare este SLC44A1_rs328006, Fig. 5.21. Desigur, diferențele sunt minore. Rezultatele complete pentru metoda *Random Forest* se găsesc în Anexa 2, Tabelul A2.1, respectiv fig. A2.1 și fig. A2.2.

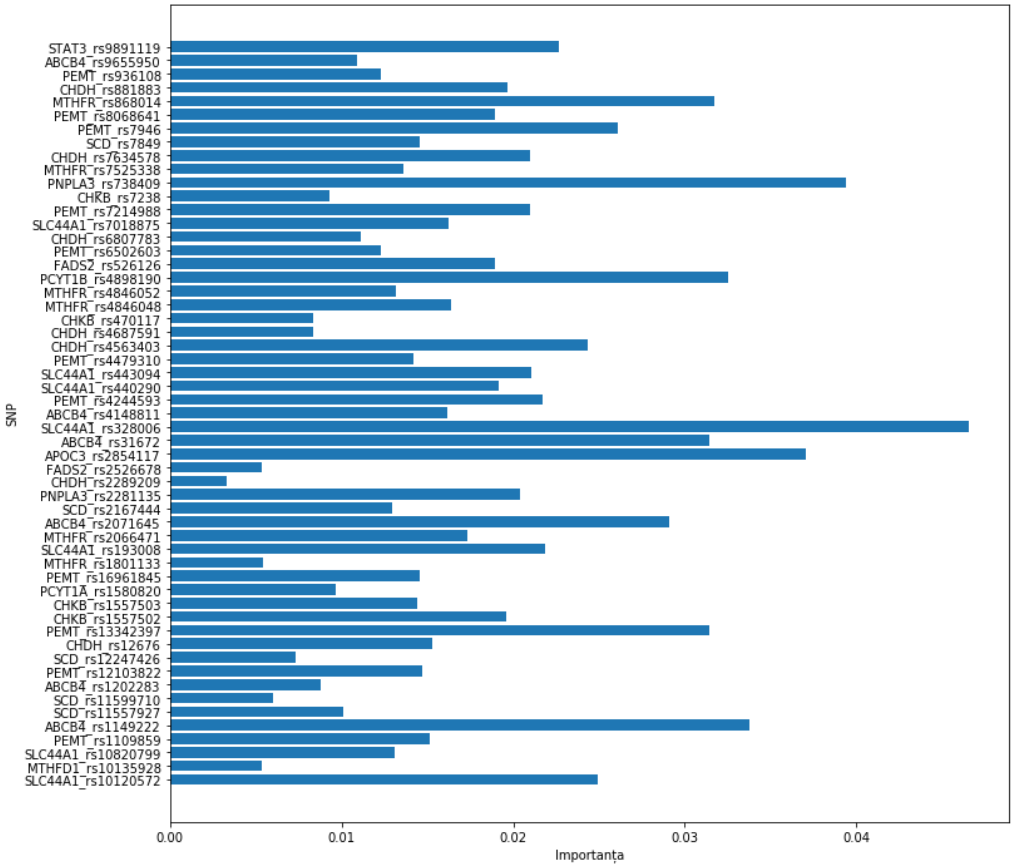


Fig. 5.21 Importanța *marker*-ilor genetici folosind metoda *Random Forest* (starea afecțiunii este binară)

5.5. Sistem de votare bazat pe diferențe dintre grupurile SNP și stadiul afecțiunii

5.5.1. Prezentarea metodei

În continuare este prezentat un alt model de predicție care are la bază o metodă de votare, asemănătoare cu o metodă *Ensemble*. Această metodă ia în considerație diferențele de frecvență dintre grupurile SNP și stadiu. Se poate spune că este o derivată a definiției date de Géron, subcapitolul 2.1.10, doar că în locul a 1.000 de persoane aleatorii se vor alege 1.000 de experți. Pașii metodei sunt descriși în Fig. 5.22.

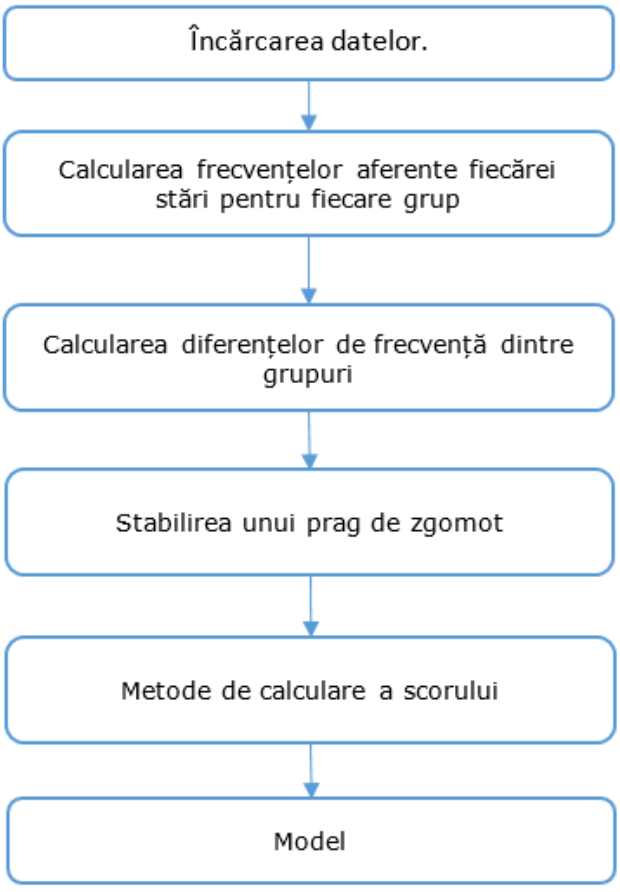


Fig. 5.22. Pașii pentru obținerea modelului de predicție

Procesul de creare a modelului începe prin încărcarea datelor și calcularea procentajului (p), care se va stoca în matricea de procentaje (P), pentru fiecare clasă (c) asociată fiecărui polimorfism (SNP), pentru fiecare stadiu al afecțiunii (s). Pentru

calcularea valorii procentajului se utilizează ecuația (5.5-1), unde *N* reprezintă o funcție de numărare condiționată de variabilele funcției.

$$P(SNP, s, c) = \frac{N(SNP, s, c)}{N(SNP, s)} \tag{5.5-1}$$

După efectuarea operației vom obține o matrice tridimensională (*P*) care va conține valorile pentru toate combinațiile dintre cei trei parametri. Această matrice va reprezenta „baza de cunoștințe" pentru generarea agenților de votare (experții). Agenții de votare se vor forma în funcție de înregistrarea care urmează a fi prezisă, folosind funcția *dif* descrisă în (5.5-2). Funcția primește ca parametri de intrare SNP-ul și clasa care este asociată acestuia. Prin parametrul *c′* se reprezintă celelalte clase asociate SNP-ului. În cazul bazei de date asociate steatozei, în urma calculării agenților de votare vom obține o matrice bidimensională.

$$dif(SNP, s, cc') = P(SNP, s, c) - P(SNP, s, c') \tag{5.5-2}$$

Desigur, performanța modelului generat este la fel de bună precum calitatea și cantitatea datelor care stau la baza acestuia. Dacă numărul înregistrărilor nu este suficient de mare, de exemplu 200 de înregistrări per stadiu-clasă, cu certitudine modelul, matricea tridimensională, va avea incorporat un oarecare nivel de zgomot, de impuritate. Acest zgomot apare datorită rezoluției statistice scăzute care rezultă din lipsa numărului de înregistrări. Pentru a rezolva această problemă, se poate introduce un prag de zgomot (ε) care va acționa ca o limită pentru funcția *dif*. Nu există un prag predefinit, acesta trebuie determinat printr-o metodă de optimizare sau o căutare exhaustivă în spațiul valorilor, care este cuprins între 0 și 1. Prin urmare, determinarea valorii funcției *dif* se realizează cu o funcție decizională, precum este redat în ecuația (5.5-3).

$$dif(SNP, s, cc') = \begin{cases} P(SNP, s, c) - P(SNP, s, c') \mid P(SNP, s, c) - P(SNP, s, c') \geq \varepsilon \\ 0 \mid P(SNP, s, c) - P(SNP, s, c') < \varepsilon \end{cases} \tag{5.5-3}$$

După introducerea nivelului de zgomot și calcularea matricei aferente înregistrării, putem determina care este stadiul afecțiunii aferent genotipului nostru. Există mai multe moduri de calculare (predicție) a stadiului. Putem aborda o strategie de vot cu acumulare sau una cu vot individual. Pentru strategia de vot cu acumulare, pentru fiecare stadiu al afecțiunii se vor însuma ponderile aferente fiecărui SNP, precum este prezentat în ecuația (5.5-4). La final, fiecare stadiu al afecțiunii va avea asociată o valoare reală. Stadiul cu valoarea cea mai mare este asociat genotipului.

$$acu(c) = \sum_{i=1}^{n} dif(SNP_i, s, cc') \tag{5.5-4}$$

Pentru utilizarea strategiei de vot individual, vom apela la funcția de transformare prezentată în (5.5-5). Această strategie presupune ca fiecărui agent,

indiferent de ponderea lui, să i se atribuie un vot pozitiv dacă ponderea este strict pozitivă, un vot negativ dacă ponderea este strict negativă, iar dacă ponderea este zero, pentru starea calculată, agentul este considerat ca fiind zgomot, acesta neprimind dreptul de vot.

$$vot\ (SNP, s, cc') = \begin{cases} 1, dacă\ dif(SNP, s, cc') > 0 \\ 0, dacă\ dif(SNP, s, cc') = 0 \\ -1, dacă\ dif(SNP, s, cc') < 0 \end{cases} \quad (5.5\text{-}5)$$

Pentru aplicarea strategiei cu vot individual vom utiliza ecuația (5.5-6). La fel ca în cazul strategiei cu acumulare, la finalul operațiilor vom avea un vector în care fiecare stadiu al afecțiunii are un număr de voturi. Stadiul cu cele mai multe voturi este asociat genotipului.

$$ind(c) = \sum_{i=1}^{n} vot(SNP_i, s, cc') \quad (5.5\text{-}6)$$

Desigur, se pot implementa și alte derivate ale acestei metode, precum calcularea ponderilor folosind o scară logaritmică, sau implementarea unei strategii hibride de vot ș.a.m.d. În cazul scării logaritmice, dacă o variantă genetică are o valoare suficient de mare, teoretic aceasta ar trebui să apară ca fiind corelată și dacă se folosesc metode statistice specifice (*Pearson*).

Dacă facem o optimizare a valorii pragului de zgomot observăm că strategia prin vot individual poate obține valori mai mari ale acurateței. În Fig. 5.23 prezentăm rezultatele aferente bazei de date când stadiul steatozei are cinci stări. Strategia votului individual reușește să obțină o performanță mai bună decât strategia prin acumulare. Se poate observa că, după ce pragul atinge valoarea de 0,35, modelul începe să indice o singură clasă, prin urmare acesta a intrat în zone de supra-potrivire. Pentru cazul când stadiile steatozei sunt în formă binară, Fig. 5.24, acuratețea maximă este atinsă când pragul este în apropierea valorii de 0,16. După ce se depășește această valoare, modelul intră în zona de *overfitting*.

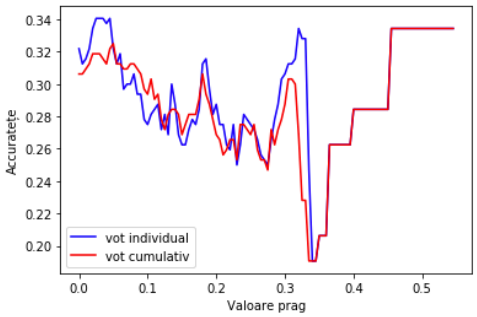


Fig. 5.23 Performanța sistemului de vot folosind cele două strategii pentru fișierul cu cinci stadii ale steatozei

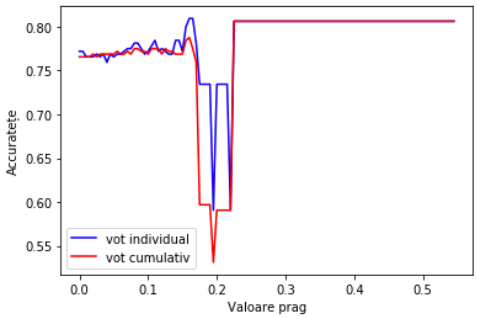


Fig. 5.24 Performanța sistemului de vot folosind cele două strategii pentru fișierul cu stadii binare

5.5.2. Interpretarea rezultatelor

Folosind această metodă putem extrage importanța fiecărui agent de votare în funcție de stadiul afecțiunii. În figurile: Fig. 5.25, Fig. 5.26 și Fig. 5.27 avem „baza de cunoștințe" generată de metodă pentru fișierul cu stadii multiple ale steatozei. Agenții redați în imagini au fost filtrați folosind un prag $\varepsilon = 0{,}16$.

În urma generării modelului sau a bazei de cunoștințe, se pot emite câteva observații. SNP-urile din gena ABCB4, în starea 1 (heterozigot), par să indice o afinitate către stadiul 2 al steatozei în defavoarea stadiilor 0 și 3. Un alt SNP care iese în evidență este FADS2_rs2526678 care în stare homozigotă *wild-type* indică afinitate pentru toate stadiile în care apare implicat stadiul 0. Dacă extrapolăm această informație, SNP-ul, de fapt în stare homozigotă (clasa 0), indică faptul că stadiul steatozei nu este 0. Dacă evaluăm starea heterozigotă (clasa 1) vedem că pentru toate stările unde este implicat stadiul 0 se indică o aversiune față de aceste stadii. Din nou extrapolând rs2526678 heterozigot pentru acest SNP ne indică afinitate către stadiul 0. Imediat, următoarea întrebare este ce se întâmplă pentru clasa 2, adică homozigot (AA, Tabelul 5.1). Din punct de vedere genetic, afinitatea ar trebui să fie cel puțin la fel de puternică precum în cazul heterozigot. Dacă se investighează Fig. 5.27 pentru acest SNP, observăm că există o afinitate către lipsa steatozei, ceea ce este de tip binar. Trebuie pusă în context această afinitate. Dacă ne întoarcem la analiza descriptivă din Fig. 5.1, observăm că acest SNP pentru starea 2 are foarte puține înregistrări. Prin urmare, eșantionul avut la dispoziție nu poate fi luat în seamă.

Un alt SNP cu o manifestare interesantă este MTHFR_rs4846048, care în stare heterozigotă indică o aversiune pentru stadiul 0. În cazul stării homozigot *wild-type* observăm că situație este inversă, indicându-se o afinitate pentru 0, cel puțin în comparație cu stadiile 1 și -1.

Dacă analizăm SNP-urile de pe gena PNPLA3, cu starea de homozigot *wild-type*, observăm că acestea indică o aversiune pentru stadiul 3 al steatozei, în timp ce în stare heterozigotă acestea indică o afinitate pentru stadiul 3. În același timp, putem observa că situația este inversă pentru stadiul -1. Desigur, acest lucru nu este o noutate. În subcapitolul 2.4.1, am prezentat o serie de argumente pentru cazul SNP-ului PNPLA3_rs738409. Rezultatele generate de model sunt în concordanță cu ceea ce indică literatura. Starea 2 confirmă ușor ceea ce indică starea 1, dar, din nou, lipsa înregistrărilor împiedică atingerea performanței maxime.

Un alt set de SNP-uri sunt SCD_rs2167444 și SCD_rs7848 care în stare homozigotă *wild-type* indică o aversiune pentru stadiul 0, iar în stare heterozigotă, SCD_rs2167444, indică o afinitate pentru stadiul 0. Acestea sunt doar câteva observații care se pot realiza pe baza reprezentărilor grafice. Desigur, dacă se modifică pragul de zgomot, atunci aceste reprezentări se vor schimba. Dacă pragul crește, atunci numărul agenților de vot vor scădea, iar dacă pragul scade numărul agenților de votare va crește.

În cazul clasei 2 a SNP-urilor, din analiza descriptivă din Fig. 5.1 ne putem da seama că volumul de înregistrări este prea mic pentru a fi relevant. Deși modelul nu este extrem de sensibil la numărul de înregistrări necesare pentru fiecare clasă, un număr minim totuși trebuie furnizat. De asemenea, granularitatea rezultatelor calculelor realizate de metodă este direct proporțional cu numărul înregistrărilor. Deci, este indicat să avem cel puțin 100 de înregistrări per clasă pentru a obține o granularitate de cel puțin un procent.

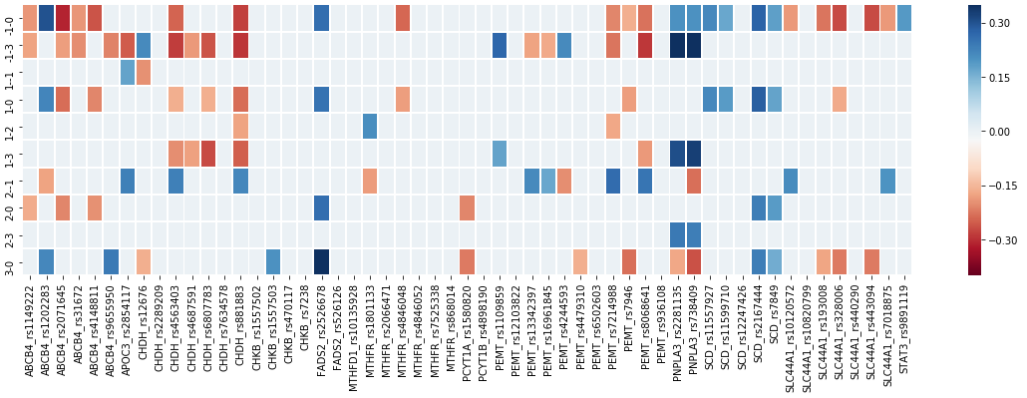


Fig. 5.25 Ponderea SNP-urilor pentru clasa 0

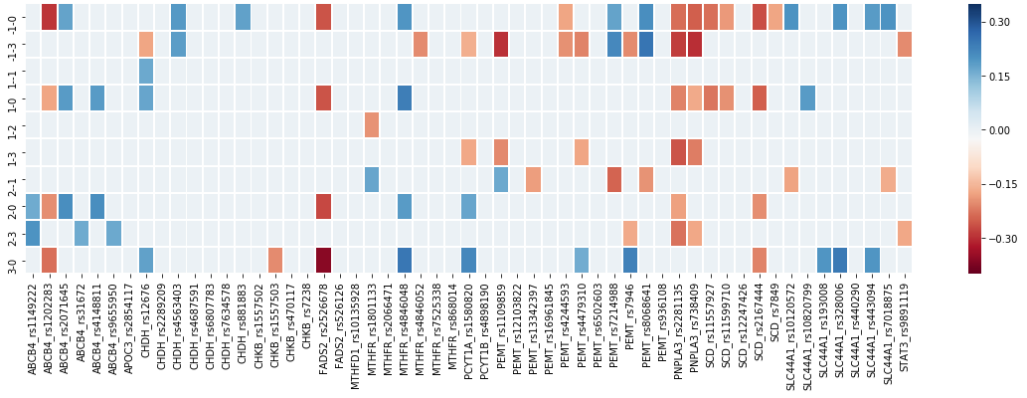


Fig. 5.26 Ponderea SNP-urilor pentru clasa 1

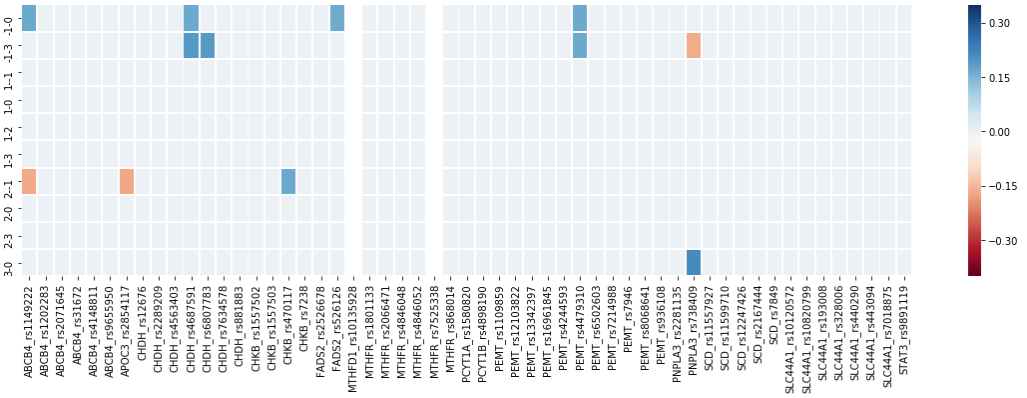


Fig. 5.27 Ponderea SNP-urilor pentru clasa 2

5.6. Discuții despre capitol

Când au fost analizate datele din fișierul original (stadiul steatozei cu cinci stări) s-a obținut o performanță generală scăzută. Acuratețea medie a fost de 32%, ceea ce înseamnă că modelele generează un rezultat greșit în majoritatea cazurilor. Mai mult de atât, având în vedere că s-a testat prin validare încrucișată, dacă baza de date nu este echilibrată, modelul bazat pe arborii decizionali tinde a se supra-specializa în predicția celei mai predominante clase a afecțiunii. În plus, faptul că *k* este sub 0,20, indică faptul că arborii sunt în acord ușor [149], [150]. Valoarea zero sau o valoare negativă a coeficientului înseamnă că nu există niciun acord asupra rezultatului. Având în vedere aceste aspecte, rezultatele pentru coeficientul *k*, în prima fază, au fost foarte scăzute, de la un acord ușor până la dezacord.

Din perspectiva eșantionării, cu excepția vârfurilor din rata de predicție (40% cu eșantionare aleatorie), nu au existat diferențe relevante. Intervalul dintre valorile minime și maxime ale lui *k* a fost în același domeniu de valori pentru toate cele 8 configurații. Acest lucru a fost valabil și pentru intervalele ratei de predicție.

Când au fost analizate metodele de fasonare, s-a observat că rezultatele au fost ușor mai bune când s-a aplicat MDL. Atunci când nu a fost aplicată fasonarea, intervalul de acurateță medie, pentru toate configurațiile, a fost cuprins între 24% și 26%. Când s-a folosit MDL, acuratețea a fost între 28% și 32%. În plus, intervalul minim de predicție, când nu s-a făcut fasonarea, era cuprins între 15% și 17%, în timp ce pentru MDL a fost între 17% la 30%. Prin urmare, în acest caz, fasonarea a mărit gradul de predicție. La coeficientul *k*, nu s-au observat diferențe notabile.

În a doua parte a studiului, când stadiile afecțiunii au fost transformate în valori binare, a existat o creștere a performanței de predicție. Acuratețea maximă obținută a fost de 91%, iar media maximă obținută a fost de 80%. Valoarea minimă a coeficientului *k* a fost -0,20 iar valoarea maximă a fost de 0,37. În comparație cu situația precedentă, diferența dintre valorile limită a crescut, câștigul fiind în mare parte pe limita superioară, de la 0,18 la 0,37. Cu toate acestea, chiar și cu valoarea maximă crescută, rezultatele au rămas în ceea ce a fost considerat zona coerentă a coeficientului [150], de la 0,21 la 0,40.

Similar cu analiza anterioară, rata de predicție de 91% a fost obținută prin prelevarea aleatorie și MDL. Atunci când analiza s-a bazat pe metoda de fasonare, similară analizei anterioare, performanța generală a acurateței a crescut. Iar când nu a fost aplicată fasonarea, intervalul mediu de performanță al acurateței a fost cuprins între 66% și 70%, iar valorile minime au fost cuprinse între 53% și 57%. Când s-a aplicat metoda MDL, performanța medie a fost cuprinsă între 79% și 80%, iar valorile minime între 68% și 76%. Chiar și valorile maxime au fost mai bune pentru MDL variind de la 82% la 91% , în loc de 75% la 82%.

În ambele cazuri, utilizarea MDL-ului a crescut acuratețea cu o valoare între 5 și 15 procente. Este posibil ca metoda de generare a modelului a supra-potrivit arborii decizionali din cauza datelor insuficiente. Atunci când a fost aplicată fasonarea, ramurile arborilor cu putere scăzută de clasificare au fost îndepărtate, ceea ce a dus la o creștere a preciziei de predicție. După cum au arătat *Răileanu* și *Stoffel* în comparația teoretică [151], nu s-a observat o diferență semnificativă între metodele *Gini Index* și *Gain Ratio*.

Folosind arborii decizionali, nu am reușit să generăm un model care să prezică relațiile dintre stadiul steatozei și SNP-uri cu o acurateță mai mare de 75%. Pe de altă parte, după ce am redus dimensiunea (anvergura) ieșirii, am putut genera modele care au putut prezice cu o acurateță mai mare de 90%.

În principiu, prezicerea unui fenotip utilizând datele genomice ar trebui să fie un proces relativ simplu, folosind metode de învățare automată. În cazul ideal, modelul ar avea ca intrare secvența ADN și fenotipul detaliat la ieșire. În realitate, studiul prezentat arată că lucrurile stau altfel, iar relațiile multiple și complexe dintr-o celulă sunt destul de greu de modelat. Desigur, acest experiment a fost realizat cu un număr de 55 de SNP-uri, aceasta fiind o limitare. Setul de date folosit a avut 400 de înregistrări pentru 55 de variante genetice, care au avut trei stări posibile pentru SNP-uri și cinci stări posibile pentru rezultat. Acestea nu sunt reprezentative pentru întregul genom, deoarece o ființă umană are un număr de ordinul 10^6 variante genetice.

Dacă analizăm metoda propusă în subcapitolul 5.5, aceasta a oferit rezultate apropiate de cele ale arborilor decizionali și ale modelelor *Random Forest*. Un avantaj față de arborii decizionali sunt hărțile de vot, care permit vizualizarea relațiilor dintre starea SNP-urilor și stadiul în care se află steatoza.

5.7. Concluzii de capitol și contribuții proprii

Primul subcapitol prezintă materialele utilizate 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 *Stochastic-Gradient Descent* (SGD). Pentru predicția individuală a stadiului, modelul SGD are o acuratețe medie de aproximativ 70%, dar pentru predicția simultană a celor cinci stări ale steatozei această valoare scade semnificativ, fiind de doar 25%. În continuare pentru predicția multiclasă s-au folosit arborii decizionali. Pentru determinarea configurației optime, o serie de parametri au fost analizați. Scorul mediu al acurateței, pentru steatoza 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 5.4 gradul complexității este redus prin eliminarea stadiilor steatozei și înlocuirea acestora cu simpla prezență a acestei afecțiuni. În cazul acesta folosind arborii decizionali s-a reușit obstinarea unui model care a reușit obținerea unui scor de 91%, scorul mediu fiind de 81%. Aceeași performanță a fost obținută și în cazul ansamblului de arbori decizionali. Așa cum a fost prezentat în literatură, 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 sau nu. În plus, nu s-a ținut cont de diferența dintre subiecții de sex masculin și de sex feminin. Este posibil ca anumite SNP-uri să fie relevante, mai mult sau mai puțin, pe baza sexului. Configurație cu cea mai bună precizie pentru arborii decizionali a fost folosirea funcției *Gini Index* cu fasonare MDL și cu metoda 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%. Această metodă are o performanță marginal mai scăzută decât a arborilor decizionali și sistemelor de predicție în ansamblu. Totuși, metoda prezintă câteva avantaje precum generarea hărților de vot care permit identificarea mult mai ușoară a relațiilor dintre SNP-uri și afecțiune. De altfel, o serie de relații au fost scoase în evidență pentru steatoză. De exemplu, SNP-urile rs2167444 și rs7848 aflate pe gena SCD, în stare heterozigotă, par să aibă o afinitate pentru stadiul 0 al steatozei, iar SNP-urile de pe gena ABCB4 par să indice o afinitate pentru starea 2 a steatozei.

Contribuțiile personale pentru acest capitol au la bază variantele genetice prezentate în Tabelul 5.1. Acestea sunt:

1. Efectuarea unui studiu pentru determinarea corelațiilor dintre variantele genetice și stadiul steatozei;
2. Generarea unor modele de clasificare folosind metoda *Stochastic Gradient Descent* (SGD), configurație binară și multiclasă, pentru determinarea stadiului steatozei pe baza unui genotip;
3. Analiza parametrilor de performanță (acuratețe, sensibilitate și precizie) pentru modelele de la punctul 2 și optimizarea performanței acestora prin analiza evoluției parametrilor oferiți de metodă;
4. Generarea unor modele de clasificare bazate pe arbori decizionali, singulari și în configurație de tip Ansamblu (*Random Forest)* pentru determinarea stadiului steatozei pe baza unui genotip;
5. Analiza parametrilor de performanță pentru modelele de la punctul 4 și optimizarea performanței acestora prin analiza evoluției parametrilor oferiți de metodă;
6. Propunerea unei metode de vot a SNP-urilor care să fie utilizată pentru determinarea legăturilor între acestea și afecțiune. Dezvoltarea unei metode pentru stabilirea pragului de zgomot pentru această metodă.
7. Hărțile de vot și modelul pentru determinarea stadiului steatozei.

6. CONCLUZII FINALE. CONTRIBUȚII PROPRII. PUBLICAȚII. PERSPECTIVE DE DEZVOLTARE.

6.1. Concluzii finale

Prezenta lucrare abordează domeniul multidisciplinar al bioinformaticii, mai precis ramura de analiză a datelor genomice. Obiectivul principal a fost determinarea unor metode și modele pentru identificarea afecțiunii pacientului, folosind date din cadrul genotipului și date fenotipice.

Lucrarea are o abordare progresivă. Aceasta urmărește fluxul de lucru după analiza secundară a datelor generate de dispozitivele de secvențiere. Contribuțiile se pot integra după diferiți pași din fluxul de lucru al analizei terțiare. În prima parte a lucrării se propun diferite strategii de identificare a afecțiunii folosind termenii HPO și variantele genetice avute la dispoziție. Această metodă poate fi folosită în mai multe moduri. Ulterior pentru rafinarea rezultatelor, se sugerează câteva metode de filtrare a variantelor genetice. Prima dintre aceste metode se bazează pe analiza grupului de pacienți și gruparea lor pe baza genotipului. Folosind această metodă, pacientul cu fenotip necunoscut poate fi analizat printr-o semnătură genetică indicativă a apropierii de unul dintre grupurile de pacienți cu afecțiuni asemănătoare. A doua metodă propusă este folosirea predictorilor *in silico* pentru identificarea variantelor, respectiv genelor care pot fi cauzatoare. În cadrul ultimei metode este prezentată o analiză statistică pentru identificarea posibilelor erori de secvențiere.

A doua parte a lucrării se concentrează asupra analizării procesului de matisare din punct de vedere genomic. Capitolul începe prin propunerea unor modele de identificarea a regiunilor de matisare, respectiv a regiunilor de prindere, situl acceptor și segmentul de pirimidine. Extragerea și validarea modelelor se realizează folosind setul de date *Homo Sapiens Splice Site Dataset*. Desigur, identificarea secvenței de ADN care reprezintă zona de matisare este doar primul pas. Pentru identificarea completă este necesar ca în apropierea regiunii de matisare să se regăsească o serie de regiuni activatoare, semnale de *splicing*. Regiunile activatoare au fost prezentate, în diverse forme, în literatura de specialitate. Folosind aceste informații a fost determinată o metodă pentru calcularea modificărilor care survin în cazul apariției unor variante genetice în cadrul acestor regiuni.

Ultima parte a lucrării este dedicată determinării unor modele matematice prin care să se facă predicția afecțiunilor complexe, exemplul prezentat fiind steatoza. Inițial sunt generate o serie de modele folosind metodele consacrate în cadrul învățării automate, precum arbori decizionali sau *Random Forest*. Motivul folosirii unui număr limitat de metode se datorează celui de-al doilea obiectiv urmărit, anume extragerea informațiilor relevante despre contribuția fiecărui SNP în cadrul afecțiunii finale. Prin urmare metodele care generează modele de tipul *black-box* (NN) nu au fost incluse în studiu. Pe lângă modelele obținute din cadrul metodelor specifice învățării automate, se mai prezintă o nouă metodă care se bazează pe diferențele dintre aparițiile SNP-urilor în cazul persoanelor afectate. De asemenea, este prezentat și modul de obținere a parametrului dedicat filtrării zgomotului. În cele din urmă sunt prezentate hărțile cu valori și metodele prin care se poate face predicția afecțiunii pe baza genotipului.

6.2. Contribuții personale

Urmărind obiectivele lucrării, în continuare enumerăm (reluăm) principalele contribuții:

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* pentru 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 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).

6.3. Direcții viitoare de cercetare

Rezultatele obținute contribuie la îmbunătățirea analizei terțiare pentru secvențierea ADN-ului și, totodată, îmbunătățesc procesul de diagnosticare, respectiv identificare a genelor și a variantelor patogene. Dezvoltarea și integrarea metodelor într-o platformă pentru analiza genotipurilor este de bun augur pentru că, odată cu creșterea volumului de informații genetice, metodele vor avea un randament superior.

În cazul metodelor prezentate în capitolului 3, acestea se pot concatena într-un flux de analiză, astfel crescând precizia rezultatelor. Pentru predictorii *in silico* se pot genera modele de detecție a patogenității folosind metode precum sunt rețelele neuronale.

Pentru metoda calculării semnalului de matisare, capitolul 4, aceasta poate fi aplicată într-un studiu mai amplu pentru identificarea variantelor sinonime care sunt

posibil patogene. De asemenea, se pot identifica și variante genetice considerate benigne din punct de vedere al modificării aduse codonului, dar care sunt, de fapt, patogene.

Pentru metoda propusă în capitolul 5, aceasta poate fi folosită pentru modelarea altor afecțiuni complexe. De astfel, această metodă a fost folosită și pentru modelarea efectului tratamentului hipolipemiant asupra anumitor genotipuri.

6.4. Publicații

A. Lucrări științifice publicate în reviste indexate *Web of Science* (WoS, ISI)

1. Florina Stoica, Adela Chirita-Emandi, Nicoleta Andreescu, Alina Stanciu, **Cristian G. Zimbru**, Maria Puiu, „Clinical relevance of retinal structure in children with laser-treated retinopathy of prematurity versus controls - using optical coherence tomography", *Acta Ophthalmologica* (FI 3.15), vol. 96, no. 2, pp 222-228, Martie 2018, WOS:000425369200019;

2. Vlad Serafim, Adela Chiriță-Emandi, Nicoleta Andreescu, Diana-Andreea Tiugan, Paul Tutac, Corina Paul, Iulian Velea, Alexandra Mihăilescu, Costela Lăcrimioara Șerban, **Cristian G. Zimbru**, Maria Puiu, Mihai Dinu Niculescu, „Single Nucleotide Polymorphisms in PEMT and MTHFR Genes are Associated with Omega 3 and 6 Fatty Acid Levels in the Red Blood Cells of Children with Obesity", *Nutrients* (FI 4.17), vol. 11, no. 11, pp 2600, Noiembrie 2019, WOS:000502274600051;

3. Adela Chirita-Emandi, Nicoleta Andreescu, **Cristian G. Zimbru**, Paul Tutac, Smaranda Arghirescu, Margit Serban, Maria Puiu, „Challenges in reporting pathogenic/potentially pathogenic variants in 94 cancer predisposing genes - in pediatric patients screened with NGS panels", *Scientific Reports* (FI 4.12), vol. 10, no. 1, pp 223, Ianuarie 2020;
DOI- https://doi.org/10.1038/s41598-019-57080-9

B. Lucrări științifice publicate în volumele unor manifestări științifice (Proceedings) indexate Web of Science-WoS (ISI) Proceedings

4. **Cristian G. Zimbru**, Nicoleta Andreescu, Adela Chirita-Emandi, Ioan Silea, Maria Puiu, Mihai D. Niculescu, „Analysis of Decision Tree Performance in Predicting the Relationship between a Scored Outcome and Multiple Single Nucleotide Polymorphisms", *2017 IEEE International Conference On E-Health And Bioengineering Conference (EHB)*, pp 57-60, 2017, WOS:000445457500015;

5. **Cristian G. Zimbru**, Nicoleta Andreescu, Adela Chirita-Emandi, Antonius Stanciu, Ioan Silea, Mihai D. Niculescu, Maria Puiu, „Splice Site Pattern Analysis and Identification of Similar Sequences in the Deep Intron Areas of Human Chromosome 21", *2017 IEEE International Conference On E-Health And Bioengineering Conference (EHB)*, pp 145-148, 2017, WOS:000445457500037;

6. Vlad-Ilie Ungureanu, Bianca-Alexandra Trutiu, Ioan Silea, Paul Negirla, **Cristian Zimbru**, Razvan-Catalin Miclea, „Automatic mapping of a room using LIDAR-based measuring sensor”, *22nd International Conference on Control Systems and Computer Science (CSCS)*, pp. 689-695, 2019, WOS:000491270300116;

C. Lucrări științifice publicate în reviste de specialitate indexate BDI

7. **Cristian G. Zimbru**, Nicoleta Andreescu, Adriana Albu, Adela Chirita-Emandi, Antonius Stanciu, Maria Puiu, „Performance Evaluation of In Silico Predictors for the Classification of Clinvar Variants”, *2019 E-Health and Bioengineering Conference (EHB)*, 2019, IEEE Xplore;
   DOI- https://doi.org/10.1109/EHB47216.2019.8969963

8. **Cristian G. Zimbru**, Adriana Albu, Nicoleta Andreescu, Adela ChiritaEmandi, Maria Puiu, „Determining Splicing Signal Variation in Humans by Analyzing the Regulatory Splicing Motifs”, *2019 E-Health and Bioengineering Conference (EHB),* 2019, IEEE Xplore;
   DOI- https://doi.org/10.1109/EHB47216.2019.8969983

9. Adriana Albu, Loredana Stanciu, Mădălina-Sofia Pașca, **Cristian G. Zimbru**, „Choosing Between Artificial Neural Networks and Bayesian Inference in Stroke Risk Prediction”, *2019 E-Health and Bioengineering Conference (EHB),* 2019, IEEE Xplore;
   DOI- https://doi.org/10.1109/EHB47216.2019.8970035

D. Lucrări științifice publicate în volumele unor manifestări științifice

10. **Cristian Zimbru**, Nicoleta Andreescu, Adela Chirita-Emandi, Maria Puiu, „Specialistul bioinformatician în analiza genomului”, *ColaboRARE Conference*, Ianuarie 2016;

11. **Cristian Zimbru**, Nicoleta Andreescu, Adela Chirita-Emandi, Antonius Stanciu, Ioan Silea, Maria Puiu, „Detection of high-risk intron areas that can cause splicing errors”, *Advanced Lecture Course on Systems Biology*, Martie 2016;

12. **Cristian Zimbru**, Nicoleta Andreescu, Adela Chirita Emandi, Ioan Silea, Maria Puiu, „Deep intron splice site-like sequences evidence for introns late theory”, *The IX Conference on Medical Genetics with International Participation*, Septembrie 2016;

13. Maria Puiu, Nicoleta Andreescu, Simona Farcas, **Cristian Zimbru**, Adela Chirita-Emandi, „Molecular genetics for the patients - prioritization criteria in the face of limited resources”, *European Society Human Genetics Confrence*, 2016;

14. Iulia Perva, Ciprian Perva, Dumitru Daniel Rusu, Simona Farcaș, Alexandra Mihăilescu, **Cristian Zimbru**, Narcis Morar, Maria Puiu, „Karyotyping Optimised Online Learning”, *European Society of Human Genetics Conference*, 2017;

15. Iulia Jurca-Simina, Adela Chirita-Emandi, Nicoleta Andreescu, Simona Farcaș, Alexandra Mihailescu, AM Popa, Paul Tutac, **Cristian Zimbru**, Andreea Dobrescu, Iulia Perva, A Murariu, Maria Puiu, „Burden Of Rare Genetic Diseases–Experience Of Timis Regional Centre Of Medical Genetics", *Revista Societății Române de Chirurgie Pediatrică*, 2019.

6.5. Granturi și premii

Premiul al 3-lea în cadrul conferinței EHB 2017 pentru lucrarea „Splice Site Pattern Analysis and Identification of Similar Sequences in the Deep Intron Areas of Human Chromosome 21" scrisă de Cristian ZIMBRU et al. (Romania).

Obținerea unui grant de tip *Young Scientist Fellowship* în cadrul sesiunii de lucru *Advanced Lecture Course on Systems Biology* prin competiție de abstracte, Innsbruck, 28 February – 5 March 2016.

BIBLIOGRAFIE

[1] "Deep learning for genomics," *Nat. Genet.*, vol. 51, no. 1, pp. 1–1, Jan. 2019, doi: 10.1038/s41588-018-0328-0.

[2] P. Aurel, *Dictionar de Genetică Moleculară și Inginerie Genetică*, 1st ed. AcademicPres, 2012.

[3] X. Yu and S. Sun, "Comparing a few SNP calling algorithms using low-coverage sequencing data," *BMC Bioinformatics*, vol. 14, p. 274, 2013, doi: 10.1186/1471-2105-14-274.

[4] S. Kim, H. Cho, D. Lee, and M. J. Webster, "Association between SNPs and gene expression in multiple regions of the human brain," *Transl. Psychiatry*, vol. 2, no. 5, p. e113, May 2012, doi: 10.1038/tp.2012.42.

[5] M. J. Hall, K. J. Ruth, D. Y. Chen, L. M. Gross, and V. N. Giri, "Interest in genomic SNP testing for prostate cancer risk: a pilot survey," *Hered. Cancer Clin. Pract.*, vol. 13, Apr. 2015, doi: 10.1186/s13053-015-0032-3.

[6] D. S. P. Tallapragada, S. Bhaskar, and G. R. Chandak, "New insights from monogenic diabetes for 'common' type 2 diabetes," *Front. Genet.*, vol. 6, Aug. 2015, doi: 10.3389/fgene.2015.00251.

[7] I. Peter *et al.*, "Association of Type 2 Diabetes Susceptibility Loci With One-Year Weight Loss in the Look AHEAD Clinical Trial," *Obes. Silver Spring Md*, vol. 20, no. 8, pp. 1675–1682, Aug. 2012, doi: 10.1038/oby.2012.11.

[8] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.

[9] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977.

[10] E. R. Mardis, "DNA sequencing technologies: 2006-2016," *Nat. Protoc.*, vol. 12, no. 2, pp. 213–218, Feb. 2017, doi: 10.1038/nprot.2016.182.

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

[12] S. Liu, Y. Wang, and F. Wang, "A fast read alignment method based on seed-and-vote for next generation sequencing," *BMC Bioinformatics*, vol. 17, no. Suppl 17, p. 466, Dec. 2016, doi: 10.1186/s12859-016-1329-6.

[13] Í. F. do Valle *et al.*, "Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data," *BMC Bioinformatics*, vol. 17, no. Suppl 12, p. 341, Nov. 2016, doi: 10.1186/s12859-016-1190-7.

[14] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinforma. Oxf. Engl.*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, doi: 10.1093/bioinformatics/btp324.

[15] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Mar. 2012, doi: 10.1038/nmeth.1923.

[16] A. Dobin *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[17] R. Poplin *et al.*, "Scaling accurate genetic variant discovery to tens of thousands of samples," *bioRxiv*, p. 201178, Nov. 2017, doi: 10.1101/201178.

[18] E. Garrison and G. Marth, "Haplotype-based variant detection from short-read sequencing," *ArXiv12073907 Q-Bio*, Jul. 2012.

[19] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010, doi: 10.1038/nmeth0410-248.
[20] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nat. Protoc.*, vol. 4, no. 7, pp. 1073–1081, Jun. 2009, doi: 10.1038/nprot.2009.86.
[21] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor," *Genome Biol.*, vol. 17, p. 122, 2016, doi: 10.1186/s13059-016-0974-4.
[22] M. K. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, "Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets," *Proc. IEEE*, vol. 104, no. 1, pp. 176–197, Jan. 2016, doi: 10.1109/JPROC.2015.2494198.
[23] M. Batterham, L. Tapsell, K. Charlton, J. O'Shea, and R. Thorne, "Using data mining to predict success in a weight loss trial," *J. Hum. Nutr. Diet.*, p. n/a-n/a, Jan. 2017, doi: 10.1111/jhn.12448.
[24] S. Alkoshi, N. Maimaiti, and M. Dahlui, "Cost-effectiveness analysis of rotavirus vaccination among Libyan children using a simple economic model," *Libyan J. Med.*, vol. 9, no. 1, p. 26236, Jan. 2014, doi: 10.3402/ljm.v9.26236.
[25] K. K, L. P G, L. Rangarajan, and A. K. K, "Effective Feature Selection for Classification of Promoter Sequences," *PloS One*, vol. 11, no. 12, p. e0167165, 2016, doi: 10.1371/journal.pone.0167165.
[26] S. Salzberg, A. L. Delcher, K. H. Fasman, and J. Henderson, "A Decision Tree System for Finding Genes in DNA," *J. Comput. Biol.*, vol. 5, no. 4, pp. 667–680, Jan. 1998, doi: 10.1089/cmb.1998.5.667.
[27] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Džeroski, "Predicting gene function using hierarchical multi-label decision tree ensembles," *BMC Bioinformatics*, vol. 11, no. 1, p. 2, Jan. 2010, doi: 10.1186/1471-2105-11-2.
[28] J.-T. Horng, L.-C. Wu, B.-J. Liu, J.-L. Kuo, W.-H. Kuo, and J.-J. Zhang, "An expert system to classify microarray gene expression data using gene selection by decision tree," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9072–9081, Jul. 2009, doi: 10.1016/j.eswa.2008.12.037.
[29] K.-H. Chen *et al.*, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," *BMC Bioinformatics*, vol. 15, no. 1, p. 49, Feb. 2014, doi: 10.1186/1471-2105-15-49.
[30] Y. Mao, X. Zhou, D. Pi, Y. Sun, and S. T. C. Wong, "Multiclass Cancer Classification by Using Fuzzy Support Vector Machine and Binary Decision Tree With Gene Selection," *BioMed Research International*, 2005. [Online]. Available: https://www.hindawi.com/journals/bmri/2005/247093/abs/. [Accessed: 30-Aug-2019].
[31] A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems," p. 718.
[32] 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.
[33] S. T. Sherry *et al.*, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001.
[34] P. D. P. Pharoah, J. Tyrer, A. M. Dunning, D. F. Easton, B. A. J. Ponder, and S. Investigators, "Association between Common Variation in 120 Candidate Genes and Breast Cancer Risk," *PLOS Genet.*, vol. 3, no. 3, p. e42, Mar. 2007, doi: 10.1371/journal.pgen.0030042.

[35] 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.
[36] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow, "MutationTaster evaluates disease-causing potential of sequence alterations," *Nat. Methods*, vol. 7, no. 8, pp. 575–576, Aug. 2010, doi: 10.1038/nmeth0810-575.
[37] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Res.*, vol. 39, no. 17, pp. e118–e118, Sep. 2011, doi: 10.1093/nar/gkr407.
[38] H. A. Shihab, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, and T. R. Gaunt, "Ranking non-synonymous single nucleotide polymorphisms based on disease concepts," *Hum. Genomics*, vol. 8, p. 11, Jun. 2014, doi: 10.1186/1479-7364-8-11.
[39] S. Kim, J.-H. Jhong, J. Lee, and J.-Y. Koo, "Meta-analytic support vector machine for integrating multiple omics data," *BioData Min.*, vol. 10, Jan. 2017, doi: 10.1186/s13040-017-0126-8.
[40] N. M. Ioannidis *et al.*, "REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants," *Am. J. Hum. Genet.*, vol. 99, no. 4, pp. 877–885, Oct. 2016, doi: 10.1016/j.ajhg.2016.08.016.
[41] K. A. Jagadeesh *et al.*, "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity," *Nat. Genet.*, vol. 48, no. 12, pp. 1581–1586, Dec. 2016, doi: 10.1038/ng.3703.
[42] 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.
[43] D. Quang, Y. Chen, and X. Xie, "DANN: a deep learning approach for annotating the pathogenicity of genetic variants," *Bioinformatics*, vol. 31, no. 5, pp. 761–763, Mar. 2015, doi: 10.1093/bioinformatics/btu703.
[44] M. Caulfield *et al.*, "The National Genomics Research and Healthcare Knowledgebase." 21-Aug-2019, doi: 10.6084/m9.figshare.4530893.v5.
[45] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, Sep. 2010, doi: 10.1093/nar/gkq603.
[46] 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.
[47] D. Smedley and P. N. Robinson, "Phenotype-driven strategies for exome prioritization of human Mendelian disease genes," *Genome Med.*, vol. 7, no. 1, Jul. 2015, doi: 10.1186/s13073-015-0199-2.
[48] P. N. Robinson *et al.*, "Improved exome prioritization of disease genes through cross-species phenotype comparison," *Genome Res.*, vol. 24, no. 2, pp. 340–348, Feb. 2014, doi: 10.1101/gr.160325.113.
[49] T. Zemojtel *et al.*, "Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome," *Sci. Transl. Med.*, vol. 6, no. 252, p. 252ra123, Sep. 2014, doi: 10.1126/scitranslmed.3009262.
[50] 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.

[51] 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.
[52] A. Javed, S. Agrawal, and P. C. Ng, "Phen-Gen: combining phenotype and genotype to analyze rare disorders," *Nat. Methods*, vol. 11, no. 9, pp. 935–937, Sep. 2014, doi: 10.1038/nmeth.3046.
[53] S. M. Berget, C. Moore, and P. A. Sharp, "Spliced segments at the 5' terminus of adenovirus 2 late mRNA," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 8, pp. 3171–3175, Aug. 1977.
[54] X. Jian, E. Boerwinkle, and X. Liu, "In silico tools for splicing defect prediction: a survey from the viewpoint of end users," *Genet. Med. Off. J. Am. Coll. Med. Genet.*, vol. 16, no. 7, pp. 497–503, Jul. 2014, doi: 10.1038/gim.2013.176.
[55] L. Cartegni, J. Wang, Z. Zhu, M. Q. Zhang, and A. R. Krainer, "ESEfinder: a web resource to identify exonic splicing enhancers," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3568–3571, Jul. 2003.
[56] L. Cartegni, M. L. Hastings, J. A. Calarco, E. de Stanchina, and A. R. Krainer, "Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2," *Am. J. Hum. Genet.*, vol. 78, no. 1, pp. 63–77, Jan. 2006, doi: 10.1086/498853.
[57] 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.
[58] 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.
[59] 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.
[60] K. Ohno, J. Takeda, and A. Masuda, "Rules and tools to predict the splicing effects of exonic and intronic mutations," *Wiley Interdiscip. Rev. RNA*, vol. 9, no. 1, p. e1451, Jan. 2018, doi: 10.1002/wrna.1451.
[61] L. Merendino, S. Guth, D. Bilbao, C. Martínez, and J. Valcárcel, "Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG," *Nature*, vol. 402, no. 6763, pp. 838–841, Dec. 1999, doi: 10.1038/45602.
[62] H. Y. Xiong *et al.*, "The human splicing code reveals new insights into the genetic determinants of disease," *Science*, p. 1254806, Dec. 2014, doi: 10.1126/science.1254806.
[63] A. Shibata *et al.*, "IntSplice: prediction of the splicing consequences of intronic single-nucleotide variations in the human genome," *J. Hum. Genet.*, vol. 61, no. 7, pp. 633–640, Jul. 2016, doi: 10.1038/jhg.2016.23.
[64] C. W. Smith, E. B. Porro, J. G. Patton, and B. Nadal-Ginard, "Scanning from an independently specified branch point defines the 3' splice site of mammalian introns," *Nature*, vol. 342, no. 6247, pp. 243–247, Nov. 1989, doi: 10.1038/342243a0.
[65] C. W. Smith, T. T. Chu, and B. Nadal-Ginard, "Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns," *Mol. Cell. Biol.*, vol. 13, no. 8, pp. 4939–4952, Aug. 1993.
[66] Y. Wang, M. Ma, X. Xiao, and Z. Wang, "Intronic Splicing Enhancers, Cognate Splicing Factors and Context Dependent Regulation Rules," *Nat. Struct. Mol. Biol.*, vol. 19, no. 10, pp. 1044–1052, Oct. 2012, doi: 10.1038/nsmb.2377.

[67] A. Anna and G. Monika, "Splicing mutations in human genetic disorders: examples, detection, and confirmation," *J. Appl. Genet.*, vol. 59, no. 3, pp. 253–268, 2018, doi: 10.1007/s13353-018-0444-7.
[68] R. Breathnach and P. Chambon, "Organization and expression of eucaryotic split genes coding for proteins," *Annu. Rev. Biochem.*, vol. 50, pp. 349–383, 1981, doi: 10.1146/annurev.bi.50.070181.002025.
[69] M. B. Shapiro and P. Senapathy, "RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression," *Nucleic Acids Res.*, vol. 15, no. 17, pp. 7155–7174, Sep. 1987.
[70] D. Brett, H. Pospisil, J. Valcárcel, J. Reich, and P. Bork, "Alternative splicing and genome complexity," *Nat. Genet.*, vol. 30, no. 1, pp. 29–30, Jan. 2002, doi: 10.1038/ng803.
[71] E. D. Lynch, M. K. Lee, J. E. Morrow, P. L. Welcsh, P. E. León, and M. C. King, "Nonsyndromic deafness DFNA1 associated with mutation of a human homolog of the Drosophila gene diaphanous," *Science*, vol. 278, no. 5341, pp. 1315–1318, Nov. 1997.
[72] S. Lefebvre *et al.*, "Identification and characterization of a spinal muscular atrophy-determining gene," *Cell*, vol. 80, no. 1, pp. 155–165, Jan. 1995.
[73] "HGMD® home page." [Online]. Available: http://www.hgmd.cf.ac.uk/ac/index.php. [Accessed: 02-Jan-2016].
[74] K. H. Lim, L. Ferraris, M. E. Filloux, B. J. Raphael, and W. G. Fairbrother, "Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 27, pp. 11093–11098, Jul. 2011, doi: 10.1073/pnas.1101135108.
[75] T. Sterne-Weiler, J. Howard, M. Mort, D. N. Cooper, and J. R. Sanford, "Loss of exon identity is a common mechanism of human inherited disease," *Genome Res.*, vol. 21, no. 10, pp. 1563–1571, Oct. 2011, doi: 10.1101/gr.118638.110.
[76] "Splice-Site Analyzer Tool." [Online]. Available: http://ibis.tau.ac.il/ssat/SpliceSiteFrame.htm. [Accessed: 07-Oct-2017].
[77] "splice view." [Online]. Available: http://bioinfo.itb.cnr.it/oriel/splice-view.html. [Accessed: 07-Oct-2017].
[78] "New GENSCAN Web Server at MIT." [Online]. Available: http://genes.mit.edu/GENSCAN.html. [Accessed: 07-Oct-2017].
[79] "NetGene2 Server." [Online]. Available: http://www.cbs.dtu.dk/services/NetGene2/. [Accessed: 07-Oct-2017].
[80] "BDGP: Splice Site Prediction by Neural Network." [Online]. Available: http://www.fruitfly.org/seq_tools/splice.html. [Accessed: 07-Oct-2017].
[81] R. I. Dogan, L. Getoor, W. J. Wilbur, and S. M. Mount, "SplicePort—An interactive splice-site analysis tool," *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. W285–W291, Jul. 2007, doi: 10.1093/nar/gkm407.
[82] G. Yeo and C. B. Burge, "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals," *J. Comput. Biol.*, vol. 11, no. 2–3, pp. 377–394, Mar. 2004, doi: 10.1089/1066527041410418.
[83] "MaxEntScan::scoresplice." [Online]. Available: http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html. [Accessed: 07-Oct-2017].
[84] L. Eng *et al.*, "Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: Maximum entropy estimates of splice junction strengths," *Hum. Mutat.*, vol. 23, no. 1, pp. 67–76, 2004, doi: 10.1002/humu.10295.

[85] "Automated Splice Site and Exon Definition Analyses." [Online]. Available: http://splice.uwo.ca/. [Accessed: 07-Oct-2017].
[86] P. Divina, A. Kvitkovicova, E. Buratti, and I. Vorechovsky, "Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping," *Eur. J. Hum. Genet.*, vol. 17, no. 6, pp. 759–765, Jun. 2009, doi: 10.1038/ejhg.2008.257.
[87] K. H. Lim and W. G. Fairbrother, "Spliceman—a computational web server that predicts sequence variations in pre-mRNA splicing," *Bioinformatics*, vol. 28, no. 7, pp. 1031–1032, Apr. 2012, doi: 10.1093/bioinformatics/bts074.
[88] F.-O. Desmet, D. Hamroun, M. Lalande, G. Collod-Béroud, M. Claustres, and C. Béroud, "Human Splicing Finder: an online bioinformatics tool to predict splicing signals," *Nucleic Acids Res.*, vol. 37, no. 9, p. e67, May 2009, doi: 10.1093/nar/gkp215.
[89] S. Schwartz, E. Hall, and G. Ast, "SROOGLE: webserver for integrative, user-friendly visualization of splicing signals," *Nucleic Acids Res.*, vol. 37, no. Web Server issue, pp. W189–W192, Jul. 2009, doi: 10.1093/nar/gkp320.
[90] C. Houdayer *et al.*, "Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants," *Hum. Mutat.*, vol. 33, no. 8, pp. 1228–1238, Aug. 2012, doi: 10.1002/humu.22101.
[91] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Comput. Biol.*, vol. 8, no. 12, Dec. 2012, doi: 10.1371/journal.pcbi.1002822.
[92] F. Camastra, M. D. Di Taranto, and A. Staiano, "Statistical and Computational Methods for Genetic Diseases: An Overview," *Comput. Math. Methods Med.*, vol. 2015, 2015, doi: 10.1155/2015/954598.
[93] M. Benedict and X. Zhang, "Non-alcoholic fatty liver disease: An expanded review," *World J. Hepatol.*, vol. 9, no. 16, pp. 715–732, Jun. 2017, doi: 10.4254/wjh.v9.i16.715.
[94] S. Romeo *et al.*, "Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease," *Nat. Genet.*, vol. 40, no. 12, pp. 1461–1465, Dec. 2008, doi: 10.1038/ng.257.
[95] A. Kotronen *et al.*, "Prediction of Non-Alcoholic Fatty Liver Disease and Liver Fat Using Metabolic and Genetic Factors," *Gastroenterology*, vol. 137, no. 3, pp. 865–872, Sep. 2009, doi: 10.1053/j.gastro.2009.06.005.
[96] L. Valenti *et al.*, "I148M patatin-like phospholipase domain-containing 3 gene variant and severity of pediatric nonalcoholic fatty liver disease," *Hepatol. Baltim. Md*, vol. 52, no. 4, pp. 1274–1280, Oct. 2010, doi: 10.1002/hep.23823.
[97] S. Romeo *et al.*, "The 148M allele of the PNPLA3 gene is associated with indices of liver damage early in life," *J. Hepatol.*, vol. 53, no. 2, pp. 335–338, Aug. 2010, doi: 10.1016/j.jhep.2010.02.034.
[98] Y. Rotman, C. Koh, J. M. Zmuda, D. E. Kleiner, T. J. Liang, and NASH CRN, "The association of genetic variability in patatin-like phospholipase domain-containing protein 3 (PNPLA3) with histological severity of nonalcoholic fatty liver disease," *Hepatol. Baltim. Md*, vol. 52, no. 3, pp. 894–903, Sep. 2010, doi: 10.1002/hep.23759.
[99] K. Kantartzis *et al.*, "Dissociation between fatty liver and insulin resistance in humans carrying a variant of the patatin-like phospholipase 3 gene," *Diabetes*, vol. 58, no. 11, pp. 2616–2623, Nov. 2009, doi: 10.2337/db09-0279.

[100] K. Hotta *et al.*, "Association of the rs738409 polymorphism in PNPLA3 with liver damage and the development of nonalcoholic fatty liver disease," *BMC Med. Genet.*, vol. 11, p. 172, Dec. 2010, doi: 10.1186/1471-2350-11-172.
[101] L. E. Wagenknecht *et al.*, "Association of PNPLA3 with non-alcoholic fatty liver disease in a minority cohort: the Insulin Resistance Atherosclerosis Family Study," *Liver Int. Off. J. Int. Assoc. Study Liver*, vol. 31, no. 3, pp. 412–416, Mar. 2011, doi: 10.1111/j.1478-3231.2010.02444.x.
[102] 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.
[103] M. Stephens and D. J. Balding, "Bayesian statistical methods for genetic association studies," *Nat. Rev. Genet.*, vol. 10, no. 10, pp. 681–690, Oct. 2009, doi: 10.1038/nrg2615.
[104] M. C. M. Garcia, E. T. Martins, and F. M. Azevedo, "Decision tree induction to prediction of prognosis in severe traumatic brain injury of Brazilian patients from Florianopolis city," in *13th IEEE International Conference on BioInformatics and BioEngineering*, 2013, pp. 1–4, doi: 10.1109/BIBE.2013.6701601.
[105] S. Uppu, A. Krishna, and R. Gopalan, "A review of machine learning and statistical approaches for detecting SNP interactions in high-dimensional genomic data," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, Dec. 2016, doi: 10.1109/TCBB.2016.2635125.
[106] G. Calcagno *et al.*, "A multilayer perceptron neural network-based approach for the identification of responsiveness to interferon therapy in multiple sclerosis patients," *Inf. Sci.*, vol. 180, no. 21, pp. 4153–4163, Nov. 2010, doi: 10.1016/j.ins.2010.07.004.
[107] M. D. Ritchie, A. A. Motsinger, W. S. Bush, C. S. Coffey, and J. H. Moore, "Genetic Programming Neural Networks: A Powerful Bioinformatics Tool for Human Genetics," *Appl. Soft Comput.*, vol. 7, no. 1, pp. 471–479, Jan. 2007, doi: 10.1016/j.asoc.2006.01.013.
[108] X. Chen and H. Ishwaran, "Pathway hunting by random survival forests," *Bioinformatics*, vol. 29, no. 1, pp. 99–105, Jan. 2013, doi: 10.1093/bioinformatics/bts643.
[109] 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)*, 2019, pp. 1–4, doi: 10.1109/EHB47216.2019.8969963.
[110] X. Liu, C. Wu, C. Li, and E. Boerwinkle, "dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs," *Hum. Mutat.*, vol. 37, no. 3, pp. 235–241, Mar. 2016, doi: 10.1002/humu.22932.
[111] X. Liu, X. Jian, and E. Boerwinkle, "dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions," *Hum. Mutat.*, vol. 32, no. 8, pp. 894–899, 2011, doi: 10.1002/humu.21517.
[112] B. Wang, L. Wan, A. Wang, and L. M. Li, "An adaptive decorrelation method removes Illumina DNA base-calling errors caused by crosstalk between adjacent clusters," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Feb. 2017, doi: 10.1038/srep41348.
[113] F. Pfeiffer *et al.*, "Systematic evaluation of error rates and causes in short samples in next-generation sequencing," *Sci. Rep.*, vol. 8, no. 1, pp. 1–14, Jul. 2018, doi: 10.1038/s41598-018-29325-6.

[114] T. Massingham and N. Goldman, "All Your Base: a fast and accurate probabilistic approach to base calling," *Genome Biol.*, vol. 13, no. 2, p. R13, Feb. 2012, doi: 10.1186/gb-2012-13-2-r13.
[115] V. S. Bondarenko and M. S. Gelfand, "Evolution of the Exon-Intron Structure in Ciliate Genomes," *PLOS ONE*, vol. 11, no. 9, p. e0161476, Sep. 2016, doi: 10.1371/journal.pone.0161476.
[116] L. F. Donaldson and N. Beazley-Long, "Alternative RNA splicing: contribution to pain and potential therapeutic strategy," *Drug Discov. Today*, doi: 10.1016/j.drudis.2016.06.017.
[117] A. J. Ward and T. A. Cooper, "The Pathobiology of Splicing," *J. Pathol.*, vol. 220, no. 2, pp. 152–163, Jan. 2010, doi: 10.1002/path.2649.
[118] M. C. Wahl, C. L. Will, and R. Lührmann, "The spliceosome: design principles of a dynamic RNP machine," *Cell*, vol. 136, no. 4, pp. 701–718, Feb. 2009, doi: 10.1016/j.cell.2009.02.009.
[119] A. Agrawal *et al.*, "An intronic ABCA3 mutation responsible for respiratory disease," *Pediatr. Res.*, vol. 71, no. 6, pp. 633–637, Jun. 2012, doi: 10.1038/pr.2012.21.
[120] C. Bouyer, L. Forestier, G. Renand, and A. Oulmouden, "Deep Intronic Mutation and Pseudo Exon Activation as a Novel Muscular Hypertrophy Modifier in Cattle," *PLoS ONE*, vol. 9, no. 5, May 2014, doi: 10.1371/journal.pone.0097399.
[121] V. Mehar *et al.*, "Congenital contractural arachnodactyly due to a novel splice site mutation in the FBN2 gene," *J. Pediatr. Genet.*, vol. 3, no. 3, pp. 163–166, Sep. 2014, doi: 10.3233/PGE-14093.
[122] R. Tang, D. O. Prosser, and D. R. Love, "Evaluation of Bioinformatic Programmes for the Analysis of Variants within Splice Site Consensus Regions," *Adv. Bioinforma.*, vol. 2016, 2016, doi: 10.1155/2016/5614058.
[123] 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)*, 2017, pp. 145–148, doi: 10.1109/EHB.2017.7995382.
[124] C. Tyner *et al.*, "The UCSC Genome Browser database: 2017 update," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D626–D634, Jan. 2017, doi: 10.1093/nar/gkw1134.
[125] E. Fröhlich and S. Salar-Behzadi, "Toxicological Assessment of Inhaled Nanoparticles: Role of in Vivo, ex Vivo, in Vitro, and in Silico Studies," *Int. J. Mol. Sci.*, vol. 15, no. 3, pp. 4795–4822, Mar. 2014, doi: 10.3390/ijms15034795.
[126] T. Heimbach, B. Xia, T. Lin, and H. He, "Case Studies for Practical Food Effect Assessments across BCS/BDDCS Class Compounds using In Silico, In Vitro, and Preclinical In Vivo Data," *AAPS J.*, vol. 15, no. 1, pp. 143–158, Jan. 2013, doi: 10.1208/s12248-012-9419-5.
[127] J. J. Ramírez-Espinosa *et al.*, "Antidiabetic activity of some pentacyclic acid triterpenoids, role of PTP–1B: In vitro, in silico, and in vivo approaches," *Eur. J. Med. Chem.*, vol. 46, no. 6, pp. 2243–2251, Jun. 2011, doi: 10.1016/j.ejmech.2011.03.005.
[128] M. Burset, I. A. Seledtsov, and V. V. Solovyev, "SpliceDB: database of canonical and non-canonical mammalian splice sites," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 255–259, Jan. 2001.
[129] M. Shionyu, A. Yamaguchi, K. Shinoda, K. Takahashi, and M. Go, "AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure,

interaction and network in human and mouse," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D305-309, Jan. 2009, doi: 10.1093/nar/gkn869.
[130] J. Vaquero-Garcia *et al.*, "A new view of transcriptome complexity and regulation through the lens of local splicing variations," *eLife*, vol. 5, Feb. 2016, doi: 10.7554/eLife.11752.
[131] J. Zhou and W.-J. Chng, "Aberrant RNA splicing and mutations in spliceosome complex in acute myeloid leukemia," *Stem Cell Investig.*, vol. 4, p. 6, 2017, doi: 10.21037/sci.2017.01.06.
[132] P. Pollastro and S. Rampone, "Hs3d, a dataset of homo sapiens splice regions, and its extraction procedure from a major public database," *Int. J. Mod. Phys. C*, vol. 13, no. 08, pp. 1105–1117, Oct. 2002, doi: 10.1142/S0129183102003796.
[133] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, Mar. 1970, doi: 10.1016/0022-2836(70)90057-4.
[134] M. Suñé-Pou *et al.*, "Targeting Splicing in the Treatment of Human Disease," *Genes*, vol. 8, no. 3, Feb. 2017, doi: 10.3390/genes8030087.
[135] B. J. Blencowe, "Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases," *Trends Biochem. Sci.*, vol. 25, no. 3, pp. 106–110, Mar. 2000.
[136] E. F. Cáceres and L. D. Hurst, "The evolution, impact and properties of exonic splice enhancers," *Genome Biol.*, vol. 14, no. 12, p. R143, 2013, doi: 10.1186/gb-2013-14-12-r143.
[137] R. Savisaar and L. D. Hurst, "Exonic splice regulation imposes strong selection at synonymous sites," *Genome Res.*, vol. 28, no. 10, pp. 1442–1454, Oct. 2018, doi: 10.1101/gr.233999.117.
[138] Z. Wang, M. E. Rolish, G. Yeo, V. Tung, M. Mawson, and C. B. Burge, "Systematic Identification and Analysis of Exonic Splicing Silencers," *Cell*, vol. 119, no. 6, pp. 831–845, Dec. 2004, doi: 10.1016/j.cell.2004.11.010.
[139] M. M. Scotti and M. S. Swanson, "RNA mis-splicing in disease," *Nat. Rev. Genet.*, vol. 17, no. 1, pp. 19–32, Jan. 2016, doi: 10.1038/nrg.2015.3.
[140] W. G. Fairbrother *et al.*, "RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons," *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W187–W190, Jul. 2004, doi: 10.1093/nar/gkh393.
[141] M. Raponi *et al.*, "Prediction of single-nucleotide substitutions that result in exon skipping: identification of a splicing silencer in BRCA1 exon 6," *Hum. Mutat.*, vol. 32, no. 4, pp. 436–444, Apr. 2011, doi: 10.1002/humu.21458.
[142] 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)*, 2019, pp. 1–4, doi: 10.1109/EHB47216.2019.8969983.
[143] M. Sironi *et al.*, "Silencer elements as possible inhibitors of pseudoexon splicing," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1783–1791, 2004, doi: 10.1093/nar/gkh341.
[144] X. H.-F. Zhang and L. A. Chasin, "Computational definition of sequence motifs governing constitutive exon splicing," *Genes Dev.*, vol. 18, no. 11, pp. 1241–1250, Jun. 2004, doi: 10.1101/gad.1195304.
[145] C. Zhang, W.-H. Li, A. R. Krainer, and M. Q. Zhang, "RNA landscape of evolution for optimal exon and intron discrimination," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 15, pp. 5797–5802, Apr. 2008, doi: 10.1073/pnas.0801692105.

[146] 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.
[147] J. C. Shafer, R. Agrawal, and M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining," in *Proceedings of the 22th International Conference on Very Large Data Bases*, San Francisco, CA, USA, 1996, pp. 544–555.
[148] M. Mehta, J. Rissanen, and R. Agrawal, "MDL-based Decision Tree Pruning," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Montréal, Québec, Canada, 1995, pp. 216–221.
[149] R. Bakeman, D. McArthur, V. Quera, and B. F. Robinson, "Detecting sequential patterns and determining their reliability with fallible observers," *Psychol. Methods*, vol. 2, no. 4, pp. 357–370, 1997, doi: 10.1037/1082-989X.2.4.357.
[150] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.
[151] L. E. Raileanu and K. Stoffel, "Theoretical Comparison between the Gini Index and Information Gain Criteria," *Ann. Math. Artif. Intell.*, vol. 41, no. 1, pp. 77–93, May 2004, doi: 10.1023/B:AMAI.0000018580.96245.c6.

Anexe

Anexa 1. Analiza regiunii de matisare

Tabelul A1.1 Distribuția bazelor azotate în cele 2.880 de secvențe validate ca fiind *splice site*

Poziție	Baze azotate și rația în care se regăsește pe această poziție							
	A	P (A)	T	P (T)	C	P (C)	G	P (G)
70	633	0.22	719	0.25	682	0.24	846	0.29
69	663	0.23	716	0.25	673	0.23	828	0.29
68	634	0.22	713	0.25	698	0.24	835	0.29
67	630	0.22	669	0.23	696	0.24	885	0.31
66	630	0.22	697	0.24	729	0.25	824	0.29
65	619	0.21	738	0.26	735	0.26	788	0.27
64	631	0.22	702	0.24	705	0.24	842	0.29
63	659	0.23	678	0.24	703	0.24	840	0.29
62	714	0.25	688	0.24	616	0.21	862	0.30
61	594	0.21	713	0.25	661	0.23	912	0.32
60	607	0.21	702	0.24	697	0.24	874	0.30
59	647	0.22	710	0.25	687	0.24	836	0.29
58	593	0.21	668	0.23	699	0.24	920	0.32
57	641	0.22	712	0.25	678	0.24	849	0.29
56	608	0.21	693	0.24	698	0.24	881	0.31
55	632	0.22	730	0.25	656	0.23	862	0.30
54	594	0.21	687	0.24	705	0.24	894	0.31
53	625	0.22	721	0.25	684	0.24	850	0.30
52	582	0.20	734	0.25	703	0.24	861	0.30
51	626	0.22	700	0.24	692	0.24	862	0.30
50	579	0.20	682	0.24	694	0.24	925	0.32
49	588	0.20	720	0.25	731	0.25	841	0.29
48	594	0.21	742	0.26	701	0.24	843	0.29
47	615	0.21	721	0.25	727	0.25	817	0.28
46	589	0.20	738	0.26	761	0.26	792	0.28
45	570	0.20	707	0.25	742	0.26	861	0.30
44	591	0.21	732	0.25	738	0.26	819	0.28
43	559	0.19	744	0.26	803	0.28	774	0.27
42	586	0.20	722	0.25	756	0.26	816	0.28
41	623	0.22	718	0.25	802	0.28	737	0.26
40	600	0.21	732	0.25	728	0.25	820	0.28
39	650	0.23	715	0.25	754	0.26	761	0.26
38	594	0.21	685	0.24	830	0.29	771	0.27
37	571	0.20	730	0.25	790	0.27	789	0.27
36	606	0.21	717	0.25	827	0.29	730	0.25
35	613	0.21	758	0.26	815	0.28	694	0.24
34	622	0.22	738	0.26	818	0.28	702	0.24

33	554	0.19	775	0.27	840	0.29	711	0.25
32	557	0.19	806	0.28	824	0.29	693	0.24
31	607	0.21	759	0.26	849	0.29	665	0.23
30	639	0.22	736	0.26	849	0.29	656	0.23
29	618	0.21	784	0.27	850	0.30	628	0.22
28	619	0.21	791	0.27	848	0.29	622	0.22
27	586	0.20	792	0.28	954	0.33	548	0.19
26	604	0.21	848	0.29	905	0.31	523	0.18
25	613	0.21	842	0.29	868	0.30	557	0.19
24	609	0.21	833	0.29	892	0.31	546	0.19
23	620	0.22	915	0.32	873	0.30	472	0.16
22	577	0.20	857	0.30	956	0.33	490	0.17
21	588	0.20	909	0.32	931	0.32	452	0.16
20	542	0.19	962	0.33	925	0.32	451	0.16
19	475	0.16	1022	0.35	910	0.32	473	0.16
18	430	0.15	982	0.34	1019	0.35	449	0.16
17	374	0.13	996	0.35	1062	0.37	448	0.16
16	340	0.12	1118	0.39	1024	0.36	398	0.14
15	317	0.11	1101	0.38	1060	0.37	402	0.14
14	291	0.10	1190	0.41	1032	0.36	367	0.13
13	232	0.08	1238	0.43	1044	0.36	366	0.13
12	224	0.08	1247	0.43	1067	0.37	342	0.12
11	214	0.07	1381	0.48	980	0.34	305	0.11
10	224	0.08	1258	0.44	1093	0.38	305	0.11
9	206	0.07	1242	0.43	1112	0.39	320	0.11
8	251	0.09	1103	0.38	1174	0.41	352	0.12
7	263	0.09	1124	0.39	1222	0.42	271	0.09
6	185	0.06	1165	0.40	1341	0.47	189	0.07
5	199	0.07	1363	0.47	1121	0.39	197	0.07
4	598	0.21	625	0.22	996	0.35	661	0.23
3	103	0.04	571	0.20	2190	0.76	16	0.01
2	2880	1.00	0	0.00	0	0.00	0	0.00
1	0	0.00	0	0.00	0	0.00	2880	1.00

Tabelul A1.2 Distribuția bazelor azotate în cele 329.374 de secvențe care nu sunt regiuni de *splicing*

Poziția relativ la exon	A	P (A)	T	P (T)	C	P (C)	G	P (G)
70	84145	0.26	86495	0.26	77109	0.23	81624	0.25
69	85150	0.26	87674	0.27	75815	0.23	80735	0.25
68	83718	0.25	88269	0.27	77647	0.24	79739	0.24
67	82830	0.25	89929	0.27	77446	0.24	79169	0.24
66	83404	0.25	88566	0.27	77642	0.24	79762	0.24
65	86030	0.26	88548	0.27	76846	0.23	77950	0.24
64	86441	0.26	86177	0.26	75879	0.23	80877	0.25
63	86904	0.26	86986	0.26	75391	0.23	80093	0.24
62	85596	0.26	88125	0.27	78097	0.24	77556	0.24
61	83983	0.25	87836	0.27	77748	0.24	79807	0.24
60	84091	0.26	89779	0.27	77501	0.24	78003	0.24
59	85990	0.26	88942	0.27	76730	0.23	77712	0.24
58	84609	0.26	88063	0.27	76519	0.23	80183	0.24
57	85968	0.26	86193	0.26	76703	0.23	80510	0.24
56	87687	0.27	86694	0.26	77712	0.24	77281	0.23
55	83786	0.25	88534	0.27	77881	0.24	79173	0.24
54	84654	0.26	88876	0.27	76486	0.23	79358	0.24
53	83949	0.25	87694	0.27	79523	0.24	78208	0.24
52	84311	0.26	88329	0.27	78245	0.24	78489	0.24
51	86847	0.26	87496	0.27	74826	0.23	80205	0.24
50	85931	0.26	87273	0.26	75456	0.23	80714	0.25
49	84029	0.26	86750	0.26	76167	0.23	82428	0.25
48	83941	0.25	87598	0.27	78727	0.24	79108	0.24
47	86115	0.26	88805	0.27	76782	0.23	77672	0.24
46	84259	0.26	88622	0.27	76074	0.23	80419	0.24
45	86507	0.26	86741	0.26	75218	0.23	80908	0.25
44	85600	0.26	88686	0.27	76302	0.23	78786	0.24
43	84595	0.26	87833	0.27	77393	0.23	79553	0.24
42	85387	0.26	88609	0.27	74405	0.23	80972	0.25
41	86464	0.26	88306	0.27	76446	0.23	78158	0.24
40	84526	0.26	86038	0.26	77087	0.23	81723	0.25
39	85579	0.26	88170	0.27	75225	0.23	80400	0.24
38	85909	0.26	89215	0.27	76444	0.23	77806	0.24
37	83515	0.25	86417	0.26	78183	0.24	81259	0.25
36	82063	0.25	89940	0.27	77855	0.24	79516	0.24
35	84230	0.26	92209	0.28	73891	0.22	79044	0.24
34	84348	0.26	88391	0.27	76347	0.23	80288	0.24
33	83564	0.25	87200	0.26	76005	0.23	82605	0.25
32	84955	0.26	88323	0.27	76741	0.23	79355	0.24
31	84718	0.26	86658	0.26	79026	0.24	78972	0.24
30	82580	0.25	89695	0.27	77765	0.24	79334	0.24
29	83161	0.25	89010	0.27	78709	0.24	78494	0.24

28	83909	0.25	87050	0.26	77300	0.23	81115	0.25
27	85296	0.26	86213	0.26	77614	0.24	80251	0.24
26	86290	0.26	87277	0.26	74333	0.23	81474	0.25
25	82511	0.25	87201	0.26	76408	0.23	83254	0.25
24	86481	0.26	86249	0.26	78050	0.24	78594	0.24
23	82900	0.25	90750	0.28	78314	0.24	77410	0.24
22	82659	0.25	91375	0.28	76333	0.23	79007	0.24
21	83927	0.25	86311	0.26	77804	0.24	81331	0.25
20	84631	0.26	90180	0.27	76540	0.23	78023	0.24
19	84045	0.26	89093	0.27	78305	0.24	77931	0.24
18	84188	0.26	90952	0.28	75796	0.23	78438	0.24
17	82081	0.25	94013	0.29	76786	0.23	76494	0.23
16	85097	0.26	88818	0.27	76007	0.23	79452	0.24
15	84229	0.26	88358	0.27	79082	0.24	77705	0.24
14	84193	0.26	93321	0.28	76447	0.23	75413	0.23
13	78685	0.24	92182	0.28	79106	0.24	79401	0.24
12	78844	0.24	92684	0.28	82190	0.25	75656	0.23
11	82023	0.25	93192	0.28	78003	0.24	76156	0.23
10	79400	0.24	93168	0.28	76845	0.23	79961	0.24
9	81271	0.25	92997	0.28	74445	0.23	80661	0.24
8	84769	0.26	92745	0.28	73340	0.22	78520	0.24
7	82546	0.25	93017	0.28	75227	0.23	78584	0.24
6	79870	0.24	95414	0.29	79748	0.24	74342	0.23
5	73100	0.22	95604	0.29	86785	0.26	73885	0.22
4	80586	0.24	90357	0.27	78547	0.24	79883	0.24
3	85537	0.26	55414	0.17	113004	0.34	75419	0.23
2	329374	1.00	0	0.00	0	0.00	0	0.00
1	0	0.00	0	0.00	0	0.00	329374	1.00

Tabel A1.3 Numărul de apariții al tuplului de baze azotate în fișierul cu secvențe de *splicing* invalide

Poziție	AA	TA	CA	GA	AT	CT	TT	GT	AC	TC	CC	GC	AG	TG	CG	GG
1	0	0	0	0	0	0	0	0	0	0	0	0	2880	0	0	0
2	103	571	2190	16	0	0	0	0	0	0	0	0	0	0	0	0
3	34	23	30	16	107	193	176	95	455	422	765	548	2	4	8	2
4	87	185	286	40	34	229	338	24	62	385	481	68	16	455	125	65
5	27	53	97	22	84	543	679	57	73	328	659	61	1	105	42	49
6	32	43	88	22	124	445	495	101	102	475	666	98	5	111	23	50
7	43	76	121	23	97	491	406	130	108	449	521	144	3	172	41	55
8	31	82	114	24	81	455	428	139	91	505	470	108	3	227	73	49
9	33	72	94	7	112	494	501	135	79	451	466	116	0	234	39	47
10	31	84	91	18	106	477	551	124	77	545	361	110	0	201	51	53
11	25	65	100	24	108	542	578	153	86	406	388	100	5	198	37	65
12	37	73	91	23	114	482	509	142	77	431	427	132	4	225	44	69
13	41	56	101	34	122	458	520	138	111	399	418	116	17	215	55	79
14	64	66	119	42	129	458	450	153	98	363	428	143	26	222	55	64
15	58	77	139	43	130	435	418	118	121	378	402	159	31	245	48	78
16	68	77	127	68	123	463	400	132	146	299	420	159	37	220	52	89
17	82	75	141	76	130	403	343	120	150	326	428	158	68	238	47	95
18	80	96	157	97	137	341	363	141	174	336	361	148	84	227	51	87
19	108	84	189	94	166	378	341	137	189	281	306	134	79	256	52	86
20	137	99	191	115	158	381	302	121	197	277	315	136	96	231	44	80
21	125	128	198	137	161	380	239	129	182	278	325	146	109	212	53	78
22	140	98	180	159	166	329	249	113	201	304	319	132	113	264	45	68
23	147	130	191	152	167	353	258	137	191	238	299	145	104	207	49	112
24	150	121	189	149	150	330	240	113	184	244	303	161	129	237	46	134
25	155	128	206	124	153	337	233	119	177	246	308	137	119	241	54	143

26	173	111	195	125	137	341	231	139	153	237	350	165	123	213	68	119
27	157	120	184	125	140	295	215	142	195	243	316	200	127	213	53	155
28	144	119	208	148	157	294	226	114	159	219	304	166	158	220	44	200
29	160	111	211	136	132	291	207	154	174	205	300	171	173	213	47	195
30	157	122	203	157	122	282	198	134	147	221	300	181	181	218	64	193
31	147	115	194	151	124	292	197	146	136	243	275	195	150	251	63	201
32	143	105	179	130	129	312	196	169	117	220	297	190	165	254	52	222
33	133	93	196	132	140	277	206	152	173	200	276	191	176	239	69	227
34	155	116	206	145	141	268	201	128	132	205	289	192	185	236	52	229
35	140	91	211	171	132	278	206	142	135	202	278	200	199	218	60	217
36	135	116	179	176	134	256	184	143	130	203	287	207	172	227	68	263
37	135	83	195	158	124	259	199	148	137	159	308	186	198	244	68	279
38	155	93	176	170	130	244	192	119	152	199	276	203	213	231	58	269
39	171	98	190	191	110	242	211	152	129	202	239	184	190	221	57	293
40	147	100	211	142	139	228	203	162	114	168	298	148	223	247	65	285
41	160	99	178	186	108	240	216	154	128	188	295	191	190	219	43	285
42	136	100	194	156	116	233	205	168	110	161	304	181	197	278	72	269
43	142	95	167	155	136	229	201	178	132	186	300	185	181	250	42	301
44	136	85	195	175	125	233	211	163	117	171	260	190	192	240	54	333
45	147	97	175	151	116	256	188	147	108	198	257	179	218	255	73	315
46	156	115	164	154	127	242	207	162	125	166	261	209	207	233	60	292
47	175	100	157	183	123	242	190	166	113	207	237	170	183	245	65	324
48	133	104	174	183	133	245	196	168	116	154	266	165	206	266	46	325
49	141	109	182	156	111	217	204	188	115	158	243	215	212	211	52	366
50	154	99	187	139	120	212	189	161	112	169	227	186	240	243	66	376

51	126	107	202	191	131	206	207	156	109	175	245	163	216	245	50	351
52	144	109	152	177	139	238	187	170	112	172	251	168	230	253	43	335
53	154	100	186	185	131	214	194	182	98	169	246	171	211	224	59	356
54	157	94	166	177	123	208	200	156	122	153	238	192	230	283	44	337
55	151	91	191	199	126	234	185	185	97	157	216	186	234	260	57	311
56	169	93	165	181	130	230	177	156	126	173	232	167	216	269	51	345
57	175	96	172	198	118	239	188	167	104	158	228	188	196	226	60	367
58	149	109	157	178	124	227	155	162	133	159	239	168	241	287	64	328
59	147	116	192	192	131	221	174	184	121	160	215	191	208	252	69	307
60	144	101	172	190	127	203	195	177	97	164	231	205	226	253	55	340
61	161	104	160	169	144	187	208	174	118	134	225	184	291	242	44	335
62	175	119	207	213	153	216	175	144	110	154	207	145	221	230	73	338
63	162	124	187	186	137	203	204	134	108	153	244	198	224	221	71	324
64	161	112	196	162	121	238	202	141	118	162	254	171	219	262	47	314
65	167	104	186	162	144	239	186	169	117	171	254	193	202	236	50	300
66	167	112	193	158	140	218	177	162	115	168	232	214	208	212	53	351
67	163	122	174	171	131	224	180	134	118	171	233	174	222	240	67	356
68	172	113	168	181	151	202	196	164	117	167	245	169	223	240	58	314
69	176	128	173	186	140	225	182	169	106	170	227	170	211	239	57	321

Tabelul A1.4 Numărul de apariții al tuplului de baze azotate în fișierul cu secvențe invalide de *splicing*

	AA	TA	CA	GA	AT	CT	TT	GT	AC	TC	CC	GC	AG	TG	CG	GG
1	0	0	0	0	0	0	0	0	0	0	0	0	329374	0	0	0
2	85537	55414	113004	75419	0	0	0	0	0	0	0	0	0	0	0	0
3	28487	15958	21732	19360	14105	13013	17539	10757	25698	27465	37252	22588	12296	29395	6550	27178
4	23295	17789	23587	15915	19197	28633	26486	16041	14615	20683	28532	14717	15993	30646	6033	27211
5	23399	14892	19222	15587	20554	28215	31454	15381	18692	22784	26194	19115	17225	26284	6117	24259
6	24915	15742	21091	18122	22826	24670	29966	17952	17652	20319	23319	18458	17153	26990	6147	24052
7	26591	15105	20944	19906	21424	25192	30418	15983	16783	20476	21843	16125	19971	26746	5361	26506
8	25043	17739	21413	20574	20748	25964	28702	17331	16324	19165	21292	16559	19156	27391	5776	26197
9	24252	16297	22548	18174	20956	25028	28326	18687	15922	18361	23362	16800	18270	30184	5907	26300
10	23928	15350	21612	18510	20635	27829	28696	16008	16778	19691	22984	17392	20682	29455	5578	24246
11	23832	16848	23614	17729	19321	28703	28654	16514	16357	20126	23818	17702	19334	27056	6055	23711
12	23276	15175	22136	18257	19531	25963	30015	17175	15226	21722	24978	20264	20652	25270	6029	23705
13	24145	15174	21546	17820	20069	26126	29907	16080	17035	20778	23231	18062	22944	27462	5544	23451
14	24951	15915	24160	19167	21404	25544	28729	17644	16700	19226	23224	17297	21174	24488	6154	23597
15	25876	15545	22305	20503	19773	23361	27595	17629	16724	20067	24218	18073	22724	25611	6123	23247

16	25349	16087	24371	19290	19337	24402	28874	16205	15477	21482	21875	17173	21918	27570	6138	23826
17	25333	16011	21834	18903	21349	25820	29113	17731	15444	20502	22203	18637	22062	25326	5939	23167
18	25744	15443	23486	19515	19183	26083	27960	17726	15851	20176	22340	17429	23267	25514	6396	23261
19	25267	17183	22469	19126	19432	25475	27239	16947	16036	20451	23322	18496	23896	25307	5274	23454
20	24458	16141	23186	20845	19836	25279	27109	17956	16277	18928	23912	17423	23356	24133	5427	25107
21	24228	15838	23415	20446	18990	23848	26455	17018	15906	21277	22950	17671	23534	27805	6120	23872
22	24632	16184	22571	19272	18835	27641	27568	17331	15366	21707	22222	17038	24067	25291	5880	23769
23	24418	15613	22541	20328	21468	25995	26242	17045	17005	20048	23824	17437	23590	24346	5690	23784
24	24950	16286	23425	21820	19436	24412	25747	16654	15284	19369	22611	20786	22841	25799	5960	23994
25	23691	15782	22717	20321	19140	23245	26956	17860	16090	18839	22471	19008	27369	25700	5900	24285
26	24327	15654	24794	21515	20276	23903	25836	17262	15852	18692	22021	17768	24841	26031	6896	23706
27	23945	15621	24018	21712	18857	24106	25940	17310	15867	19987	23354	18406	25240	25502	5822	23687
28	24518	16582	23398	19411	18765	25859	26260	16166	15620	20367	23855	17458	24258	25801	5597	25459
29	24149	15844	23090	20078	18976	24453	27537	18044	15835	21290	24292	17292	23620	25024	5930	23920
30	23546	15491	24243	19300	20045	25917	26544	17189	16369	19276	22804	19316	24758	25347	6062	23167
31	25418	15743	23195	20362	19059	23898	27513	16188	16439	19360	23964	19263	24039	25707	5684	23542
32	24930	16001	22466	21558	19002	25400	27123	16798	16350	18307	22395	19689	23282	25769	5744	24560
33	23992	15659	23362	20551	18888	22987	27329	17996	15911	19081	23702	17311	25557	26322	6296	24430
34	24279	16374	22688	21007	19163	23528	28263	17437	16319	20801	22019	17208	24469	26771	5656	23392

35	24467	15641	23480	20642	19486	26506	28242	17975	15203	19471	21273	17944	22907	26586	6596	22955
36	24257	15204	22611	19991	20186	25653	26732	17369	15167	19545	24353	18790	23905	24936	5566	25109
37	24556	16274	23263	19422	19820	23344	26818	16435	15745	20225	23703	18510	25788	25898	6134	23439
38	24921	15942	23351	21695	20780	23890	26709	17836	16062	19915	22203	18264	23816	25604	5781	22605
39	25382	15355	23455	21387	19479	24760	26486	17445	15412	18887	23138	17788	24253	25310	5734	25103
40	25407	15302	24032	19785	19327	23270	26904	16537	16211	19463	23183	18230	25519	26637	5961	23606
41	25373	16510	23796	20785	19277	23479	27599	17951	16783	19206	21040	19416	23954	25294	6090	22820
42	25531	15682	24163	20011	18909	24900	27620	17180	15494	18418	22146	18347	24661	26113	6184	24014
43	25062	16950	23135	19448	18938	24468	27208	17219	16190	19565	23065	18573	25410	24963	5634	23546
44	24750	16139	23888	20823	19914	24064	27055	17653	16351	19479	21783	18689	25492	24068	5483	23743
45	25012	16260	24484	20751	19500	24029	26357	16855	15382	20682	22070	17084	24365	25323	5491	25729
46	25093	16345	22825	19996	19268	25577	27186	16591	16259	19749	23049	17017	25495	25525	5331	24068
47	25517	16422	23727	20449	18968	25818	26619	17400	16083	19774	23245	17680	23373	24783	5937	23579
48	25301	16295	22597	19748	18960	25204	25917	17517	16024	19442	23007	20254	23744	25096	5359	24909
49	24721	16091	23203	20014	19514	24240	26146	16850	15839	19025	22527	18776	25857	26011	5486	25074
50	25592	16233	23057	21049	19357	24326	26427	17163	16549	18781	21490	18636	25349	26055	5953	23357
51	25197	16986	24354	20310	18837	25776	26024	16859	15323	19449	22474	17580	24954	25870	5641	23740

52	24097	16250	23053	20911	18908	26293	26215	16913	15668	20271	24655	17651	25276	24958	5522	22733
53	24273	16638	22902	20136	19769	25156	26086	16683	17062	20855	23056	18550	23550	25297	5372	23989
54	24648	16581	23657	19768	19431	25576	26541	17328	15456	19733	22787	18510	24251	25679	5861	23567
55	24854	16248	22910	19774	19671	26053	25705	17105	16936	19851	23343	17751	26226	24890	5406	22651
56	25984	16427	23747	21529	19235	24678	25909	16872	16628	19308	22418	19358	24121	24549	5860	22751
57	25205	16569	23971	20223	18854	23300	26496	17543	15937	19374	22833	18559	24613	25624	6415	23858
58	25390	16350	23357	19512	19524	24371	26960	17208	15789	20236	23022	17472	25287	25396	5980	23520
59	25262	16172	24030	20526	19659	24379	27552	17352	15644	20352	23040	17694	23526	25703	6052	22431
60	25162	15833	22737	20359	19579	25443	27384	17373	15565	19904	23807	18225	23677	24715	5761	23850
61	25196	15728	22880	20179	19419	25200	25974	17243	15757	20368	24327	17296	25224	26055	5690	22838
62	25673	16377	23222	20324	19848	23679	26028	18570	17204	19541	22687	18665	24179	25040	5803	22534
63	26711	16376	22933	20884	19890	23915	25941	17240	15403	19078	23041	17869	24437	24782	5990	24884
64	25370	16275	24521	20275	19212	24061	26067	16837	16055	19892	22595	17337	25393	26314	5669	23501
65	25001	16151	24605	20273	19175	24978	26285	18110	15558	20533	22103	18652	23670	25597	5956	22727
66	23835	15838	23561	20170	19329	24951	27001	17285	15758	20401	22892	18591	23908	26689	6042	23123
67	24460	15731	23371	19268	18812	26086	26860	18171	15801	20229	22466	18949	24645	25449	5724	23351
68	24498	15952	23082	20186	19313	24477	26824	17655	16188	20224	22694	18541	25151	24674	5562	24352
69	24341	16046	24069	20694	19096	24611	26237	17730	15400	19381	22572	18461	25308	24831	5857	24739

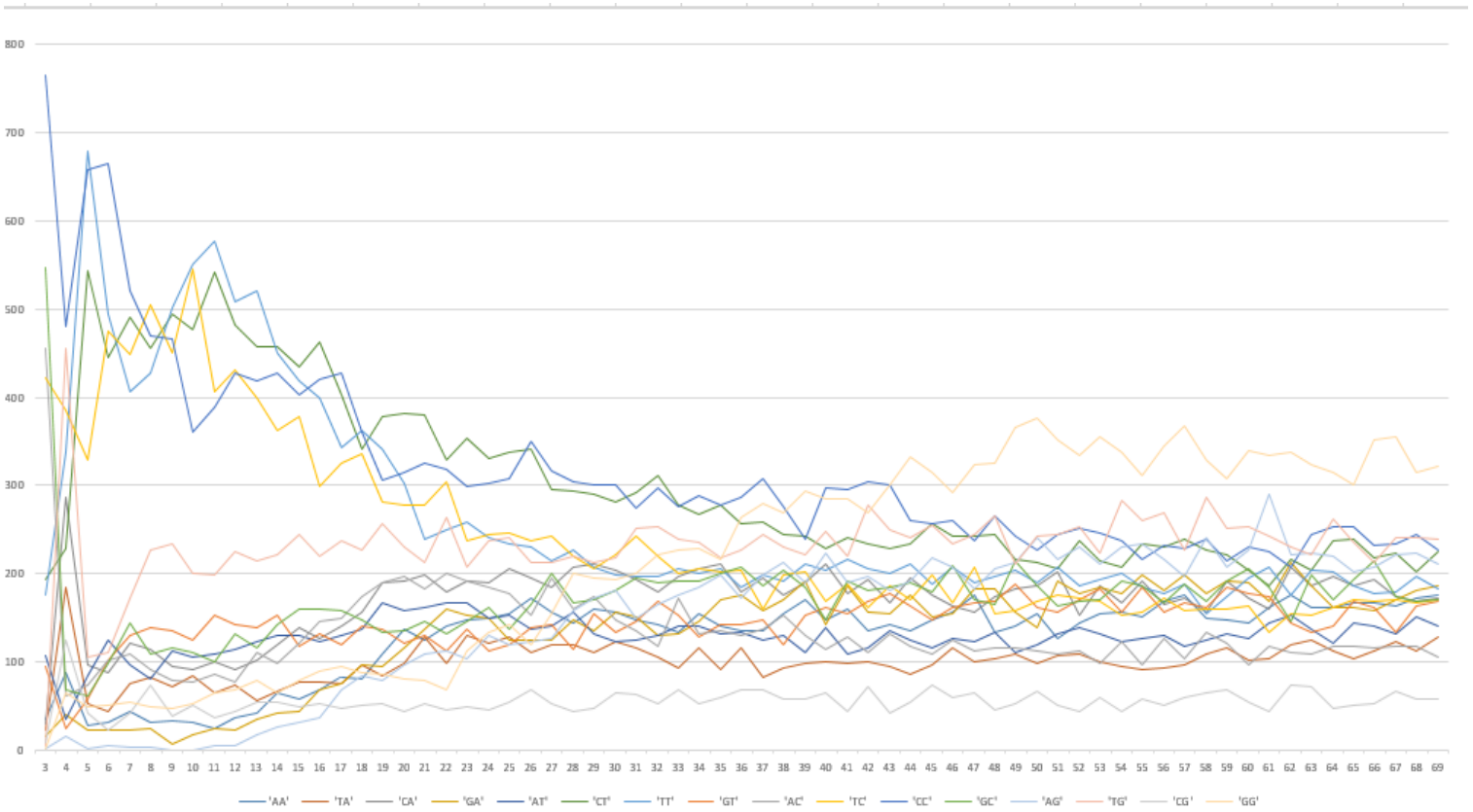


Fig. A1.1 Numărul de apariții al fiecărui tuplu pentru fiecare poziție din lista secvențelor valide de *splicing*

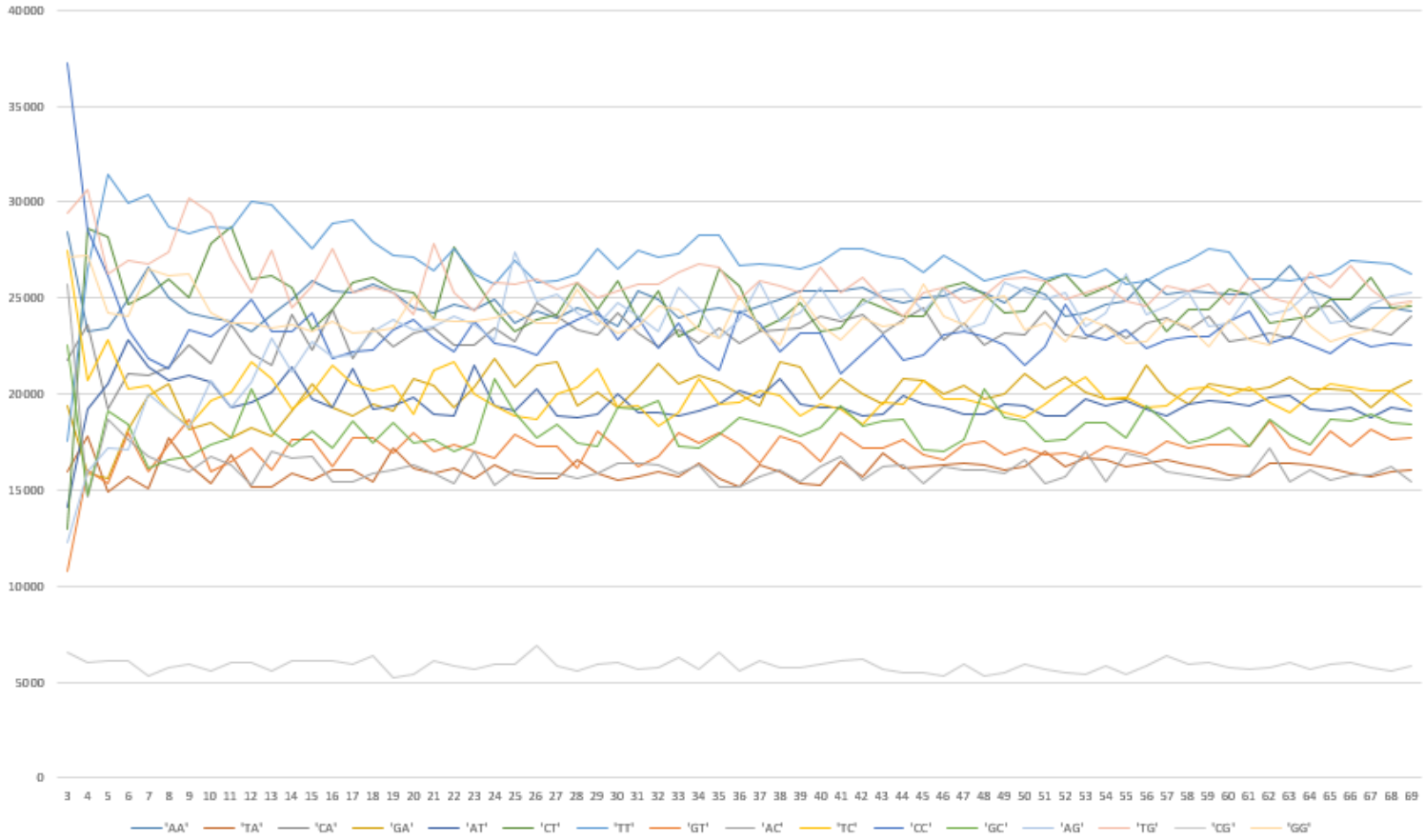


Fig. A1.2 Numărul de apariții al fiecărui tuplu pentru fiecare poziție din lista secvențelor invalide de *splicing*

Tabelul A1.5 Diferența dintre raporturile de reprezentare a tuplilor între secvențele valide și cele invalide de *splicing*

Poz	AA	TA	CA	GA	AT	CT	TT	GT	AC	TC	CC	GC	AG	TG	CG	GG
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	-0.22	0.03	0.42	-0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	-0.07	-0.04	-0.06	-0.05	-0.01	0.03	0.01	0.00	0.08	0.06	0.15	0.12	-0.04	-0.09	-0.02	-0.08
4	-0.04	0.01	0.03	-0.03	-0.05	-0.01	0.04	-0.04	-0.02	0.07	0.08	-0.02	-0.04	0.06	0.03	-0.06
5	-0.06	-0.03	-0.02	-0.04	-0.03	0.10	0.14	-0.03	-0.03	0.04	0.15	-0.04	-0.05	-0.04	0.00	-0.06
6	-0.06	-0.03	-0.03	-0.05	-0.03	0.08	0.08	-0.02	-0.02	0.10	0.16	-0.02	-0.05	-0.04	-0.01	-0.06
7	-0.07	-0.02	-0.02	-0.05	-0.03	0.09	0.05	0.00	-0.01	0.09	0.11	0.00	-0.06	-0.02	0.00	-0.06
8	-0.07	-0.03	-0.03	-0.05	-0.03	0.08	0.06	0.00	-0.02	0.12	0.10	-0.01	-0.06	0.00	0.01	-0.06
9	-0.06	-0.02	-0.04	-0.05	-0.02	0.10	0.09	-0.01	-0.02	0.10	0.09	-0.01	-0.06	-0.01	0.00	-0.06
10	-0.06	-0.02	-0.03	-0.05	-0.03	0.08	0.10	-0.01	-0.02	0.13	0.06	-0.01	-0.06	-0.02	0.00	-0.06
11	-0.06	-0.03	-0.04	-0.05	-0.02	0.10	0.11	0.00	-0.02	0.08	0.06	-0.02	-0.06	-0.01	-0.01	-0.05
12	-0.06	-0.02	-0.04	-0.05	-0.02	0.09	0.09	0.00	-0.02	0.08	0.07	-0.02	-0.06	0.00	0.00	-0.05
13	-0.06	-0.03	-0.03	-0.04	-0.02	0.08	0.09	0.00	-0.01	0.08	0.07	-0.01	-0.06	-0.01	0.00	-0.04
14	-0.05	-0.03	-0.03	-0.04	-0.02	0.08	0.07	0.00	-0.02	0.07	0.08	0.00	-0.06	0.00	0.00	-0.05
15	-0.06	-0.02	-0.02	-0.05	-0.01	0.08	0.06	-0.01	-0.01	0.07	0.07	0.00	-0.06	0.01	0.00	-0.04
16	-0.05	-0.02	-0.03	-0.03	-0.02	0.09	0.05	0.00	0.00	0.04	0.08	0.00	-0.05	-0.01	0.00	-0.04
17	-0.05	-0.02	-0.02	-0.03	-0.02	0.06	0.03	-0.01	0.01	0.05	0.08	0.00	-0.04	0.01	0.00	-0.04
18	-0.05	-0.01	-0.02	-0.03	-0.01	0.04	0.04	0.00	0.01	0.06	0.06	0.00	-0.04	0.00	0.00	-0.04
19	-0.04	-0.02	0.00	-0.03	0.00	0.05	0.04	0.00	0.02	0.04	0.04	-0.01	-0.05	0.01	0.00	-0.04
20	-0.03	-0.01	0.00	-0.02	-0.01	0.06	0.02	-0.01	0.02	0.04	0.04	-0.01	-0.04	0.01	0.00	-0.05
21	-0.03	0.00	0.00	-0.01	0.00	0.06	0.00	-0.01	0.01	0.03	0.04	0.00	-0.03	-0.01	0.00	-0.05
22	-0.03	-0.02	-0.01	0.00	0.00	0.03	0.00	-0.01	0.02	0.04	0.04	-0.01	-0.03	0.01	0.00	-0.05
23	-0.02	0.00	0.00	-0.01	-0.01	0.04	0.01	0.00	0.01	0.02	0.03	0.00	-0.04	0.00	0.00	-0.03
24	-0.02	-0.01	-0.01	-0.01	-0.01	0.04	0.01	-0.01	0.02	0.03	0.04	-0.01	-0.02	0.00	0.00	-0.03
25	-0.02	0.00	0.00	-0.02	0.00	0.05	0.00	-0.01	0.01	0.03	0.04	-0.01	-0.04	0.01	0.00	-0.02
26	-0.01	-0.01	-0.01	-0.02	-0.01	0.05	0.00	0.00	0.00	0.03	0.05	0.00	-0.03	-0.01	0.00	-0.03
27	-0.02	-0.01	-0.01	-0.02	-0.01	0.03	0.00	0.00	0.02	0.02	0.04	0.01	-0.03	0.00	0.00	-0.02
28	-0.02	-0.01	0.00	-0.01	0.00	0.02	0.00	-0.01	0.01	0.01	0.03	0.00	-0.02	0.00	0.00	-0.01
29	-0.02	-0.01	0.00	-0.01	-0.01	0.03	-0.01	0.00	0.01	0.01	0.03	0.01	-0.01	0.00	0.00	0.00
30	-0.02	0.00	0.00	0.00	-0.02	0.02	-0.01	-0.01	0.00	0.02	0.03	0.00	-0.01	0.00	0.00	0.00
31	-0.03	-0.01	0.00	-0.01	-0.01	0.03	-0.02	0.00	0.00	0.03	0.02	0.01	-0.02	0.01	0.00	0.00

32	-0.03	-0.01	-0.01	-0.02	-0.01	0.03	-0.01	0.01	-0.01	0.02	0.04	0.01	-0.01	0.01	0.00	0.00
33	-0.03	-0.02	0.00	-0.02	-0.01	0.03	-0.01	0.00	0.01	0.01	0.02	0.01	-0.02	0.00	0.00	0.00
34	-0.02	-0.01	0.00	-0.01	-0.01	0.02	-0.02	-0.01	0.00	0.01	0.03	0.01	-0.01	0.00	0.00	0.01
35	-0.03	-0.02	0.00	0.00	-0.01	0.02	-0.01	-0.01	0.00	0.01	0.03	0.01	0.00	-0.01	0.00	0.01
36	-0.03	-0.01	-0.01	0.00	-0.01	0.01	-0.02	0.00	0.00	0.01	0.03	0.01	-0.01	0.00	0.01	0.02
37	-0.03	-0.02	0.00	0.00	-0.02	0.02	-0.01	0.00	0.00	-0.01	0.03	0.01	-0.01	0.01	0.00	0.03
38	-0.02	-0.02	-0.01	-0.01	-0.02	0.01	-0.01	-0.01	0.00	0.01	0.03	0.02	0.00	0.00	0.00	0.02
39	-0.02	-0.01	-0.01	0.00	-0.02	0.01	-0.01	0.00	0.00	0.01	0.01	0.01	-0.01	0.00	0.00	0.03
40	-0.03	-0.01	0.00	-0.01	-0.01	0.01	-0.01	0.01	-0.01	0.00	0.03	0.00	0.00	0.00	0.00	0.03
41	-0.02	-0.02	-0.01	0.00	-0.02	0.01	-0.01	0.00	-0.01	0.01	0.04	0.01	-0.01	0.00	0.00	0.03
42	-0.03	-0.01	-0.01	-0.01	-0.02	0.01	-0.01	0.01	-0.01	0.00	0.04	0.01	-0.01	0.02	0.01	0.02
43	-0.03	-0.02	-0.01	-0.01	-0.01	0.01	-0.01	0.01	0.00	0.01	0.03	0.01	-0.01	0.01	0.00	0.03
44	-0.03	-0.02	0.00	0.00	-0.02	0.01	-0.01	0.00	-0.01	0.00	0.02	0.01	-0.01	0.01	0.00	0.04
45	-0.02	-0.02	-0.01	-0.01	-0.02	0.02	-0.01	0.00	-0.01	0.01	0.02	0.01	0.00	0.01	0.01	0.03
46	-0.02	-0.01	-0.01	-0.01	-0.01	0.01	-0.01	0.01	-0.01	0.00	0.02	0.02	-0.01	0.00	0.00	0.03
47	-0.02	-0.02	-0.02	0.00	-0.01	0.01	-0.01	0.00	-0.01	0.01	0.01	0.01	-0.01	0.01	0.00	0.04
48	-0.03	-0.01	-0.01	0.00	-0.01	0.01	-0.01	0.01	-0.01	-0.01	0.02	0.00	0.00	0.02	0.00	0.04
49	-0.03	-0.01	-0.01	-0.01	-0.02	0.00	-0.01	0.01	-0.01	0.00	0.02	0.02	0.00	-0.01	0.00	0.05
50	-0.02	-0.01	-0.01	-0.02	-0.02	0.00	-0.01	0.00	-0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.06
51	-0.03	-0.01	0.00	0.00	-0.01	-0.01	-0.01	0.00	-0.01	0.00	0.02	0.00	0.00	0.01	0.00	0.05
52	-0.02	-0.01	-0.02	0.00	-0.01	0.00	-0.01	0.01	-0.01	0.00	0.01	0.00	0.00	0.01	0.00	0.05
53	-0.02	-0.02	0.00	0.00	-0.01	0.00	-0.01	0.01	-0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.05
54	-0.02	-0.02	-0.01	0.00	-0.02	-0.01	-0.01	0.00	0.00	-0.01	0.01	0.01	0.01	0.02	0.00	0.05
55	-0.02	-0.02	0.00	0.01	-0.02	0.00	-0.01	0.01	-0.02	-0.01	0.00	0.01	0.00	0.01	0.00	0.04
56	-0.02	-0.02	-0.01	0.00	-0.01	0.00	-0.02	0.00	-0.01	0.00	0.01	0.00	0.00	0.02	0.00	0.05
57	-0.02	-0.02	-0.01	0.01	-0.02	0.01	-0.02	0.00	-0.01	0.00	0.01	0.01	-0.01	0.00	0.00	0.05
58	-0.03	-0.01	-0.02	0.00	-0.02	0.00	-0.03	0.00	0.00	-0.01	0.01	0.01	0.01	0.02	0.00	0.04
59	-0.03	-0.01	-0.01	0.00	-0.01	0.00	-0.02	0.01	-0.01	-0.01	0.00	0.01	0.00	0.01	0.01	0.04
60	-0.03	-0.01	-0.01	0.00	-0.02	-0.01	-0.02	0.01	-0.01	0.00	0.01	0.02	0.01	0.01	0.00	0.05
61	-0.02	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	0.01	-0.01	-0.02	0.00	0.01	0.02	0.00	0.00	0.05
62	-0.02	-0.01	0.00	0.01	-0.01	0.00	-0.02	-0.01	-0.01	-0.01	0.00	-0.01	0.00	0.00	0.01	0.05
63	-0.02	-0.01	0.00	0.00	-0.01	0.00	-0.01	-0.01	-0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.04
64	-0.02	-0.01	-0.01	-0.01	-0.02	0.01	-0.01	0.00	-0.01	0.00	0.02	0.01	0.00	0.01	0.00	0.04
65	-0.02	-0.01	-0.01	-0.01	-0.01	0.01	-0.02	0.00	-0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.04
66	-0.01	-0.01	0.00	-0.01	-0.01	0.00	-0.02	0.00	-0.01	0.00	0.01	0.02	0.00	-0.01	0.00	0.05
67	-0.02	-0.01	-0.01	0.00	-0.01	0.00	-0.02	-0.01	-0.01	0.00	0.01	0.00	0.00	0.01	0.01	0.05
68	-0.01	-0.01	-0.01	0.00	-0.01	0.00	-0.01	0.00	-0.01	0.00	0.02	0.00	0.00	0.01	0.00	0.04

Tabelul A1.6 Raportul dintre perechile de baze azotate posibile și tuplul care are numărul maxim de apariții pentru o anumită poziție

Poz.	AA	TA	CA	GA	AT	CT	TT	GT	AC	TC	CC	GC	AG	TG	CG	GG
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
2	0.05	0.26	1.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.04	0.03	0.04	0.02	0.14	0.25	0.23	0.12	0.59	0.55	1.00	0.72	0.00	0.01	0.01	0.00
4	0.18	0.38	0.59	0.08	0.07	0.48	0.70	0.05	0.13	0.80	1.00	0.14	0.03	0.95	0.26	0.14
5	0.04	0.08	0.14	0.03	0.12	0.80	1.00	0.08	0.11	0.48	0.97	0.09	0.00	0.15	0.06	0.07
6	0.05	0.06	0.13	0.03	0.19	0.67	0.74	0.15	0.15	0.71	1.00	0.15	0.01	0.17	0.03	0.08
7	0.08	0.15	0.23	0.04	0.19	0.94	0.78	0.25	0.21	0.86	1.00	0.28	0.01	0.33	0.08	0.11
8	0.06	0.16	0.23	0.05	0.16	0.90	0.85	0.28	0.18	1.00	0.93	0.21	0.01	0.45	0.14	0.10
9	0.07	0.14	0.19	0.01	0.22	0.99	1.00	0.27	0.16	0.90	0.93	0.23	0.00	0.47	0.08	0.09
10	0.06	0.15	0.17	0.03	0.19	0.87	1.00	0.23	0.14	0.99	0.66	0.20	0.00	0.36	0.09	0.10
11	0.04	0.11	0.17	0.04	0.19	0.94	1.00	0.26	0.15	0.70	0.67	0.17	0.01	0.34	0.06	0.11
12	0.07	0.14	0.18	0.05	0.22	0.95	1.00	0.28	0.15	0.85	0.84	0.26	0.01	0.44	0.09	0.14
13	0.08	0.11	0.19	0.07	0.23	0.88	1.00	0.27	0.21	0.77	0.80	0.22	0.03	0.41	0.11	0.15
14	0.14	0.14	0.26	0.09	0.28	1.00	0.98	0.33	0.21	0.79	0.93	0.31	0.06	0.48	0.12	0.14
15	0.13	0.18	0.32	0.10	0.30	1.00	0.96	0.27	0.28	0.87	0.92	0.37	0.07	0.56	0.11	0.18
16	0.15	0.17	0.27	0.15	0.27	1.00	0.86	0.29	0.32	0.65	0.91	0.34	0.08	0.48	0.11	0.19
17	0.19	0.18	0.33	0.18	0.30	0.94	0.80	0.28	0.35	0.76	1.00	0.37	0.16	0.56	0.11	0.22
18	0.22	0.26	0.43	0.27	0.38	0.94	1.00	0.39	0.48	0.93	0.99	0.41	0.23	0.63	0.14	0.24
19	0.29	0.22	0.50	0.25	0.44	1.00	0.90	0.36	0.50	0.74	0.81	0.35	0.21	0.68	0.14	0.23
20	0.36	0.26	0.50	0.30	0.41	1.00	0.79	0.32	0.52	0.73	0.83	0.36	0.25	0.61	0.12	0.21
21	0.33	0.34	0.52	0.36	0.42	1.00	0.63	0.34	0.48	0.73	0.86	0.38	0.29	0.56	0.14	0.21
22	0.43	0.30	0.55	0.48	0.50	1.00	0.76	0.34	0.61	0.92	0.97	0.40	0.34	0.80	0.14	0.21
23	0.42	0.37	0.54	0.43	0.47	1.00	0.73	0.39	0.54	0.67	0.85	0.41	0.29	0.59	0.14	0.32
24	0.45	0.37	0.57	0.45	0.45	1.00	0.73	0.34	0.56	0.74	0.92	0.49	0.39	0.72	0.14	0.41
25	0.46	0.38	0.61	0.37	0.45	1.00	0.69	0.35	0.53	0.73	0.91	0.41	0.35	0.72	0.16	0.42
26	0.49	0.32	0.56	0.36	0.39	0.97	0.66	0.40	0.44	0.68	1.00	0.47	0.35	0.61	0.19	0.34
27	0.50	0.38	0.58	0.40	0.44	0.93	0.68	0.45	0.62	0.77	1.00	0.63	0.40	0.67	0.17	0.49
28	0.47	0.39	0.68	0.49	0.52	0.97	0.74	0.38	0.52	0.72	1.00	0.55	0.52	0.72	0.14	0.66

29	0.53	0.37	0.70	0.45	0.44	0.97	0.69	0.51	0.58	0.68	1.00	0.57	0.58	0.71	0.16	0.65
30	0.52	0.41	0.68	0.52	0.41	0.94	0.66	0.45	0.49	0.74	1.00	0.60	0.60	0.73	0.21	0.64
31	0.50	0.39	0.66	0.52	0.42	1.00	0.67	0.50	0.47	0.83	0.94	0.67	0.51	0.86	0.22	0.69
32	0.46	0.34	0.57	0.42	0.41	1.00	0.63	0.54	0.38	0.71	0.95	0.61	0.53	0.81	0.17	0.71
33	0.48	0.34	0.71	0.48	0.51	1.00	0.74	0.55	0.62	0.72	1.00	0.69	0.64	0.86	0.25	0.82
34	0.54	0.40	0.71	0.50	0.49	0.93	0.70	0.44	0.46	0.71	1.00	0.66	0.64	0.82	0.18	0.79
35	0.50	0.33	0.76	0.62	0.47	1.00	0.74	0.51	0.49	0.73	1.00	0.72	0.72	0.78	0.22	0.78
36	0.47	0.40	0.62	0.61	0.47	0.89	0.64	0.50	0.45	0.71	1.00	0.72	0.60	0.79	0.24	0.92
37	0.44	0.27	0.63	0.51	0.40	0.84	0.65	0.48	0.44	0.52	1.00	0.60	0.64	0.79	0.22	0.91
38	0.56	0.34	0.64	0.62	0.47	0.88	0.70	0.43	0.55	0.72	1.00	0.74	0.77	0.84	0.21	0.97
39	0.58	0.33	0.65	0.65	0.38	0.83	0.72	0.52	0.44	0.69	0.82	0.63	0.65	0.75	0.19	1.00
40	0.49	0.34	0.71	0.48	0.47	0.77	0.68	0.54	0.38	0.56	1.00	0.50	0.75	0.83	0.22	0.96
41	0.54	0.34	0.60	0.63	0.37	0.81	0.73	0.52	0.43	0.64	1.00	0.65	0.64	0.74	0.15	0.97
42	0.45	0.33	0.64	0.51	0.38	0.77	0.67	0.55	0.36	0.53	1.00	0.60	0.65	0.91	0.24	0.88
43	0.47	0.32	0.55	0.51	0.45	0.76	0.67	0.59	0.44	0.62	1.00	0.61	0.60	0.83	0.14	1.00
44	0.41	0.26	0.59	0.53	0.38	0.70	0.63	0.49	0.35	0.51	0.78	0.57	0.58	0.72	0.16	1.00
45	0.47	0.31	0.56	0.48	0.37	0.81	0.60	0.47	0.34	0.63	0.82	0.57	0.69	0.81	0.23	1.00
46	0.53	0.39	0.56	0.53	0.43	0.83	0.71	0.55	0.43	0.57	0.89	0.72	0.71	0.80	0.21	1.00
47	0.54	0.31	0.48	0.56	0.38	0.75	0.59	0.51	0.35	0.64	0.73	0.52	0.56	0.76	0.20	1.00
48	0.41	0.32	0.54	0.56	0.41	0.75	0.60	0.52	0.36	0.47	0.82	0.51	0.63	0.82	0.14	1.00
49	0.39	0.30	0.50	0.43	0.30	0.59	0.56	0.51	0.31	0.43	0.66	0.59	0.58	0.58	0.14	1.00
50	0.41	0.26	0.50	0.37	0.32	0.56	0.50	0.43	0.30	0.45	0.60	0.49	0.64	0.65	0.18	1.00
51	0.36	0.30	0.58	0.54	0.37	0.59	0.59	0.44	0.31	0.50	0.70	0.46	0.62	0.70	0.14	1.00
52	0.43	0.33	0.45	0.53	0.41	0.71	0.56	0.51	0.33	0.51	0.75	0.50	0.69	0.76	0.13	1.00
53	0.43	0.28	0.52	0.52	0.37	0.60	0.54	0.51	0.28	0.47	0.69	0.48	0.59	0.63	0.17	1.00
54	0.47	0.28	0.49	0.53	0.36	0.62	0.59	0.46	0.36	0.45	0.71	0.57	0.68	0.84	0.13	1.00
55	0.49	0.29	0.61	0.64	0.41	0.75	0.59	0.59	0.31	0.50	0.69	0.60	0.75	0.84	0.18	1.00
56	0.49	0.27	0.48	0.52	0.38	0.67	0.51	0.45	0.37	0.50	0.67	0.48	0.63	0.78	0.15	1.00
57	0.48	0.26	0.47	0.54	0.32	0.65	0.51	0.46	0.28	0.43	0.62	0.51	0.53	0.62	0.16	1.00
58	0.45	0.33	0.48	0.54	0.38	0.69	0.47	0.49	0.41	0.48	0.73	0.51	0.73	0.88	0.20	1.00
59	0.48	0.38	0.63	0.63	0.43	0.72	0.57	0.60	0.39	0.52	0.70	0.62	0.68	0.82	0.22	1.00
60	0.42	0.30	0.51	0.56	0.37	0.60	0.57	0.52	0.29	0.48	0.68	0.60	0.66	0.74	0.16	1.00
61	0.48	0.31	0.48	0.50	0.43	0.56	0.62	0.52	0.35	0.40	0.67	0.55	0.87	0.72	0.13	1.00

62	0.52	0.35	0.61	0.63	0.45	0.64	0.52	0.43	0.33	0.46	0.61	0.43	0.65	0.68	0.22	1.00
63	0.50	0.38	0.58	0.57	0.42	0.63	0.63	0.41	0.33	0.47	0.75	0.61	0.69	0.68	0.22	1.00
64	0.51	0.36	0.62	0.52	0.39	0.76	0.64	0.45	0.38	0.52	0.81	0.54	0.70	0.83	0.15	1.00
65	0.56	0.35	0.62	0.54	0.48	0.80	0.62	0.56	0.39	0.57	0.85	0.64	0.67	0.79	0.17	1.00
66	0.48	0.32	0.55	0.45	0.40	0.62	0.50	0.46	0.33	0.48	0.66	0.61	0.59	0.60	0.15	1.00
67	0.46	0.34	0.49	0.48	0.37	0.63	0.51	0.38	0.33	0.48	0.65	0.49	0.62	0.67	0.19	1.00
68	0.55	0.36	0.54	0.58	0.48	0.64	0.62	0.52	0.37	0.53	0.78	0.54	0.71	0.76	0.18	1.00
69	0.55	0.40	0.54	0.58	0.44	0.70	0.57	0.53	0.33	0.53	0.71	0.53	0.66	0.74	0.18	1.00

Anexa 2. Rezultate *Random Forest*

Tabel A2.1 Rezultatele configurațiilor metodei Random Forest

Etapa	Criteriul pentru despărțire	Numărul maxim de niveluri	Eșantionare	Acuratețe [%] (media/min/ max/stdev)	Coeficinetul Cohen (k)*10^2 (media/min/max /stddev)
Etapa 1: fișierul inițial	*Information Gain*	Automat	Stratificat	32 / 18 / 45 / 4	4 / (-14) / 23 / 5
	Information Gain ratio	Automat	Stratificat	32 / 21 / 45 / 4	6 / (-7) / 21 / 6
	Gini Index	Automat	Stratificat	33 / 22 / 43 / 4	5 / (-8) / 19 / 6
	Information Gain	Automat	Aleatoriu	30 / 20 / 40 / 4	4 / (-12) / 17 / 5
	Information Gain Ratio	Automat	Aleatoriu	31 / 17 / 47 / 5	5 / (-10) / 23 / 6
	Gini index	Automat	Aleatoriu	33 / 22 / 45 / 4	6 / (-6) / 23 / 6
	Gini index	3	Aleatoriu	32 / 20 / 46 / 5	4 / (-8) / 20 / 6
	Gini index	5	Aleatoriu	32 / 23 / 45 / 5	3 / (-11) / 17 / 6
	Gini index	50	Aleatoriu	32 / 20 / 43 / 4	5 / (-7) / 22 / 5
Etapa 2: variantele genetice în stadiul binar (prezent – absent) indiferent de zigozitate	*Information Gain*	Automat	Stratificat	31 / 22 / 42 / 4	3 / (-8) / 19 / 6
	Information Gain Ratio	Automat	Stratificat	31 / 21 / 42 / 4	3 / (-10) / 20 / 6
	Gini Index	Automat	Stratificat	31 / 21 / 42 / 4	3 / (-10) / 20 / 6
	Information Gain	Automat	Aleatoriu	30 / 20 / 41 / 4	3 / (-13) / 14 / 5
	Information Gain Ratio	Automat	Aleatoriu	30 / 21 / 42 / 5	3 / (-10) / 20 / 6
	Gini Index	Automat	Aleatoriu	30 / 21 / 46 / 5	3 / (11) / 24 / 6
	Gini Index	3	Aleatoriu	33 / 22 / 43 / 4	4 / (-4) / 16 / 4
	Gini Index	5	Aleatoriu	32 / 22 / 46 / 4	4 / (-10) / 22 / 5
	Gini Index	50	Aleatoriu	31 / 20 / 45 / 5	4 / (-12) / 20 / 6
Etapa 3: Proiectul valori binare	*Information Gain*	Automat	Stratificat	81 / 75 / 85 / 2	6 / (-10) / 28 / 8
	Information Gain Ratio	Automat	Stratificat	81 / 76 / 85 / 2	8 / (-8) / 33 / 10
	Gini Index	Automat	Stratificat	81 / 76 / 83 / 2	7 / (-8) / 25 / 8
	Information Gain	Automat	Aleatoriu	81 / 72 / 91 / 4	6 / (-4) / 27 / 8
	Information Gain Ratio	Automat	Aleatoriu	80 / 70 / 87 / 4	7 / (-6) / 30 / 8
	Gini Index	Automat	Aleatoriu	81 / 72 / 88 / 4	6 / (-6) / 31 / 8
	Gini Index	3	Aleatoriu	81 / 72 / 91 / 4	1 / (-4) / 17 / 4
	Gini Index	5	Aleatoriu	81 / 68 / 91 / 4	4 / (-6) / 27 / 6
	Gini Index	50	Aleatoriu	81 / 66 / 90 / 4	6 / (-8) / 30 / 9
Etapa 4: Reducerea	*Information Gain*	Automat	Stratificat	80 / 76 / 84 / 2	3 / (-8) / 25 / 6
	Information Gain Ratio	Automat	Stratificat	80 / 76 / 85 / 2	3 / (-8) / 29 / 7
	Gini Index	Automat	Stratificat	80 / 73 / 85 / 2	2 / (-8) / 29 / 7

	Information Gain	Automat	Aleatoriu	80 / 70 / 91 / 2	2 / (-10) / 19 / 6
	Inf. GR	Automat	Aleatoriu	79 / 71 / 88 / 4	2 / (-11) / 22 / 6
	Gini Index	Automat	Aleatoriu	80 / 70 / 89 / 4	3 / (-13) / 20 / 7
	Gini Index	3	Aleatoriu	81 / 64 / 91 / 5	0 / (-2) / 0 / 0
	Gini Index	5	Aleatoriu	81 / 71 / 91 / 4	1 / (-5) / 13 / 4
	Gini Index	50	Aleatoriu	81 / 73 / 90 / 4	4 / (-11) / 32 / 9

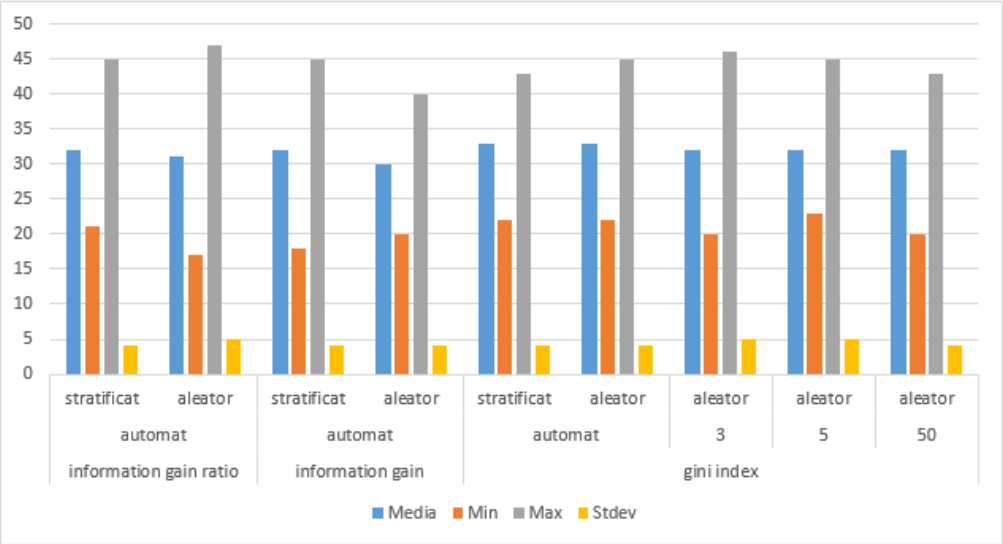


Fig. A2.1 Rezultatele obținute în urma aplicării metodei *Random Forest* pentru steatoza cu stări multiple

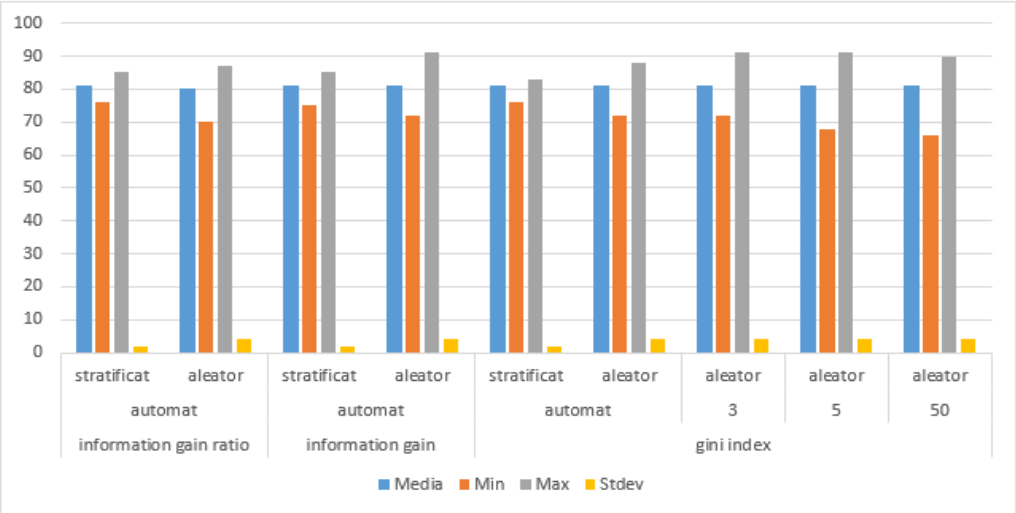


Fig. A2.2 Rezultatele obținute în urma aplicării metodei *Random Forest* pentru steatoza cu stare binară